Statistical Modelling & Probability basics

Sofia Olhede

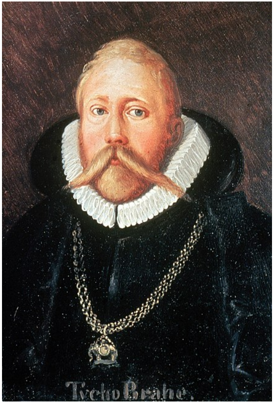
EPFL

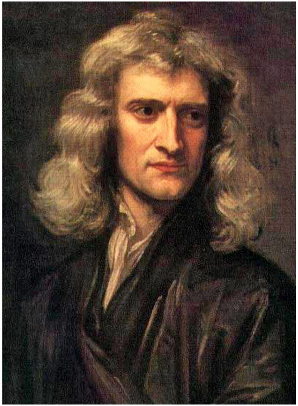
September 15, 2020

1. Lecture MA 413 – Statistics for Data Science
2. Basics of Modelling
3. Statistical Modelling
4. Conditional Probability and Dependence

- This course is taught by Sofia Olhede (me).
- The schedule is virtual lectures on Mondays (at 12 noon), and Tuesdays (at 2 pm) with a problem class on Wednesdays (at 1 pm).
- The recommended texts are Davison, A.C. (2003). Statistical Models, Cambridge, Panaretos, V.M. (2016). Statistics for Mathematicians. Birkhäuser and Wasserman, L. (2004). All of Statistics. Springer.
- There are two frameworks for statistical modelling; the explanatory model framework, the predictive framework. There are two goals in when extracting structure from data: 1) prediction, e.g predict future responses given future inputs; 2) extract information on how the response variable relate to any input.
- The explanatory framework starts from assuming a model to describe the observations.
- The predictive framework starts from assuming you can find a function $f(x)$ which maps from input x to an output $f(x)$. Predictions are usually implemented by an algorithm, e.g. set of rules followed in problem-solving operations.

- What are examples of this? Johannes Kepler modelled the laws of planetary motion from observations by Tycho Brahe.
- But following their work, Isaac Newton formulated the three Laws of Motion.



- We cannot always model the resolution of the observed data, or indeed all variation.
- This prompts us to introduce stochasticity in our model.
- Why is data stochastic?
- i) Measurement error, ii) chaos, iii) intrinsic stochasticity, iv) sampled data or v) fundamental limit of a process.
- How does probability fit in?

  * Process of interest conceptualised as a probability model;

  * Use model to learn about the probability of outcomes.
- What is the role of statistics?

  * Process of interest instantiated from a mathematical model;

  * The data is viewed as observations from that model.

- Example: Coin flipping

  The variables $Y_1, \ldots, Y_n \in \{0,1\}^n$ are outcomes from flipping a coin 10 times. We might model

$$Y_i \overset{iid}{\sim} \mathrm{Bernoulli}(\theta).$$

Bernoulli Distribution

$$Y \sim \mathrm{Bernoulli}(\theta),$$

if

$$Y = \begin{cases} 1 & \mathrm{wp} \quad \theta \\ 0 & \mathrm{wp} \quad 1-\theta \end{cases}.$$

- Say we observe $(0,0,0,1,0,1,1,1,1,1)$.

- Probability Qns:

  What is the probability of k-long run?
  If we keep tossing, how many k-long runs?

- Statistics Qns:

  If the coin fair? $(\theta = 1/2?)$
  What is a good value of θ given Y?
  How large an error are we likely to make guessing θ from Y?

- Model the distribution $F(y_1, \ldots, y_n; \theta)$ where $y \in \mathcal{Y}^n$ and $y_i \in \mathcal{Y}$.
- Usually we assume that $F(y_1, \ldots, y_n; \theta)$ is known, but θ is unknown.
- Observe realisation of $Y = (Y_1, \ldots, Y_n)^T \in \mathcal{Y}^n$.
- Use the realisation in order to make assertions concerning the true value of θ, and quantify the uncertainty.
- When $F(\cdot; \boldsymbol{\theta})$ is known then we have a parametric problem, when $F(\cdot)$ is unknown the problem is non-parametric. (In between is the semi-parametric framework).
- The first problem is <u>parametric</u>, the second <u>non–parametric</u>. Sometimes we speak of <u>finite dimensional</u> and <u>infinite dimensional</u> problems.

- Typical Statistics problems include:

  Prediction;
  Model fit assessment;
  Estimation;
  Hypothesis testing;
  Confidence intervals;
  Marginal Inference;
  Regression.

- Algebra of events. Experiment: a process whose outcome is uncertain.
- Outcomes are normally understood using set theory.

Basics of Probability I

- We shall model outcomes of experiments. A possible outcome ω is called an elementary event.
- The set of outcomes will be written as Ω.
- We always assume $\Omega \neq \emptyset$.
- An event is a subset $F \subset \Omega$ of Ω. An event F is "realised" whenever the outcome of the experiment is an element of F.
- The union of two events F_1 and F_2 written as $F_1 \cup F_2$ occurs if and only if either of F_1 or F_2 occurs. Equivalently

$$F_1 \cup F_2 = \{\omega \in \Omega : \ \omega \in F_1 \,\mathrm{or} \omega \in F_2\}.$$

- The intersection of two events F_1 and F_2 written as $F_1 \cap F_2$ occurs if and only if both of F_1 or F_2 occurs. Equivalently

$$F_1 \cap F_2 = \{\omega \in \Omega : \ \omega \in F_1 \,\mathrm{and} \omega \in F_2\}.$$

- Union and intersection of several events $F_1 \cup \cdots \cup F_n$ and $F_1 \cap \cdots \cap F_n$ are defined iteratively.

Basics of Probability II

- The complement of an event F written as F^c contains all the elements in Ω that are not in F or

$$F^c = \{\omega \in \Omega : \omega \notin F\}.$$

- Two events F_1 and F_2 are disjoint if they have no elements in common, or $F_1 \cap F_2 = \emptyset$.
- A partition $\{F_n\}_{n \geq 1}$ is a collection of events such that $F_i \cap F_j = \emptyset$ for all $i \neq j$ and $\cup_{n \geq 1} F_n = \Omega$.
- The difference between two elements F_1 and F_2 is defined as $F_1 \backslash F_2 = F_1 \cap F_2^c$. Notice that the difference is NOT symmetric.

Basics of Probability III

- The following properties hold:

$$
\begin{aligned}
(i) &\quad (F_1 \cup F_2) \cup F_3 = F_1 \cup (F_2 \cup F_3) = F_1 \cup F_2 \cup F_3 : \ \text{associativity} \\
(ii) &\quad (F_1 \cap F_2) \cap F_3 = F_1 \cap (F_2 \cap F_3) = F_1 \cap F_2 \cap F_3 : \ \text{associativity} \\
(iii) &\quad \underline{F_1 \cap (F_2 \cup F_3) = (F_1 \cap F_2) \cup (F_1 \cap F_3)} : \quad \text{distributivity} \\
(iv) &\quad \underline{F_1 \cup (F_2 \cap F_3) = (F_1 \cup F_2) \cap (F_1 \cup F_3)} : \quad \text{distributivity} \\
(v) &\quad \underline{(F_1 \cup F_2)^c = F_1^c \cap F_2^c \quad \text{and} \quad (F_1 \cap F_2)^c = F_1^c \cup F_2^c}, \\
&\quad \text{De Morgan's Laws.}
\end{aligned}
$$

Basics of Probability IV

- Probability measures (without measure theory!!!)
- A Probability measure $\mathbb{P}$: is a real function defined over the events in Ω. This is assigning a probability to an event.
- This measure is interpreted as a measure of certainty: how certain are we that an event will happen?
- The measure is assumed to follow the following three constraints
  1. $\mathbb{P}(F) \geq 0$ for all $F \subset \Omega$.
  2. $\mathbb{P}(\Omega) = 1$.
  3. If an event G is a countable union $G = \cup_{n \geq 1} F_n$ of disjoint events $\{F_n\}$ then

$$\mathbb{P}(G) = \sum_{n \geq 1} \mathbb{P}(F_n).$$

Basics of Probability IV

- Having restated the **three axioms of probability**,
  1. $\mathbb{P}(F) \geq 0$ for all $F \subset \Omega$.
  2. $\mathbb{P}(\Omega) = 1$.
  3. If an event G is a countable union $G = \cup_{n \geq 1} F_n$ of disjoint events $\{F_n\}$ then
  $$\mathbb{P}(G) = \sum_{n \geq 1} \mathbb{P}(F_n).$$

  we can now establish other properties of of probability.

Basics of Probability V

- We seek to show that $\Pr(F_1 \cup F_2) = \Pr(F_1) - \Pr(F_1 \cap F_2) + \Pr(F_2)$. First we note that $\underline{F_1 \cup F_2 = (F_1 \backslash F_2) \cup (F_2 \backslash F_1) \cup (F_1 \cap F_2)}$. We note that the intersection of these three is zero. Secondly we use the third axiom of probability to say that as $\underline{F_1 = (F_1 \backslash F_2) \cup (F_1 \cap F_2)}$ and the latter two sets do not intersect

$$\begin{aligned}
\Pr(F_1 \cup F_2) &= \Pr(F_1 \backslash F_2) + \Pr(F_2 \backslash F_1) + \Pr(F_1 \cap F_2) \\
&= \Pr(F_1) - \Pr(F_1 \cap F_2) + \Pr(F_2) - \Pr(F_1 \cap F_2) \\
&+ \Pr(F_1 \cap F_2) \\
&= \Pr(F_1) - \Pr(F_1 \cap F_2) + \Pr(F_2). \qquad (1)
\end{aligned}$$

□

Basics of Probability VI

- Secondly we seek to show that $\Pr(F_1 \cap F_2) \leq \min\{\Pr(F_1), \Pr(F_2)\}$. We recall that as
$$F_1 = (F_1 \cap F_2) \cup (F_1 \backslash F_2).$$
As the latter two do not intersect we can yet again use axiom 3 and so arrive at
$$\Pr(F_1) = \Pr(F_1 \cap F_2) + \Pr(F_1 \backslash F_2).$$
As the last quantity is non-negative we have
$$\Pr(F_1 \cap F_2) \leq \Pr(F_1).$$
We can repeat the argument for F_2 and so arrive at
$$\Pr(F_1 \cap F_2) \leq \min\{\Pr(F_1), \Pr(F_2)\}.$$
- Finally we note that be definition $F \cup F^c = \Omega$. By the third axiom: $1 = \Pr(\Omega) = \Pr(F) + \Pr(F^c)$. From this we deduce $\Pr(F^c) = 1 - \Pr(F)$.

Conditional Probability and Independence I

- Suppose that we do not know that a precise outcome $\omega \in \Omega$ has occured, but we do know that $\omega \in F_2$ for some event F_2, and we want to understand the probability that $\omega \in F_1$.
- For any pair of events F_1 and F_2 such that $\Pr(F_2) > 0$ then we define the conditional probability of F_1 given F_2 to be

$$\underline{\Pr(F_1|F_2)} = \frac{\Pr(F_1 \cap F_2)}{\Pr(F_2)}.$$

- A partition of Ω is a collection of disjoint sets $\{F_j\}$ such that $\cup_j F_j = \Omega$.
- Let G be an event and $\{F_n\}_{n \geq 1}$ be a partition of Ω such that $\Pr(F_n) > 0$ for all n. We then have

  - Law of total probability:
  $\Pr(G) = \sum_{n=1}^{\infty} \Pr(G|F_n)\Pr(F_n)$.
  - Bayes' theorem: $\Pr(F_j|G) = \frac{\Pr(F_j \cap G)}{\Pr(G)} = \frac{\Pr(F_j \cap G)}{\sum_n \Pr(F_n \cap G)}$.

Conditional Probability and Independence II

- The events $\{G_n\}_{n\geq 1}$ are called independent if and only if for any sub-collection $\{G_{i_1}, \dots, G_{i_K}\}$, $K < \infty$, we have:

$$\Pr(G_{i_1} \cap \cdots \cap G_{i_K}) = \Pr(G_{i_1}) \times \Pr(G_{i_2}) \times \cdots \times \Pr(G_{i_K}).$$

- Random varibles, numerical summaries of the outcome of a random experiment.
- We can concerntrate on range of random variable, rather than look at Ω.
- - A random variable is a real function $X : \Omega \to \mathbb{R}$.
- We write $\{a \leq X \leq b\}$ to denote the event

$$\{\omega \in \Omega : a \leq X(\omega) \leq b\}.$$

- More generally, if $A \subset \mathbb{R}$ is a more general subset, we write $\{X \in A\}$ to denote the event

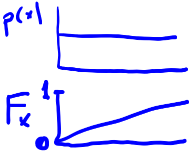
$$\{\omega \in \Omega : X(\omega) \in A\}.$$

Conditional Probability and Independence III

- If we have a probability measure defined on the events of Ω then X induces a new probability measure on subsets of the real line. This is described by the distribution function (or cumulative distribution function) $F_X : \mathbb{R} \to [0, 1]$ of a random variable X (or the law of X).

$$F_X(x) = \Pr(X \leq x).$$

- By its definition, a distribution function satisfies the following properties:



(i) $x \leq y \Rightarrow F_X(x) \leq F_X(y)$
(ii) $\lim_{x\to\infty} F_X(x) = 1$, $\lim_{x\to-\infty} F_X(x) = 0$.
(iii) $F_X(x)$ is right continuous
(iv) F_X is left limited
(v) $\Pr(a < X \leq b) = F_X(b) - F_X(a)$.
(vi) $\Pr(X > a) = 1 - F(a)$.

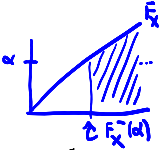
Conditional Probability and Independence IV

- Given a probability $\alpha \in (0,1)$ which is the (smallest) real number x such that $\Pr(X \leq x) = \alpha$?

- Let X be a random variable and F_X be its distribution function. We define the quantile function of X to be the function

$$F_X^- : (0,1) \to \mathbb{R} \tag{2}$$

$$F_X^-(\alpha) = \inf\{t \in \mathbb{R} : F_X(t) \geq \alpha\}. \tag{3}$$



- If F_X is strictly increasing and continuous, then $F_X^- = F_X^{-1}$.
- Given an $\alpha(0,1)$ the α-quantile of X is the real number

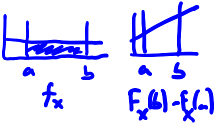
$$q_\alpha = F_X^-(\alpha).$$

- Let $Y \sim \mathrm{Unif}(0,1)$ and let F be a distribution function. Then the distribution function of the random variable $X = F^-(Y)$ is given precisely by F.

Conditional Probability and Independence V

- Can be used to generate realisations from any distribution:
- Provided we can generate realisations from uniform on $[0, 1]$.
- Can do this with binary expressions and Bernoulli draws.
- Reduces to the problem to infinite coin flipping.
- Let X be a random variable with strictly increasing and distribution function F_X. Then $F_X(X) \sim \mathrm{Unif}(0, 1)$.
- A continuous random variable X has probability density function f_X if

$$F_X(b) - F_X(a) = \int_a^b f_X(t)\, dt.$$



- By its definition a pdf satisfies

f ≡ density

(i) $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$,

(ii) $f_X(x) = F'(x)$ whenever $f_X(x)$ is continuous,

F ≡ probability

(iii) Note that $f_X(x) \neq \Pr(X = x) = 0$. Note that $f_X(x) > 1$ may be possible and $f_X(x)$ can even be unbounded.

$P(A < x \leq B) = F_x(B) - F_x(A)$

Conditional Probability and Independence VI

- For a discrete random variable X we may define its probability mass function (PMF) to be

PMF → discrete int
PDF → continuous int

$$f_X(x) = \Pr(X = x).$$

- The PMF satisfied the following three constraints

  (i) $\Pr(X \in A) = \sum_{t \in A \cap t \in \mathcal{X}} f_X(t)$, where $A \subseteq \mathcal{X}$ and $\mathcal{X} = \{x \in \mathbb{R} : \ f_X(x) > 0\}$.
  (ii) $F_X(x) = \sum_{t \in (-\infty, x) \cap \mathcal{X}} f_X(t)$ for all $x \in \mathbb{R}$ and $\mathcal{X} = \{x \in \mathbb{R} : \ f_X(x) > 0\}$.
  (iii) An immediate corollary is that $F_X(x)$ is piecewise constant with jumps at the points in $\mathcal{X}$.

Conditional Probability and Independence VII

- Examples: Bernoulli RVs

$$\begin{aligned} \mathcal{X} &= \{0, 1\}, \quad 1 > \theta > 0 \\ \Pr(X = 0) &= 1 - \theta \\ \Pr(X = 1) &= \theta. \end{aligned} \tag{4}$$

- Poisson RVs

$$\mathcal{X} = \{0, 1, 2, 3, \dots\}, \quad \mu > 0 \tag{5}$$

$$\Pr(X = x) = \frac{e^{-\mu}\mu^x}{x!}. \tag{6}$$

Conditional Probability and Independence VIII

instead of using random vars, we use the output of a function applied to random vars

- Transformed Mass Functions: let X be discrete taking values in $\mathcal{X}$ and let $Y = g(X)$. Then Y takes values in $\mathcal{Y} = g(\mathcal{X})$. Furthermore

$$F_Y(y) = \Pr(g(X) \le y) = \sum_{x \in \mathcal{X}} f_X(x) I\{g(x) \le y\}, \quad y \in \mathcal{Y} \tag{7}$$

$$f_Y(y) = \Pr(g(X) = y) = \sum_{x \in \mathcal{X}} f_X(x) I\{g(x) = y\}, \quad y \in \mathcal{Y} \tag{8}$$

- Let X be continuous taking values in $\mathcal{X} \subseteq \mathbb{R}$ and let $g : \mathcal{X} \to \mathbb{R}$ a transformation that is 1) monotone, 2) continuously differentiable, and 3) with non-vanishing derivative.

so that there's only 1 point corresponding in inverse:

- If $Y = g(X)$ then Y takes values in $\mathcal{Y} = g(\mathcal{X})$ and

$$f_Y(y) = \left| \frac{\partial}{\partial y} g^{-1}(y) \right| f_X(g^{-1}(y)), \quad y \in \mathcal{Y}.$$

inverse function, not $\frac{1}{g}$

NB the absolute value is necessary a transformation can be both non-decreasing and non-increasing. NB densities are always ≥ 0.

Random Vectors

- Random vectors: A random vector for a fixed positive integer d is $\mathrm{X} = \left(X_1 \quad \ldots X_d\right)^T$ is a finite collection of random variables.
- We want to understand the joint distribution of these random variables.
- The joint distribution of the random vector $\mathrm{X} = \left(X_1 \quad \ldots X_d\right)^T$ is defined as

$$F_{\mathrm{X}}(x_1, \ldots, x_d) = \Pr(X_1 \leq x_1, \ldots, X_d \leq x_d).$$

- Correspondingly one defines

  - A joint mass function if $\{X_i\}$ are all discrete, e.g.

$$f_{\mathrm{X}}(x_1, \ldots, x_d) = \Pr(X_1 = x_1, \ldots, X_d = x_d).$$

Random Vectors II

- - A joint density function if there exists $f_{\mathrm{X}} : \mathbb{R}^d \to [0, \infty)$ such that

$$F_{\mathrm{X}}(x_1, \ldots, x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f_{\mathrm{X}}(u_1, \ldots, u_d)\, du_1 \ldots du_d.$$

  In the latter case when $f_{\mathrm{X}}(x_1, \ldots, x_d)$ is continuous at x

$$f_{\mathrm{X}}(x_1, \ldots, x_d) = \frac{\partial^d}{\partial x_1 \ldots \partial x_d} F_X(x_1, \ldots, x_d).$$

- Given the joint distribution of X we can isolate the distribution of X_i.
- In the discrete case the marginal mass function of X_i is given by

$$f_{X_i}(x_i) = \Pr(X_i = x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_d} f_{\mathrm{X}}(x_1, \ldots, x_d).$$

no x_i

- In the continuous case, the marginal density function of X_i is given by

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathrm{X}}(y_1, \ldots, y_{i-1}, x_i, y_{i+1}, y_d)\, dy_1 \ldots dy_{i-1} dy_{i+1} \ldots dy_d.$$