Statistical Theory: Introduction

Statistical Theory

Guillaume Dehaene
Ecole polytechnique fédérale de Lausanne

What is This Course About?

What is This Course About

Statistics $\longrightarrow$ Extracting Information from Data

- Age of Universe (Astrophysics)
- Microarrays (Genetics)
- Stock Markets (Finance)
- Pattern Recognition (Artificial Intelligence)
- Climate Reconstruction (Paleoclimatology)
- Quality Control (Mass Production)
- Random Networks (Internet)
- Inflation (Economics)
- Phylogenetics (Evolution)
- Molecular Structure (Structural Biology)
- Seal Tracking (Marine Biology)
- Disease Transmission (Epidemics)

- Variety of different forms of data are bewildering
- But concepts involved in their analysis show fundamental similarities
- Imbed and rigorously study in a framework
- Is there a unified mathematical theory?


big
data

What is This Course About?

We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression.

Ronald A. Fisher

The object of rigor is to sanction and legitimize the the conquests of intuition, and there was never any other object for it.

Jacques Hadamard

What is This Course About?

Statistical Theory: What and How?

- What? The rigorous study of the procedure of extracting information from data using the formalism and machinery of mathematics.
- How? Thinking of data as outcomes of probability experiments

- Probability offers a natural language to describe uncertainty or partial knowledge
- Deep connections between probability and formal logic [Jaynes]
- Can break down phenomenon into *systematic* and *random* parts.

What can Data be?

To do probability we simply need a *measurable space* $(\Omega, \mathcal{F})$. Hence, almost anything that can be mathematically expressed can be thought as data (numbers, functions, graphs, shapes,...)

What is This Course About?

The Job of the Probabilist

Given a probability model $\mathbb{P}$ on a measurable space $(\Omega, \mathcal{F})$ find the probability $\mathbb{P}[A]$ that the outcome of the experiment is $A \in \mathcal{F}$.

The Job of the Statistician

Given an outcome of $A \in \mathcal{F}$ (the data) of a probability experiment on $(\Omega, \mathcal{F})$, tell me something *interesting** about the (unknown / partially unknown) probability model $\mathbb{P}$ that generated the outcome.
(*something in addition to what I knew before observing the outcome A)

The three questions of statistics

1. **Estimation**: what is the best sub-model inside a model class?
2. **Comparison**: of two/multiple models, which is the best?
3. **Prediction**: can I predict new values of the data?

A Probabilist and a Statistician Flip a Coin

Example

Let $X_1, ..., X_{10}$ denote the results of flipping a coin ten times, with

$$X_i = \begin{cases} 0 & \text{if heads} \\ 1 & \text{if tails} \end{cases}, \quad i = 1, ..., 10.$$

A plausible model is $X_i \overset{iid}{\sim} \text{Bernoulli}(\theta)$. We record the outcome

$$\mathbf{X} = (0, 0, 0, 1, 0, 1, 1, 1, 1, 1).$$

Probabilist Asks:

- Probability of outcome as function of θ?
- Probability of k-long run?
- If keep tossing, how many k-long runs? How long until k-long run?

A Probabilist and a Statistician Flip a Coin

Example (cont'd)

Statistician Asks:

- Is the coin fair?
- What is the true value of θ given $\mathbf{X}$?
- How much error do we make when trying to decide the above from $\mathbf{X}$?
- How does our answer change if $\mathbf{X}$ is perturbed?
- Is there a "best" solution to the above problems?
- How sensitive are our answers to departures from $X_i \overset{iid}{\sim} \text{Bernoulli}(\theta)$
- How do our "answers" behave as # tosses $\longrightarrow \infty$?
- How many tosses would we need until we can get "accurate answers"?
- Does our model agree with the data?

Statistical Theory (MATH442): Technicalities

- Web: `http://smat.epfl.ch/courses/theory.php`
- Course:
  - Tuesday, 9h15 – 11h00
  - Me
- Exercises:
  - Monday, 10h15 – 12h00
  - Tomas Rubin, `tomas.rubin@epfl.ch`
- **Only a final exam.**

Probability Review

Algebra of Events

Random experiment: process whose outcome is uncertain.
Outcomes and any statement involving them must be expressed via **set theory**.

- A possible outcome ω of a random experiment is called an elementary event.
- The set of all possible outcomes, say Ω is assumed non-empty, $\Omega \neq \emptyset$.
- An event is a subset $F \subset \Omega$ of Ω. An event F "is realised" (or "occurs") whenever the outcome of the experiment is an element of F.
- The union of two events F_1 and F_2, written $F_1 \cup F_2$ occurs if and only if either of F_1 or F_2 occurs. Equivalently, $\omega \in F_1 \cup F_2$ if and only if $\omega \in F_1$ or $\omega \in F_2$,

$$F_1 \cup F_2 = \{\omega \in \Omega : \omega \in F_1 \text{ or } \omega \in F_2\}$$

- The intersection of two events F_1 and F_2, written $F_1 \cap F_2$ occurs if and only both F_1 and F_2 occur. Equivalently, $\omega \in F_1 \cap F_2$ if and only if $\omega \in F_1$ and $\omega \in F_2$,

$$F_1 \cap F_2 = \{\omega \in \Omega : \omega \in F_1 \text{ and } \omega \in F_2\}$$

- Unions and intersections of several events, $F_1 \cup \ldots \cup F_n$ and $F_1 \cap \ldots \cap F_n$ are defined iteratively from the definition for unions and intersections of pairs.

Algebra of Events

- The complement of an event F, denoted F^c, contains all the elements of Ω that are not contained in F,
$$F^c = \{\omega \in \Omega : \omega \notin F\}.$$
- Two events F_1 and F_2 are called disjoint if the contain no common elements, that is $F_1 \cap F_2 = \emptyset$.
- A partition $\{F_n\}_{n\geq 1}$ of Ω 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 of two events F_1 and F_2 is defined as $F_1 \setminus F_2 = F_1 \cap F_2^c$. It contains all the elements of F_1 that are not contained in F_2. Notice that the difference is not symmetric: $F_1 \setminus F_2 \neq F_2 \setminus F_1$.
- It can be checked that the following properties hold true

(i) $(F_1 \cup F_2) \cup F_3 = F_1 \cup (F_2 \cup F_3) = F_1 \cup F_2 \cup F_3$
(ii) $(F_1 \cap F_2) \cap F_3 = F_1 \cap (F_2 \cap F_3) = F_1 \cap F_2 \cap F_3$
(iii) $F_1 \cap (F_2 \cup F_3) = (F_1 \cap F_2) \cup (F_1 \cap F_3)$
(iv) $F_1 \cup (F_2 \cap F_3) = (F_1 \cup F_2) \cap (F_1 \cup F_3)$
(v) $(F_1 \cup F_2)^c = F_1^c \cap F_2^c$ and $(F_1 \cap F_2)^c = F_1^c \cup F_2^c$

Probability Measures

Probability measure $\mathbb{P}$: real function defined over the events of Ω, assigning a probability to any event.

- Interpreted as a measure the long-run relative frequency of the event.
- Interpreted as a measure of how certain we are that the event will occur.

Postulated to satisfy the following properties:

1. $\mathbb{P}(F) \geq 0$, for all events F.
2. $\mathbb{P}(\Omega) = 1$.
3. If $\{F_n\}_{n \geq 1}$ are disjoint events, and $F = \cup_{n \geq 1} F_n$ is an event given by their union, then

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

Probability Measures

The following properties are immediate consequences of the probability axioms:

- $\mathbb{P}(F^c) = 1 - \mathbb{P}(F)$.
- $\mathbb{P}(F_1 \cap F_2) \leq \min\{\mathbb{P}(F_1), \mathbb{P}(F_2)\}$.
- $\mathbb{P}(F_1 \cup F_2) = \mathbb{P}(F_1) + \mathbb{P}(F_2) - \mathbb{P}(F_1 \cap F_2)$.
- Continuity from below: let $\{F_n\}_{n\geq 1}$ be nested events, such that $F_j \subseteq F_{j+1}$ for all j, and let F be an event given by $F = \cup_{n\geq 1} F_n$. Then $\mathbb{P}(F_n) \overset{n\to\infty}{\longrightarrow} \mathbb{P}(F)$.
- Continuity from above: let $\{F_n\}_{n\geq 1}$ be nested events, such that $F_j \supseteq F_{j+1}$ for all j, and let F be an event given by $F = \cap_{n\geq 1} F_n$. Then $\mathbb{P}(F_n) \overset{n\to\infty}{\longrightarrow} \mathbb{P}(F)$.
- If $\Omega = \{\omega_1, ..., \omega_K\}$, $K < \infty$, is a finite set, then for any event $F \subseteq \Omega$, we have $\mathbb{P}(F) = \sum_{j:\omega_j \in F} \mathbb{P}(\omega_j)$.

Conditional Probability and Independence

Suppose we don't know the precise outcome $\omega \in \Omega$ that has occurred, but we are told that $\omega \in F_2$ for some event F_2, and are asked to now calculate the probability that $\omega \in F_1$ also, for some other event F_1, we need conditional probability.

- For any pair of events F_1, F_2 such that $\mathbb{P}(F_2) > 0$, we define the conditional probability of F_1 given F_2 to be

$$\mathbb{P}(F_1|F_2) = \frac{\mathbb{P}(F_1 \cap F_2)}{\mathbb{P}(F_2)}.$$

- Let G be an event and $\{F_n\}_{n\geq 1}$ be a partition of Ω such that $\mathbb{P}(F_n) > 0$ for all n. We then have:
  - Law of total probability: $\mathbb{P}(G) = \sum_{n=1}^{\infty} \mathbb{P}(G|F_n)\mathbb{P}(F_n)$
  - Bayes' theorem: $\mathbb{P}(F_j|G) = \frac{\mathbb{P}(F_j \cap G)}{\mathbb{P}(G)} = \frac{\mathbb{P}(G|F_j)\mathbb{P}(F_j)}{\sum_{n=1}^{\infty} \mathbb{P}(G|F_n)\mathbb{P}(F_n)}$
- The events $\{G_n\}_{n\geq 1}$ are called independent if and only if for any finite sub-collection $\{G_{i_1}, \ldots, G_{i_K}\}$, $K < \infty$, we have:

$$\mathbb{P}(G_{i_1} \cap \cdots \cap G_{i_K}) = \mathbb{P}(G_{i_1}) \times \mathbb{P}(G_{i_2}) \times \ldots \times \mathbb{P}(G_{i_K})$$

Random Variables and Distribution Functions

Random variables: numerical summaries of the outcome of a random experiment.

They allow us to not worry too much about precise structure of outcome $\omega \in \Omega$

We can concentrate on range of a random variable, rather than consider 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

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

- 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). This is defined as

$$F_X(x) = \mathbb{P}(X \leq x).$$

Random Variables and Distribution Functions

- 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) $\lim_{y\downarrow x} F_X(y) = F_X(x)$, that is, F_X is right-continuous.
  - (iv) $\lim_{y\uparrow x} F_X(y)$ exists, that is, F_X is left-limited.
  - (v) $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$.
  - (vi) $\mathbb{P}(X > a) = 1 - F(a)$.
  - (vii) Let $D_X := \{x \in \mathbb{R} : F_X(x) - \lim_{y\uparrow x} F_X(y) > 0\}$ be the set of points where F_X is not continuous.
    - D_X is a countable set.
    - If $\mathbb{P}(\{X \in D_F\}) = 1$ then X is called a *discrete* random variable (equivalently, X has a finite or countable range, with probability 1).
    - If $D_X = \emptyset$ then X is called a *continuous* random variable (the distribution function F_X is continuous).
    - It may very well happen that a random variable may be neither discrete nor continuous.

Probability Mass Functions

The probability mas function (or frequency function) $f_X : \mathbb{R} \to [0, 1]$ of a discrete random variable X is defined as

$$f_X(x) = \mathbb{P}(X = x).$$

By its definition, a probability mass function satisfies

(i) $\mathbb{P}(X \in A) = \sum_{t \in A \cap \mathcal{X}} f_X(t)$, for $A \subseteq \mathbb{R}$ 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 $\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}$.

Probability Density Functions

A continuous random variable X has probability density function $f_X : \mathbb{R} \to [0, +\infty)$ if

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

for all real numbers $a < b$. By its definition, a probability density satisfies

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

(ii) $f_X(x) = F_X'(x)$, whenever f_X is continuous at x.

(iii) Note that $f_X(x) \neq \mathbb{P}(X = x) = 0$. In fact, it can be $f(x) > 1$ for some x. It can even happen that f is unbounded.

Random Vectors and Joint Distributions

A random vector $\mathbf{X} = (X_1, \ldots, X_d)^\top$ is a finite collection of random variables (arranged as the coordinates of a vector)

We want to make probabilistic statements on the joint behaviour of all variables.

- The joint distribution function of a random vector $\mathbf{X} = (X_1, \ldots, X_d)^\top$ is defined as:

$$F_{\mathbf{X}}(x_1, \ldots, x_d) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d).$$

- Correspondingly, one defines the
  - joint frequency function, if the $\{X_i\}_{i=1}^d$ are all discrete,

$$f_{\mathbf{X}}(x_1, \ldots, x_d) = \mathbb{P}(X_1 = x_1, \ldots, X_d = x_d).$$

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

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

    In this case, when $f_{\mathbf{X}}$ is continuous at the point $\mathbf{x}$,

$$f_{\mathbf{X}}(x_1, \ldots, x_d) = \frac{\partial^d}{\partial x_1 \ldots \partial x_d} F_{\mathbf{X}}(x_1, \ldots, x_d)$$

Marginal Distributions

Given the joint distribution of the random vector $\mathbf{X} = (X_1, \ldots, X_d)^\top$, we can isolate the distribution of a single coordinate, say X_i.

- discrete case, the marginal frequency function of X_i is given by

$$f_{X_i}(x_i) = \mathbb{P}(X_i = x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_d} f_{\mathbf{X}}(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)$$

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

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

- More generally, we can define the joint frequency/density of a random vector formed by a subset of the coordinates of $\mathbf{X} = (X_1, \ldots, X_d)^\top$, say the first k
  - Discrete case:
    $f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \sum_{x_{k+1}} \cdots \sum_{x_d} f_{\mathbf{X}}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_d)$.
  - Continuous case
    $f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) =$
    $\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_{\mathbf{X}}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_d) dx_{k+1} \ldots dx_d$.
- I.e. to marginalise we integrate/sum out the remaining random variables from the overall joint density/frequency.
- Marginals do not uniquely determine the joint distribution.

Conditional Distributions

We may wish to make probabilistic statements about the potential outcomes of one random variable, if we already know the outcome of another.

For this we need the notion of a conditional density/frequency function.

If $(X_1, ..., X_d)$ is a continuous/discrete random vector, we define the conditional probability density/frequency function of $(X_1, ..., X_k)$ given $\{X_{k+1} = x_{k+1}, ..., X_d = x_d\}$ as

$$f_{X_1,...,X_k|X_{k+1},...,X_d}(x_1, ..., x_k|x_{k+1}, ..., x_d) = \frac{f_{X_1,...,X_d}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_d)}{f_{X_{k+1},...,X_d}(x_{k+1}, ..., x_d)}$$

provided that $f_{X_{k+1},...,X_d}(x_{k+1}, ..., x_d) > 0$.

Independent Random Variables

The random variables $X_1, \ldots, X_d$ are called independent, denoted if and only if, for all $x_1, \ldots, x_d \in \mathbb{R}$

$$F_{X_1,\ldots,X_d}(x_1, \ldots, x_d) = F_{X_1}(x_1) \times \ldots \times F_{X_d}(x_d).$$

Equivalently, $X_1, \ldots, X_d$ are independent if and only if, for all $x_1, \ldots, x_d \in \mathbb{R}$

$$f_{X_1,\ldots,X_d}(x_1, \ldots, x_d) = f_{X_1}(x_1) \times \ldots \times f_{X_d}(x_d).$$

Note that when random variables are independent, conditional distributions reduce to the corresponding marginal distributions.

When they are independent, knowing the value of one of the random variables gives us no information about the distribution of the rest.

Conditionally Independent Random Variables

The random vector X in $\mathbb{R}^d$ is called conditionally independent of the random vector Y given the random vector Z, written

$$X \perp\!\!\!\perp_Z Y \quad \text{or} \quad X \perp\!\!\!\perp Y \mid Z,$$

if and only if, for all $x_1, \ldots, x_d \in \mathbb{R}$

$$F_{X_1,\ldots,X_d|Y,Z}(x_1, \ldots, x_d) = F_{X_1,\ldots,X_d|Z}(x_1, \ldots, x_d).$$

Equivalently, if and only if, for all $x_1, \ldots, x_d \in \mathbb{R}$

$$f_{X_1,\ldots,X_d|Y,Z}(x_1, \ldots, x_d) = f_{X_1,\ldots,X_d|Z}(x_1, \ldots, x_d).$$

Knowing Y in addition to knowing Z gives us no more information about X.

Consequence: if X is conditionally independent of Y given Z, then

$$F_{X,Y|Z} = F_{X|Y,Z} F_{Y|Z} = F_{X|Z} F_{Y|Z}$$

Consequence: $X \perp\!\!\!\perp_Z Y \iff Y \perp\!\!\!\perp_Z X$

Expectation

The expectation (or expected value) of a random variable X formalises the notion of the "average" value taken by that random variable.

- For continuous variables:

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x\, f_X(x) dx.$$

- For discrete variables:

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x\, f_X(x), \qquad \mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}.$$

The expectation satisfies the following properties:

- Linearity: $\mathbb{E}[X_1 + \alpha X_2] = \mathbb{E}[X_1] + \alpha \mathbb{E}[X_2]$.
- $\mathbb{E}[h(x)] = \sum_{x \in \mathcal{X}} h(x) f_X(x)$ (discrete case)
  or
  $\mathbb{E}[h(x)] = \int_{-\infty}^{+\infty} h(x) f(x) dx$ (continuous case).

Variance, Covariance, Correlation

The variance of a random variable X expresses how disperse the realisations of X are around its expectation.

$$\mathrm{var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] \qquad (\text{if } \mathbb{E}[X^2] < \infty).$$

Furthermore, the covariance of a random variable X_1 with another random variable X_2 expresses the degree of linear dependency between the two.

$$\mathrm{cov}(X_1, X_2) = \mathbb{E}\left[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))\right] \qquad (\text{if } \mathbb{E}[X_i^2] < \infty).$$

The correlation between X_1 and X_2 is defined as

$$\mathrm{cor}(X_1, X_2) = \frac{\mathrm{cov}(X_1, X_2)}{\sqrt{\mathrm{var}(X_1)\mathrm{var}(X_2)}}.$$

It also expresses the degree of linear dependency. Its advantage is that it is invariant to changes of units of measurement, and moreover can be be understood in absolute terms (it belongs to ranges in $[-1, 1]$), as a result of the correlation inequality (itself a consequence of the Cauchy-Schwarz inequality):

$$|\mathrm{corr}(X_1, X_2)| \leq \sqrt{\mathrm{var}(X_1)\mathrm{var}(X_2)}.$$

Variance, Covariance, Correlation

Some useful formulas relating expectations, variance, and covariances are:

- $\mathrm{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathrm{cov}(X, X)$
- $\mathrm{var}(aX + b) = a^2 \mathrm{var}(X)$
- $\mathrm{var}(\sum_i X_i) = \sum_i \mathrm{var}(X_i) + \sum_{i \neq j} \mathrm{cov}(X_i, X_j)$
- $\mathrm{cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2]$
- $\mathrm{cov}(aX_1 + bX_2, Y) = a \cdot \mathrm{cov}(X_1, Y) + b \cdot \mathrm{cov}(X_2, Y)$
- if $\mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] < \infty$, then the following are equivalent:
  - (i) $\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2]$
  - (ii) $\mathrm{cov}(X_1, X_2) = 0$
  - (iii) $\mathrm{var}(X_1 \pm X_2) = \mathrm{var}(X_1) + \mathrm{var}(X_2)$

  Independence will imply these three last properties, but none of these properties imply independence.

Conditional expectation and variance

- $\mathbf{X} = (\mathbf{Y}^\top, \mathbf{Z}^\top)^\top$
- conditional expectation of $S = S(\mathbf{X}) = S(\mathbf{Y}, \mathbf{Z})$ given $\mathbf{Z} = \mathbf{z}$

$$\mathbb{E}[S|\mathbf{Z} = \mathbf{z}] = \int_{\mathbb{R}^{d-r}} S(\mathbf{y}, \mathbf{z}) f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z}) d\mu_{\mathbf{Y}}(\mathbf{y})$$

- $\mathbb{E}[S|\mathbf{Z}]$ is a random variable (a function of $\mathbf{Z}$)
- $\mathbb{E}\{\mathbb{E}[S|\mathbf{Z}]\} = \mathbb{E}[S]$ (expectation of conditional expectation is marginal expectation)
- $\mathbb{E}[g(\mathbf{Z})S|\mathbf{Z}] = g(\mathbf{Z})\mathbb{E}[S|\mathbf{Z}]$ (taking out what is known)
- $\mathbb{E}\{\mathbb{E}[S|\mathbf{Z}]\,|g(\mathbf{Z})\} = \mathbb{E}[S|g(\mathbf{Z})]$ (tower property)
- If S is independent of $\mathbf{Z}$, then $\mathbb{E}[S|\mathbf{Z}] = \mathbb{E}[S]$ (independence)
  - More generally, if $\mathbf{W}$ is independent of both S and $\mathbf{Z}$, then

$$\mathbb{E}[S|(\mathbf{W}^\top, \mathbf{Z}^\top)^\top] = \mathbb{E}[S|\mathbf{Z}].$$

- conditional variance $\mathrm{var}[S|\mathbf{Z}] = \mathbb{E}[(S - \mathbb{E}[S|\mathbf{Z}])^2|\mathbf{Z}]$
- $\mathrm{var}S = \mathrm{var}(\mathbb{E}[S|\mathbf{Z}]) + \mathbb{E}(\mathrm{var}[S|\mathbf{Z}])$
- **general definition**: $\mathbb{E}[X|Z]$ is a function of $\mathbf{Z}$ satisfying that $\mathbb{E}\{\mathbf{1}_{\{\mathbf{Z}\in A\}}\mathbb{E}[X|\mathbf{Z}]\} = \mathbb{E}\{\mathbf{1}_{\{\mathbf{Z}\in A\}}X\}$ for all Borel A.

Some Important Inequalities

Given X be a non-negative random variable. Then, given any $\epsilon > 0$,

$$\mathbb{P}[X \geq \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon} \qquad \text{[Markov]}$$

Let X be a random variable with finite mean $\mathbb{E}[X] < \infty$. Then, given any $\epsilon > 0$,

$$\mathbb{P}\Big[|X - \mathbb{E}[X]| \geq \epsilon\Big] \leq \frac{\mathrm{Var}[X]}{\epsilon^2} \qquad \text{[Chebyschev]}$$

For any convex[1] function $\varphi : \mathbb{R} \to \mathbb{R}$, if $\mathbb{E}|\varphi(X)| + \mathbb{E}|X| < \infty$, then

$$\varphi\Big(\mathbb{E}[X]\Big) \leq \mathbb{E}[\varphi(X)] \qquad \text{[Jensen]}$$

Let X be a real random variable with $\mathbb{E}[X^2] < \infty$. Let $g : \mathbb{R} \to \mathbb{R}$ be a non decreasing function such that $\mathbb{E}[g^2(X)] < \infty$. Then,

$$\mathrm{Cov}[X, g(X)] \geq 0 \qquad \text{[Monotonicity and Covariance]}$$

[1] Recall that a function φ is convex if $\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y)$ for all x, y, and $\lambda \in [0, 1]$.

Moment Generating Functions

Let X be a random variable taking values in $\mathbb{R}$. The moment generating function (MGF) of X is defined as

$$M_X(t) : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$$

$$M_X(t) = \mathbb{E}\left[e^{tX}\right], \qquad t \in \mathbb{R}.$$

When $M_X(t), M_Y(t)$ exist (are finite) for $t \in I \ni 0$, then:

- $\mathbb{E}[|X|^k] < \infty$ and $\mathbb{E}[X^k] = \frac{d^k M_X}{dt^k}(0)$, for all $k \in \mathbb{N}$.
- $M_X = M_Y$ on I if and only if $F_X = F_Y$
- $M_{X+Y} = M_X M_Y$

Similarly, for a random vector X in $\mathbb{R}^d$, we define the MGF (with analogous properties)

$$M_X(u) : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$$

$$M_X(u) = \mathbb{E}\left[e^{u^\top X}\right], \qquad u \in \mathbb{R}^d.$$

Bernoulli Distribution

A random variable X is said to follow the Bernoulli distribution with parameter $p \in (0, 1)$, denoted $X \sim \text{Bern}(p)$, if

1. $\mathcal{X} = \{0, 1\}$,
2. $f(x; p) = p\mathbf{1}\{x = 1\} + (1 - p)\mathbf{1}\{x = 0\}$.

The mean, variance and moment generating function of $X \sim \text{Bern}(p)$ are given by

$$\mathbb{E}[X] = p, \qquad \text{Var}[X] = p(1 - p), \qquad M(t) = 1 - p + pe^t.$$

Binomial Distribution

A random variable X is said to follow the Binomial distribution with parameters $p \in (0,1)$ and $n \in \mathbb{N}$, denoted $X \sim \text{Binom}(n,p)$, if

1. $\mathcal{X} = \{0,1,2,...,n\}$,
2. $f(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x}$.

The mean, variance and moment generating function of $X \sim \text{Binom}(n,p)$ are given by

$$\mathbb{E}[X] = np, \qquad \text{Var}[X] = np(1-p), \qquad M(t) = (1-p+pe^t)^n.$$

- If $X = \sum_{i=1}^{n} Y_i$ where $Y_i \overset{iid}{\sim} Bern(p)$, then $X \sim \text{Binom}(n,p)$.

Geometric Distribution

A random variable X is said to follow the Geometric distribution with parameter $p \in (0, 1)$, denoted $X \sim \text{Geom}(p)$, if

1. $\mathcal{X} = \{0\} \cup \mathbb{N}$,
2. $f(x; p) = (1 - p)^x p$.

The mean, variance and moment generating function of $X \sim \text{Geom}(p)$ are given by

$$\mathbb{E}[X] = \frac{1-p}{p}, \qquad \text{Var}[X] = \frac{(1-p)}{p^2}, \qquad M(t) = \frac{p}{1-(1-p)e^t},$$

the latter for $t < -\log(1-p)$.

- Let $\{Y_i\}_{i \geq 1}$ be an infinite collection of random variables, where $Y_i \overset{iid}{\sim} Bern(p)$. Let $T = \min\{k \in \mathbb{N} : Y_k = 1\} - 1$. Then $T \sim Geom(p)$.

Negative Binomial Distribution

A random variable X is said to follow the Negative Binomial distribution with parameters $p \in (0,1)$ and $r > 0$, denoted $X \sim \text{NegBin}(r,p)$, if

1. $\mathcal{X} = \{0\} \cup \mathbb{N}$,
2. $f(x;p,r) = \binom{x+r-1}{x}(1-p)^x p^r$.

The mean, variance and moment generating function of $X \sim \text{NegBin}(r,p)$ are given by

$$\mathbb{E}[X] = r\frac{1-p}{p}, \qquad \text{Var}[X] = r\frac{(1-p)}{p^2}, \qquad M(t) = \frac{p^r}{[1-(1-p)e^t]^r},$$

the latter for $t < -\log(1-p)$.

- If $X = \sum_{i=1}^{r} Y_i$ where $Y_i \overset{iid}{\sim} Geom(p)$, then $X \sim \text{NegBin}(r,p)$.

Poisson Distribution

A random variable X is said to follow the Poisson distribution with parameters $\lambda > 0$, denoted $X \sim \text{Poisson}(\lambda)$, if

1. $\mathcal{X} = \{0\} \cup \mathbb{N}$,
2. $f(x; \lambda) = e^{-\lambda} \dfrac{\lambda^x}{x!}$.

The mean, variance and moment generating function of $X \sim \text{Poisson}(\lambda)$ are given by

$$\mathbb{E}[X] = \lambda, \qquad \text{Var}[X] = \lambda, \qquad M(t) = \exp\{\lambda(e^t - 1)\}.$$

- Let $\{X_n\}_{n \geq 1}$ be a sequence of $Binom(n, p_n)$ random variables, such that $p_n = \lambda/n$, for some constant $\lambda > 0$. Then $f_{X_n} \overset{n \to \infty}{\longrightarrow} f_Y$, where $Y \sim Poisson(\lambda)$.
- Let $X \sim Poisson(\lambda)$ and $Y \sim Poisson(\mu)$ be independent. The conditional distribution of X given $X + Y = k$ is $Binom(k, \lambda/(\lambda + \mu))$ (useful in contingency tables).

Uniform Distribution

A random variable X is said to follow the uniform distribution with parameters $-\infty < \theta_1 < \theta_2 < \infty$, denoted $X \sim \text{Unif}(\theta_1, \theta_2)$, if

$$f_X(x;\theta) = \begin{cases} (\theta_2 - \theta_1)^{-1} & \text{if } x \in (\theta_1, \theta_2), \\ 0 & \text{otherwise.} \end{cases}$$

The mean, variance and moment generating function of $X \sim \text{Unif}(\theta_1, \theta_2)$ are given by

$$\mathbb{E}[X] = (\theta_1 + \theta_2)/2, \qquad \text{Var}[X] = (\theta_2 - \theta_1)^2/12$$

$$M(t) = \frac{e^{t\theta_2} - e^{t\theta_1}}{t(\theta_2 - \theta_1)}, \; t \neq 0, \qquad M(0) = 1.$$

Exponential Distribution

A random variable X is said to follow the exponential distribution with parameter $\lambda > 0$, denoted $X \sim \text{Exp}(\lambda)$, if

$$f_X(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

The mean, variance and moment generating function of $X \sim \text{Exp}(\lambda)$ are given by

$$\mathbb{E}[X] = \lambda^{-1}, \qquad \text{Var}[X] = \lambda^{-2}, \qquad M(t) = \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$

If X, Y are independent exponential random variables with rates λ_1 and λ_2, then $Z = \min\{X, Y\}$ is also exponential with rate $\lambda_1 + \lambda_2$.

Lack of memory characterisation:

1. Let $X \sim Exp(\lambda)$. Then $\mathbb{P}[X \geq x + t | X \geq t] = \mathbb{P}[X \geq x]$.
2. Conversely: if X is a random variable such that $\mathbb{P}(X > 0) > 0$ and

$$\mathbb{P}(X > t + s | X > t) = \mathbb{P}(X > s), \qquad \forall t, s \geq 0,$$

Gamma Distribution

A random variable X is said to follow the gamma distribution with parameters $r > 0$ and $\lambda > 0$ (the *shape* and *scale* parameters, respectively), denoted $X \sim \text{Gamma}(r, \lambda)$, if

$$f_X(x; r, \lambda) = \begin{cases} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

The mean, variance and moment generating function of $X \sim \text{Gamma}(r, \lambda)$ are given by

$$\mathbb{E}[X] = r/\lambda, \qquad \text{Var}[X] = r/\lambda^2, \qquad M(t) = \left(\frac{\lambda}{\lambda - t}\right)^r, \quad t < \lambda.$$

- If $X_1, \ldots, X_r \overset{iid}{\sim} Exp(\lambda)$, then $Y = \sum_{i=1}^{r} X_i \sim Gamma(r, \lambda)$

Normal (Gaussian) Distribution

A random variable X is said to follow the normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ (the *mean* and *variance* parameters, respectively), denoted $X \sim \mathrm{N}(\mu, \sigma^2)$, if

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad x \in \mathbb{R}.$$

The mean, variance and moment generating function of $X \sim \mathrm{N}(\mu, \sigma^2)$ are given by

$$\mathbb{E}[X] = \mu, \qquad \mathrm{Var}[X] = \sigma^2, \qquad M(t) = \exp\{t\mu + t^2\sigma^2/2\}.$$

In the special case $Z \sim N(0, 1)$, we use the notation $\varphi(z) = f_Z(z)$ and $\Phi(z) = F_Z(z)$, and call these the *standard normal density* and *standard normal CDF*, respectively.

Standardisation

Lemma

Let $X_1, \ldots, X_n$ independent random variables such that $X_i \sim N(\mu_i, \sigma_i^2)$, and let $S_n = \sum_{i=1}^n X_i$. Then,

$$S_n \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$

Lemma

$X \sim N(\mu, \sigma^2)$ if and only if there exists $Z \sim N(0,1)$ such that $X = \sigma Z + \mu$

Consequently, if $X \sim N(\mu, \sigma^2)$, then

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

where Φ is the standard normal CDF,

$$\Phi(u) = \int_{-\infty}^{u} (2\pi)^{-1/2} \exp\{-z^2/2\} dz,$$

that is, the distribution function of $Z \sim N(0,1)$.

Gaussian Sampling

Theorem (Gaussian Sampling)

Let $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, *and define*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \& \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Then

1. *The joint distribution of* $X_1, ..., X_n$ *has probability density function,*

$$f_{X_1,...,X_n}(x_1, ..., x_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}.$$

2. *The sample mean is distributed as* $\bar{X} \sim N(\mu, \sigma^2/n)$.
3. *The random variables* $\bar{X}$ *and* S^2 *are independent.*
4. *The random variable* S^2 *satisfies* $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$.

Sampling Distributions: Chi-square Distribution

A random variable X is said to follow the chi-square distribution with parameter $k \in \mathbb{N}$ (called the number of degrees of freedom), denoted $X \sim \chi^2_k$, if it holds that $X \sim Gamma(k/2, 1/2)$. In other words,

$$f_X(x; k) = \begin{cases} \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

The mean, variance and moment generating function of $X \sim \chi^2_k$ are given by

$$\mathbb{E}[X] = k, \qquad \mathrm{Var}[X] = 2k, \qquad M(t) = (1-2t)^{-k/2}, \quad t < \frac{1}{2}.$$

Theorem

Let $\{Z_1, ..., Z_k\}$ be i.i.d. $N(0,1)$ random variables. Then,

$$Z_1^2 + \ldots Z_k^2 \sim \chi^2_k.$$

Sampling Distributions: Student's t-distribution

A random variable X is said to follow Student's t distribution with parameter $k \in \mathbb{N}$ (called the number of degrees of freedom), denoted $X \sim \mathrm{t}_k$, if,

$$f_X(x;k) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}},$$

Assuming $k > 2$, the mean and variance of $X \sim \mathrm{t}_k$ are given by

$$\mathbb{E}[X] = 0, \qquad \mathrm{Var}[X] = \frac{k}{k-2}.$$

The mean is undefined for $k = 1$ and the variance is undefined for $k \leq 2$. The moment generating function is undefined for any $k \in \mathbb{N}$.

Theorem (Student's Statistic and its Sampling Distribution)

Let $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. *Then,* $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim \mathrm{t}_{n-1}$.

Sampling Distributions: Snedecor-Fisher F distribution

A random variable X is said to follow the F distribution with parameters $d_1 \in \mathbb{N}$ and $d_2 \in \mathbb{N}$, denoted $X \sim \mathrm{F}_{d_1,d_2}$, if

$$f_X(x; d_1, d_2) = \begin{cases} \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{d_1/2} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2} x\right)^{-\frac{d_1+d_2}{2}}, & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

The mean, variance of $X \sim \mathrm{F}_{d_1,d_2}$ are given by

$\mathbb{E}[X] = \frac{d_2}{d_2 - 2}$, provided $d_2 > 2$, $\mathrm{Var}[X] = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2 - 4)(d_2 - 2)^2}$ provided d_2

The moment generating function does not exist.

Theorem

Let $X_1 \sim \chi^2_{d_1}$ and $X_2 \sim \chi^2_{d_2}$ be independent random variables. Then,

$$\frac{X_1/d_1}{X_2/d_2} \sim \mathrm{F}_{d_1,d_2}.$$

Quantile Function and Quantiles

Given a probability $\alpha \in (0,1)$, which is the (smallest) real number x such that $\mathbb{P}[X \leq x] = \alpha$?
We need to **invert** the distribution function.

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

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

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

Given an $\alpha \in (0,1)$, we call the real number

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

the α-quantile of X (or, equivalently, of F_X).

Transformations of random vectors

- continuous $\mathbf{X} = (X_1, \ldots, X_d)^\top$ with density $f_{\mathbf{X}}$
- $h : \mathbb{R}^d \to \mathbb{R}^d$
- $\mathbf{Y} = h(\mathbf{X}) = h(X_1, \ldots, X_d)$
- $\mathbb{P}(\mathbf{X} \in A) = 1$ for some open set $A \subset \mathbb{R}^d$
- $h : A \to h(A)$ is one-to-one, has continuous partial derivatives and $|\mathbf{J}_h(\mathbf{x})| \neq 0$ for all $\mathbf{x} \in A$
- then the density of $\mathbf{Y} = h(X_1, \ldots, X_d)$ is

$$f_{\mathbf{Y}}(\mathbf{y}) = \begin{cases} f_{\mathbf{X}}(h^{-1}(\mathbf{y})) \frac{1}{|\mathbf{J}_h(h^{-1}(\mathbf{y}))|^{-1}} \\ \quad = f_{\mathbf{X}}(h^{-1}(\mathbf{y}))|\mathbf{J}_{h^{-1}}(\mathbf{y})|, & \mathbf{y} \in h(A), \\ 0 & \text{otherwise} \end{cases}$$

Elements of a Statistical Model

Back To Statistics: The Basic Setup

Elements of a Statistical Model:

- Have a random experiment with sample space Ω.
- $\mathbf{X} : \Omega \to \mathbb{R}^n$ is a random variable, $\mathbf{X} = (X_1, ..., X_n)$, defined on Ω
- When outcome of experiment is $\omega \in \Omega$, we observe $\mathbf{X}(\omega)$ and call it the *data* (usually ω omitted).
- Probability experiment of observing a realisation of $\mathbf{X}$ completely determined by distribution F of $\mathbf{X}$.
- F assumed to be member of family $\mathcal{F}$ of distributions on $\mathbb{R}^n$.

Goal

Learn about $F \in \mathcal{F}$ given data $\mathbf{X}$.

The Basic Setup: An Illustration

Example (Coin Tossing)

Consider the following probability space:

- $\Omega = [0,1]^n$ with elements $\omega = (\omega_1, ..., \omega_n) \in \Omega$
- $\mathcal{F}$ are Borel subsets of Ω (product σ-algebra)
- $\mathbb{P}$ is the uniform probability measure (Lebesge measure) on $[0,1]^n$

Now we can define the experiment of n coin tosses as follows:

- Let $\theta \in (0,1)$ be a constant
- For $i = 1, ..., n$ let $X_i = \mathbf{1}\{\omega_i > \theta\}$
- Let $\mathbf{X} = (X_1, ..., X_n)$, so that $\mathbf{X} : \Omega \to \{0,1\}^n$
- Then $F_{X_i}(\mathbf{x}_i) = \mathbb{P}[X_i \leq x_i] = \begin{cases} 0 & \text{if } x_i \in (-\infty, 0), \\ \theta & \text{if } x_i \in [0,1), \\ 1 & \text{if } x_i \in [1, +\infty). \end{cases}$
- And $F_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^n F_{X_i}(x_i)$

Parameters and Parametrizations

Describing Families of Distributions: Parametric Models

Definition (Parametrization)

Let Θ be a set, $\mathcal{F}$ be a family of distributions and $g : \Theta \to \mathcal{F}$ an onto mapping. The pair (Θ, g) is called a *parametrization* of $\mathcal{F}$.

Definition (Parametric Model)

A *parametric model* with parameter space $\Theta \subseteq \mathbb{R}^d$ is a family of probability models $\mathcal{F}$ parametrized by Θ, $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$.

Example (IID Normal Model)

$$\mathcal{F} = \left\{ \prod_{i=1}^{n} \int_{-\infty}^{x_i} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} dy_i : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+ \right\}$$

- When Θ is not Euclidean, we call $\mathcal{F}$ *non-parametric*
- When Θ is a product of a Euclidean and a non-Eucidean space, we call $\mathcal{F}$ *semi-parametric*

Parametric Models

Example (Geometric Distribution)

Let $X_1, ..., X_n$ be iid geometric(p) distributed: $\mathbb{P}[X_i = k] = p(1-p)^k$, $k \in \mathbb{N} \cup \{0\}$. Two possible parametrizations are:

1. $[0, 1] \ni p \mapsto$ geometric(p)
2. $[1, \infty) \ni \mu \mapsto$ geometric with mean μ

Example (Poisson Distribution)

Let $X_1, ..., X_n$ be Poisson(λ) distributed: $\mathbb{P}[X_i = k] = e^{-\lambda}\frac{\lambda^k}{k!}$, $k \in \mathbb{N} \cup \{0\}$. Three possible parametrizations are:

1. $[0, \infty) \ni \lambda \mapsto$ Poisson(λ)
2. $[0, \infty) \ni \mu \mapsto$ Poisson with mean μ
3. $[0, \infty) \ni \sigma^2 \mapsto$ Poisson with variance σ^2

Example (Non-Parametric Regression)

Let $t_i = iT/n$ and $C_0 \ni f : [0, T] \to \mathbb{R}$, and $\mathrm{var}epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Let,

$$Y_i = f(t_i) + \mathrm{var}epsilon_i.$$

Then,

$$(Y_1, ..., Y_n)^\top = \mathbf{Y} \sim \mathcal{N}_n \left((f(t_1), ..., f(t_n))^\top, \sigma^2 I_n \right)$$

and the parametrisation is

$$(f, \sigma^2) \mapsto \mathcal{N}_n \left((f(t_1), ..., f(t_n))^\top, \sigma^2 I_n \right)$$

Identifiability

- Parametrization often suggested from phenomenon we are modelling
- But any set Θ and surjection $g : \Theta \to \mathcal{F}$ give a parametrization.
- Many parametrizations possible! Is *any* parametrization sensible?

Definition (Identifiability)

A parametrization (Θ, g) of a family of models $\mathcal{F}$ is called *identifiable* if $g : \Theta \to \mathcal{F}$ is a bijection (i.e. if g is injective on top of being surjective).

When a parametrization is not identifiable:

- Have $\theta_1 \neq \theta_2$ but $F_{\theta_1} = F_{\theta_2}$.
- Even with ∞ amounts of data we could not distinguish θ_1 from θ_2.

Definition (Parameter)

A parameter is a function $\nu : F_\theta \to \mathcal{N}$, where $\mathcal{N}$ is arbitrary.

- A parameter is a *feature* of the distribution F_θ
- When $\theta \mapsto F_\theta$ is identifiable, then $\nu(F_\theta) = q(\theta)$ for some $q : \Theta \to \mathcal{N}$.

Identifiability

Example (Binomial Thinning)

Let $\{B_{i,j}\}$ be an infinite iid array of Bernoulli(ψ) variables and $\xi_1, ..., \xi_n$ be an iid sequence of geometric(p) random variables with probability mass function $\mathbb{P}[\xi_i = k] = p(1-p)^k, k \in \mathbb{N} \cup \{0\}$. Let $X_1, ..., X_n$ be iid random variables defined by

$$X_j = \sum_{i=1}^{\xi_j} B_{i,j}, \quad j = 1, .., n$$

Any $F_X \in \mathcal{F}$ is completely determined by (ψ, p), so $[0,1]^2 \ni (\psi, q) \mapsto F_X$ is a parametrization of $\mathcal{F}$. Can show (how?)

$$X \sim \text{geometric}\left(\frac{p}{\psi(1-p)+p}\right)$$

However (ψ, p) is not identifiable (why?).

Parametric Inference for Regular Models

Will focus on parametric families $\mathcal{F}$. The aspects we will wish to learn about will be *parameters* of $F \in \mathcal{F}$.

Regular Models

Assume from now on that in any parametric model we consider either:

1. All of the F_θ are continuous with densities $f(\mathbf{x}, \theta)$
2. All of the F_θ are discrete with frequency functions $p(\mathbf{x}, \theta)$ and there exists a countable set A that is independent of θ such that $\sum_{\mathbf{x} \in A} p(\mathbf{x}, \theta) = 1$ for all $\theta \in \Theta$.

Will be considering the mathematical aspects of problems such as:

1. Estimating which $\theta \in \Theta$ (i.e. which $F_\theta \in \mathcal{F}$) generated $\mathbf{X}$
2. Deciding whether some hypothesized values of θ are consistent with $\mathbf{X}$
3. The performance of methods and the existence of optimal methods
4. What happens when our model is wrong?