Statistical Theory (Week 1): Introduction

Erwan Koch

Ecole Polytechnique Fédérale de Lausanne (Institute of Mathematics)

EPFL

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)

- The variety of different forms of data is bewildering.
- Can we formulate a unified mathematical theory?

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/statistics and logic [Jaynes].
- One 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 main questions of statistics:

1. **Estimation**: adjusting the parameters of a model to fit data.
2. **Comparison**: of two/multiple models; which one is the best?
3. **Prediction**: can I predict new values of the data?

A Probabilist and a Statistician Flip a Coin

Example

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

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

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

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

Probabilist Asks:

- Probability of that outcome as a function of θ?
- Probability of a k-long run?
- If one keeps tossing, how many k-long runs? How long until a 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 $\boldsymbol{X}$?
- How much error do we make when trying to decide the above from $\boldsymbol{X}$?
- How does our answer change if $\boldsymbol{X}$ is perturbed?
- Is there a "best" solution to the above problems?
- How sensitive are our answers to departures from $X_i \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$?
- 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?

The three aspects of statistics

In order to do good statistics, we need to worry about the following three different problems:

- **Mathematical rigour**.
  Statisticians want to draw rigorous conclusions from a dataset. In order to do so, they must possess a perfect understanding of the probabilistic underpinnings of statistical analysis.
- **Correct modelling of the data**.
  In order to rigorously analyze a dataset, we need to formulate a **model** of how it was generated. This choice is extremely important and difficult. This is why mathematicians often do not like statistics.
- **Computational efficiency**.
  In order to be useful, a statistical analysis must run in a **short amount of time** on any standard computer. It must thus be:
  - Efficiently computable (P vs NP).
  - Correctly implemented.

The three themes of this course

In practice, this course focuses on three important topics:

- Giving a general framework for statistical inference: maximum-likelihood methods.
- Analyzing the behaviour of statistical methods when the number of data points tends to ∞: asymptotic results.
- Analyzing the efficiency of various approaches to statistics: is there an optimal way to do statistics?

Statistical Theory (MATH-442): Technicalities

- Course:
  - Tuesday, 08h15 – 10h00
  - Me
- Exercises:
  - Tuesday, 10h15 – 12h00
  - Leonardo Santoro, `leonardo.santoro@epfl.ch`
- All the material (course description, reference, slides, exercises and solutions) is on Moodle.
- Evaluation: **only a final exam** (only a non-programmable calculator will be allowed).

General advice

- Statistics is not extremely challenging from a mathematical point of view. It is challenging because of the **conceptual effort to match mathematics and reality**.
- Even though this is a theoretical course, you should try to work on the other two aspects of statistics:
  - Implement the methods of the course in simple examples.
  - We will briefly mention model choice here and there. Try to think about it on your own.
- Go to exercise sessions, it will help you a lot!
- Work in groups.
- Everyone in the class should ask **at least two questions** at each lecture.
  THERE IS NO SUCH THING AS A BAD QUESTION!!

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 (note that $F \in \mathcal{F}$). An event F "is realized" (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 they 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-valued function defined over the events of Ω, assigning a probability to any event.

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

Postulated to satisfy the following axioms:

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, 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 $\mathbb{P}(F_n) \stackrel{n\to\infty}{\longrightarrow} \mathbb{P}(F)$.
- 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 $\mathbb{P}(F_n) \stackrel{n\to\infty}{\longrightarrow} \mathbb{P}(F)$.
- If $\Omega = \{\omega_1, \ldots, \omega_K\}$, $K < \infty$, 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 (mutually) 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 the precise structure of the outcome $\omega \in \Omega$. We can concentrate on the range of a random variable rather than consider Ω.

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

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

Let $\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}$. By definition, we have

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

(ii) $F_X(x) = \sum_{t \in (-\infty, x] \cap \mathcal{X}} f_X(t)$, for all $x \in \mathbb{R}$.

(iii) An immediate corollary is that $F_X(x)$ is piecewise constant with jumps at the points in $\mathcal{X}$.

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.

- In the 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 marginalize 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 $\perp\!\!\!\perp$, 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 Vectors

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

$$\mathbf{X} \perp\!\!\!\perp_{\mathbf{Z}} \mathbf{Y} \quad \text{or} \quad \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z},$$

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

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

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

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

It means that knowing $\mathbf{Y}$ in addition to knowing $\mathbf{Z}$ does not give us more information about $\mathbf{X}$.

Consequence: if $\mathbf{X}$ is conditionally independent of $\mathbf{Y}$ given $\mathbf{Z}$, then

$$F_{\mathbf{X},\mathbf{Y}|\mathbf{Z}} = F_{\mathbf{X}|\mathbf{Y},\mathbf{Z}} F_{\mathbf{Y}|\mathbf{Z}} = F_{\mathbf{X}|\mathbf{Z}} F_{\mathbf{Y}|\mathbf{Z}}.$$

Consequence: $\mathbf{X} \perp\!\!\!\perp_{\mathbf{Z}} \mathbf{Y} \iff \mathbf{Y} \perp\!\!\!\perp_{\mathbf{Z}} \mathbf{X}$.

Expectation

The expectation (or expected value) of a random variable X formalizes 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
  $\mathbb{E}[h(X)] = \int_{-\infty}^{+\infty} h(x) f_X(x) dx$ (continuous case).

Variance, Covariance, Correlation

The variance of a random variable X expresses how scattered the realizations 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

$$\mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)\mathrm{Var}(X_2)}}.$$

It also expresses the degree of linear dependency. Its advantage is that it is invariant to changes of units of measurement, and moreover it can 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 implies the three last properties, but none of these properties implies independence.

Conditional expectation and variance

- Let s be a function from $\mathbb{R}^2$ to $\mathbb{R}$. The conditional expectation of $S = s(X, Y)$ given $Y = y$ is defined as

$$\mathbb{E}[S|Y = y] = \int_{\mathbb{R}} s(x, y) f_{X|Y}(x|y) dx.$$

- $\mathbb{E}[S|Y]$ is a random variable (a function of Y)!
- $\mathbb{E}\{\mathbb{E}[S|Y]\} = \mathbb{E}[S]$ (expectation of conditional expectation is marginal expectation).
- $\mathbb{E}[g(Y)S|Y] = g(Y)\mathbb{E}[S|Y]$ (taking out what is known).
- $\mathbb{E}\{\mathbb{E}[S|Y]\,|g(Y)\} = \mathbb{E}[S|g(Y)]$ (tower property).
- If S is independent of Y, then $\mathbb{E}[S|Y] = \mathbb{E}[S]$ (independence).
- If W is independent of both S and Y, then $\mathbb{E}[S|W, Y] = \mathbb{E}[S|Y]$.
- The conditional variance is defined by $\mathrm{Var}[S|Y] = \mathbb{E}[(S - \mathbb{E}[S|Y])^2|Y]$.
- $\mathrm{Var}(S) = \mathrm{Var}(\mathbb{E}[S|Y]) + \mathbb{E}(\mathrm{Var}[S|Y])$.
- General definition: $\mathbb{E}[X|Y]$ is a function of Y satisfying $\mathbb{E}\{\mathbf{1}_{\{Y \in A\}}\mathbb{E}[X|Y]\} = \mathbb{E}\{\mathbf{1}_{\{Y \in A\}}X\}$ for all Borel set A.

Some Important Inequalities

- Let X be a non-negative random variable with finite expectation. Then, for 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 first and second moments. Then, for any $\epsilon > 0$,
$$\mathbb{P}\Big[|X - \mathbb{E}[X]| \geq \epsilon\Big] \leq \frac{\mathrm{Var}[X]}{\epsilon^2} \qquad \text{[Chebyshev]}.$$
- 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 real-valued random variable. The moment generating function (MGF) of X is defined as

$$M_X : \begin{array}{ccc} \mathbb{R} & \mapsto & \mathbb{R} \cup \{\infty\} \\ t & \to & \mathbb{E}\left[e^{tX}\right]. \end{array}$$

- Let I be an open interval around 0. If $M_X(t), M_Y(t)$ exist (are finite) for any $t \in I$, 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 $\mathbf{X}$ in $\mathbb{R}^d$, we define the MGF (with analogous properties) by

$$M_{\mathbf{X}} : \begin{array}{ccc} \mathbb{R}^d & \mapsto & \mathbb{R} \cup \{\infty\} \\ \mathbf{u} & \to & \mathbb{E}\left[e^{\mathbf{u}^\top \mathbf{X}}\right]. \end{array}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and

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

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

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

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 $X \sim \text{Exp}(\lambda)$. 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

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

then there exits a $\lambda > 0$ such that $X \sim \text{Exp}(\lambda)$.

Gamma Distribution

A random variable X is said to follow the gamma distribution with parameters $r > 0$ and $\lambda > 0$ (the *shape* and *rate* 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

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

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

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

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

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.

Standardization

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 $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, *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* $X_1, \dots, X_n$ *has probability density function,*

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

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* $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$.

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 $X \sim \text{Gamma}(k/2, 1/2)$. 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 $Z_1, \dots, Z_k$ be iid $N(0,1)$ random variables. Then,

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

A random variable X is said to follow the Student t distribution with parameter $k \in \mathbb{N}$ (called the number of degrees of freedom), denoted $X \sim \mathrm{t}_k$, if

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

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: Fisher-Snedecor F distribution

A random variable X is said to follow the Fisher-Snedecor F distribution with parameters $d_1, d_2 \in \mathbb{N}$, 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

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

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 $\alpha \in (0,1)$ (so-called confidence interval), what 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. The quantile function of X is defined by

$$F_X^{-} : \begin{array}{ccc} (0,1) & \mapsto & \mathbb{R} \\ \alpha & \rightarrow & \inf\{t \in \mathbb{R} : F_X(t) \geq \alpha\}. \end{array}$$

- 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

- Let $\mathbf{X} = (X_1, \ldots, X_d)^\top$ be a continuous random vector with density $f_\mathbf{X}$.
- Let $h : \mathbb{R}^d \to \mathbb{R}^d$ and $\mathbf{Y} = h(\mathbf{X}) = h(X_1, \ldots, X_d)$.
- Assume that $\mathbb{P}(\mathbf{X} \in A) = 1$ for some open set $A \subset \mathbb{R}^d$
- Assume that $h : A \to h(A)$ 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 $\mathbf{Y}$ 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

Elements of a Statistical Model:

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

Goal

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

The Basic Setup: An Ilustration

Example (Coin Tossing)

Consider the following probability space:

- $\Omega = [0,1]^n$ with elements $\omega = (\omega_1, ..., \omega_n) \in \Omega$.
- $\mathcal{F}$ the set of Borel subsets of Ω (product σ-algebra).
- $\mathbb{P}$ is the uniform probability measure (Lebesgue 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 $\boldsymbol{X} = (X_1, ..., X_n)^\top$, so that $\boldsymbol{X} : \Omega \to \{0,1\}^n$.
- Then $F_{X_i}(\boldsymbol{x}_i) = \mathbb{P}[X_i \leq x_i] = \begin{cases} 0 & \text{if } x_i \in (-\infty, 0), \\ \theta & \text{if } x_i \in [0,1), \\ 1 & \text{if } x_i \in [1, +\infty). \end{cases}$
- And $F_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^n F_{X_i}(x_i)$.

Parameters and Parametrizations

Describing Families of Distributions: Parametric Models

Definition (Parametrization)

Let Θ be a set, $\mathcal{F}$ be a family of distributions and $g : \Theta \to \mathcal{F}$ a surjective 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-Euclidean space, we call $\mathcal{F}$ *semi-parametric*.

Parametric Models

Example (Geometric Distribution)

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

1. $[0,1] \ni p \mapsto \text{Geom}(p)$
2. $[1, \infty) \ni \mu \mapsto$ Geom with mean μ

Example (Poisson Distribution)

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

1. $[0, \infty) \ni \lambda \mapsto \text{Poisson}(\lambda)$
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)

For $i = 1, \dots, n$, let $t_i = iT/n$ and $C_0 \ni f : [0, T] \to \mathbb{R}$, and $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Let,

$$Y_i = f(t_i) + \varepsilon_i.$$

Then,

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

and the parametrization 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., g is injective on top of being surjective).

When a parametrization is not identifiable:

- We can have $\theta_1 \neq \theta_2$ but $F_{\theta_1} = F_{\theta_2}$.
- Even with an ∞ amount 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 Bern(ψ) variables and $\xi_1, ..., \xi_n$ be an iid sequence of Geom(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

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

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 $\mathcal{F}$. We can show (how?) that

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

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 are *parameters* of $F \in \mathcal{F}$.

Regular Models

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

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

We will consider the mathematical aspects of problems such as:

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

Statistical Theory: Explanation to slide 59 of week 1

Fall Semester 2020

Tomas Rubin, tomas.rubin@epfl.ch

During the exercise session I was asked about the example on slide 59 of week 1 (of the lecture) where it is claimed that the sum of random number of Bernoulli random variables (where the random number of summands is geometrically distributed) is again a geometric distribution with given parameter.

It can be shown by calculating the moment generating functions. By the answer to this Stackexchange question (and after adjusting the notation), we have

$$M_X(t) = M_\xi(\log M_B(t))$$

where

$$M_\xi(t) = \frac{p}{1-(1-p)e^t}, \qquad \text{(geometric distribution with the parameter } p\text{)},$$
$$M_b(t) = 1 - \psi + \psi e^t \qquad \text{(Bernoulli distribution with the parameter } \psi\text{)}.$$

Hence

$$M_X(t) = M_\xi(\log M_B(t)) = \frac{p}{1-(1-p)(1-\psi+\psi e^t)}$$

which can be manipulated to the form

$$M_X(t) = \frac{\frac{p}{\psi(1-p)+p}}{1 - \frac{\psi(1-p)}{\psi(1-p)+p} e^t}$$

where we recognise the moment generating function of the geometric distribution with the parameter $\frac{p}{\psi(1-p)+p}$.

Statistical Theory (Week 2): Overview of Stochastic Convergence

Erwan Koch

Ecole Polytechnique Fédérale de Lausanne (Institute of Mathematics)

Motivation: Functions of Random Variables

Functions of Random Variables

Let $X_1, \ldots, X_n$ be identically distributed with $\mathbb{E}[X_i] = \mu$ and $\text{var}[X_i] = \sigma^2$, and consider

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- If the X_i are independent and $X_i \sim \mathcal{N}(\mu, \sigma^2)$ or $X_i \sim \exp(\lambda = 1/\mu)$ then we know $\text{dist}[\bar{X}_n]$.
- But the X_i may be from some more general distribution.
- The joint distribution of X_i may not even be completely understood/known.

We would like to be able to say something about $\bar{X}_n$ even in those cases!

Perhaps this is not easy for fixed n, but what about letting $n \to \infty$?
$\hookrightarrow$(a very common approach in mathematics).

Functions of Random Variables

- Once we assume that $n \to \infty$ we start understanding dist$[\bar{X}_n]$ more:
  - At a crude level $\bar{X}_n$ becomes concentrated around μ:

$$\mathbb{P}[|\bar{X}_n - \mu| < \epsilon] \approx 1, \quad \forall\, \epsilon > 0, \text{ as } n \to \infty.$$

  - Perhaps more informative is to look at the "magnified difference":

$$\mathbb{P}[\sqrt{n}(\bar{X}_n - \mu) \leq x] \overset{n\to\infty}{\approx} \; ? \quad \to \quad \text{could yield } \mathbb{P}[\bar{X}_n \leq x].$$

- More generally $\longrightarrow$ We want to understand distribution of $Y_n = g(X_1, \dots, X_n)$ for some general g:
  - Often infeasible.
  - Thus, we resort to asymptotic approximations to understand the behaviour of Y_n.
- Such approximations are appropriate in many situations but be careful to the fact that asymptotics are often abused (used for n very small!).

Stochastic Convergence

Convergence of Random Variables

- Need to make precise what we mean by Y_n is "concentrated" around μ as $n \to \infty$.
- More generally what does "Y_n behaves like Y" for large n mean?
- $\text{dist}[g(X_1, ..., X_n)] \overset{n\to\infty}{\approx}$?

$\hookrightarrow$ We need appropriate notions of convergence for random variables.

Recall that random variables are *functions* between *measurable spaces*.

$\Longrightarrow$ Convergence of random variables can be defined in various ways:

- Convergence in probability (convergence in measure).
- Convergence in distribution (weak convergence).
- Convergence with probability 1 (almost sure convergence).
- Convergence in L^p (convergence in the p-th moment).

All these notions are qualitatively different. Some modes of convergence are stronger than others.

Convergence in Probability

Definition (Convergence in Probability)

Let $\{X_n\}_{n\geq 1}$ and X be random variables defined on the same probability space. We say that X_n converges in probability to X as $n \to \infty$ (and write $X_n \xrightarrow{p} X$) if for any $\epsilon > 0$,

$$\mathbb{P}[|X_n - X| > \epsilon] \xrightarrow{n\to\infty} 0.$$

Intuitively, if $X_n \xrightarrow{p} X$, then for large n, $X_n \approx X$ with probability close to 1.

Example

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0,1)$, and define $M_n = \max\{X_1, ..., X_n\}$. Then,

$$F_{M_n}(x) = x^n \implies \mathbb{P}[|M_n - 1| > \epsilon] = \mathbb{P}[M_n < 1 - \varepsilon]$$
$$= (1-\epsilon)^n \xrightarrow{n\to\infty} 0$$

for any $0 < \epsilon < 1$. Hence $M_n \xrightarrow{p} 1$.

Convergence in Probability

Lemma (Ky-Fan definition of convergence in probability)

$X_n \xrightarrow{p} X$ *if and only if there exists some sequence* $\alpha_n \downarrow 0$ *such that*

$$\mathbb{P}[|X_n - X| > \alpha_n] \leq \alpha_n, \qquad \forall\, n \geq 1.$$

Proof.

Suppose that there exists such an α_n. Then for any $\epsilon > 0$, there exists $N_\epsilon \in \mathbb{N}$ such that for all $n \geq N_\epsilon$, $\alpha_n < \epsilon$. It follows that, for any $n \geq N_\epsilon$,

$$\mathbb{P}[|X_n - X| > \epsilon] \leq \mathbb{P}[|X_n - X| > \alpha_n] \leq \alpha_n,$$

which gives $\mathbb{P}[|X_n - X| > \epsilon] \stackrel{n\to\infty}{\longrightarrow} 0$ since $\alpha_n \stackrel{n\to\infty}{\longrightarrow} 0$. For the converse, suppose that $X_n \xrightarrow{p} X$. Then, there exists $\{n_k\}_{k\geq 1}$ such that

$$n_k < n_{k+1}, \quad \& \quad \mathbb{P}[|X_n - X| > 1/k] \leq \frac{1}{k}, \forall\, n \geq n_k.$$

Define $\alpha_n = \sum_{k=1}^{\infty} \frac{1}{k} \mathbf{1}\{n_k \leq n < n_{k+1}\}$. We have $\mathbb{P}[|X_n - X| > \alpha_n] \leq \alpha_n$ for all $n \geq 1$ and $\alpha_n \downarrow 0$, which completes the proof. □

Convergence in Probability

Exercise

Knowledge of the sequence α_n can be used to characterize the speed at which the convergence occurs.
Indeed, if, for all n, $\alpha_n \geq \alpha'_n$ are two sequences controlling the convergence respectively of $X_n \xrightarrow{p} X$ and $X'_n \xrightarrow{p} X$, then the convergence of X'_n is faster than that of X_n.

Convergence in Distribution

Definition (Convergence in Distribution)

Let $\{X_n\}$ and X be random variables (not necessarily defined on the same probability space). We say that X_n converges in distribution to X as $n \to \infty$ (and write $X_n \xrightarrow{d} X$) if

$$\mathbb{P}[X_n \leq x] \xrightarrow{n\to\infty} \mathbb{P}[X \leq x],$$

at every continuity point of $F_X(x) = \mathbb{P}[X \leq x]$.

Example

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0,1)$, $M_n = \max\{X_1, ..., X_n\}$, and $Q_n = n(1 - M_n)$.

$$\mathbb{P}[Q_n \leq x] = \mathbb{P}[M_n \geq 1 - x/n] = 1 - \left(1 - \frac{x}{n}\right)^n \xrightarrow{n\to\infty} 1 - e^{-x}$$

for all $x \geq 0$. Hence $Q_n \xrightarrow{d} Q$, with $Q \sim \text{Exp}(1)$.

Some Comments on "$\xrightarrow{p}$" and "$\xrightarrow{d}$"

- "$\xrightarrow{p}$" involves the *random variables themselves*.
- "$\xrightarrow{d}$" relates their *distribution functions*.
  - ↪ Can be used to approximate distributions (approximation error?).
- Both notions of convergence are *metrizable*.
  - ↪ I.e., there exist metrics on the space of random variables and distribution functions that are compatible with these notions of convergence.
  - ↪ Hence can use things such as the triangle inequality, ...
- Convergence in probability implies convergence in distribution.
- Convergence in distribution does NOT imply convergence in probability.
  - ↪ E.g., if $X \sim \mathcal{N}(0,1)$, then $-X + \frac{1}{n} \xrightarrow{d} X$ but $-X + \frac{1}{n} \xrightarrow{p} -X$.
- "$\xrightarrow{d}$" is also known as "weak convergence".

Equivalent definition: $X_n \xrightarrow{d} X \iff \mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all continuous and bounded functions f.

Useful Theorems

Some Basic Results

Theorem

(a) $X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X.$

(b) *For any* $c \in \mathbb{R}$, $X_n \xrightarrow{d} c \implies X_n \xrightarrow{p} c.$

Proof

(a) Let x be a continuity point of F_X. Then, for any $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}[X_n \le x] &= \mathbb{P}[X_n \le x, |X_n - X| \le \epsilon] + \mathbb{P}[X_n \le x, |X_n - X| > \epsilon] \\ &\le \mathbb{P}[X \le x + \epsilon] + \mathbb{P}[|X_n - X| > \epsilon],\end{aligned}$$

using $\{X_n \le x, |X_n - X| \le \epsilon\} \subset \{X \le x + \epsilon\}$. Similarly,

$$\begin{aligned}\mathbb{P}[X \le x - \epsilon] &\le \mathbb{P}[X \le x - \epsilon, |X_n - X| \le \epsilon] + \mathbb{P}[X \le x - \epsilon, |X_n - X| > \epsilon] \\ &\le \mathbb{P}[X_n \le x] + \mathbb{P}[|X_n - X| > \epsilon],\end{aligned}$$

as $\{X \le x - \epsilon, |X_n - X| \le \epsilon\} \subset \{X_n \le x\}$.

(proof cont'd).

The previous inequality yields

$$\mathbb{P}[X \leq x - \epsilon] - \mathbb{P}[|X_n - X| > \epsilon] \leq \mathbb{P}[X_n \leq x].$$

Therefore,

$$\mathbb{P}[X \leq x - \epsilon] - \mathbb{P}[|X_n - X| > \epsilon] \leq \mathbb{P}[X_n \leq x] \leq \mathbb{P}[X \leq x + \epsilon] + \mathbb{P}[|X_n - X| > \epsilon].$$

Hence, letting n tend to infinity and then ϵ tend to 0 leads that $\mathbb{P}(X_n \leq x) \stackrel{n \to \infty}{\longrightarrow} \mathbb{P}(X \leq x)$.

(b) Let F be the distribution function of the degenerate random variable taking the single value c. We have

$$F(x) = \mathbb{P}[c \leq x] = \begin{cases} 1 & \text{if } x \geq c, \\ 0 & \text{if } x < c. \end{cases}$$

Now,

$$\begin{aligned} \mathbb{P}[|X_n - c| > \epsilon] &= \mathbb{P}[\{X_n - c > \epsilon\} \cup \{X_n < c - \epsilon\}] \\ &= \mathbb{P}[X_n > c + \epsilon] + \mathbb{P}[X_n < c - \epsilon] \\ &\leq 1 - \mathbb{P}[X_n \leq c + \epsilon] + \mathbb{P}[X_n \leq c - \epsilon] \\ &\stackrel{n \to \infty}{\longrightarrow} 1 - F(\underbrace{c + \epsilon}_{\geq c}) + F(\underbrace{c - \epsilon}_{< c}) = 0. \end{aligned}$$

Theorem (Continuous Mapping Theorem)

Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then,

(a) $X_n \xrightarrow{p} X \implies g(X_n) \xrightarrow{p} g(X)$.

(b) $Y_n \xrightarrow{d} Y \implies g(Y_n) \xrightarrow{d} g(Y)$.

Exercise

Prove part (a). You may assume without proof the *Subsequence Lemma*: $X_n \xrightarrow{p} X$ if and only if every subsequence X_{n_m} of X_n, has a further subsequence $X_{n_{m(k)}}$ such that $\mathbb{P}[X_{n_{m(k)}} \xrightarrow{k \to \infty} X] = 1$.

Theorem (Slutsky's Theorem)

Assume that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c \in \mathbb{R}$. Then

(a) $X_n + Y_n \xrightarrow{d} X + c$.

(b) $X_n Y_n \xrightarrow{d} cX$.

Proof of Slutsky's Theorem.

(a) We assume without loss of generality that $c = 0$. Let x be a continuity point of F_X. We have, for any $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}[X_n + Y_n \leq x] &= \mathbb{P}[X_n + Y_n \leq x, |Y_n| \leq \epsilon] + \mathbb{P}[X_n + Y_n \leq x, |Y_n| > \epsilon] \\ &\leq \mathbb{P}[X_n \leq x + \epsilon] + \mathbb{P}[|Y_n| > \epsilon],\end{aligned}$$

as $\{X_n + Y_n \leq x, |Y_n| \leq \epsilon\} \subset \{X_n \leq x + \epsilon\}$. Similarly,

$$\begin{aligned}\mathbb{P}[X_n \leq x - \epsilon] &= \mathbb{P}[X_n \leq x - \epsilon, |Y_n| \leq \epsilon] + \mathbb{P}[X_n \leq x - \epsilon, |Y_n| > \epsilon] \\ &\leq \mathbb{P}[X_n + Y_n \leq x] + \mathbb{P}[|Y_n| > \epsilon],\end{aligned}$$

since $\{X_n \leq x - \epsilon, |Y_n| \leq \epsilon\} \subset \{X_n + Y_n \leq x\}$. Therefore,

$$\mathbb{P}[X_n \leq x - \epsilon] - \mathbb{P}[|Y_n| > \epsilon] \leq \mathbb{P}[X_n + Y_n \leq x] \leq \mathbb{P}[X_n \leq x + \epsilon] + \mathbb{P}[|Y_n| > \epsilon].$$

Choosing ϵ such that $x - \epsilon$ and $x + \epsilon$ are continuity points of F_X and letting n tend to infinity, and then letting ϵ tend to 0 gives $\mathbb{P}(X_n + Y_n \leq x) \overset{n \to \infty}{\longrightarrow} \mathbb{P}(X \leq x)$. □

Proof of Slutsky's Theorem.

(b) We assume again without loss of generality that $c = 0$. Let $\epsilon, M > 0$:

$$\begin{aligned}\mathbb{P}[|X_n Y_n| > \epsilon] &= \mathbb{P}[|X_n Y_n| > \epsilon, |Y_n| \leq 1/M] + \mathbb{P}[|X_n Y_n| > \epsilon, |Y_n| > 1/M] \\ &\leq \mathbb{P}[|X_n| > \epsilon M] + \mathbb{P}[|Y_n| > 1/M] \\ &\overset{n\to\infty}{\longrightarrow} \mathbb{P}[|X| > \epsilon M] + 0.\end{aligned}$$

Choosing ϵ and M such that ϵM and $-\epsilon M$ are continuity points of F_X and letting n tend to infinity, and then letting M tend to infinity, leads $\mathbb{P}[|X_n Y_n| > \epsilon] \overset{n\to\infty}{\longrightarrow} 0$ for any $\epsilon > 0$, and thus the result. □

Theorem (General Version of Slutsky's Theorem)

Let $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be continuous and suppose that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c \in \mathbb{R}$. Then, $g(X_n, Y_n) \xrightarrow{d} g(X, c)$ as $n \to \infty$.

$\hookrightarrow$Notice that the general version of Slutsky's theorem does not follow immediately from the continuous mapping theorem.

- The multivariate version (see later) of the continuous mapping theorem would be applicable if (X_n, Y_n) weakly converged jointly in distribution (i.e., convergence of the joint distributions) to (X, c).
- But here we assume only marginal convergence (i.e., $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ separately, but their joint behaviour is unspecified).
- The key of the proof is that in the special case where $Y_n \xrightarrow{d} c$ where c is a constant, then marginal convergence $\iff$ joint convergence.
- However if $X_n \xrightarrow{d} X$ where X is non-degenerate, and $Y_n \xrightarrow{d} Y$ where Y is non-degenerate, then the theorem fails.
- Note that even the special cases (addition and multiplication) of Slutsky's theorem fail if both X and Y are non-degenerate.

Theorem (The Delta Method)

Let $Z_n := a_n(X_n - \theta) \xrightarrow{d} Z$ where $a_n, \theta \in \mathbb{R}$ for all n and $a_n \uparrow \infty$. Let $g(\cdot)$ be continuously differentiable at θ. Then, $a_n(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)Z$.

Proof

By a Taylor expansion around θ, we have

$$g(X_n) = g(\theta) + g'(\theta_n^*)(X_n - \theta),$$

where θ_n^* lies between X_n and θ and hence satisfies $|\theta_n^* - \theta| \leq |X_n - \theta|$. Moreover, $|X_n - \theta| = a_n^{-1} \cdot |a_n(X_n - \theta)| = a_n^{-1} Z_n \xrightarrow{p} 0$ by Slutsky's theorem. Therefore, $\theta_n^* \xrightarrow{p} \theta$ and, by the continuous mapping theorem, $g'(\theta_n^*) \xrightarrow{p} g'(\theta)$. Finally,

$$\begin{aligned} a_n(g(X_n) - g(\theta)) &= a_n(g(\theta) + g'(\theta_n^*)(X_n - \theta) - g(\theta)) \\ &= g'(\theta_n^*)a_n(X_n - \theta) \xrightarrow{d} g'(\theta)Z, \end{aligned}$$

using Slutsky's Theorem.

Note that the Delta Method is applicable even when $g'(\theta)$ is not continuous (the proof uses Skorokhod representation).

Exercise: Give a counterexample showing that neither $X_n \xrightarrow{p} X$ or $X_n \xrightarrow{d} X$ ensure that $\mathbb{E}[X_n] \to \mathbb{E}[X]$ as $n \to \infty$.

Theorem (Convergence of Expectations)

If $|X_n| < M < \infty$ and $X_n \xrightarrow{d} X$, then $\mathbb{E}[X]$ exists and $\mathbb{E}[X_n] \xrightarrow{n\to\infty} \mathbb{E}[X]$.

Proof.

Assume first that X_n are non-negative for any n. Then,

$$\begin{aligned} |\mathbb{E}[X_n] - \mathbb{E}[X]| &= \left| \int_0^\infty (\mathbb{P}[X_n > x] - \mathbb{P}[X > x]) dx \right| \\ &= \left| \int_0^M (\mathbb{P}[X_n > x] - \mathbb{P}[X > x]) dx \right| \\ &\leq \int_0^M |\mathbb{P}[X_n > x] - \mathbb{P}[X > x]| \, dx \overset{n\to\infty}{\to} 0, \end{aligned}$$

since $\mathbb{P}[X_n > x] \overset{n\to\infty}{\to} \mathbb{P}[X > x]$ for all but a countable number of points and the integration domain is bounded. □

Exercise: Generalize the proof to the case of less restrictive assumptions.

Remarks on Weak Convergence

- Often difficult to establish weak convergence directly (from definition).
- Indeed, in most interesting cases, F_n is not specified exactly.
- We need other more "handy" sufficient conditions.

Scheffé's Theorem

Let X_n have density functions (or mass functions) f_n, and let X have density function (or mass function) f. Then

$$f_n \overset{n\to\infty}{\longrightarrow} f \text{ (a.e.)} \implies X_n \overset{d}{\to} X.$$

- The converse to Scheffé's theorem is NOT true (why?).

Continuity Theorem

Let X_n and X have characteristic functions (cf) $\varphi_n(t) = \mathbb{E}[e^{itX_n}]$, and $\varphi(t) = \mathbb{E}[e^{itX}]$, respectively. Then,

(a) $X_n \overset{d}{\to} X \Leftrightarrow \varphi_n \to \varphi$ pointwise.

(b) If $\varphi_n(t)$ converges pointwise to some limit function $\psi(t)$ that is continuous at zero, then:

(i) $\exists$ a measure ν with cf ψ.

(ii) $F_{X_n} \overset{d}{\to} \nu$.

Weak Convergence of Random Vectors

Definition

Let $\{\mathbf{X}_n\}$ be a sequence of random vectors of $\mathbb{R}^d$, and $\mathbf{X}$ a random vector of $\mathbb{R}^d$ with $\mathbf{X}_n = (X_n^{(1)}, \dots, X_n^{(d)})^\mathsf{T}$ and $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})^\mathsf{T}$. Define the distribution functions $F_{\boldsymbol{X}_n}(\boldsymbol{x}) = \mathbb{P}[X_n^{(1)} \leq x^{(1)}, \dots, X_n^{(d)} \leq x^{(d)}]$ and $F_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{P}[X^{(1)} \leq x^{(1)}, \dots, X^{(d)} \leq x^{(d)}]$, for $\boldsymbol{x} = (x^{(1)}, \dots, x^{(d)})^\mathsf{T} \in \mathbb{R}^d$. We say that $\mathbf{X}_n$ converges in distribution to $\boldsymbol{X}$ as $n \to \infty$ (and write $\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{X}$) if for every continuity point of $F_{\boldsymbol{X}}$ we have

$$F_{\boldsymbol{X}_n}(\boldsymbol{x}) \stackrel{n\to\infty}{\longrightarrow} F_{\boldsymbol{X}}(\boldsymbol{x}).$$

There is a link between univariate and multivariate weak convergence.

Theorem (Cramér-Wold Device)

Let $\{\mathbf{X}_n\}$ be a sequence of random vectors of $\mathbb{R}^d$, and $\mathbf{X}$ a random vector of $\mathbb{R}^d$. Then,

$$\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{X} \Leftrightarrow \boldsymbol{\theta}^\mathsf{T}\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{\theta}^\mathsf{T}\boldsymbol{X}, \ \forall \boldsymbol{\theta} \in \mathbb{R}^d.$$

Stronger Notions of Convergence

Almost Sure Convergence and Convergence in L^p

There are also two stronger convergence concepts (that do not compare).

Definition (Almost Sure Convergence)

Let $\{X_n\}_{n\geq 1}$ and X be random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $A := \{\omega \in \Omega : X_n(\omega) \overset{n\to\infty}{\to} X(\omega)\}$. We say that X_n converges almost surely to X as $n \to \infty$ (and write $X_n \xrightarrow{a.s.} X$) if $\mathbb{P}[A] = 1$.

More plainly, we say that $X_n \xrightarrow{a.s.} X$ if $\mathbb{P}[X_n \to X] = 1$.

Definition (Convergence in L^p)

Let $\{X_n\}_{n\geq 1}$ and X be random variables defined on the same probability space. We say that X_n converges to X in L^p as $n \to \infty$ (and write $X_n \overset{L^p}{\to} X$) if

$$\mathbb{E}\left[|X_n - X|^p\right] \overset{n\to\infty}{\longrightarrow} 0.$$

Note that $\|X\|_{L^p} := (\mathbb{E}|X|^p)^{1/p}$ defines a complete norm (when finite).

Relationship Between Different Types of Convergence

- $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$.
- $X_n \xrightarrow{L^p} X$, for $p > 0 \implies X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$.
- for $p \geq q$, $X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{L^q} X$.
- There is no implicative relationship between "$\xrightarrow{a.s.}$" and "$\xrightarrow{L^p}$".

Theorem (Skorokhod's Representation Theorem)

Let $\{X_n\}_{n\geq 1}, X$ be random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $X_n \overset{d}{\to} X$. Then, there exist random variables $\{Y_n\}_{n\geq 1}, Y$ defined on some probability space $(\Omega', \mathcal{G}, \mathbb{Q})$ such that:

(i) $Y \overset{d}{=} X$ & $Y_n \overset{d}{=} X_n$, $\forall n \geq 1$.

(ii) $Y_n \xrightarrow{a.s.} Y$.

Exercise

Prove part (b) of the continuous mapping theorem.

The Two "Big" Theorems

Recalling two basic Theorems

Theorem (Strong Law of Large Numbers)

Let $\{X_n\}$ *be iid random variables with* $\mathbb{E}[X_k] = \mu$ *and* $\mathbb{E}[|X_k|] < \infty$ *for all* $k \geq 1$. *Then,*

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{a.s.} \mu.$$

- "Strong" is as opposed to the "weak" law which gives "$\xrightarrow{p}$" instead of "$\xrightarrow{a.s.}$".
- This is insanely strong: $\mathbb{E}[|X_k|] < \infty$ is the weakest condition for it to have an expected value. The theorem reads: if there is an expected value, we can find it with the empirical mean.
- The strong law says **nothing useful** about the **size** of the error.

Recalling two basic theorems

Theorem (Central Limit Theorem)

Let $\{\mathbf{X}_n\}$ be an iid sequence of random vectors in $\mathbb{R}^d$ with mean $\boldsymbol{\mu}$ and covariance Σ and define $\bar{\mathbf{X}}_n := \sum_{m=1}^{n} \mathbf{X}_m / n$. Then,

$$\sqrt{n}\Sigma^{-\frac{1}{2}}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{d} \boldsymbol{Z} \sim \mathcal{N}_d(0, I_d).$$

- Insanely strong theorem: as soon as the covariance exists, we are in business.
- Once more, no control about the size of the error.
- There are many variants of this basic CLT.

Convergence Rates

The mathematician rarely cares about convergence speed. The statistician does (should?) because **data is money**.

- Law of Large Numbers: assuming finite variance, L^2 rate of $n^{-1/2}$. Optimal because of the CLT.
- What about the Central Limit Theorem?

The Berry-Esseen theorem

Theorem (Berry-Esseen {Bentkus, 2005, Theory Prob Appl})

Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be iid random vectors taking values in $\mathbb{R}^d$ and such that $\mathbb{E}[\boldsymbol{X}_i] = 0$, $cov[\boldsymbol{X}_i] = I_d$ and $\mathbb{E}\left[\|\boldsymbol{X}_i\|^3\right] < \infty$. Define

$$\boldsymbol{S}_n = \frac{1}{\sqrt{n}}(\boldsymbol{X}_1 + \ldots + \boldsymbol{X}_n).$$

If $\mathcal{A}$ denotes the class of convex subsets of $\mathbb{R}^d$, then for $\boldsymbol{Z} \sim \mathcal{N}_d(\mathbf{0}, I_d)$,

$$\sup_{A \in \mathcal{A}} |\mathbb{P}[\boldsymbol{S}_n \in A] - \mathbb{P}[\boldsymbol{Z} \in A]| \leq C \frac{d^{1/4}\mathbb{E}\left[\|\boldsymbol{X}_i\|^3\right]}{\sqrt{n}},$$

where $\|.\|$ denotes the Euclidean norm. The constant C is universal and satisfies $C \leq 4$.

It allows one to quantify the approximation error in the CLT and to build confidence regions with guaranteed coverage.

Statistical Theory (Week 3): Principles of Data Reduction

Erwan Koch

Ecole Polytechnique Fédérale de Lausanne (Institute of Mathematics)

Statistical Models and The Problem of Inference

Recall our setup:

- A random vector $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$.
- A family of distributions $\mathcal{F}$ parametrized by $\Theta \subseteq \mathbb{R}^d$, i.e., $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$.
- $\boldsymbol{X} \sim F_\theta \in \mathcal{F}$.

The Problem of Point Estimation

1. Assume that F_θ is known up to the parameter θ which is unknown.
2. Let $(x_1, \ldots, x_n)^\top$ be a realization of $\boldsymbol{X} \sim F_\theta$ which is available to us.
3. Estimate the value of θ that generates $\boldsymbol{X}$, given $(x_1, \ldots, x_n)^\top$.

The only guide (apart from knowledge of $\mathcal{F}$) at hand is the data $(x_1, \ldots, x_n)^\top$:

↪ We would like to summarize the information in $(x_1, \ldots, x_n)^\top$ without loosing too much information.

↪ Anything we will use is a function of the data $g(x_1, \ldots, x_n)$.

↪ We need to study the properties of such functions and the corresponding potential information loss.

The data-processing inequality

- **Key idea**: whatever we do with the data, it cannot increase our information.
- By transforming the data / projecting it down onto the value of a statistic, at best we preserve the information that is in the data; any function of $x_1, \ldots, x_n$ carries at most the same information but usually less.
- Only new data brings new information.

Statistics of the data

Statistics

Definition (Statistic)

Let $\boldsymbol{X} \sim F_\theta$. A *statistic* T is a (measurable) function of $\boldsymbol{X}$ that does not depend on θ. Thus, $T = T(\boldsymbol{X})$. Note that T is not necessarily real-valued.

$\hookrightarrow$ Intuitively, any function of $\boldsymbol{X}$ alone is a statistic.
$\hookrightarrow$ Any statistic is itself a random variable (or vector) with its own distribution.

Example

$T(\boldsymbol{X}) = n^{-1} \sum_{i=1}^{n} X_i$ is a statistic (since n, the sample size, is known).

Example

$T(\boldsymbol{X}) = (X_{(1)}, \dots, X_{(n)})^\top$ where $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ are the order statistics of $\boldsymbol{X}$. Since T depends only on the values of $\boldsymbol{X}$, T is a statistic.

Example

$T(\boldsymbol{X}) = c$, where c is a known constant, is a statistic.

Ancillarity

Statistics and Information About θ

- Evident from previous examples: some statistics are more informative and others are less informative regarding the true value of θ.
- Any $T(\boldsymbol{X})$ that is not "1–1" with $\boldsymbol{X}$ carries less information about θ than $\boldsymbol{X}$.
- Which are "good" and which are "bad" statistics?

Definition (Ancillary Statistic)

A statistic T is an *ancillary statistic* (for θ) if its distribution does not functionally depend θ.

$\hookrightarrow$ So an ancillary statistic has the same distribution for any $\theta \in \Theta$.

Ancillarity example

Example

Suppose that $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$ (only the mean μ is unknown).

Let $T(X_1, \ldots, X_n) = X_1 - X_2$.

Then $T \sim N(0, 2)$, giving that T is ancillary for the unknown parameter μ. Nevertheless, if both μ and σ^2 were unknown, T would not be ancillary for $\theta = (\mu, \sigma^2)$.

Statistics and Information about θ

- If T is ancillary for θ then T contains no information about θ.
- In order to contain any useful information about θ, the distribution of T must depend explicitly on θ.
- Intuitively, the amount of information that T gives on θ increases as the dependence of dist(T) on θ increases.

Example

Let $X_1, ..., X_n \overset{iid}{\sim} \text{Unif}(0, \theta)$, $S = \min(X_1, \dots, X_n)$ and $T = \max(X_1, \dots, X_n)$. Then:

- $f_S(x; \theta) = \frac{n}{\theta} \left(1 - \frac{x}{\theta}\right)^{n-1}, \quad 0 \leq x \leq \theta.$
- $f_T(x; \theta) = \frac{n}{\theta} \left(\frac{x}{\theta}\right)^{n-1}, \quad 0 \leq x \leq \theta.$

$\hookrightarrow$ Neither S nor T are ancillary for θ.

$\hookrightarrow$ As $n \uparrow \infty$, f_S becomes concentrated around 0.

$\hookrightarrow$ As $n \uparrow \infty$, f_T becomes concentrated around θ.

$\hookrightarrow$ Indicates that T provides more information about θ than does S.

Sufficiency

Statistics and Information about θ

- Let $\boldsymbol{X} = (X_1, \ldots, X_n)^\top \sim F_\theta$ and $T(\boldsymbol{X})$ be a statistic.

- The *level sets* (also called *fibres* or *contours*) of T are the sets

$$A_t = \{\boldsymbol{x} \in \mathbb{R}^n : T(\boldsymbol{x}) = t\}, \quad t \in \text{Range}(T).$$

  For a given t, A_t is the set of all potential realizations that lead to the value t for T.

$\hookrightarrow$ T is constant when restricted to a level set.

- Any realization of $\boldsymbol{X}$ that falls in a given level set is equivalent as far as T is concerned.
- Any inference drawn through T will be the same within level sets.
- Now, look at dist($\boldsymbol{X}$) on a level set A_t: $f_{\boldsymbol{X}|T=t}(\boldsymbol{x})$.

Statistics and Information about θ

- Suppose that $f_{\boldsymbol{X}|T=t}$ changes depending on θ: we are losing information when using T.
- Suppose $f_{\boldsymbol{X}|T=t}$ is functionally independent of θ:

$\Longrightarrow$ $\boldsymbol{X}$ contains no information about θ on the set A_t.

$\Longrightarrow$ In other words, $\boldsymbol{X}$ is ancillary for θ on A_t.

- If this is true for each $t \in \text{Range}(T)$ then $T(\boldsymbol{X})$ contains the same information about θ as $\boldsymbol{X}$ does.
  - $\hookrightarrow$ It does not matter whether we observe $\boldsymbol{X} = (X_1, ..., X_n)$ or just $T(\boldsymbol{X})$.
  - $\hookrightarrow$ Knowing the exact value $\boldsymbol{X}$ in addition to knowing $T(\boldsymbol{X})$ does not give us any additional information — $\boldsymbol{X}$ is irrelevant if we already know $T(\boldsymbol{X})$.

Definition (Sufficient Statistic)

A statistic $T = T(\boldsymbol{X})$ is said to be *sufficient* for the parameter θ if, for all (Borel) sets B, $\mathbb{P}[\boldsymbol{X} \in B | T(\boldsymbol{X}) = t]$ does not depend on θ for all $t \in \text{Range}(T)$.

Sufficient Statistics

Example (Bernoulli Trials)

Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Bern}(\theta)$ and $T(\boldsymbol{X}) = \sum_{i=1}^n X_i$. For any $\boldsymbol{x} \in \{0,1\}^n$ and $t = \Sigma_{i=1}^n x_i$,

$$\mathbb{P}[\boldsymbol{X} = \boldsymbol{x} | T = t] = \frac{\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}, T = t]}{\mathbb{P}[T = t]} = \frac{\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}]}{\mathbb{P}[T = t]} = \frac{\theta^{\Sigma_{i=1}^n x_i}(1-\theta)^{n-\Sigma_{i=1}^n x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1},$$

which is independent of θ.

$\Longrightarrow$ T is sufficient for $\theta \rightarrow$ Given the number of tosses that came heads, knowing *which tosses* came heads is irrelevant in deciding if the coin is fair. E.g., with $n = 7$ and $t = 4$, we do not care whether we obtained 0 0 1 1 1 0 1, 1 0 0 0 1 1 1 or 1 0 1 0 1 0 1.

Sufficient Statistics

- Definition hard to verify (especially for continuous variables).
- Definition does not allow easy identification of sufficient statistics.

Theorem (Fisher-Neyman Factorization Theorem)

Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ has a joint density or frequency function $f(\boldsymbol{x};\theta)$, $\theta \in \Theta$. A statistic $T = T(\boldsymbol{X})$ is sufficient for θ if and only if

$$f(\boldsymbol{x};\theta) = g(T(\boldsymbol{x});\theta)h(\boldsymbol{x}).$$

Example

Let $X_1, ..., X_n \overset{iid}{\sim} \text{Unif}(0,\theta)$ with density $f(x;\theta) = \mathbf{1}\{x \in [0,\theta]\}/\theta$. Then,

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \frac{1}{\theta^n}\mathbf{1}\{\boldsymbol{x} \in [0,\theta]^n\} = \frac{\mathbf{1}\{\max[x_1, ..., x_n] \leq \theta\}\mathbf{1}\{\min[x_1, ..., x_n] \geq 0\}}{\theta^n}.$$

Therefore, $T(\boldsymbol{X}) = X_{(n)} = \max[X_1, ..., X_n]$ is sufficient for θ.

Proof of Neyman-Fisher Theorem - Discrete Case.

Suppose first that T is sufficient. Then

$$\begin{aligned} f(\boldsymbol{x};\theta) = \mathbb{P}[\boldsymbol{X}=\boldsymbol{x}] &= \sum_t \mathbb{P}[\boldsymbol{X}=\boldsymbol{x}, T=t] \\ &= \mathbb{P}[\boldsymbol{X}=\boldsymbol{x}, T=T(\boldsymbol{x})] \\ &= \mathbb{P}[T=T(\boldsymbol{x})]\mathbb{P}[\boldsymbol{X}=\boldsymbol{x}|T=T(\boldsymbol{x})]. \end{aligned}$$

Since T is sufficient, $\mathbb{P}[\boldsymbol{X}=\boldsymbol{x}|T=T(\boldsymbol{x})]$ is independent of θ and so $f(x;\theta)=g(T(\boldsymbol{x});\theta)h(\boldsymbol{x})$.

Now suppose that $f(\boldsymbol{x};\theta)=g(T(\boldsymbol{x});\theta)h(\boldsymbol{x})$. Then if $t=T(\boldsymbol{x})$,

$$\begin{aligned} \mathbb{P}[\boldsymbol{X}=\boldsymbol{x}|T=t] = \frac{\mathbb{P}[\boldsymbol{X}=\boldsymbol{x}]}{\mathbb{P}[T=t]} &= \frac{g(T(\boldsymbol{x});\theta)h(\boldsymbol{x})}{\sum_{\boldsymbol{y}:T(\boldsymbol{y})=t} g(T(\boldsymbol{y});\theta)h(\boldsymbol{y})} \\ &= \frac{h(\boldsymbol{x})}{\sum_{T(\boldsymbol{y})=t} h(\boldsymbol{y})}, \end{aligned}$$

which does not depend upon θ. □

Minimal Sufficiency

Minimally Sufficient Statistics

- We saw that a sufficient statistic keeps what is important about the parameter. But it can also contain useless information.
- How much information can we throw away? Is there a "smallest" sufficient statistic?

Definition (Minimally Sufficient Statistic)

A statistic $T = T(\boldsymbol{X})$ is said to be *minimally sufficient* for the parameter θ if it is sufficient for θ and, for any other sufficient statistic $S = S(\boldsymbol{X})$, there exists a function g such that

$$T(\boldsymbol{X}) = g(S(\boldsymbol{X})).$$

Lemma

If T and S are minimally sufficient statistics for the parameter θ, then there exist injective functions g and h such that $S = g(T)$ and $T = h(S)$.

Theorem

Let $\boldsymbol{X} = (X_1, ..., X_n)^\top$ have joint density or frequency function $f(\boldsymbol{x};\theta)$ and $T = T(\boldsymbol{X})$ be a statistic. Suppose that $f(\boldsymbol{x};\theta)/f(\boldsymbol{y};\theta)$ is independent of θ if and only if $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Then T is minimally sufficient for θ.

Proof.

Assume for simplicity that $f(\boldsymbol{x};\theta) > 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$ and $\theta \in \Theta$.
[**Sufficiency part**] Let A_t, $t \in \text{Range}(T)$, be the level sets of T. For each t, we denote by $\boldsymbol{y}_t \in A_t$ a representative element of the level set A_t. For any $\boldsymbol{x}$, $\boldsymbol{y}_{T(\boldsymbol{x})}$ is in the same level set as $\boldsymbol{x}$, entailing by assumption that

$$f(\boldsymbol{x};\theta)/f(\boldsymbol{y}_{T(\boldsymbol{x})};\theta)$$

does not depend on θ. Introducing $g(t;\theta) := f(\boldsymbol{y}_t;\theta)$, we have

$$f(\boldsymbol{x};\theta) = \frac{f(\boldsymbol{y}_{T(\boldsymbol{x})};\theta)f(\boldsymbol{x};\theta)}{f(\boldsymbol{y}_{T(\boldsymbol{x})};\theta)} = g(T(\boldsymbol{x});\theta)h(\boldsymbol{x}).$$

It follows from the factorization theorem that T is a sufficient statistic.

(proof cont'd).

[**Minimality part**] Let T' be any other sufficient statistic. By the factorization theorem, there exist g' and h' such that

$$f(\boldsymbol{x};\theta) = g'(T'(\boldsymbol{x});\theta)h'(\boldsymbol{x}).$$

Let $\boldsymbol{x}, \boldsymbol{y}$ be such that $T'(\boldsymbol{x}) = T'(\boldsymbol{y})$. Then

$$\frac{f(\boldsymbol{x};\theta)}{f(\boldsymbol{y};\theta)} = \frac{g'(T'(\boldsymbol{x});\theta)h'(\boldsymbol{x})}{g'(T'(\boldsymbol{y});\theta)h'(\boldsymbol{y})} = \frac{h'(\boldsymbol{x})}{h'(\boldsymbol{y})}.$$

Since this ratio does not depend on θ, we have by assumption that $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Hence, the level sets of T' are subsets of the level sets of T, which implies that T is a function of T'. Thus, T is minimal as this is true for any sufficient statistic T'. □

Example (Bernoulli Trials)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bern}(\theta)$. Let $\boldsymbol{x}, \boldsymbol{y} \in \{0,1\}^n$ be two possible realizations. Then

$$\frac{f(\boldsymbol{x};\theta)}{f(\boldsymbol{y};\theta)} = \frac{\theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}}{\theta^{\Sigma y_i}(1-\theta)^{n-\Sigma y_i}},$$

which is constant if and only if $T(\boldsymbol{x}) = \sum x_i = \sum y_i = T(\boldsymbol{y})$, so that T is minimally sufficient.

Exercise

Prove that the likelihood $f(\boldsymbol{X};\theta)$ (which is a **random function**) is a sufficient statistic. Let θ_0 be some arbitrary value such that for all $\boldsymbol{x}$, $f(\boldsymbol{x};\theta_0) \neq 0$. Prove that the normalized likelihood $f(\boldsymbol{X};\theta)/f(\boldsymbol{X};\theta_0)$ is minimally sufficient.

This exercise shows that a “minimal” statistic can be quite big.

Completeness

Complete Statistics

- Ancillary Statistic $\rightarrow$ Contains no information on θ.
- Minimally Sufficient Statistic $\rightarrow$ Contains all the relevant information about θ and as little irrelevant as possible.
- Should they be mutually independent?
- Is it possible to remove the totality of the irrelevant information?

Definition (Complete Statistic)

Let $\{g(t;\theta) : \theta \in \Theta\}$ be a family of densities (or frequencies) corresponding to a statistic $T(\boldsymbol{X})$. The statistic T is called *complete* if given any measurable function h, it holds that

$$\int h(t)g(t;\theta)dt = 0 \quad \forall\theta \in \Theta \implies \mathbb{P}[h(T) = 0] = 1 \quad \forall\theta \in \Theta.$$

Not clear why the term "complete" was chosen – one reason might be the resemblance to the notion of *complete system* in a Hilbert space (whose orthogonal complement is the zero space), in reference to $\{g(\cdot;\theta)\}_{\theta\in\Theta}$.

Complete Statistics

Example (Bernoulli Trials)

Let $X_1, ..., X_n \overset{iid}{\sim} \text{Bern}(\theta)$, $\theta \in (0,1)$, and $T = \sum X_i$. Let h be an arbitrary and measurable function. We have

$$\mathbb{E}[h(T)] = \sum_{t=0}^{n} h(t) \binom{n}{t} \theta^t (1-\theta)^{n-t} = (1-\theta)^n \sum_{t=0}^{n} h(t) \binom{n}{t} \left(\frac{\theta}{1-\theta}\right)^t.$$

As θ ranges in $(0,1)$, the ratio $\theta/(1-\theta)$ ranges in $(0,\infty)$. Thus, $\mathbb{E}[h(T)] = 0$ for all $\theta \in (0,1)$ implies that, for all $x > 0$,

$$P(x) = \sum_{t=0}^{n} h(t) \binom{n}{t} x^t = 0,$$

i.e., the polynomial $P(x)$ is uniformly zero over the entire positive real line. Hence, its coefficients must be all zero, so $h(t) = 0$, $t = 1, ..., n$. Thus, $\mathbb{P}[h(T) = 0] = 1$ for all $\theta \in (0,\infty)$.

Complete Statistics

$\hookrightarrow$ Why is completeness relevant to data reduction?

Lemma

If T is complete, then $h(T)$ is ancillary for θ if and only if $h(T) = c$ a.s.

Proof.

Let T be a complete statistic. If $h(T) = c$ a.s., $h(T)$ is obviously ancillary for θ. Conversely, let now $h(T)$ be ancillary. Then its distribution does not depend on θ, which implies that $\mathbb{E}[h(T)] = c$, for some constant c, regardless of θ. Equivalently, $\mathbb{E}[h(T) - c] = 0$ for any θ. By completeness of T, $\mathbb{P}[h(T) = c] = 1$, i.e., $h(T) = c$ a.s. $\square$

- It means that only the trivial (i.e., constant) functions of T are ancillary.
- In other words, a complete statistic contains no ancillary information.
- Contrast to a sufficient statistic:
  - A sufficient statistic keeps all the relevant information.
  - A complete statistic throws away all the irrelevant information.

Complete Statistics

Theorem (Basu's Theorem)

A complete sufficient statistic is independent of every ancillary statistic.

Proof.

We consider the discrete case only. Let T and S be complete sufficient and ancillary statistics, respectively. It suffices to show that, for any $s \in \text{Range}(S)$ and $t \in \text{Range}(T)$,

$$\mathbb{P}[S(\boldsymbol{X}) = s | T(\boldsymbol{X}) = t] = \mathbb{P}[S(\boldsymbol{X}) = s].$$

Define

$$h(t) = \mathbb{P}[S(\boldsymbol{X}) = s | T(\boldsymbol{X}) = t] - \mathbb{P}[S(\boldsymbol{X}) = s].$$

We have that:

1. $\mathbb{P}[S(\boldsymbol{x}) = s]$ does not depend on θ (by ancillarity).
2. $\mathbb{P}[S(\boldsymbol{X}) = s | T(\boldsymbol{X}) = t] = \mathbb{P}[\boldsymbol{X} \in \{x : S(x) = s\} | T = t]$ does not depend on θ (by sufficiency).

Thus, h does not depend on θ, which is necessary for $h(T)$ to be a statistic.

(proof cont'd).

Now, for any $\theta \in \Theta$,

$$\begin{aligned}\mathbb{E}[h(T)] &= \sum_t (\mathbb{P}[S(\boldsymbol{X}) = s | T(\boldsymbol{X}) = t] - \mathbb{P}[S(\boldsymbol{X}) = s])\mathbb{P}[T(\boldsymbol{X}) = t] \\ &= \sum_t \mathbb{P}[S(\boldsymbol{X}) = s | T(\boldsymbol{X}) = t]\mathbb{P}[T(\boldsymbol{X}) = t] \\ &\quad - \mathbb{P}[S(\boldsymbol{X}) = s] \sum_t \mathbb{P}[T(\boldsymbol{X}) = t] \\ &= \mathbb{P}[S(\boldsymbol{X}) = s] - \mathbb{P}[S(\boldsymbol{X}) = s] = 0.\end{aligned}$$

Since T is complete, it follows that $h(t) = 0$ a.s. for all $t \in \text{Range}(T)$. □

Basu's Theorem is useful for deducing independence of two statistics:

- No need to determine their joint distribution.
- Need to show completeness (usually hard analytical problem).
- We will see models for which completeness is easy to check.

Completeness and Minimal Sufficiency

Theorem (Lehmann-Scheffé)

Let $\boldsymbol{X}$ have density $f(\boldsymbol{x};\theta)$. If $T(\boldsymbol{X})$ is sufficient and complete for θ then T is minimally sufficient.

Proof.

First we show that a minimally sufficient statistic exists. We define an equivalence relation, denoted by $\equiv$, as $\boldsymbol{x} \equiv \boldsymbol{x}'$ if and only if $f(\boldsymbol{x};\theta)/f(\boldsymbol{x}';\theta)$ is independent of θ. Let S be a function such that $S(\boldsymbol{z}) = c_{\boldsymbol{x}}$ for any $\boldsymbol{z}$ belonging to the class with representative $\boldsymbol{x}$ (S is constant on that class), and such that $\boldsymbol{x}^{(1)} \not\equiv \boldsymbol{x}^{(2)} \Rightarrow c_{\boldsymbol{x}^{(1)}} \neq c_{\boldsymbol{x}^{(2)}}$. Then, $f(\boldsymbol{x};\theta)/f(\boldsymbol{y};\theta)$ is independent of θ if and only if $S(\boldsymbol{x}) = S(\boldsymbol{y})$, giving that S is minimally sufficient. This establishes the existence.

Note that to be perfectly rigorous, we should check that S is measurably constructible; see the proof by Lehmann–Scheffé (1950) for corresponding details.

(proof cont'd).

Therefore, as T is sufficient, there exists a function g_1 such that $S = g_1(T)$. Let $g_2(S) = \mathbb{E}[T|S]$ (which does not depend on θ since S is sufficient) and consider

$$g(T) = T - g_2(S).$$

We have

$$\mathbb{E}[g(T)] = \mathbb{E}[T] - \mathbb{E}\left\{\mathbb{E}[T|S]\right\} = \mathbb{E}[T] - \mathbb{E}[T] = 0.$$

for all θ. By completeness of T, it follows that $g(T) = 0$, i.e., $g_2(S) = T$ a.s. The function g_2 has to be injective since otherwise it would contradict the minimal sufficiency of S. As moreover $S = g_1(T)$, there is a bijective relationship between S and T, yielding that T is minimally sufficient. □

The log-likelihood is minimally sufficient (if normalized), but not necessarily complete!

Exercise

Consider the following situation:

- We pick a random number $\mathbb{N} \ni N \sim F_n$
- We gather N iid random variables $X_1 \dots X_N \sim \mathcal{N}(\mu, 1)$.

1. Write down the normalized log-likelihood function $\mu \to LL(\mu) - LL(0)$ as a function of N and $\boldsymbol{X}$. This is a **function-valued random variable**.
2. Prove that it is minimally sufficient. Note that the log-likelihood $\mu \to LL(\mu)$ is only sufficient, not minimally sufficient.
3. Prove that it is not complete.

Summary

We looked at how to "summarize" the data by computing the value of a statistic $S(\boldsymbol{X})$, where $\boldsymbol{X} \sim F_\theta$:

- Ancillarity: S carries no information on θ.
- Sufficiency: S does not lose information on θ.
- Minimal sufficiency: S does not lose information on θ and carries as little ancillary information as possible.
- Completeness: S carries no ancillary information.

Most of the time, a minimally sufficient statistic exists: the normalized log-likelihood. A complete sufficient statistic may, however, not exist.

Statistical Theory (Week 4): Special Families of Models

Erwan Koch

Ecole Polytechnique Fédérale de Lausanne (Institute of Mathematics)

Focus on Parametric Families

Focus on Parametric Families

Recall our setup:

- A random vector $\boldsymbol{X} = (X_1, ..., X_n)^\top$.
- A family of distributions $\mathcal{F}$ parametrized by $\Theta \subseteq \mathbb{R}^d$, i.e., $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$.
- $\boldsymbol{X} \sim F_\theta \in \mathcal{F}$.

The Problem of Point Estimation

1. Assume that F_θ is known up to the parameter θ which is unknown.
2. Let $(x_1, ..., x_n)^\top$ be a realization of $\boldsymbol{X} \sim F_\theta$ which is available to us.
3. Estimate the value of θ that generates $\boldsymbol{X}$, given $(x_1, ..., x_n)^\top$.

The only guide (apart from knowledge of $\mathcal{F}$) at hand is the data $(x_1, ..., x_n)^\top$:

$\hookrightarrow$ Anything we will use is a function of the data $g(x_1, ..., x_n)$.

- So far we have focused on the aspects: approximation of the distributions of $g(X_1, \ldots, X_n)$ + data reduction (how to find the best possible function g?)
- But what about $\mathcal{F}$?

Focus on Parametric Families

We describe $\mathcal{F}$ by a *parametrization* $\Theta \ni \theta \mapsto F_\theta$.

Definition (Parametrization)

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

$\hookrightarrow$ It assigns a label $\theta \in \Theta$ to each member 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, $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$.

So far we have seen a number of examples of distributions and have shown some properties of each distribution individually.

Question

Are there general families of distributions that contain the standard ones as special cases and for which a general and abstract study can be performed?

Exponential Families of Distributions

Exponential Families of Distributions

Definition (Exponential Family)

Let $\boldsymbol{X} = (X_1, ..., X_n)^\top$ have joint distribution F_θ with parameter $\theta \in \mathbb{R}^p$. We say that the family of distributions F_θ is a k-parameter exponential family if the joint density or joint frequency function of $(X_1, ..., X_n)^\top$ admits the form

$$f(\boldsymbol{x}; \theta) = \exp\left\{\sum_{i=1}^{k} c_i(\theta) T_i(\boldsymbol{x}) - d(\theta) + S(\boldsymbol{x})\right\}, \quad \boldsymbol{x} \in \mathcal{X}, \theta \in \Theta,$$

with $\text{supp}\{f(\cdot; \theta)\} = \mathcal{X}$ independent of θ.

- k need not equal p, although they coincide in many cases.
- Frequently, it is more convenient to re-parametrize this model by introducing $\phi_i = c_i(\theta)$, $i = 1, \ldots, k$. The vector $\phi = (\phi_1, \ldots, \phi_k)^\top$ is called the **natural parameter**.
- The value of k may be reduced if the ϕ_i or T_i satisfy linear constraints.
- We will assume that the representation above is minimal in the sense that neither the T_i nor the ϕ_i satisfy a linear constraint.

Motivation: Maximum Entropy Under Constraints

Consider the following variational (i.e., optimization) problem:

Determine the probability distribution f supported on $\mathcal{X}$ which maximizes the entropy

$$H(f) = -\int_{\mathcal{X}} f(\boldsymbol{x}) \log f(\boldsymbol{x}) d\boldsymbol{x},$$

under the linear (moment) constraints

$$\int_{\mathcal{X}} T_i(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} = \alpha_i, \qquad i = 1, \ldots, k.$$

Philosophy:

- Question: how to choose a probability model for a given situation?
- Solution: maximum entropy approach. In any given situation, the idea is to choose the distribution that gives the *highest uncertainty* while satisfying situation–specific required constraints.

Proposition

When a solution to the constrained optimization problem exists, it is unique and has the form

$$f(\boldsymbol{x}) = Q(\lambda_1, \ldots, \lambda_k) \exp\left\{ \sum_{i=1}^{k} \lambda_i T_i(\boldsymbol{x}) \right\}.$$

Proof.

Let f be written as above and g be a density also satisfying the constraints. Then,

$$\begin{aligned}
H(g) &= -\int_{\mathcal{X}} g(\boldsymbol{x}) \log g(\boldsymbol{x}) d\boldsymbol{x} = -\int_{\mathcal{X}} g(\boldsymbol{x}) \log \left[\frac{g(\boldsymbol{x})}{f(\boldsymbol{x})} f(\boldsymbol{x}) \right] d\boldsymbol{x} \\
&= -\int_{\mathcal{X}} g(\boldsymbol{x}) \log \left[\frac{g(\boldsymbol{x})}{f(\boldsymbol{x})} \right] d\boldsymbol{x} - \int_{\mathcal{X}} g(\boldsymbol{x}) \log f(\boldsymbol{x}) d\boldsymbol{x} \\
&= -\underbrace{KL(g \| f)}_{\geq 0} \qquad - \int_{\mathcal{X}} g(\boldsymbol{x}) \log f(\boldsymbol{x}) d\boldsymbol{x} \\
&\leq -\log Q(\lambda_1, \ldots, \lambda_k) \underbrace{\int_{\mathcal{X}} g(\boldsymbol{x}) d\boldsymbol{x}}_{=1} - \int_{\mathcal{X}} g(\boldsymbol{x}) \left(\sum_{i=1}^{k} \lambda_i T_i(\boldsymbol{x}) \right) d\boldsymbol{x}.
\end{aligned}$$

(proof cont'd).

As g also satisfies the moment constraints, the last term is

$$\begin{aligned} &= -\log Q(\lambda_1,\dots,\lambda_k) - \int_{\mathcal{X}} f(\mathbf{x}) \left(\sum_{i=1}^{k} \lambda_i T_i(\mathbf{x}) \right) d\mathbf{x} = -\int_{\mathcal{X}} f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x} \\ &= H(f). \end{aligned}$$

The uniqueness of the solution follows from the fact that strict equality can only occur when $KL(g\|f) = 0$, which happens if and only if $g = f$. □

- The λ_i's are the Lagrange multipliers derived by the Lagrange form of the optimization problem.
- These are derived so that the constraints are satisfied.
- They give us the $c_i(\theta)$ in our definition of exponential families.
- Note that the presence of $S(\mathbf{x})$ in our definition is compatible: $S(\mathbf{x}) = c_{k+1} T_{k+1}(\mathbf{x})$, where c_{k+1} *does not* depend on θ.

  (provision for a multiplier that may not depend on parameter)

Example (Binomial Distribution)

Let $X \sim$Binom(n, θ) with n known. Then, for $x = 1, \dots, n$,

$$f(x;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \exp\left[\log\left(\frac{\theta}{1-\theta}\right)x + n\log(1-\theta) + \log\binom{n}{x}\right],$$

and so dist(X) belongs to a one-parameter exponential family.

Example (Gamma Distribution)

Let $X_1, ..., X_n \overset{iid}{\sim}$Gamma with unknown shape parameter α and unknown rate parameter λ. Then, provided $x_1, \dots, x_n > 0$,

$$\begin{aligned} f(\boldsymbol{x};\alpha,\lambda) &= \prod_{i=1}^{n} \frac{\lambda^\alpha x_i^{\alpha-1}\exp(-\lambda x_i)}{\Gamma(\alpha)} \\ &= \exp\left[(\alpha-1)\sum_{i=1}^{n}\log x_i - \lambda\sum_{i=1}^{n} x_i + n\alpha\log\lambda - n\log\Gamma(\alpha)\right]. \end{aligned}$$

Hence dist$(\boldsymbol{X})$ belongs to a two-parameter exponential family.

Example (Heteroskedastic Gaussian Distribution)

Let $X_1, ..., X_n \overset{iid}{\sim} N(\theta, \theta^2)$, where $\theta > 0$. Then, for any $\boldsymbol{x} \in \mathbb{R}^n$,

$$\begin{aligned} f(\boldsymbol{x}; \theta) &= \prod_{i=1}^{n} \frac{1}{\theta\sqrt{2\pi}} \exp\left[-\frac{1}{2\theta^2}(x_i - \theta)^2\right] \\ &= \exp\left[-\frac{1}{2\theta^2}\sum_{i=1}^{n} x_i^2 + \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{n}{2}\left\{(1 + 2\log\theta) + \log(2\pi)\right\}\right]. \end{aligned}$$

Notice that even though $k = 2$ here, the dimension of the parameter space is 1. This is an example of a *curved exponential family*.

Example (Uniform Distribution)

Let $X \sim \text{Unif}(0, \theta)$. Then,

$$f(x; \theta) = \frac{\mathbf{1}\{x \in [0, \theta]\}}{\theta}.$$

Since the support of f, $\mathcal{X}$, depends on θ, $\text{dist}(X)$ *does not belong* to an exponential family.

Exponential Families of Distributions

Proposition

Suppose that $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ has a one-parameter exponential family distribution with density or frequency function

$$f(\boldsymbol{x};\theta) = \exp\left[c(\theta)T(\boldsymbol{x}) - d(\theta) + S(\boldsymbol{x})\right]$$

for $\boldsymbol{x} \in \mathcal{X}$, where

(a) the parameter space Θ is open;

(b) $c(\cdot)$ is twice continuously differentiable with non vanishing derivative.

Then, d is twice differentiable and

$$\mathbb{E}[T(\boldsymbol{X})] = \frac{d'(\theta)}{c'(\theta)} \quad \& \quad \mathrm{Var}[T(\boldsymbol{X})] = \frac{d''(\theta)c'(\theta) - d'(\theta)c''(\theta)}{[c'(\theta)]^3}.$$

Proof.

Define $\phi = c(\theta)$ the *natural parameter* of the exponential family. Since $c \in C^2$ and $c' \neq 0$, the inverse function theorem states that there exists an open neighbourhood U of ϕ such that $c^{-1}(\phi)$ exists and is continuously differentiable on U, with derivative

$$\frac{d}{d\phi} c^{-1}(\phi) = \frac{1}{c'(c^{-1}(\phi))}.$$

Since U is open, there exists s sufficiently small so that $\phi + s \in U$. Letting $\gamma(\phi) = d(c^{-1}(\phi))$ on U, the MGF of $T(\boldsymbol{X})$ is

$$\begin{aligned}
\mathbb{E}[\exp[sT(\boldsymbol{X})]] &= \int e^{sT(\boldsymbol{x})} e^{\phi T(\boldsymbol{x}) - \gamma(\phi) + S(\boldsymbol{x})} d\boldsymbol{x} \\
&= e^{\gamma(\phi+s) - \gamma(\phi)} \underbrace{\int e^{(\phi+s)T(\boldsymbol{x}) - \gamma(\phi+s) + S(\boldsymbol{x})} d\boldsymbol{x}}_{=1} \\
&= \exp[\gamma(\phi + s) - \gamma(\phi)].
\end{aligned}$$

(proof cont'd).

It follows that $M_T(s) < \infty$ for s sufficiently small, and thus that

- all moments of T exist;
- $M_T(s)$ is infinitely differentiable on an open neighbourhood of 0.

Therefore, $\gamma(s+\phi)$ is infinitely differentiable for s small enough, i.e., γ is infinitely differentiable in an open neighbourhood of ϕ. Now, differentiating the MGF wrt s and setting $s = 0$, we get

$$\mathbb{E}[T(\boldsymbol{X})] = \gamma'(\phi) \quad \& \quad \text{Var}[T(\boldsymbol{X})] = \gamma''(\phi).$$

To complete the proof, we recall that $\gamma(\phi) = d(c^{-1}(\phi))$. Using the fact that $c \in C^2$ and $\gamma \in C^\infty$, easy computations using the inverse function theorem yield

$$\gamma'(\phi) = d'(\theta)/c'(\theta) \quad \text{and} \quad \gamma''(\phi) = [d''(\theta)c'(\theta) - d'(\theta)c''(\theta)]/[c'(\theta)]^3.$$

□

Exercise

Extend the result to the means, variances and covariances of the random variables $T_1(\boldsymbol{X}), ..., T_k(\boldsymbol{X})$ in a k-parameter exponential family.

Exponential Families and Sufficiency

Lemma

Suppose that $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ *has a k-parameter exponential family distribution with density or frequency function*

$$f(\boldsymbol{x}; \theta) = \exp\left[\sum_{i=1}^{k} c_i(\theta) T_i(\boldsymbol{x}) - d(\theta) + S(\boldsymbol{x})\right]$$

for $\boldsymbol{x} \in \mathcal{X}$. *Then, the statistic* $(T_1(\boldsymbol{X}), \dots, T_k(\boldsymbol{X}))^\top$ *is sufficient for* θ.

The statistic $(T_1(\boldsymbol{X}), \dots, T_k(\boldsymbol{X}))^\top$ is sometimes called the natural sufficient statistic.

Proof.

Let $\boldsymbol{T}(\boldsymbol{X}) = (T_1(\boldsymbol{X}), \dots, T_k(\boldsymbol{X}))^\top$. We have

$$f(\boldsymbol{x}; \theta) = g(\boldsymbol{T}(\boldsymbol{x}); \theta) h(\boldsymbol{x}),$$

where $g(\boldsymbol{T}(\boldsymbol{x}); \theta) = \exp\left\{\sum_i c_i(\theta) T_i(\boldsymbol{x}) - d(\theta)\right\}$ and $h(\boldsymbol{x}) = \exp\{S(\boldsymbol{x})\}\mathbf{1}\{\mathbf{x} \in \mathcal{X}\}$. The factorization theorem yields the result. □

Sampling Exponential Families

- The families of distributions obtained by sampling from exponential families are themselves exponential families.
- Let $X_1, ..., X_n$ be iid according to a k-parameter exponential family. The density (or frequency function) of $\boldsymbol{X} = (X_1, ..., X_n)^\top$ is

$$
\begin{aligned}
f(\boldsymbol{x};\theta) &= \prod_{j=1}^{n} \exp\left[\sum_{i=1}^{k} c_i(\theta)\, T_i(x_j) - d(\theta) + S(x_j)\right] \\
&= \exp\left[\sum_{i=1}^{k} c_i(\theta)\tau_i(\boldsymbol{x}) - nd(\theta) + \sum_{j=1}^{n} S(x_j)\right],
\end{aligned}
$$

  where $\tau_i(\boldsymbol{X}) = \sum_{j=1}^{n} T_i(X_j)$, $i = 1, ..., k$. The latter are called the *natural statistics*.
- Note that the natural sufficient statistic is k-dimensional for any n.
- What about the distribution of $\boldsymbol{\tau} = (\tau_1(\boldsymbol{X}), ..., \tau_k(\boldsymbol{X}))^\top$?

The Natural Statistics

Lemma

The distribution of $\boldsymbol{\tau} = (\tau_1(\boldsymbol{X}), \ldots, \tau_k(\boldsymbol{X}))^\top$ *is of exponential family form with natural parameters* $c_1(\theta), \ldots, c_k(\theta)$.

Proof. (discrete case).

Let $\mathcal{T}_{\boldsymbol{y}} = \{\boldsymbol{x} : \tau_1(\boldsymbol{x}) = y_1, \ldots, \tau_k(\boldsymbol{x}) = y_k\}$ be the level set of $\boldsymbol{y} \in \mathbb{R}^k$. We have

$$\begin{aligned}
\mathbb{P}[\boldsymbol{\tau} = \boldsymbol{y}] &= \sum_{\boldsymbol{x} \in \mathcal{T}_{\boldsymbol{y}}} \mathbb{P}[\boldsymbol{X} = \boldsymbol{x}] = \delta(\theta) \sum_{\boldsymbol{x} \in \mathcal{T}_{\boldsymbol{y}}} \exp\left[\sum_{i=1}^{k} c_i(\theta)\tau_i(\boldsymbol{x}) + \sum_{j=1}^{n} S(x_j)\right] \\
&= \delta(\theta) \exp\left[\sum_{i=1}^{k} c_i(\theta) y_i\right] \sum_{\boldsymbol{x} \in \mathcal{T}_{\boldsymbol{y}}} \exp\left[\sum_{j=1}^{n} S(x_j)\right] \\
&= \delta(\theta) \mathbb{S}(\boldsymbol{y}) \exp\left[\sum_{i=1}^{k} c_i(\theta) y_i\right],
\end{aligned}$$

where $\delta(\theta) = \exp(-nd(\theta))$. □

The Natural Statistics

Lemma

For any $A \subseteq \{1, ..., k\}$, the joint distribution of $\{\tau_i(\boldsymbol{X}); i \in A\}$ conditional on $\{\tau_i(\boldsymbol{X}); i \in A^c\}$ is of exponential family form, and depends only on $\{c_i(\theta); i \in A\}$.

Proof. (discrete case).

Let $\mathcal{T}_i = \tau_i(\boldsymbol{X})$, $i = 1, \ldots, k$, $\mathcal{T}_A = \{\tau_i(\boldsymbol{X}) : i \in A\}$ and $\boldsymbol{y}_A = \{y_i : i \in A\}$. Recall that we have $\mathbb{P}[\boldsymbol{\tau} = \boldsymbol{y}] = \delta(\theta)\mathcal{S}(\boldsymbol{y}) \exp\left[\sum_{i=1}^{k} c_i(\theta) y_i\right]$. Thus,

$$\begin{aligned}
&\mathbb{P}[\mathcal{T}_A = \boldsymbol{y}_A | \mathcal{T}_{A^c} = \boldsymbol{y}_{A^c}] \\
&= \frac{\mathbb{P}[\mathcal{T}_A = \boldsymbol{y}_A, \mathcal{T}_{A^c} = \boldsymbol{y}_{A^c}]}{\sum_{\boldsymbol{w} \in \mathbb{R}^{\#(A)}} \mathbb{P}[\mathcal{T}_A = \boldsymbol{w}, \mathcal{T}_{A^c} = \boldsymbol{y}_{A^c}]} \\
&= \frac{\delta(\theta)\mathcal{S}((\boldsymbol{y}_A, \boldsymbol{y}_{A^c})) \exp\left[\sum_{i \in A} c_i(\theta) y_i\right] \exp\left[\sum_{i \in A^c} c_i(\theta) y_i\right]}{\delta(\theta) \exp\left[\sum_{i \in A^c} c_i(\theta) y_i\right] \sum_{\boldsymbol{w} \in \mathbb{R}^{\#(A)}} \mathcal{S}((\boldsymbol{w}, \boldsymbol{y}_{A^c})) \exp\left[\sum_{i \in A} c_i(\theta) w_i\right]} \\
&= \Delta(\{c_i(\theta) : i \in A\}) \mathcal{S}(\boldsymbol{y}_A, \boldsymbol{y}_{A^c}) \exp\left[\sum_{i \in A} c_i(\theta) y_i\right].
\end{aligned}$$

□

The Natural Statistics and Sufficiency

Look at the previous results through the prism of the canonical parametrization:

- We already know that $\boldsymbol{\tau}$ is sufficient for $\phi = (c_1(\theta), \ldots, c_k(\theta))^\top$.
- But the previous result tells us something even stronger:

> Each τ_i, $i = 1, \ldots, k$, gives information about $\phi_i = c_i(\theta)$ ("conditionally sufficient").

- In fact any $\boldsymbol{\tau}_A$ gives information about ϕ_A ("conditionally sufficient"), $\forall$ $A \subseteq \{1, ..., k\}$.
- Therefore, each natural statistic contains relevant information about each natural parameter.
- A useful result that is by no means true for any distribution.

Exponential Families and Completeness

Theorem

Suppose that $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ *has a* k*-parameter exponential family distribution with density or frequency function*

$$f(\boldsymbol{x}; \theta) = \exp\left[\sum_{i=1}^{k} c_i(\theta) T_i(\boldsymbol{x}) - d(\theta) + S(\boldsymbol{x})\right]$$

for $\boldsymbol{x} \in \mathcal{X}$. *Define* $C = \{(c_1(\theta), \dots, c_k(\theta))^\top : \theta \in \Theta\}$. *If the set* C *contains an open set (i.e., a* k*-dimensional rectangle), then the statistic* $(T_1(\boldsymbol{X}), \dots, T_k(\boldsymbol{X}))^\top$ *is complete for* θ*, and so minimally sufficient.*

A k-parameter exponential family satisfying the condition on C is said to be of full rank.

Intuitively, this result says that a k-dimensional sufficient statistic in a k-parameter exponential family will also be complete for θ provided that the effective dimension of C is k.

Proof. (Case $k = 1$)

Recall that T also has a 1-parameter exponential family distribution, with natural parameter $c(\theta)$ and density

$$f_T(t) = \delta(\theta)\mathcal{S}(t)\exp\{c(\theta)t\}.$$

Let $g(\cdot)$ be such that $\mathbb{E}_\theta[g(T)] = 0$ for all $\theta \in \Theta$. This translates into

$$\delta(\theta)\int_{\mathbb{R}} g(t)\mathcal{S}(t)\exp\{c(\theta)t\}dt = 0, \qquad \forall\theta \in \Theta.$$

We write $g = g^+ - g^- = g(t)\mathbf{1}\{g(t) \geq 0\} - |g(t)|\mathbf{1}\{g(t) < 0\}$, i.e., we decompose g into its positive and negative parts. This yields

$$\int_{\mathbb{R}} g^+(t)\mathcal{S}(t)\exp\{c(\theta)t\}dt = \int_{\mathbb{R}} g^-(t)\mathcal{S}(t)\exp\{c(\theta)t\}dt, \qquad \forall\theta \in \Theta.$$

Since $\mathbb{E}_\theta[g(T)]$ exists for all θ, the two terms above are finite $\forall\theta$.

(proof cont'd)

Our trick will be to view the two previous integrands as probability densities, which is possible as $\mathcal{S}(t) \geq 0$. Let θ_0 be such that $c(\theta_0)$ is in the interior of C (such a θ_0 exists by our assumption that C contains an open set). Let us define r by the value of either side when $\theta = \theta_0$, i.e.,

$$r = \int_{\mathbb{R}} g^+(t)\mathcal{S}(t)\exp\{c(\theta_0)t\}dt.$$

Then,

$$F(u) = \int_{-\infty}^{u} \frac{1}{r} g^+(t)\mathcal{S}(t)\exp\{c(\theta_0)t\}dt \quad \& \quad G(u) = \int_{-\infty}^{u} \frac{1}{r} g^-(t)\mathcal{S}(t)\exp\{c(\theta_0)t\}dt$$

define two probability distribution functions, with densities given by the integrands. Using this definition and dividing both sides of our previous equality by r, we obtain

$$\mathbb{E}[\exp\{[c(\theta) - c(\theta_0)]Z\}] = \mathbb{E}[\exp\{[c(\theta) - c(\theta_0)]W\}],$$

where $Z \sim F$ and $W \sim G$. These equalities are valid for all θ, and so for an open neighbourhood of $\phi = c(\theta) - c(\theta_0)$ containing zero. By the characterization property of the MGFs, we obtain that $F = G$, and so $g^+ = g^-$ almost everywhere (a.e.), i.e., $g = 0$ a.e. Thus, T is complete.

Summary on exponential families

- An exponential family gives a max-entropy model of the data.
- The statistic $\boldsymbol{T}(\boldsymbol{X}) = (T_1(\boldsymbol{X}), ..., T_k(\boldsymbol{X}))^\top$ is sufficient for θ.
- If the exponential family is full rank, then $\boldsymbol{T}(\boldsymbol{X})$ is also complete for θ. The conjunction of "sufficient" and "complete" almost never occurs outside of exponential families.
- The natural sufficient statistic is k-dimensional whatever the sample size n.

BUT, KEY LESSON: For our data, it's better to have a good model which has drawbacks from a mathematical viewpoint than a bad one which has great mathematical properties!!

Transformation Families

Groups Acting on the Data Space

Basic Idea

Often we can generate a family of distributions of the same form (but with different parameters) by letting a group act on our data space $\mathcal{X}$.

Recall: a group is a set G along with a binary operator $\circ$ such that:

1. $g, g' \in G \implies g \circ g' \in G$.
2. $(g \circ g') \circ g'' = g \circ (g' \circ g''), \ \forall g, g', g'' \in G$.
3. $\exists \ e \in G : e \circ g = g \circ e = g, \ \forall g \in G$.
4. $\forall g \in G \ \exists \ g^{-1} \in G : g \circ g^{-1} = g^{-1} \circ g = e$.

Often, groups are sets of transformations and the binary operator is the composition operator (e.g., $SO(2)$, the group of rotations of $\mathbb{R}^2$):

$$\begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} = \begin{bmatrix} \cos(\phi+\psi) & -\sin(\phi+\psi) \\ \sin(\phi+\psi) & \cos(\phi+\psi) \end{bmatrix}.$$

Groups Acting on the Data Space

- Let $(G, \circ)$ be a group of transformations, with $G \ni g : \mathcal{X} \to \mathcal{X}$.
- $gX := g(X)$ and $(g_2 \circ g_1)X := g_2(g_1(X))$.
- Obviously $\text{dist}(gX)$ changes as g ranges in G.
- Is this change completely arbitrary or are there situations where it has a simple structure?

Definition (Transformation Family)

Let G be a group of transformations acting on $\mathcal{X}$ and let $\{f_\theta(x); \theta \in \Theta\}$ be a parametric family of densities on $\mathcal{X}$. If there exists a bijection $h : G \to \Theta$ then the family $\{f_\theta\}_{\theta \in \Theta}$ will be called a *(group) transformation family* if

$$X \sim f_\theta \Rightarrow g(X) \sim f_{h(g)*\theta},$$

where $*$ is a binary operator on Θ.

Hence Θ admits a group structure $\bar{G} := (\Theta, *)$ via

$$\theta_1 * \theta_2 := h(h^{-1}(\theta_1) \circ h^{-1}(\theta_2)).$$

Usually we write $g_\theta = h^{-1}(\theta)$, so $g_\theta \circ g_{\theta'} = g_{\theta * \theta'}$

Invariance and Equivariance

Define an equivalence relation on $\mathcal{X}$ via G, by

$$x \overset{G}{\equiv} x' \iff \exists\, g \in G : x' = g(x).$$

This partitions $\mathcal{X}$ into equivalence classes called the *orbits* of $\mathcal{X}$ under G.

Definition (Invariant Statistic)

A statistic T that is constant on the orbits of $\mathcal{X}$ under G is called an *invariant statistic*. That is, T is invariant with respect to G if, for any arbitrary $x \in \mathcal{X}$, we have $T(x) = T(gx)$ for any $g \in G$.

Notice that it may be that $T(x) = T(y)$ but x, y are not in the same orbit, i.e., in general the orbits under G are subsets of the level sets of an invariant statistic T. When orbits and level sets coincide, we have:

Definition (Maximal Invariant)

A statistic T will be called a *maximal invariant* for G when

$$T(x) = T(y) \iff x \overset{G}{\equiv} y.$$

Invariance and Equivariance

- Intuitively, a maximal invariant is a reduced version of the data that represent it as closely as possible, under the requirement of remaining invariant with respect to G.
- If T is an invariant statistic with respect to the group defining a transformation family, then it is ancillary.

Definition (Equivariance)

A statistic $S : \mathcal{X} \to \Theta$ will be called equivariant for a transformation family if $S(g_\theta x) = \theta * S(x), \quad \forall\ g_\theta \in G \ \&\ x \in \mathcal{X}$.

- Equivariance may be a natural property to require if S is used as an *estimator* of the true parameter $\theta \in \Theta$, as it suggests that a transformation of a sample by g_ψ would yield an estimator that is the original one transformed by ψ.

Lemma (Constructing Maximal Invariants)

Let $S : \mathcal{X} \to \Theta$ be an equivariant statistic for a transformation family with parameter space Θ and transformation group G. Then, $T(X) = g^{-1}_{S(X)}X$ defines a maximally invariant statistic.

Proof.

$$T(g_\theta x) \overset{def}{=} (g^{-1}_{S(g_\theta x)} \circ g_\theta)x \overset{eqv}{=} (g^{-1}_{\theta * S(x)} \circ g_\theta)x = [(g^{-1}_{S(x)} \circ g^{-1}_\theta) \circ g_\theta]x = T(x)$$

so that T is invariant. To show maximality, notice that

$$T(x) = T(y) \implies g^{-1}_{S(x)}x = g^{-1}_{S(y)}y \implies y = \underbrace{g_{S(y)} \circ g^{-1}_{S(x)}}_{=g \in G}x$$

so that $\exists g \in G$ with $y = gx$ which completes the proof. □

Location-Scale Families

An important transformation family is the *location-scale* model:

- Let $X = \eta + \tau\varepsilon$ with $\varepsilon \sim f$ completely known.
- Parameter is $\theta = (\eta, \tau) \in \Theta = \mathbb{R} \times \mathbb{R}_+$.
- Define set of transformations on $\mathcal{X}$ by $g_\theta x = g_{(\eta,\tau)}x = \eta + \tau x$.
- We have
    - $g_{(\eta,\tau)} \circ g_{(\mu,\sigma)}x = \eta + \tau\mu + \tau\sigma x = g_{(\eta+\tau\mu,\tau\sigma)}x$, giving that the set of transformations is closed under composition.
    - $g_{(\mu,\sigma)} \circ g_{(\eta,\tau)}x = g_{(\eta,\tau)} \circ g_{(\mu,\sigma)}x$;
    - $g_{(0,1)} \circ g_{(\eta,\tau)} = g_{\eta,\tau} \circ g_{(0,1)} = g_{(\eta,\tau)}$ (so $\exists$ identity);
    - $g(-\eta/\tau, \tau^{-1}) \circ g_{(\eta,\tau)} = g_{(\eta,\tau)} \circ g(-\eta/\tau, \tau^{-1}) = g_{(0,1)}$ (so $\exists$ inverse).

  Hence $G = (\{g_\theta : \theta \in \mathbb{R} \times \mathbb{R}_+\}, \circ)$ is a group.
- The action of G on random sample $\boldsymbol{X} = \{X_i\}_{i=1}^n$ is $g_{(\eta,\tau)}\boldsymbol{X} = \eta\mathbf{1}_n + \tau\boldsymbol{X}$.
- The (unique) induced group action on Θ is $(\eta,\tau) * (\mu,\sigma) = (\eta + \tau\mu, \tau\sigma)$.

- The sample mean and sample variance are equivariant, because with $S(\boldsymbol{X}) = (\bar{X}, V^{1/2})$, where $V = \frac{1}{n-1}\sum(X_j - \bar{X})^2$, we have

$$\begin{aligned} S(g_{(\eta,\tau)}\boldsymbol{x}) &= \left(\overline{\eta + \tau\boldsymbol{X}}, \left\{\frac{1}{n-1}\sum(\eta + \tau X_j - \overline{(\eta + \tau X)})^2\right\}^{1/2}\right) \\ &= \left(\eta + \tau\bar{X}, \left\{\frac{1}{n-1}\sum(\eta + \tau X_j - \eta - \tau\bar{X})^2\right\}^{1/2}\right) \\ &= (\eta + \tau\bar{X}, \tau V^{1/2}) = (\eta, \tau) * S(\boldsymbol{X}). \end{aligned}$$

- A maximal invariant is given by $A = g^{-1}_{S(\boldsymbol{X})}\boldsymbol{X}$ the corresponding parameter being $(-\bar{X}/V^{1/2}, V^{-1/2})$. Hence the vector of residuals is a maximal invariant:

$$A = \frac{(\boldsymbol{X} - \bar{X}\mathbf{1}_n)}{V^{1/2}} = \left(\frac{X_1 - \bar{X}}{V^{1/2}}, \ldots, \frac{X_n - \bar{X}}{V^{1/2}}\right).$$

Transformation Families

Example (The Multivariate Gaussian Distribution)

- Let $\boldsymbol{Z} \sim \mathcal{N}_d(0, I)$ and consider $\boldsymbol{X} = \boldsymbol{\mu} + \Omega \boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \Omega\Omega^{\mathsf{T}})$.
- The parameter is $(\boldsymbol{\mu}, \Omega) \in \mathbb{R}^d \times \mathrm{GL}(d)$.
- It holds that
  - The set of transformations is closed under $\circ$.
  - $g_{(0,I)} \circ g_{(\boldsymbol{\mu},\Omega)} = g_{\boldsymbol{\mu},\Omega} \circ g_{(0,I)} = g_{(\boldsymbol{\mu},\Omega)}$.
  - $g_{(-\Omega^{-1}\boldsymbol{\mu},\Omega^{-1})} \circ g_{(\boldsymbol{\mu},\Omega)} = g_{(\boldsymbol{\mu},\Omega)} \circ g_{(-\Omega^{-1}\boldsymbol{\mu},\Omega^{-1})} = g_{(0,I)}$.

  Hence $G = (\{g_\theta : \theta \in \mathbb{R}^d \times \mathrm{GL}(d)\}, \circ)$ is a group (affine group).
- The action of G on $\boldsymbol{X}$ is $g_{(\boldsymbol{\mu},\Omega)}\boldsymbol{X} = \boldsymbol{\mu} + \Omega\boldsymbol{X}$.
- The induced group action on Θ is $(\boldsymbol{\mu}, \Omega) * (\boldsymbol{\nu}, \Psi) = (\boldsymbol{\nu} + \Psi\mu, \Psi\Omega)$.

Summary

We have presented two useful types of parametric models for data:

- The exponential families: defined from a max-entropy principle. Most often, $\boldsymbol{T}(\boldsymbol{X})$ is a complete and minimally sufficient statistic.
- The transformation families, most often of the form $\boldsymbol{X} = \mu + \sigma \boldsymbol{Y}$.

We will further study these two types of models in the remainder of the course. We will focus on exponential families.

Statistical Theory (Week 5): Basic Principles of Point Estimation

Erwan Koch

Ecole Polytechnique Fédérale de Lausanne (Institute of Mathematics)

EPFL

The Problem of Point Estimation

Point Estimation for Parametric Families

Recall our setup:

- A random vector $\boldsymbol{X} = (X_1, ..., X_n)^\top$.
- A family of distributions $\mathcal{F}$ parametrized by $\Theta \subseteq \mathbb{R}^d$, i.e., $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$.
- $\boldsymbol{X} \sim F_\theta \in \mathcal{F}$.

The Problem of Point Estimation

1. Assume that F_θ is known up to the parameter θ which is unknown.
2. Let $(x_1, ..., x_n)^\top$ be a realization of $\boldsymbol{X} \sim F_\theta$ which is available to us.
3. Estimate the value of θ that generates $\boldsymbol{X}$, given $(x_1, ..., x_n)^\top$.

Aspects considered so far in link with point estimation:

- Approximation of the distribution of $g(X_1, ..., X_n)$ by letting $n \uparrow \infty$.
- Appropriate data reduction by studying the information on θ carried by $g(X_1, .., X_n)$.
- Study of general parametric models.

Today: How do we estimate θ in general? Presentation of some general recipes.

Point Estimators

Definition (Point Estimator)

Let $\{F_\theta\}$ be a parametric model with parameter space $\Theta \subseteq \mathbb{R}^d$ and let $\boldsymbol{X} = (X_1, \ldots, X_n)^\top \sim F_{\theta_0}$ for some $\theta_0 \in \Theta$. A point estimator $\hat{\theta}$ of θ_0 is a statistic $T : \mathbb{R}^n \to \Theta$, whose primary purpose is to estimate θ_0.

Therefore any statistic $T : \mathbb{R}^n \to \Theta$ is a candidate estimator!

$\hookrightarrow$ Harder to answer what a *good* estimator is!

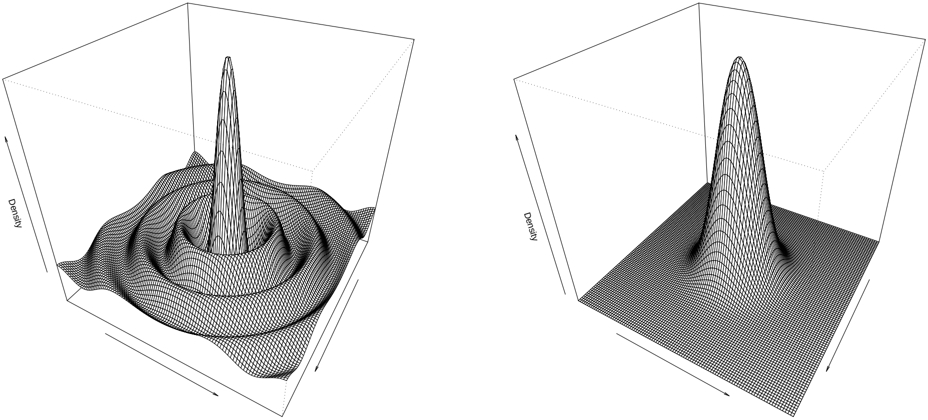
- Any estimator is of course a random variable.
- Hence as a general principle, good should mean:

$$\text{dist}(\hat{\theta}) \text{ concentrated around } \theta.$$

  $\hookrightarrow$ An infinite-dimensional description of quality.

- Look at some simpler measures of quality?

Concentration around a Parameter



Bias, Variance and Mean Squared Error

Bias and Mean Squared Error

Definition (Bias)

The *bias* of an estimator $\hat{\theta}$ of $\theta \in \Theta$ is defined to be

$$\text{bias}(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta}] - \theta.$$

Describes how "off" we are from the target on average when employing $\hat{\theta}$.

Definition (Unbiasedness)

An estimator $\hat{\theta}$ of $\theta \in \Theta$ is *unbiased* if $\mathbb{E}_\theta[\hat{\theta}] = \theta$, i.e., $\text{bias}(\hat{\theta}) = 0$.

We will see that not too much weight should be placed on unbiasedness.

Definition (Mean Squared Error)

The *mean squared error* (MSE) of an estimator $\hat{\theta}$ of $\theta \in \Theta \subseteq \mathbb{R}$ is defined to be

$$\text{MSE}(\hat{\theta}) = \mathbb{E}_\theta\left[(\hat{\theta} - \theta)^2\right].$$

Bias and Mean Squared Error

Bias and MSE combined provide a coarse but simple description of concentration around θ:

- Bias gives us an indication of the location of dist($\hat{\theta}$) relative to θ (somehow assumes that the mean is a good measure of location).
- MSE gives us a measure of spread/dispersion of dist($\hat{\theta}$) around θ.
- If $\hat{\theta}$ is unbiased for $\theta \in \mathbb{R}$ then $\mathsf{MSE}(\hat{\theta}) = \mathsf{Var}(\hat{\theta})$.
- For $\Theta \subseteq \mathbb{R}^d$, $\mathsf{MSE}(\hat{\theta}) := \mathbb{E}[\|\hat{\theta} - \theta\|^2]$, where $\|.\|$ denotes the Euclidean norm.

Example

Let $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ and let $\hat{\mu} := \overline{X}$. Then

$$\mathbb{E}[\hat{\mu}] = \mu \quad \text{and} \quad \mathsf{MSE}(\hat{\mu}) = \mathsf{Var}(\hat{\mu}) = \frac{\sigma^2}{n}.$$

In this case bias and MSE yield a complete description of the concentration of dist($\hat{\mu}$) around μ, since $\hat{\mu}$ is Gaussian and hence completely determined by its mean and its variance.

The Bias-Variance Decomposition of MSE

Bias-Variance Decomposition for $\Theta \subseteq \mathbb{R}$

$$\mathsf{MSE}(\hat{\theta}) = \mathsf{Var}(\hat{\theta}) + \mathsf{bias}^2(\hat{\theta}).$$

Proof.

We have

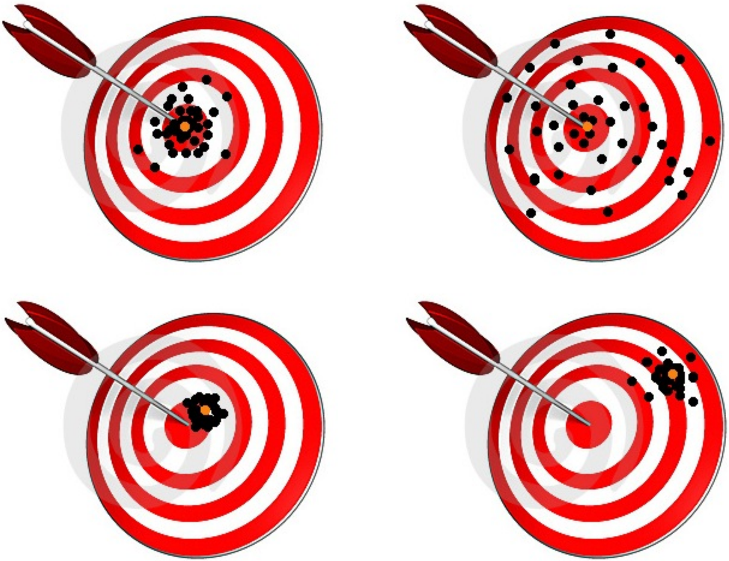
$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right]+\mathbb{E}\left[\hat{\theta}\right]-\theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right]\right)^2+\left(\mathbb{E}\left[\hat{\theta}\right]-\theta\right)^2+2\left(\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right]\right)\left(\mathbb{E}\left[\hat{\theta}\right]-\theta\right)\right] \\
&= \mathbb{E}\left[\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right]\right]^2+\left(\mathbb{E}\left[\hat{\theta}\right]-\theta\right)^2.
\end{aligned}$$

□

The Bias-Variance Decomposition of MSE

- A simple yet fundamental relationship.
- Requiring a small MSE does not necessarily require unbiasedeness.
- Unbiasedeness is a sensible property, but sometimes biased estimators perform better than unbiased ones.
- Sometimes, better to have a bias/variance tradeoff (e.g., in non-parametric regression).

Bias–Variance Tradeoff



Consistency

We can also consider the quality of an estimator not for a given sample size, but as the sample size increases.

Consistency

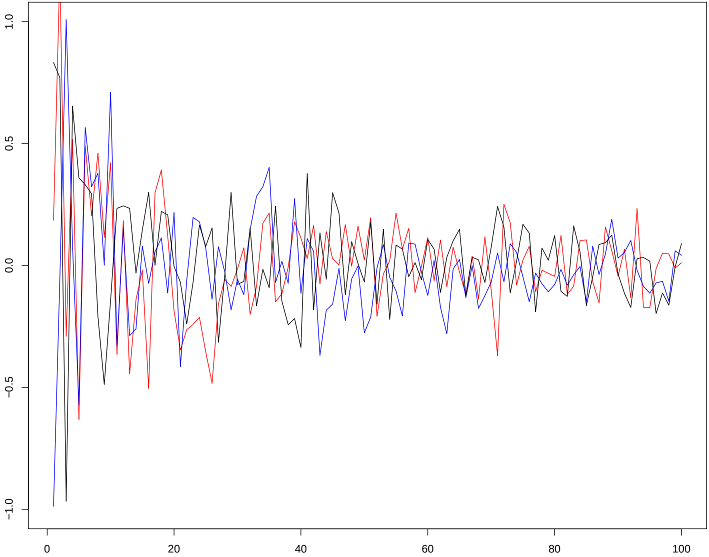
A sequence of estimators $\{\hat{\theta}_n\}_{n\geq 1}$ of $\theta \in \Theta$ is said to be *consistent* if

$$\hat{\theta}_n \xrightarrow{p} \theta.$$

- A consistent estimator becomes increasingly concentrated around the true value θ as the sample size grows (usually, $\hat{\theta}_n$ is an estimator based on n random variables $X_1, \ldots, X_n$).
- Often considered as a "must have" property, but ...
- A more detailed understanding of the "asymptotic quality" of $\hat{\theta}$ requires the study of dist$[\hat{\theta}_n]$ as $n \uparrow \infty$.

Consistency

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(0,1)$. Plots of $\bar{X}_n$ wrt n for 3 different samples.



The Plug-In Principle

Plug-In Estimators

We want to find general procedures for constructing estimators. $\hookrightarrow$ Here we use the definition of a general parameter: a parameter is a function $\nu : \mathcal{F} \to \mathcal{N}$. Under identifiability $\nu(F_\theta) = q(\theta)$, for some $q : \Theta \to \mathcal{N}$.

The Plug-In Principle

Let $\nu(F_\theta)$ be a parameter of interest for a parametric model $\{F_\theta\}_{\theta \in \Theta}$. If we can construct an estimator $\hat{F}$ of F_θ using our sample $\boldsymbol{X}$, then we can use $\nu(\hat{F})$ as an estimator of $\nu(F_\theta)$. Such an estimator is called a *plug-in estimator*.

- In practice such a principle is useful when we can explicitly describe the mapping $F_\theta \mapsto \nu(F_\theta)$.
- In the case of θ, we are essentially "reversing" our point of view: viewing θ as a function of F_θ instead of F_θ as a function of θ, and estimating F_θ instead of θ.
- Note here that $\nu(F_\theta) = \theta = \theta(F_\theta)$ if q is taken to be the identity.

Parameters as Functionals of F

Examples of "functional parameters":

- The mean: $\mu(F) := \int_{-\infty}^{+\infty} x dF(x)$.
- The variance: $\sigma^2(F) := \int_{-\infty}^{+\infty} [x - \mu(F)]^2 dF(x)$.
- The median: $\text{med}(F) := \inf\{x : F(x) \geq 1/2\}$.
- An indirectly defined parameter $\theta(F)$ such that

$$\int_{-\infty}^{+\infty} \psi(x - \theta(F)) dF(x) = 0.$$

- The density (when it exists) at x_0: $\theta(F) := \left. \frac{d}{dx} F(x) \right|_{x=x_0}$.

The Empirical Distribution Function

Plug-in Principle

We need to estimate F. In the case of θ, this principle converts the problem of estimating θ into the problem of estimating F. But how to estimate F?

Consider the case when $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ has iid components. Let F be the distribution function of each X_i. We may define the empirical version of F as

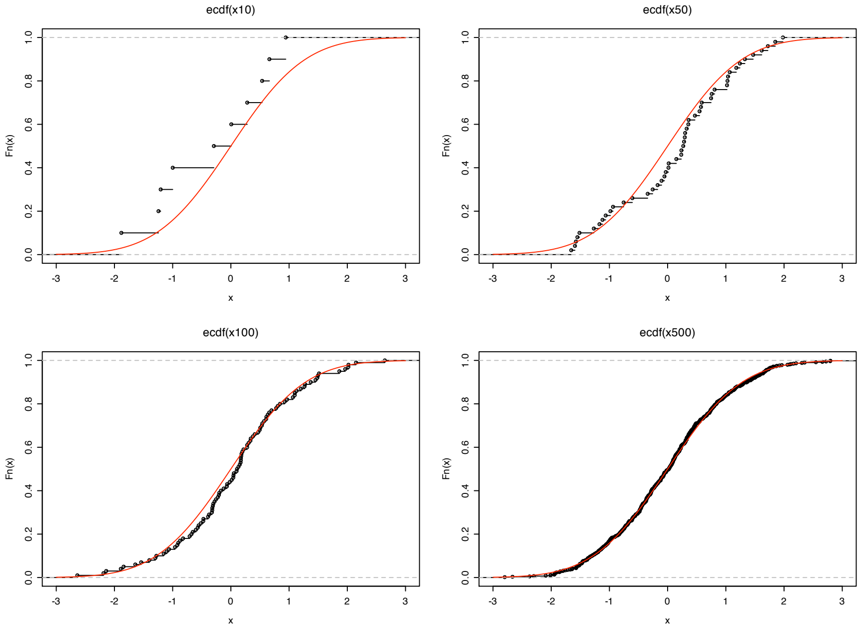
$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\},$$

called the empirical distribution function (edf).

- It places mass $1/n$ on each observation.
- For any $x \in \mathbb{R}$, letting $Y_i = \mathbf{1}\{X_i \leq x\}$, $i = 1 \dots, n$, we have $Y_1, \dots, Y_n \overset{iid}{\sim} \text{Bern}(F(x))$. Thus, the SLLN gives, for any $x \in \mathbb{R}$,

$$\hat{F}_n(x) \xrightarrow{a.s.} F(x).$$

Suggests using $\nu(\hat{F}_n)$ as estimator of $\nu(F)$.


ecdf(x10)
ecdf(x50)
ecdf(x100)
ecdf(x500)
Fn(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3

The Empirical Distribution Function

We are actually doing better than just pointwise convergence!

Theorem (Glivenko-Cantelli)

Let $X_1, \ldots, X_n$ be independent random variables, distributed according to F. Then, $\hat{F}_n(x) = n^{-1}\sum_{i=1}^n \mathbf{1}\{X_i \leq x\}$ converges uniformly to F with probability 1, i.e.,

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{a.s.} 0.$$

Proof.

Assume first that $F(x) = x\mathbf{1}\{0 \leq x \leq 1\}$, i.e, $X_i \sim \text{Unif}(0,1)$. Fix a regular finite partition $0 = x_1 \leq x_2 \leq \ldots \leq x_m = 1$ of $[0,1]$; for any $k = 1, \ldots, m$, $x_{k+1} - x_k = 1/(m-1)$. Using the monotonicity of F and $\hat{F}_n$, it is not too difficult to see that

$$\sup_x |\hat{F}_n(x) - F(x)| < \max_k |\hat{F}_n(x_k) - F(x_{k+1})| + \max_k |\hat{F}_n(x_k) - F(x_{k-1})|.$$

(proof cont'd)

Adding and subtracting $F(x_k)$ within each absolute value and applying the triangle inequality, we can upper-bound the previous expression by

$$2\max_k |\hat{F}_n(x_k) - F(x_k)| + \underbrace{\max_k |F(x_k) - F(x_{k+1})| + \max_k |F(x_k) - F(x_{k-1})|}_{=\max_k |x_k - x_{k+1}| + \max_k |x_k - x_{k-1}| = \frac{2}{m-1}}.$$

Letting $n \uparrow \infty$, the SLLN implies that the first term vanishes a.s. Since m is arbitrary, we have for any $\epsilon > 0$

$$\lim_{n\to\infty} \left[\sup_x |\hat{F}_n(x) - F(x)|\right] < \epsilon \quad a.s.,$$

which gives the result when F is the uniform df.

Let now $X_1, \ldots, X_n \overset{iid}{\sim} F$, where F is a general df (here assumed strictly increasing for simplicity). For $i = 1, \ldots, n$, let $U_i = F(X_i)$. It is clear that $U_1, \ldots, U_n \overset{iid}{\sim} \text{Unif}(0, 1)$.

(proof cont'd).

Letting $\hat{G}_n$ be the edf of $U_1, \ldots, U_n$, we have

$$\hat{F}_n(x) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\} = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{U_i \leq F(x)\} = \hat{G}_n(F(x)), \quad \text{a.s.}$$

In other words, $$\hat{F}_n = \hat{G}_n \circ F, \text{ a.s.}$$

Now let $A = F(\mathbb{R}) \subseteq [0,1]$. From the first part of the proof,

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| = \sup_{t \in A} |\hat{G}_n(t) - t| \leq \sup_{t \in [0,1]} |\hat{G}_n(t) - t| \xrightarrow{a.s.} 0$$

since obviously $A \subseteq [0,1]$. $\square$

Example (Mean of a function)

Consider $\mu_h(F) = \int_{-\infty}^{+\infty} h(x) dF(x)$. A plug-in estimator based on the edf is

$$\hat{\mu}_h := \mu_h(\hat{F}_n) = \int_{-\infty}^{+\infty} h(x) d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} h(X_i).$$

Example (Variance)

Consider now $\sigma^2(F) = \int_{-\infty}^{+\infty} (x - \mu(F))^2 dF(x)$. Plugging in $\hat{F}_n$ gives

$$\sigma^2(\hat{F}_n) = \int_{-\infty}^{+\infty} x^2 d\hat{F}_n(x) - \left(\int_{-\infty}^{+\infty} x d\hat{F}_n(x)\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Exercise

Show that $\sigma^2(\hat{F}_n)$ is a biased but consistent estimator for any F.

Example (Density Estimation)

Let $\theta(F) = f(x_0)$, where f is the density of F. The latter satisfies

$$F(t) = \int_{-\infty}^{t} f(x)dx.$$

If we tried to plug-in $\hat{F}_n$ then our estimator would require differentiation of $\hat{F}_n$ at x_0. Clearly, the edf plug-in estimator does not exist since $\hat{F}_n$ is a step function. We will need a "smoother" estimate of F to plug in, e.g.,

$$\tilde{F}_n(x) := \int_{-\infty}^{\infty} G(x-y)d\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} G(x - X_i)$$

for some continuous df G concentrated closely around 0.

- We saw that plug-in estimators are usually easy to obtain via $\hat{F}_n$.
- But such estimators are not necessarily as "innocent" as they seem.

The Moment Principle

The Method of Moments

Prof. Panaretos: “Perhaps the oldest estimation method (K. Pearson)”.

Method of Moments

Let $X_1, \dots, X_n$ be an iid sample from F_θ, $\theta \in \mathbb{R}^p$. The *method of moments* (MoM) estimator $\hat{\theta}$ of θ is the solution wrt θ to the p random equations

$$\int_{-\infty}^{+\infty} x^{k_j} d\hat{F}_n(x) = \int_{-\infty}^{+\infty} x^{k_j} dF_\theta(x), \quad \{k_j\}_{j=1}^{p} \subset \mathbb{N}.$$

- In some sense this is a plug-in estimator — we estimate the theoretical moments by the sample moments in order to then estimate θ.
- Useful when exact functional form of $\theta(F)$ unavailable.
- While the initially introduced method involves equating moments, it may be generalized to equating p theoretical functionals to their empirical analogues. The choice of the functionals can be important.

Motivational Diversion: The Moment Problem

Theorem

Suppose that F is a distribution determined by its moments. Let $\{F_n\}$ be a sequence of distributions such that $\int x^k dF_n(x) < \infty$ for all n and k. Then,

$$\lim_{n\to\infty} \int x^k dF_n(x) = \int x^k dF(x), \quad \forall\ k \geq 1 \implies F_n \xrightarrow{d} F.$$

BUT: Not all distributions are determined by their moments!

Lemma

The distribution of X is determined by its moments, provided that there exists an open neighbourhood A containing zero such that

$$M_X(u) = \mathbb{E}\left[e^{uX}\right] < \infty, \quad \forall\ u \in A.$$

Example (Exponential Distribution)

Suppose $X_1, ..., X_n \overset{iid}{\sim} \text{Exp}(\lambda)$. Then, for any $r > 0$, $\mathbb{E}[X_i^r] = \lambda^{-r}\Gamma(r+1)$. Hence, we may define a class of estimators of λ depending on r,

$$\hat{\lambda} = \left[\frac{1}{n\Gamma(r+1)} \sum_{i=1}^{n} X_i^r\right]^{-1/r}.$$

Then, we need to tune the value of r to get a "best estimator" (will see later ...).

Example (Gamma Distribution)

Let $X_1, ..., X_n \overset{iid}{\sim} \text{Gamma}(\alpha, \lambda)$. The first two moment equations are

$$\frac{\alpha}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \quad \text{and} \quad \frac{\alpha}{\lambda^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

yielding the estimators $\hat{\alpha} = \bar{X}^2/\hat{\sigma}^2$ and $\hat{\lambda} = \bar{X}/\hat{\sigma}^2$.

Example (Discrete Uniform Distribution)

Let $X_1, ..., X_n \overset{iid}{\sim} \text{Unif}\{1, 2, ..., \theta\}$, for $\theta \in \mathbb{N}$. Using the first moment of the distribution we obtain the equation

$$\bar{X} = \frac{1}{2}(\theta + 1)$$

yielding the MoM estimator $\hat{\theta} = 2\bar{X} - 1$.

A nice feature of MoM estimators is that they generalize to non-iid data.
$\rightarrow$ if $\boldsymbol{X} = (X_1, ..., X_n)^\top$ has distribution depending on $\theta \in \mathbb{R}^p$, one can choose statistics $T_1, ..., T_p$ whose expectations depend on θ:

$$\mathbb{E}_\theta[T_k] = g_k(\theta),$$

and then equate

$$T_k(\boldsymbol{X}) = g_k(\theta), \quad k = 1, \ldots, p.$$

$\rightarrow$ Important here that T_k is a reasonable estimator of $\mathbb{E}[T_k]$.

Comments on Plug-In and MoM Estimators

- Usually easy to compute and can be valuable as preliminary estimates for algorithms that attempt to compute better (but not easily computable) estimates.
- Can give a starting point to search for better estimators in situations where simple intuitive estimators are not available.
- Often these estimators are consistent $\implies$ corresponding estimates likely to be close to the true parameter value for large sample size. Methods of proof for consistency:
  - $\hookrightarrow$ Use empirical process theory for plug-in estimators.
  - $\hookrightarrow$ Estimating equation theory for MoM's.
- Can lead to biased estimators, or even completely ridiculous estimators (see later).

Comments on Plug-In and MoM Estimators

- The estimate provided by an MoM estimator may $\notin \Theta$! (Exercise: show that this can happen with the binomial distribution, with both n and p unknown).
- We will later discuss optimality in estimation, and appropriateness (or inappropriateness) will become clearer.
- Many of these estimators do not depend solely on sufficient statistics.
  $\hookrightarrow$ Sufficiency seems to play an important role in optimality — and it does (more later).
- We now see a method where estimator depends *only* on a sufficient statistic, when such a statistic exists.

The Likelihood Principle

The Likelihood Function

A central theme in statistics. Introduced by Ronald Fisher.

Definition (The Likelihood Function)

Let $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ be a random vector with density (or frequency function) $f(\boldsymbol{x}; \theta)$, $\theta \in \Theta \subseteq \mathbb{R}^p$. The likelihood function $L(\theta)$ is the random function

$$L(\theta) = f(\boldsymbol{X}; \theta).$$

- Notice that we consider L as a function of θ and NOT of $\boldsymbol{X}$.
- Interpretation: Most easily interpreted in the discrete case $\rightarrow$ How likely does the value θ make what we observed? In the the continuous case: how likely does θ make a value in a small neighbourhood of what we observed?
- When $\boldsymbol{X}$ has iid coordinates with density $f(\cdot; \theta)$, then the likelihood is

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta).$$

Maximum Likelihood Estimators

Definition (Maximum Likelihood Estimators)

Let $\boldsymbol{X} = (X_1, ..., X_n)^\top$ be a random vector from F_θ, and suppose that $\hat{\theta}$ is such that

$$L(\hat{\theta}) \geq L(\theta), \quad \forall\, \theta \in \Theta.$$

Then $\hat{\theta}$ is called *a maximum likelihood estimator (MLE) of* θ.

We call $\hat{\theta}$ *the* maximum likelihood estimator, when it is the unique maximum of $L(\theta)$,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta).$$

Intuitively, a maximum likelihood estimator chooses that value of θ which is the most compatible with our observation in the sense that *it makes what we observed most probable*. In not-so-mathematical terms, $\hat{\theta}$ is the value of θ that is most likely to have produced the data.

Comments on MLEs

Saw that MoM and Plug-In estimators often do not depend only on sufficient statistics

$\hookrightarrow$ they also use too much "irrelevant" information.

- If T is a sufficient statistic for θ then the Factorization theorem implies that

$$L(\theta) = g(T(\boldsymbol{X}); \theta)h(\boldsymbol{X}) \propto g(T(\boldsymbol{X}); \theta),$$

i.e., <u>any</u> MLE depends on the data ONLY through the sufficient statistic.

- MLEs are also invariant. If $g : \Theta \to \Theta'$ is a bijection, and if $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.

Comments on MLEs

- When the support of a distribution depends on a parameter, maximization is usually performed by direct inspection.
- For a very broad class of statistical models, the likelihood can be maximized via differential calculus. If Θ is open, the support of the distribution does not depend on θ and the likelihood is differentiable, then the MLE satisfies the log-likelihood equations

$$\nabla_\theta \log L(\theta) = 0.$$

- Maximizing $\log L(\theta)$ is equivalent to maximizing $L(\theta)$.
- When Θ is not open, likelihood equations can be used provided that we verify that the maximum is not reached on the boundary of Θ.

Example (Uniform Distribution)

Let $X_1, ..., X_n \overset{iid}{\sim} \text{Unif}(0, \theta)$. The likelihood is

$$L(\theta) = \theta^{-n} \prod_{i=1}^{n} \mathbf{1}\{0 \leq X_i \leq \theta\} = \theta^{-n} \mathbf{1}\{\theta \geq X_{(n)}\}.$$

Hence if $\theta < X_{(n)}$ the likelihood equals zero and, in the domain $[X_{(n)}, \infty)$, it is a decreasing function of θ. Thus, $\hat{\theta} = X_{(n)}$.

Example (Poisson Distribution)

Let $X_1, ..., X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$. Then,

$$L(\lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{X_i}}{X_i!} e^{-\lambda} \right\}, \text{ giving } \log L(\lambda) = -n\lambda + \log \lambda \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!).$$

Therefore, $\nabla_\lambda \log L(\lambda) = -n + \lambda^{-1} \sum X_i = 0$ we obtain $\hat{\lambda} = \bar{X}$ since $\nabla^2_\lambda \log L(\lambda) = -\lambda^{-2} \sum X_i < 0$.

Statistical Theory (Week 6): Maximum Likelihood Estimation

Erwan Koch

Ecole Polytechnique Fédérale de Lausanne (Institute of Mathematics)

EPFL

The Problem of Point Estimation

Point Estimation for Parametric Families

Recall our setup:

- A random vector $\boldsymbol{X} = (X_1, ..., X_n)^\top$.
- A family of distributions $\mathcal{F}$ parametrized by $\Theta \subseteq \mathbb{R}^d$, i.e., $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$.
- $\boldsymbol{X} \sim F_\theta \in \mathcal{F}$.

The Problem of Point Estimation

1. Assume that F_θ is known up to the parameter θ which is unknown.
2. Let $(x_1, ..., x_n)^\top$ be a realization of $\boldsymbol{X} \sim F_\theta$ which is available to us.
3. Estimate the value of θ that generates $\boldsymbol{X}$, given $(x_1, ..., x_n)^\top$.

Last week, we saw three estimation methods:

- The plug-in method.
- The method of moments.
- The maximum likelihood method.

Today: focus on maximum likelihood. Why does it make sense? What are the properties of the maximum likelihood estimator?

Maximum Likelihood Estimators

Maximum Likelihood Estimators

Recall our definition of a maximum likelihood estimator:

Definition (Maximum Likelihood Estimators)

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ be a random vector from F_θ, and suppose that $\hat{\theta}$ is such that

$$L(\hat{\theta}) \geq L(\theta), \quad \forall\, \theta \in \Theta.$$

Then $\hat{\theta}$ is called *a maximum likelihood estimator (MLE) of* θ.

We call $\hat{\theta}$ *the* maximum likelihood estimator, when it is the unique maximum of $L(\theta)$. We have

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta).$$

$\rightarrow$ $\hat{\theta}$ makes what we observed *most probable*, or, "most likely" $\rightarrow$ Makes sense intuitively. But why should it make sense mathematically?

Kullback-Leibler Divergence

Definition (Kullback-Leibler Divergence)

Let $p(x)$ and $q(x)$ be two probability density (or frequency) functions on $\mathbb{R}$. The *Kullback-Leibler divergence* of q with respect to p is defined as

$$KL(p\|q) := \int_{-\infty}^{+\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx = \mathbb{E}\left[\log\left(\frac{p(X)}{q(X)}\right)\right],$$

where X has $p(x)$ as density (or frequency) function.

- We have $KL(p\|p) = 0$.
- Let $X \sim p(\cdot)$. By Jensen's inequality and using the fact that q integrates to 1, we have

$$KL(p\|q) = \mathbb{E}\{-\log[q(X)/p(X)]\} \geq -\log\left\{\mathbb{E}\left[\frac{q(X)}{p(X)}\right]\right\} = 0.$$

- $p \neq q$ implies that $KL(p\|q) > 0$.

$\Longrightarrow$ KL is, in a sense, a distance between probability distributions.
But KL is not a metric: no symmetry and no triangle inequality!

Relationship with Kullback-Leibler Divergence

Likelihood through KL-divergence

Lemma (Maximum Likelihood as Minimum KL-Divergence)

An estimator $\hat{\theta}$ *based on an iid sample* $X_1, \ldots, X_n$ *is a MLE if and only if* $KL(\hat{F}_n \| F_{\hat{\theta}}) \leq KL(\hat{F}_n \| F_\theta)$ *for all* $\theta \in \Theta$*.*

Proof (discrete case).

Let δ_y be the Dirac measure at y. We recall that $\int h(x) d\hat{F}_n(x) = n^{-1} \sum h(X_i)$, which yields

$$
\begin{aligned}
KL(\hat{F}_n \| F_\theta) = \int_{-\infty}^{+\infty} \log\left(\frac{\sum_{i=1}^n \delta_{X_i}(x)/n}{f(x;\theta)}\right) d\hat{F}_n(x) &= \frac{1}{n}\sum_{i=1}^n \log\left(\frac{n^{-1}}{f(X_i;\theta)}\right) \\
&= -\frac{1}{n}\sum_{i=1}^n \log n - \frac{1}{n}\sum_{i=1}^n \log f(X_i;\theta) \\
&= -\log n - \frac{1}{n}\log\left(\prod_{i=1}^n f(X_i;\theta)\right) \\
&= -\log n - \frac{1}{n}\log L(\theta),
\end{aligned}
$$

which is minimized wrt to θ iff $L(\theta)$ is maximized wrt θ. □

Likelihood through KL-divergence

$\rightarrow$ Therefore, maximizing the likelihood is equivalent to choosing the element of the parametric family $\{F_\theta\}_{\theta\in\Theta}$ that minimizes the KL-divergence with the empirical distribution function.

Intuition:

- $\hat{F}_n$ is (with probability 1) a uniformly good approximation of F_{θ_0}, where θ_0 the true parameter, for large n.
  $\implies$ So F_{θ_0} is "very close" to $\hat{F}_n$ for n large.
- So taking the MLE is equivalent to take the "projection" of $\hat{F}_n$ into $\{F_\theta\}_{\theta\in\Theta}$ as the estimator of F_{θ_0}. The "projection" is with respect to the KL-divergence.

Advanced remarks on KL-divergence:

- $KL(p\|q)$ measures how likely it would be to distinguish if an observation X came from q or p given that it came from p.
- A related quantity is the *entropy* of p, defined as $-\int \log(p(x))p(x)dx$ which measures the "inherent randomness" of p (how "surprising" an outcome from p is on average).

Asymptotic Properties of the MLE

Asymptotic theory for MLEs

- Under what conditions is an MLE consistent?
- How does the distribution of $\hat{\theta}_{MLE}$ concentrate around θ as $n \to \infty$?

In many cases (e.g., when the MLE coincides with an MoM estimator), this can be seen directly.

Example (Geometric distribution)

Let $X_1, \dots, X_n$ be iid Geometric random variables with frequency function

$$f(x;\theta) = \theta(1-\theta)^x, \quad x = 0, 1, 2, \dots$$

It is easy to see that the MLE of θ is

$$\hat{\theta}_n = \frac{1}{\bar{X}_n + 1}.$$

By the central limit theorem, $\sqrt{n}\left[\bar{X}_n - (\theta^{-1} - 1)\right] \xrightarrow{d} N(0, \theta^{-2}(1-\theta))$.

Example (Geometric distribution)

Now applying the delta method with $g(x) = 1/(1+x)$ and thus $g'(x) = -1/(1+x)^2$, we get

$$\sqrt{n}\left[g(\bar{X}_n) - g(\theta^{-1} - 1)\right] \xrightarrow{d} g'(\theta^{-1} - 1)N(0, \theta^{-2}(1-\theta)),$$

and therefore

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \theta^2(1-\theta)).$$

Example (Uniform distribution)

Suppose that $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0, \theta)$. The MLE of θ is

$$\hat{\theta}_n = X_{(n)} = \max\{X_1, \ldots, X_n\}$$

and its df is

$$\mathbb{P}[\hat{\theta}_n \le x] = (x/\theta)^n \mathbf{1}\{x \in [0, \theta]\}.$$

Thus for any $\epsilon > 0$,

$$\mathbb{P}[|\hat{\theta}_n - \theta| > \epsilon] = \mathbb{P}[\hat{\theta}_n < \theta - \epsilon] = \left(\frac{\theta - \epsilon}{\theta}\right)^n \xrightarrow{n \to \infty} 0,$$

so that the MLE is a consistent estimator.

Example (Uniform distribution)

To determine the asymptotic concentration of $\text{dist}(\hat{\theta}_n)$ around θ, we study the magnified difference $n(\theta - \hat{\theta}_n)$. We have

$$\begin{aligned}\mathbb{P}[n(\theta - \hat{\theta}_n) \leq x] &= \mathbb{P}\left[\hat{\theta}_n \geq \theta - \frac{x}{n}\right] \\ &= 1 - \left(1 - \frac{x}{\theta n}\right)^n \\ &\stackrel{n\to\infty}{\longrightarrow} 1 - \exp(-x/\theta),\end{aligned}$$

so that $n(\theta - \hat{\theta}_n)$ weakly converges to an exponential random variable. Thus we understand the concentration of $\text{dist}(\theta - \hat{\theta}_n)$ around zero for large n as that of an exponential distribution with variance θ^2/n^2.

Asymptotic theory for the MLE

From now on, assume that $X_1, \ldots, X_n$ are iid with density (frequency) $f(x;\theta)$, $\theta \in \mathbb{R}$. Notations:

- $\ell(x;\theta) = \log f(x;\theta)$.
- $\ell'(x;\theta)$, $\ell''(x;\theta)$ and $\ell'''(x;\theta)$ are partial derivatives wrt θ.

Regularity Conditions

(A1) Θ is an open subset of $\mathbb{R}$.

(A2) The support of f, supp $f = \{x : f(x;\theta) > 0\}$, is independent of θ.

(A3) f is thrice continuously differentiable wrt θ for all $x \in$ supp f.

(A4) $\mathbb{E}_\theta[\ell'(X_i;\theta)] = 0 \ \forall\theta$ and $\text{Var}_\theta[\ell'(X_i;\theta)] = I(\theta) \in (0,\infty) \ \forall\theta$.

(A5) $-\mathbb{E}_\theta[\ell''(X_i;\theta)] = J(\theta) \in (0,\infty) \ \forall\theta$.

(A6) $\exists\ M(x) > 0$ and $\delta > 0$ such that $\mathbb{E}_{\theta_0}[M(X_i)] < \infty$ and

$$|\theta - \theta_0| < \delta \implies |\ell'''(x;\theta)| \leq M(x).$$

Asymptotic theory for the MLE

- The fact that Θ is open allows any estimator $\hat{\theta}$ to have a symmetric distribution around the true parameter θ (e.g., Gaussian).
- Under (A2) we have, for all $\theta \in \Theta$,

$$\frac{d}{d\theta} \int_{\operatorname{supp} f} f(x;\theta) dx = 0,$$

so that, if we can interchange integration and differentiation,

$$0 = \int \frac{\partial}{\partial\theta} f(x;\theta) dx = \int \ell'(x;\theta) f(x;\theta) dx = \mathbb{E}_\theta[\ell'(X_i;\theta)].$$

Hence, if (A2) is satisfied, (A4) can be seen as a condition that enables one to differentiate once under the integral and states that the random variable $\ell'(X_i;\theta)$ has a finite second moment for any $\theta \in \Theta$.

Asymptotic theory for the MLE

- Similarly, (A5) requires that $\ell''(X_i;\theta)$ has a first moment for all θ.
- (A2) and (A6) are smoothness conditions that will make the "linearization" of the problem useful, while (A4) and (A5) will allow us to "control" the random linearization.
- Furthermore, if we can differentiate twice under the integral, we have

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial\theta}[\ell'(x;\theta)f(x;\theta)]dx \\ &= \int \ell''(x;\theta)f(x;\theta)dx + \int(\ell'(x;\theta))^2 f(x;\theta)dx, \end{aligned}$$

which gives $I(\theta) = J(\theta)$.

Example (Exponential Family)

Let $X_1, \ldots, X_n$ be iid random variables distributed according to a one-parameter exponential family

$$f(x;\theta) = \exp\{c(\theta)T(x) - d(\theta) + S(x)\}, \quad x \in \operatorname{supp} f.$$

It follows that

$$\begin{aligned} \ell'(x;\theta) &= c'(\theta)T(x) - d'(\theta), \\ \ell''(x;\theta) &= c''(\theta)T(x) - d''(\theta). \end{aligned}$$

On the other hand, recall that

$$\begin{aligned} \mathbb{E}_\theta[T(X_i)] &= \frac{d'(\theta)}{c'(\theta)}, \\ \mathrm{Var}_\theta[T(X_i)] &= \frac{1}{[c'(\theta)]^2}\left(d''(\theta) - c''(\theta)\frac{d'(\theta)}{c'(\theta)}\right). \end{aligned}$$

Hence $\mathbb{E}_\theta[\ell'(X_i;\theta)] = c'(\theta)\mathbb{E}_\theta[T(X_i)] - d'(\theta) = 0$.

Example (Exponential Family)

Furthermore,

$$\begin{aligned} I(\theta) &= [c'(\theta)]^2 \text{Var}_\theta[T(X_i)] \\ &= d''(\theta) - c''(\theta)\frac{d'(\theta)}{c'(\theta)}, \end{aligned}$$

and

$$\begin{aligned} J(\theta) &= d''(\theta) - c''(\theta)\mathbb{E}_\theta[T(X_i)] \\ &= d''(\theta) - c''(\theta)\frac{d'(\theta)}{c'(\theta)}, \end{aligned}$$

so that $I(\theta) = J(\theta)$.

Asymptotic Normality of the MLE

Regularity Conditions

(A1) Θ is an open subset of $\mathbb{R}$.

(A2) The support of f, supp f, is independent of θ.

(A3) f is thrice continuously differentiable wrt θ for all $x \in$ supp f.

(A4) $\mathbb{E}_\theta[\ell'(X_i;\theta)] = 0 \ \forall\theta$ and $\text{Var}_\theta[\ell'(X_i;\theta)] = I(\theta) \in (0,\infty) \ \forall\theta$.

(A5) $-\mathbb{E}_\theta[\ell''(X_i;\theta)] = J(\theta) \in (0,\infty) \ \forall\theta$.

(A6) $\exists \ M(x) > 0$ and $\delta > 0$ such that $\mathbb{E}_{\theta_0}[M(X_i)] < \infty$ and

$$|\theta - \theta_0| < \delta \implies |\ell'''(x;\theta)| \leq M(x),$$

where θ_0 is the true value of the parameter.

Asymptotic Normality of the MLE

Theorem (Asymptotic Distribution of the MLE)

Let $X_1, \ldots, X_n$ be iid random variables with density (frequency) $f(x;\theta)$ (θ is the true value of the parameter) and satisfying conditions (A1)-(A6). Suppose that the sequence of MLEs $\hat{\theta}_n$ satisfies $\hat{\theta}_n \xrightarrow{p} \theta$ where

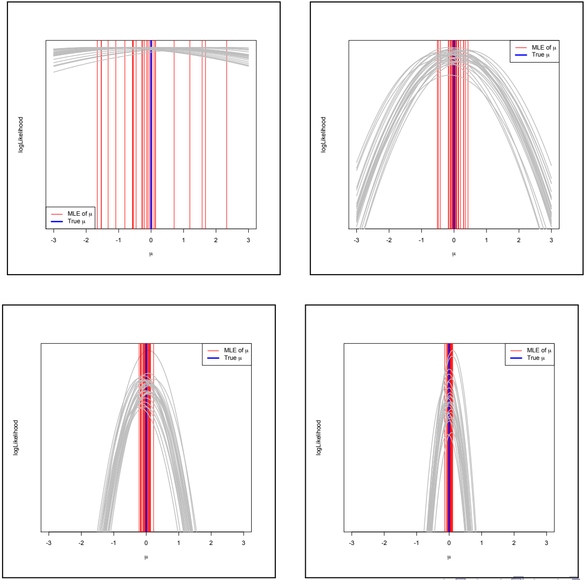
$$\sum_{i=1}^{n} \ell'(X_i; \hat{\theta}_n) = 0, \quad n = 1, 2, \ldots$$

Then,

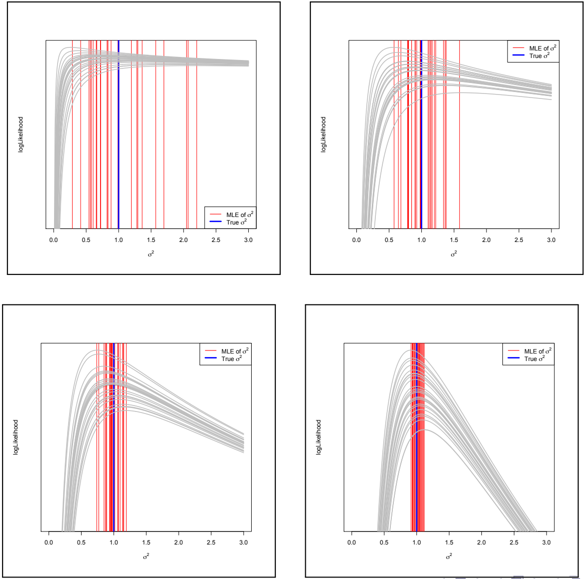
$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{I(\theta)}{J^2(\theta)}\right).$$

When $I(\theta) = J(\theta)$, we have of course $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, 1/I(\theta))$.

Why $I^{-1}(\theta)$? Curvature!



Why $I^{-1}(\theta)$? Curvature!



Proof.

Under Conditions (A1)–(A3), if $\hat{\theta}_n$ maximizes the likelihood, then

$$\sum_{i=1}^{n} \ell'(X_i; \hat{\theta}_n) = 0.$$

Expanding this equation in a Taylor series (centered on the true parameter θ), we get

$$0 = \sum_{i=1}^{n} \ell'(X_i; \hat{\theta}_n) = \sum_{i=1}^{n} \ell'(X_i; \theta) + (\hat{\theta}_n - \theta) \sum_{i=1}^{n} \ell''(X_i; \theta) + \frac{1}{2}(\hat{\theta}_n - \theta)^2 \sum_{i=1}^{n} \ell'''(X_i; \theta_n^*),$$

with θ_n^* lying between θ and $\hat{\theta}_n$.

(proof cont'd)

Dividing accross by $\sqrt{n}$ yields

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell'(X_i;\theta) + \sqrt{n}(\hat{\theta}_n - \theta)\frac{1}{n}\sum_{i=1}^{n}\ell''(X_i;\theta) \\ &\quad + \frac{1}{2}\sqrt{n}(\hat{\theta}_n - \theta)^2\frac{1}{n}\sum_{i=1}^{n}\ell'''(X_i;\theta_n^*), \end{aligned}$$

which gives that $\sqrt{n}(\hat{\theta}_n - \theta)$ equals

$$\frac{-n^{-1/2}\sum_{i=1}^{n}\ell'(X_i;\theta)}{n^{-1}\sum_{i=1}^{n}\ell''(X_i;\theta) + (\hat{\theta}_n - \theta)(2n)^{-1}\sum_{i=1}^{n}\ell'''(X_i;\theta_n^*)}.$$

Now, from (A4) and the CLT, it follows that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell'(X_i;\theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)).$$

(proof cont'd)

Next, the WLLN along with (A5) implies

$$\frac{1}{n}\sum_{i=1}^{n}\ell''(X_i;\theta) \xrightarrow{p} -J(\theta).$$

Now we show that the remainder vanishes in probability, i.e.,

$$R_n = (\hat{\theta}_n - \theta)\frac{1}{2n}\sum_{i=1}^{n}\ell'''(X_i;\theta_n^*) \xrightarrow{p} 0.$$

Since $\hat{\theta}_n - \theta \xrightarrow{p} 0$, this only requires us to prove that $\frac{1}{2n}\sum_{i=1}^{n}\ell'''(X_i;\theta_n^*)$ is bounded.

(proof cont'd)

We want to use condition (A6), which only holds if $|\theta_n^* - \theta| \leq \delta$. First, $|\theta_n^* - \theta| \leq |\hat{\theta}_n - \theta| \xrightarrow{p} 0$, we have $\mathbb{P}(|\theta_n^* - \theta| < \delta) \xrightarrow{n\to\infty} 1$. It easily follows from (A6) that

$$\mathbb{P}\left(\sum_{i=1}^{n} \left|\ell'''(X_i;\theta_n^*)\right| \leq \sum_{i=1}^{n} M(X_i)\right) \xrightarrow{n\to\infty} 1.$$

By the WLLN,

$$\frac{1}{2n}\sum_{i=1}^{n} M(X_i) \xrightarrow{p} E_\theta[M(x)]/2 < \infty.$$

At this point, we would like to use Slutsky's theorem to conclude that

$$R_n = (\hat{\theta}_n - \theta)\frac{1}{2n}\sum_{i=1}^{n} \ell'''(X_i;\theta_n^*) \xrightarrow{p} 0 \times E_\theta[M(x)]/2 = 0,$$

but we cannot really do that because we only have that the second term is bounded with probability tending to one.

(proof cont'd).

Instead, we use the facts that

$$\mathbb{P}\left(|R_n| \leq |\hat{\theta}_n - \theta| \frac{1}{2n} \sum_{i=1}^{n} M(X_i)\right) \overset{n\to\infty}{\longrightarrow} 1,$$

and, from Slutsky's theorem, that

$$|\hat{\theta}_n - \theta| \frac{1}{2n} \sum_{i=1}^{n} M(X_i) \overset{p}{\to} 0.$$

Now, observe that if Y_n and Z_n are sequences of random variables such that $\mathbb{P}(|Y_n| \leq Z_n) \overset{n\to\infty}{\longrightarrow} 1$ and $Z_n \overset{p}{\to} 0$, then $Y_n \overset{p}{\to} 0$. Indeed, for $\epsilon > 0$, we have

$$\begin{aligned}
\mathbb{P}(|Y_n| > \epsilon) &= \mathbb{P}(|Y_n| > \epsilon, |Y_n| \leq Z_n) + \mathbb{P}(|Y_n| > \epsilon, |Y_n| > Z_n) \\
&\leq \mathbb{P}(|Y_n| > \epsilon, |Y_n| \leq Z_n) + \mathbb{P}(|Y_n| > Z_n) \\
&\leq \mathbb{P}(Z_n > \epsilon) + \mathbb{P}(|Y_n| > Z_n) \overset{n\to\infty}{\longrightarrow} 0.
\end{aligned}$$

Consequently, we conclude that $R_n \overset{p}{\to} 0$.
Finally, applying Slutsky's theorem, the continuous mapping theorem and again Slutsky's theorem, yields the result. □

Consistency of the MLE

CRITICALLY!!! The previous theorem assumes that the MLE is consistent and proves that it is then asymptotically Gaussian. Proving consistency can be very hard/frustrating!

Consider the random function

$$\phi_n(t) = \frac{1}{n}\sum_{i=1}^{n}[\log f(X_i; t) - \log f(X_i; \theta)],$$

which is maximized at $t = \hat{\theta}_n$. By the WLLN, for each $t \in \Theta$,

$$\phi_n(t) \xrightarrow{p} \phi(t) = \mathbb{E}\left[\log\left(\frac{f(X_i; t)}{f(X_i; \theta)}\right)\right],$$

which is minus the KL-divergence $KL(f(\cdot; \theta)\|f(\cdot; t))$.

- The latter is minimized when $t = \theta$ and so $\phi(t)$ is maximized at $t = \theta$. Furthermore, $\phi(\theta) = 0$.
- Moreover, unless $f(x; t) = f(x; \theta)$ for all $x \in \operatorname{supp} f$, we have $\phi(t) < 0$.
- Since we are assuming identifiability, it follows that ϕ is uniquely maximized at θ.

Consistency of the MLE

Does the fact that $\phi_n(t) \xrightarrow{p} \phi(t)\ \forall t$, with ϕ_n maximized at $\hat{\theta}_n$ and ϕ maximized uniquely at θ, imply that $\hat{\theta}_n \xrightarrow{p} \theta$? Unfortunately, the answer is in general no.

Example (A Deterministic Example)

Define $\phi_n(t) = \begin{cases} 1 - n|t - n^{-1}| & \text{for } 0 \le t \le 2/n, \\ 1/2 - |t-2| & \text{for } 3/2 \le t \le 5/2, \\ 0 & \text{otherwise.} \end{cases}$

It is easy to see that $\phi_n \to \phi$ pointwise, with

$$\phi(t) = \left[\tfrac{1}{2} - |t-2|\right] \mathbf{1}\{3/2 \le t \le 5/2\}.$$

But now note that ϕ_n is maximized at $t_n = n^{-1}$ with $\phi_n(t_n) = 1$ for all n. On the other hand, ϕ is maximized at $t_0 = 2$.

More assumptions are needed on the $\phi_n(t)$!

Theorem

Suppose that $\{\phi_n(t)\}$ and $\phi(t)$ are real-valued random functions defined on the real line. Suppose that

1. *For each $M > 0$, $\sup_{|t|\leq M} |\phi_n(t) - \phi(t)| \xrightarrow{p} 0$.*
2. *T_n maximizes $\phi_n(t)$ and T_0 is the unique maximizer of $\phi(t)$.*
3. *For any $\epsilon > 0$, there exists M_ϵ such that $\mathbb{P}[|T_n| > M_\epsilon] < \epsilon$ for all n.*

Then, $T_n \xrightarrow{p} T_0$.

If all the ϕ_n and ϕ are concave, we can considerably weaken the assumptions.

Theorem

Suppose that $\{\phi_n(t)\}$ and $\phi(t)$ are random concave functions defined on the real line. Suppose that

1. *$\phi_n(t) \xrightarrow{p} \phi(t)$ for all t.*
2. *T_n maximizes ϕ_n and T_0 is the unique maximizer of ϕ.*

Then, $T_n \xrightarrow{p} T_0$.

Example (Exponential Families)

Let $X_1, \ldots, X_n$ be iid random variables from a one-parameter exponential family

$$f(x;\theta) = \exp\{c(\theta)T(x) - d(\theta) + S(x)\}, \quad x \in \text{supp} f.$$

The MLE of θ maximizes

$$\phi_n(t) = \frac{1}{n}\sum_{i=1}^{n}[c(t)T(X_i) - d(t)].$$

If $c(\cdot)$ is continuous and 1-1 with inverse $c^{-1}(\cdot)$, we can define $u = c(t)$ and consider

$$\phi_n^*(u) = \frac{1}{n}\sum_{i=1}^{n}[uT(X_i) - d_0(u)],$$

where $d_0(u) = d(c^{-1}(u))$. For any n, ϕ_n^* is concave since $(\phi_n^*)''(u) = -d_0''(u)$, which is negative (as $d_0''(u)$ can be written as a variance, see Week 4).

Example (Exponential Families)

Now, by the WLLN, for each u, we have

$$\phi_n^*(u) \xrightarrow{p} u\mathbb{E}[T(X_1)] - d_0(u) = \phi^*(u).$$

Furthermore, $\phi^*(\cdot)$ is concave and $\phi^*(u)$ is maximized when $d_0'(u) = \mathbb{E}[T(X_1)]$. But since (see Week 4)

$$\mathbb{E}[T(X_1)] = d_0'(c(\theta)),$$

ϕ^* is maximized when $d_0'(u) = d_0'(c(\theta))$. The condition holds if we set $u = c(\theta)$, so $c(\theta)$ is a maximizer of ϕ^*. By concavity, it is its unique maximizer.

Now, as $\hat{\theta}_n$ maximizes ϕ_n, $c(\hat{\theta}_n)$ maximizes ϕ_n^*. Hence, the previous theorem yields that $c(\hat{\theta}_n) \xrightarrow{p} c(\theta)$. But as $c(\cdot)$ is 1-1 and continuous, $c^{-1}(\cdot)$ is continuous and thus the continuous mapping theorem implies

$$\hat{\theta}_n \xrightarrow{p} \theta.$$

Summary

- We studied the sampling distribution of the MLE in detail.
- Under some fairly mild assumptions, if the MLE is consistent, then it is asymptotically Gaussian.
- Provided $I(\theta) = J(\theta)$ (which happens very frequently), its asymptotic variance depends on the inverse of the Fisher information $I(\theta)$. We will see later why we distinguished between $I(\theta)$ and $J(\theta)$.
- The asymptotic variance decreases in $1/n$.
- The most difficult problem is to prove the consistency of the MLE. A sufficient condition is the log-likelihood being concave. This typically occurs in exponential families if we work with the natural parameters.

Statistical Theory (Week 7): More on Maximum Likelihood Estimation

Erwan Koch

Ecole Polytechnique Fédérale de Lausanne (Institute of Mathematics)

EPFL

Maximum Likelihood Estimators

Recall our definition of a maximum likelihood estimator:

Definition (Maximum Likelihood Estimators)

Let $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ be a random sample from F_θ, and suppose that $\hat{\theta}$ is such that

$$L(\hat{\theta}) \geq L(\theta), \quad \forall\, \theta \in \Theta.$$

Then $\hat{\theta}$ is called *a maximum likelihood estimator (MLE) of* θ.

Last week, we saw that, under regularity conditions, the distribution of a consistent sequence of MLEs converges weakly to the normal distribution centred around the true parameter value. Today, we focus on the following issues:

- Consistent likelihood equation roots.
- Newton-Raphson and "one-step" estimators.
- The multivariate parameter case.
- What happens if the model has been mis-specified?

Consistent Roots of the Likelihood Equations

Consistent Likelihood Roots

Theorem

Let $\{f(\cdot;\theta)\}_{\theta\in\mathbb{R}}$ be an identifiable parametric class of densities (frequencies) and let $X_1, ..., X_n$ be iid random variables each having density $f(x;\theta_0)$. If the support of $f(\cdot;\theta)$ is independent of θ,

$$\mathbb{P}[L(\theta_0|X_1,\ldots,X_n) > L(\theta|X_1,\ldots,X_n)] \stackrel{n\to\infty}{\longrightarrow} 1$$

for any fixed $\theta \neq \theta_0$.

- Therefore, with high probability, the likelihood of the true parameter exceeds the likelihood of any other choice of parameter, provided that the sample size is large.
- This indicates that extrema of $L(\theta;\boldsymbol{X})$ should have something to do with θ_0 (even though we saw that without further assumptions, a maximizer of L is not necessarily consistent).

Proof.

We introduce the notation $\boldsymbol{X}_n = (X_1, \ldots, X_n)^\top$. We have

$$L(\theta_0 | \boldsymbol{X}_n) > L(\theta | \boldsymbol{X}_n) \iff \frac{1}{n} \sum_{i=1}^{n} \log \left[\frac{f(X_i; \theta)}{f(X_i; \theta_0)} \right] < 0.$$

Now, by the WLLN,

$$\frac{1}{n} \sum_{i=1}^{n} \log \left[\frac{f(X_i; \theta)}{f(X_i; \theta_0)} \right] \xrightarrow{p} \mathbb{E} \left\{ \log \left[\frac{f(X; \theta)}{f(X; \theta_0)} \right] \right\} = -KL(f_{\theta_0} \| f_\theta),$$

which is zero only at θ_0 and negative everywhere else. $\square$

Consistent Sequences of Likelihood Roots

Theorem (Cramér)

Let $\{f(\cdot;\theta)\}_{\theta\in\mathbb{R}}$ be an identifiable parametric class of densities (or frequencies), and Θ open. Let $X_1,\ldots,X_n$ be iid random variables each having density $f(x;\theta_0)$. Assume that the support of $f(\cdot;\theta)$ is independent of θ and that $f(x;\theta)$ is differentiable with respect to θ for (almost) all x. Then, there exists a sequence of random variables ξ_n such that

$$\ell'(X_1,\ldots,X_n;\xi_n)=0,\quad \forall\, n\geq 1,$$

and

$$\xi_n \xrightarrow{p} \theta_0.$$

- In other words, there exists a sequence of roots of the likelihood equations that is consistent for θ_0.
- In general ξ_n is not a statistic (and so not an estimator), since $\xi_n = g(X_1,...,X_n;\theta_0)$ — we need to know the true θ_0 in order to choose which of the likelihood roots to select as our ξ_n for a given sample $(X_1,\ldots,X_n)^\top$.

Proof.

Let $\alpha > 0$ be sufficiently small so that $(\theta_0 - \alpha, \theta_0 + \alpha) \subset \Theta$, and define the set

$$S_n(\alpha, \theta_0) := \left\{ \mathbf{x} \in \mathbb{R}^n : \ell(\mathbf{x}; \theta_0) > \ell(\mathbf{x}; \theta_0 - \alpha) \ \& \ \ell(\mathbf{x}; \theta_0) > \ell(\mathbf{x}; \theta_0 + \alpha) \right\}.$$

If $\mathbf{x} \in S_n(\alpha, \theta_0)$, by continuity of ℓ there exists at least one local maximum of $\ell(\boldsymbol{x}; \theta)$ in $(\theta_0 - \alpha, \theta_0 + \alpha)$, and hence at least one $t \in (\theta_0 - \alpha, \theta_0 + \alpha)$ such that $\ell'(\mathbf{x}; t) = 0$. Define $\tilde{\xi}(\mathbf{x}, \alpha, \theta_0)$ to be the closest local maximum to θ_0 when $\mathbf{x} \in S_n(\alpha, \theta_0)$ and 0 if $\mathbf{x} \notin S_n(\alpha, \theta_0)$.

Now, by our previous theorem, there exists[a] $\alpha_n \downarrow 0$ such that $\mathbb{P}_{\theta_0}[\boldsymbol{X} \in S_n(\alpha_n, \theta_0)] \stackrel{n \to \infty}{\longrightarrow} 1$. Set $\xi_n = \tilde{\xi}(\mathbf{x}, \alpha_n, \theta_0)$ and take $\delta > 0$. Then, for n sufficiently large (so that $\alpha_n < \delta$), we have

$$\mathbb{P}_{\theta_0}[|\xi_n - \theta_0| < \delta] \geq \mathbb{P}_{\theta_0}[|\xi_n - \theta_0| < \alpha_n] \geq \mathbb{P}_{\theta_0}[\boldsymbol{X} \in S_n(\alpha_n, \theta_0)],$$

as $\boldsymbol{X} \in S_n(\alpha_n, \theta_0) \implies |\xi_n - \theta_0| < \alpha_n$. This completes the proof as $\mathbb{P}_{\theta_0}[\boldsymbol{X} \in S_n(\alpha_n, \theta_0)] \stackrel{n \to \infty}{\longrightarrow} 1$. □

[a] Exercise: show this using the same trick as with the Ky-Fan definition of $\stackrel{p}{\to}$.

Corollary (Consistency of Unique Solutions)

Under the assumptions of the previous theorem, if the likelihood equation has a unique root ξ_n for each n and all $\mathbf{x}$, then ξ_n is a valid estimator and is consistent for θ_0.

- The statement remains true if the uniqueness requirement is substituted with the requirement that the probability of multiple roots tends to zero as $n \to \infty$.
- The statement does not claim that the root corresponds to a maximum: it merely requires that we have a root.
- On the other hand, even when the root is unique, the corollary says nothing about its properties for finite n.

Example (Minimum Likelihood Estimation)

Let X take the values $0, 1, 2$ with probabilities $6\theta^2 - 4\theta + 1$, $\theta - 2\theta^2$ and $3\theta - 4\theta^2$ ($\theta \in (0, 1/2)$). Then, the likelihood equation has a unique root for all x, which is a minimum for $x = 0$ and a maximum for $x = 1, 2$.

Consistent Sequences of Likelihood Roots

- Cramér's theorem does not tell us *which* root to choose, so not useful in practice.
- The easiest case is when the root is unique!
- Otherwise, we need some "external help" (non-MLE help)...

Fortunately, if some "good" estimator is already available, then ...

Lemma

Let α_n be any consistent sequence of estimators of the true parameter θ. For each n, let θ_n^* denote the root of the likelihood equations that is the closest to α_n. Then, under the assumptions of Cramér's theorem, $\theta_n^* \xrightarrow{p} \theta$.

Exercise: prove the lemma.

- Therefore, when the likelihood equations do not have a single root, we may still choose a root based on some estimator that is readily available.
  - $\hookrightarrow$ Only requires that the estimator used is consistent.
  - $\hookrightarrow$ Often the case with Plug-In or MoM estimators.

Very often, the roots are not available in closed form. In these cases, an iterative approach is required to approximate them.

Approximate Solution of the Likelihood Equations

The Newton-Raphson Algorithm

We wish to solve the equation

$$\ell'(\theta) = 0.$$

Assuming that $\tilde{\theta}$ is close to a root $\hat{\theta}$ (which is perhaps a consistent estimator), a second-order Taylor expansion yields

$$0 = \ell'(\hat{\theta}) \simeq \ell'(\tilde{\theta}) + (\hat{\theta} - \tilde{\theta})\ell''(\tilde{\theta}),$$

which gives

$$\hat{\theta} \simeq \tilde{\theta} - \frac{\ell'(\tilde{\theta})}{\ell''(\tilde{\theta})}.$$

The procedure can then be iterated by replacing $\tilde{\theta}$ by the right hand side of the above relation. In principle, each iteration improves the finite sample accuracy of the estimator. But in terms of asymptotic behaviour, a single iteration suffices!

Construction of Asymptotically MLE-like Estimators

Theorem

Suppose that Assumptions (A1)–(A6) hold and let $\tilde{\theta}_n$ be a consistent estimator of θ_0 such that $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ is bounded in probability (i.e., $\tilde{\theta}_n$ is a $\sqrt{n}-$consistent estimator). Then, the sequence of estimators

$$\delta_n = \tilde{\theta}_n - \ell'(\tilde{\theta}_n)/\ell''(\tilde{\theta}_n)$$

satisfies

$$\sqrt{n}(\delta_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta)/J(\theta)^2).$$

- With a single Newton-Raphson step, we may obtain an estimator (the so-called "one-step" estimator) that, asymptotically, behaves like a consistent MLE (provided that we start with a $\sqrt{n}-$consistent estimator).
- The "one-step" estimator does not necessarily behave like an MLE for finite n!
- The one-step δ_n satisfies the conditions of the theorem (i.e., is consistent and bounded in probability). Hence iterating to get $\zeta_n = \delta_n - \ell'(\delta_n)/\ell''(\delta_n)$ also leads to the same conclusion.

Proof.

A Taylor expansion around the true value, θ_0, yields

$$\ell'(\tilde{\theta}_n) = \ell'(\theta_0) + (\tilde{\theta}_n - \theta_0)\ell''(\theta_0) + \tfrac{1}{2}(\tilde{\theta}_n - \theta_0)^2\ell'''(\theta_n^*),$$

where θ_n^* between θ_0 and $\tilde{\theta}_n$. Substituting this expression into the definition of δ_n yields

$$\begin{aligned}\sqrt{n}(\delta_n - \theta_0) = {} & \frac{(1/\sqrt{n})\ell'(\theta_0)}{-(1/n)\ell''(\tilde{\theta}_n)} + \sqrt{n}(\tilde{\theta}_n - \theta_0) \\ & \times \left[1 - \frac{\ell''(\theta_0)}{\ell''(\tilde{\theta}_n)} - \frac{1}{2}(\tilde{\theta}_n - \theta_0)\frac{\ell'''(\theta_n^*)}{\ell''(\tilde{\theta}_n)}\right].\end{aligned}$$

Exercise

Use CLT/LLN/Slutsky to complete the proof. Hint: by Taylor expansion,

$$\tfrac{1}{n}\ell''(\tilde{\theta}_n) = \tfrac{1}{n}\textstyle\sum_i \ell''(X_i;\tilde{\theta}_n) = \tfrac{1}{n}\sum_i \ell''(X_i;\theta_0) + (\tilde{\theta}_n - \theta_0)\tfrac{1}{n}\sum_i \ell'''(X_i;\theta_0).$$

The Multiparameter Case

The Multiparameter Case

$\rightarrow$ Extension of asymptotic results to multiparameter models easy under similar assumptions, but notationally cumbersome. $\rightarrow$ Same ideas: the MLE will be a zero of the likelihood equations

$$\sum_{i=1}^{n} \nabla \ell(X_i; \boldsymbol{\theta}) = 0$$

A Taylor expansion can be formed

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla \ell(X_i; \boldsymbol{\theta}) + \left(\frac{1}{n} \sum_{i=1}^{n} \nabla^2 \ell(X_i; \boldsymbol{\theta}_n^*) \right) \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}).$$

Under regularity conditions we should have

- $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla \ell(X_i; \boldsymbol{\theta}) \xrightarrow{d} N_p(0, \text{Cov}[\nabla \ell(X_i; \boldsymbol{\theta})])$.
- $\frac{1}{n} \sum_{i=1}^{n} \nabla^2 \ell(X_i; \boldsymbol{\theta}_n^*) \xrightarrow{p} \mathbb{E}[\nabla^2 \ell(X_i; \boldsymbol{\theta})]$.

The Multiparameter Case

Regularity Conditions

(B1) The parameter space $\Theta \in \mathbb{R}^p$ is open.

(B2) The support of $f(\cdot;\boldsymbol{\theta})$, supp $f(\cdot;\boldsymbol{\theta})$, is independent of $\boldsymbol{\theta}$.

(B3) All mixed partial derivatives of ℓ wrt $\boldsymbol{\theta}$ up to degree 3 exist and are continuous.

(B4) $\mathbb{E}[\nabla\ell(X_i;\boldsymbol{\theta})] = 0\ \forall\boldsymbol{\theta}$ and $\mathrm{Cov}[\nabla\ell(X_i;\boldsymbol{\theta})] =: I(\boldsymbol{\theta}) \succ 0\ \forall\boldsymbol{\theta}$.

(B5) $-\mathbb{E}[\nabla^2\ell(X_i;\boldsymbol{\theta})] =: J(\boldsymbol{\theta}) \succ 0\ \forall\boldsymbol{\theta}$.

(B6) $\exists\delta > 0$ s.t. $\forall\boldsymbol{\theta} \in \Theta$ and for all $1 \leq j, k, l \leq p$,

$$\left|\frac{\partial}{\partial\theta_j\partial\theta_k\partial\theta_l}\ell(x;\boldsymbol{u})\right| \leq M_{jkl}(x)$$

for $\|\boldsymbol{\theta} - \boldsymbol{u}\| \leq \delta$ with M_{jkl} such that $\mathbb{E}[M_{jkl}(X_i)] < \infty$.

The interpretation of these conditions is the same as in the one-dimensional case.

Theorem (Asymptotic Normality of the MLE)

Let $X_1, \ldots, X_n$ be iid random variables with density (frequency) $f(x; \boldsymbol{\theta})$, satisfying conditions (B1)-(B6). If $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}(X_1, \ldots, X_n)$ is a consistent sequence of MLEs, then

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}_p(\mathbf{0}, J^{-1}(\boldsymbol{\theta}) I(\boldsymbol{\theta}) J^{-1}(\boldsymbol{\theta})).$$

- The theorem remains true if each X_i is a random vector.
- The proof mimics that of the one-dimensional case.

Misspecified Models and Likelihood

Misspecification of Models

- Statistical models are typically merely approximations to reality.
- George P. Box: "*All models are wrong, but some are useful*."

As worrying as this may seem, it may not be a problem in practice.

- Often the model is wrong, but is "close enough" to the true situation.
- Even if the model is wrong, the parameters often admit a fruitful interpretation in the context of the problem.

Example

Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Exp}(\lambda)$. However, assume that we decide that the appropriate model for our data is given by the two-parameter family of densities

$$f(x; \alpha, \theta) = \frac{\alpha}{\theta}\left(1 + \frac{x}{\theta}\right)^{-(\alpha+1)}, \quad x > 0,$$

where α and θ are positive unknown parameters to be estimated.

Example (cont'd)

- Notice that the exponential distribution is not a member of this parametric family.
- However, letting $\alpha, \theta \to \infty$ such that $\alpha/\theta \to \lambda$, we have

$$f(x; \alpha, \theta) \to \lambda \exp(-\lambda x).$$

  Thus, we may *approximate* the true model from within this class. Reasonable $\hat{\alpha}$ and $\hat{\theta}$ will yield a density "close" to the true density.

Example

Let $X_1, \ldots, X_n$ be independent random variables with variance σ^2 and mean

$$\mathbb{E}[X_i] = \alpha + \beta t_i.$$

If we assume that the X_i are normal when they are in fact not, the MLEs of the parameters α, β, σ^2 remain good (in fact optimal in a sense) for the true parameters (Gauss-Markov theorem).

Misspecified Models and Likelihood

The Framework

- $X_1, \ldots, X_n$ are iid random variables with distribution function F and density (or frequency) function g.
- We build a MLE assuming that the X_i admit a density in $\{f(x;\theta)\}_{\theta \in \Theta}$.
- The true density g does not correspond to any of the $\{f_\theta\}$.

Let $\hat{\theta}_n$ be a root of the likelihood equation,

$$\sum_{i=1}^{n} \ell'(X_i; \hat{\theta}_n) = 0,$$

where the log-likelihood $\ell(\theta)$ is wrt $f(\cdot;\theta)$.

- What exactly is $\hat{\theta}_n$ estimating?
- What is the behaviour of the sequence $\{\hat{\theta}_n\}_{n \geq 1}$ as $n \to \infty$?

Misspecified Models and Likelihood

Consider the functional parameter $\theta(F)$ defined by

$$\int_{-\infty}^{+\infty} \ell'(x; \theta(F)) dF(x) = 0.$$

Then, the plug-in estimator of $\theta(F)$ when using the edf $\hat{F}_n$ as an estimator of F is given by solving

$$\int_{-\infty}^{+\infty} \ell'(x; \theta(\hat{F}_n)) d\hat{F}_n(x) = 0 \iff \sum_{i=1}^{n} \ell'(X_i; \hat{\theta}_n) = 0,$$

so that the MLE is a plug-in estimator of $\theta(F)$.

Model Misspecification and the Likelihood

Theorem

Let $X_1, ..., X_n \overset{iid}{\sim} F$ *and let* $\hat{\theta}_n$ *be a random variable solving the equation* $\sum_{i=1}^{n} \ell'(X_i; \theta) = 0$ *for* θ *in the open set* Θ*. Assume that*

(a) ℓ' *is a strictly monotone function on* Θ *for each* x*.*

(b) $\int_{-\infty}^{+\infty} \ell'(x;\theta) dF(x) = 0$ *has a unique solution* $\theta = \theta(F)$ *on* Θ*.*

(c) $I(F) := \int_{-\infty}^{+\infty} [\ell'(x;\theta(F))]^2 dF(x) < \infty$*.*

(d) $J(F) := -\int_{-\infty}^{+\infty} \ell''(x;\theta(F)) dF(x) < \infty$*.*

(e) $|\ell'''(x;t)| \leq M(x)$ *for* $t \in (\theta(F) - \delta, \theta(F) + \delta)$*, some* $\delta > 0$ *and* $\int_{-\infty}^{+\infty} M(x) dF(x) < \infty$*.*

Then

$$\hat{\theta}_n \xrightarrow{p} \theta(F)$$

and

$$\sqrt{n}(\hat{\theta}_n - \theta(F)) \xrightarrow{d} \mathcal{N}(0, I(F)/J^2(F)).$$

Proof.

Assume without loss of generality that $\ell'(x;\theta)$ is strictly decreasing in θ. Let $\epsilon > 0$ and observe that

$$\begin{aligned}\mathbb{P}[|\hat{\theta}_n - \theta(F)| > \epsilon] &= \mathbb{P}\left[\left\{\hat{\theta}_n - \theta(F) > \epsilon\right\} \cup \left\{\theta(F) - \hat{\theta}_n > \epsilon\right\}\right] \\ &\leq \mathbb{P}\left[\left\{\hat{\theta}_n - \theta(F) > \epsilon\right\}\right] + \mathbb{P}\left[\left\{\theta(F) - \hat{\theta}_n > \epsilon\right\}\right].\end{aligned}$$

By our monotonicity assumption, we have

$$\hat{\theta}_n - \theta(F) > \epsilon \implies \theta(F) + \epsilon < \hat{\theta}_n \implies \frac{1}{n}\sum_{i=1}^{n} \ell'(X_i;\theta(F)+\epsilon) > 0$$

because $\hat{\theta}_n$ is the solution to the equation $\frac{1}{n}\sum_{i=1}^{n} \ell'(X_i;\theta) = 0$. Similarly,

$$\theta(F) - \hat{\theta}_n > \epsilon \implies \theta(F) - \epsilon > \hat{\theta}_n \implies \frac{1}{n}\sum_{i=1}^{n} \ell'(X_i;\theta(F)-\epsilon) < 0.$$

Hence $$\mathbb{P}[|\hat{\theta}_n - \theta(F)| > \epsilon] \le \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \ell'(X_i;\theta(F)+\epsilon) > 0\right] + \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \ell'(X_i;\theta(F)-\epsilon) < 0\right].$$

We may re-write the first term on the right-hand side as

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \ell'(X_i;\theta(F)+\epsilon) > 0\right] = \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \ell'(X_i;\theta(F)+\epsilon) - \int_{-\infty}^{\infty} \ell'(x;\theta(F)+\epsilon)dF(x) > -\int_{-\infty}^{\infty} \ell'(x;\theta(F)+\epsilon)dF(x)\right].$$

We will show that this probability converges to zero. Define

$$W_n = \frac{1}{n}\sum_{i=1}^{n} \ell'(X_i;\theta(F)+\epsilon) - \int_{-\infty}^{\infty} \ell'(x;\theta(F)+\epsilon)dF(x)$$

$$\kappa = -\int_{-\infty}^{\infty} \ell'(x;\theta(F)+\epsilon)dF(x).$$

First of all, we claim that $\kappa > 0$. To see this, note that (a) implies that

$$\begin{aligned} -\ell'(x;\theta(F)) &< -\ell'(x;\theta(F)+\epsilon), \qquad \forall x \\ \Longrightarrow -\int_{-\infty}^{\infty} \ell'(x;\theta(F))dF(x) &< -\int_{-\infty}^{\infty} \ell'(x;\theta(F)+\epsilon)dF(x). \end{aligned}$$

since $\theta(F) < \theta(F)+\epsilon$. So $\kappa > 0$ since LHS is zero by assumption (b). By assumption (c) we can use the WLLN to conclude that

$$\frac{1}{n}\sum_{i=1}^{n} \ell'(X_i;\theta(F)+\epsilon) \xrightarrow{p} \int_{-\infty}^{\infty} \ell'(x;\theta(F)+\epsilon)dF(x).$$

and, by Slutsky's theorem we conclude that

$$W_n \xrightarrow{p} 0.$$

By definition of convergence in probability, and since $\kappa > 0$, we conclude

$$\mathbb{P}[W_n > \kappa] \leq \mathbb{P}\left[\{W_n > \kappa\} \cup \{-W_n > \kappa\}\right] = \mathbb{P}[|W_n| > \kappa] \xrightarrow{n\to\infty} 0.$$

Similar arguments give

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\ell'(X_i;\theta(F)-\epsilon)<0\right]\to 0$$

and thus

$$\hat{\theta}_n \xrightarrow{p} \theta(F).$$

Expanding the equation that defines the estimator in a Taylor series, gives

$$\begin{aligned}0=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell'(X_i;\hat{\theta}_n) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell'(X_i;\theta(F))+ \\ &\quad +\sqrt{n}(\hat{\theta}_n-\theta(F))\frac{1}{n}\sum_{i=1}^{n}\ell''(X_i;\theta(F)) \\ &\quad +\sqrt{n}(\hat{\theta}_n-\theta(F))^2\frac{1}{2n}\sum_{i=1}^{n}\ell'''(X_i;\theta_n^*).\end{aligned}$$

Here, θ_n^* lies between $\theta(F)$ and $\hat{\theta}_n$.

Exercise: complete the proof by mimicking the proof of asymptotic normality of MLEs. □

- The result extends immediately to the multivariate parameter case.
- Notice that the proof is essentially identical to MLE asymptotics proof.
- The difference is the first part, where we show consistency.
- This is where assumptions (a) and (b) come in.
- These can be replaced by any set of assumptions yielding consistency.

Model Misspecification and the Likelihood

What is the interpretation of the parameter $\theta(F)$ in the misspecified setup?

Suppose that F has density (frequency) g and assume that integration/differentiation may be interchanged:

$$\int_{-\infty}^{+\infty} \frac{d}{d\theta} \log f(x;\theta) dF(x) = 0 \iff \frac{d}{d\theta} \int_{-\infty}^{+\infty} \log f(x;\theta) dF(x) = 0$$

$$\iff \frac{d}{d\theta} \left[\int_{-\infty}^{+\infty} \log f(x;\theta) dF(x) - \int_{-\infty}^{+\infty} \log g(x) dF(x) \right] = 0$$

$$\iff \frac{d}{d\theta} KL(g(x) \| f(x;\theta)) = 0$$

- We are minimizing the KL-distance between the true model F and our model.
- Hence we may intuitively think of the $\theta(F)$ as the element of Θ for which f_θ is "closest" to g in the KL-sense.

Summary

- Last week, we talked about the MLE which is asymptotically Gaussian if it is consistent. Consistency proved slightly hard to study.
- This week, we showed that by adding a small Newton-Raphson correction to a $\sqrt{n}$-consistent estimator $\hat{\theta}$, we obtain a true estimator that is $\sqrt{n}$-consistent and asymptotically Gaussian.
- We also considered what happens when the true model is not inside our parametric family:
  - We are trying to infer the best approximation of the truth inside our model class, given by $\theta(F)$.
  - Up to possible issues of consistency, the MLE correctly recovers $\theta(F)$ and is asymptotically Gaussian.

Statistical Theory (Week 8): The Decision Theory Framework

Erwan Koch

Ecole Polytechnique Fédérale de Lausanne (Institute of Mathematics)

Statistics as a Random Game

Statistics as a Random Game?

Nature and a statistician decide to play a game. What's in the box?

- A *family of distributions* $\mathcal{F}$, usually assumed to admit densities (or frequencies). This is the variant of the game we decide to play.
- A *parameter space* $\Theta \subseteq \mathbb{R}^p$ which parametrizes the family, i.e., $\mathcal{F} = \{F_\theta\}_{\theta \in \Theta}$. This represents the space of possible *plays/moves* available to Nature.
- A *data space* $\mathcal{X}$, on which the parametric family is supported. This represents the space of possible outcomes following a play by Nature.
- An *action space* $\mathcal{A}$, which represents the space of possible *actions* or *decisions* or *plays/moves* available to the statistician.
- A *loss function* $\mathcal{L} : \Theta \times \mathcal{A} \to \mathbb{R}^+$. This represents how much the statistician has to pay nature when losing.
- A set $\mathcal{D}$ of *decision rules*. Any $\delta \in \mathcal{D}$ is a (measurable) function $\delta : \mathcal{X} \to \mathcal{A}$. All these decision rules represent the possible strategies available to the statistician.

Statistics as a Random Game?

How the game is played:

- First we agree on the rules of the game:
  1. We fix a parametric family $\{F_\theta\}_{\theta \in \Theta}$.
  2. We fix an action space $\mathcal{A}$.
  3. We fix a loss function $\mathcal{L}$.
- Then we play:
  1. Nature selects (plays) $\theta_0 \in \Theta$.
  2. The statistician observes $\boldsymbol{X} \sim F_{\theta_0}$.
  3. The statistician plays $\delta(\boldsymbol{X}) \in \mathcal{A}$ in response.
  4. The statistician has to pay Nature $\mathcal{L}(\theta_0, \delta(\boldsymbol{X}))$.

Framework proposed by A. Wald in 1939. Encompasses three basic statistical problems:

- Point estimation.
- Interval estimation.
- Hypothesis testing.

Point Estimation as a Game

In the problem of point estimation we have:

1. A fixed parametric family $\{F_\theta\}_{\theta \in \Theta}$.
2. A fixed action space $\mathcal{A} = \Theta$.
3. A fixed loss function $\mathcal{L}(\theta, \alpha)$; e.g., $\|\theta - \alpha\|^2$.

The game now evolves simply as:

1. Nature picks $\theta_0 \in \Theta$.
2. The statistician observes $\boldsymbol{X} \sim F_{\theta_0}$.
3. The statistician plays $\delta(\boldsymbol{X}) \in \mathcal{A} = \Theta$.
4. The statistician loses $\mathcal{L}(\theta_0, \delta(\boldsymbol{X}))$.

Notice that in this setup, δ is an *estimator* (it is a statistic $\mathcal{X} \to \Theta$).

The statistician always loses.

$\hookrightarrow$ Is there a good strategy $\delta \in \mathcal{D}$ for the statistician to restrict his losses?

$\hookrightarrow$ Is there an optimal strategy?

Risk of a Decision Rule

Risk of a Decision Rule

The statistician would like to choose a strategy δ so as to minimize his losses. But losses are random since they depend on $\boldsymbol{X}$.

Definition (Risk)

Given a parameter $\theta \in \Theta$, the *risk* of a decision rule $\delta : \mathcal{X} \to \mathcal{A}$ is the expected loss incurred when employing δ: $R(\theta, \delta) = \mathbb{E}_\theta \left[\mathcal{L}(\theta, \delta(\boldsymbol{X})) \right]$.

Key notion of decision theory

Decision rules should be compared by comparing their risk functions.

Example (Mean Squared Error)

In point estimation, the mean squared error

$$\mathrm{MSE}_\theta(\delta(\boldsymbol{X})) = \mathbb{E}_\theta \left[\|\theta - \delta(\boldsymbol{X})\|^2 \right]$$

is the risk corresponding to a squared error loss function.

Coin Tossing Revisited

Consider the "coin tossing game" with squared error loss:

- Nature picks $\theta \in [0, 1]$.
- We observe n variables $X_i \overset{iid}{\sim} \text{Bern}(\theta)$.
- The action space is $\mathcal{A} = [0, 1]$.
- The loss function is $\mathcal{L}(\theta, \alpha) = (\theta - \alpha)^2$.

We consider 3 different decision rules $\{\delta_j\}_{j=1,2,3}$:

1. $\delta_1(\boldsymbol{X}) = \frac{1}{n} \sum_{i=1}^{n} X_i$.
2. $\delta_2(\boldsymbol{X}) = X_1$.
3. $\delta_3(\boldsymbol{X}) = 1/2$.

Let us compare these using their associated risks as benchmarks.

Coin Tossing Revisited

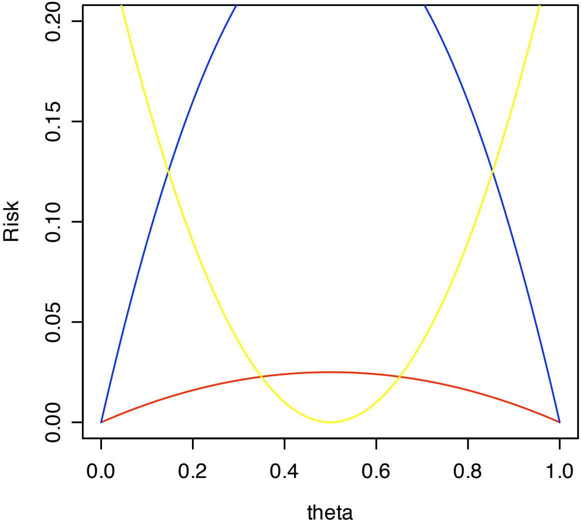
We consider the risks associated with the different decision rules:

$$R_j(\theta) = R(\theta, \delta_j(\boldsymbol{X})) = \mathbb{E}_\theta \left[(\theta - \delta_j(\boldsymbol{X}))^2 \right], \quad j = 1, 2, 3.$$

We easily obtain

- $R_1(\theta) = \frac{1}{n}\theta(1 - \theta)$.
- $R_2(\theta) = \theta(1 - \theta)$.
- $R_3(\theta) = \left(\theta - \frac{1}{2}\right)^2$.

Coin Tossing Revisited – Every dog has its day



$R_1(\theta)$, $R_2(\theta)$, $R_3(\theta)$

Admissibility and Inadmissibility

Inadmissible Decision Rules

Definition (Inadmissible Decision Rule)

Let δ be a decision rule for the experiment $(\{F_\theta\}_{\theta\in\Theta}, \mathcal{L})$. If there exists a decision rule δ^* that strictly dominates δ, i.e.,

$$R(\theta,\delta^*) \le R(\theta,\delta),\ \forall\theta\in\Theta \quad \& \quad \exists\, \theta'\in\Theta : R(\theta',\delta^*) < R(\theta',\delta),$$

then δ is called an *inadmissible decision rule*.

- An inadmissible decision rule is a "silly" strategy since we can find a strategy that always does at least as well and sometimes better.
- However "silly" is with respect to $\mathcal{L}$ and Θ. It may be that our choice of $\mathcal{L}$ is "silly"!!!
- If we change the rules of the game (i.e., different loss function or different parameter space) then domination may break down.

For example, $R_2(\theta)$ is inadmissible as $R_2(\theta) > R_1(\theta)$ for any $\theta\in(0,1)$, $R_2(0) = R_1(0) = 0$ and $R_2(1) = R_1(1) = 0$.

Inadmissible Decision Rules

Example (Exponential Distribution)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\lambda)$, $n \geq 2$. It is easy to see that the MLE of λ is

$$\hat{\lambda} = 1/\bar{X}_n,$$

where $\bar{X}_n$ is the empirical mean. It can be shown that

$$\mathbb{E}_\lambda[\hat{\lambda}] = \frac{n\lambda}{n-1},$$

which yields that $\tilde{\lambda} = (n-1)\hat{\lambda}/n$ is an unbiased estimator of λ. Observe now that

$$\text{MSE}_\lambda(\tilde{\lambda}) < \text{MSE}_\lambda(\hat{\lambda})$$

since $\tilde{\lambda}$ is unbiased and $\text{Var}_\lambda(\tilde{\lambda}) < \text{Var}_\lambda(\hat{\lambda})$. Hence the MLE is an inadmissible rule for the squared error loss.

Inadmissible Decision Rules

Example (Exponential Distribution)

The parameter space in this example is $(0, \infty)$, in which case a quadratic loss tends to penalize over-estimation more heavily than under-estimation (the maximum possible under-estimation is bounded!). Taking a different loss function might change the result! Now, instead, we consider the loss function

$$\mathcal{L}(a, b) = a/b - 1 - \log(a/b),$$

which satisfies, for each fixed a, $\lim_{b \to 0} \mathcal{L}(a, b) = \lim_{b \to \infty} \mathcal{L}(a, b) = \infty$. Now, using the fact that

$$\frac{n\lambda\bar{X}_n}{n-1} = \lambda\bar{X}_n + \frac{\lambda\bar{X}_n}{n-1},$$

we obtain, for $n > 1$,

$$\begin{aligned} R(\lambda, \tilde{\lambda}) &= \mathbb{E}_\lambda \left[\frac{n\lambda\bar{X}_n}{n-1} - 1 - \log\left(\frac{n\lambda\bar{X}_n}{n-1} \right) \right] \\ &= \underbrace{\mathbb{E}_\lambda \left[\lambda\bar{X}_n - 1 - \log(\lambda\bar{X}_n) \right]}_{R(\lambda, \hat{\lambda})} + \underbrace{\frac{\mathbb{E}_\lambda(\lambda\bar{X}_n)}{n-1} - \log\left(\frac{n}{n-1} \right)}_{g(n)}. \end{aligned}$$

Example (Exponential Distribution)

As $\mathbb{E}_\lambda[\bar{X}_n] = \lambda^{-1}$, we have

$$g(n) = \frac{1}{n-1} - \log\left(\frac{n}{n-1}\right).$$

We claim that $g(n) > 0$ for $n \geq 2$. Indeed, this is true if, for any $x \geq 1$,

$$\frac{1}{x} > \log(x+1) - \log x, \quad \text{i.e.,} \quad \frac{1}{x} > \int_x^{x+1} \frac{1}{t} dt,$$

which obviously holds as, for $t \in (x, x+1)$, $1/x > 1/t$. Consequently, $R(\lambda, \tilde{\lambda}) > R(\lambda, \hat{\lambda})$ and $\hat{\lambda}$ strictly dominates $\hat{\lambda}$.

Criteria for Choosing Decision Rules

Definition (Admissible Decision Rule)

A decision rule δ is *admissible* for the experiment $(\{F_\theta\}_{\theta\in\Theta}, \mathcal{L})$ if it is not strictly dominated by any other decision rule.

- In non-trivial problems, it may not be easy at all to decide whether a given decision rule is admissible.
  $\hookrightarrow$ E.g., Stein's paradox ("one of the most striking post-war results in mathematical statistics"-Brad Efron).
- Admissibility is a minimal requirement — what about the opposite end (optimality)?
- In almost any non-trivial experiment, there is no decision rule that makes risk uniformly smallest over θ.
  $\hookrightarrow$ Solutions:
  - Narrow down the class of possible decision rules by unbiasedness/symmetry/...considerations, and try to find *uniformly dominating* rules of all other rules (next week!).
  - Use global rather than local criteria (with respect to θ).

Minimax Rules

Minimax Decision Rules

Rather than look at risk at every θ, concentrate on maximum risk.

Definition (Minimax Decision Rule)

Let $\mathcal{D}$ be a class of decision rules for an experiment $(\{F_\theta\}_{\theta\in\Theta}, \mathcal{L})$. If $\delta \in \mathcal{D}$ is such that

$$\sup_{\theta\in\Theta} R(\theta, \delta) \leq \sup_{\theta\in\Theta} R(\theta, \delta'), \quad \forall\ \delta' \in \mathcal{D},$$

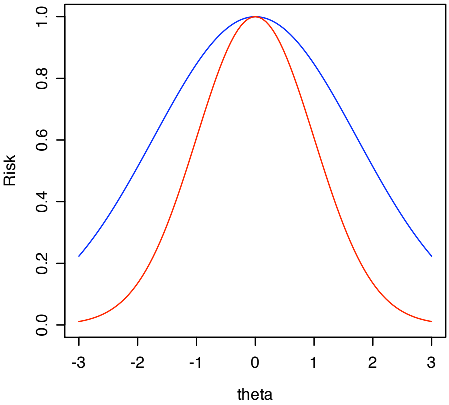
then δ is called a minimax decision rule.

- A minimax rule δ satisfies $\sup_{\theta\in\Theta} R(\theta, \delta) = \inf_{\kappa\in\mathcal{D}} \sup_{\theta\in\Theta} R(\theta, \kappa)$.
- In the minimax setup, a rule is *preferable* to another if it has smaller maximum risk.

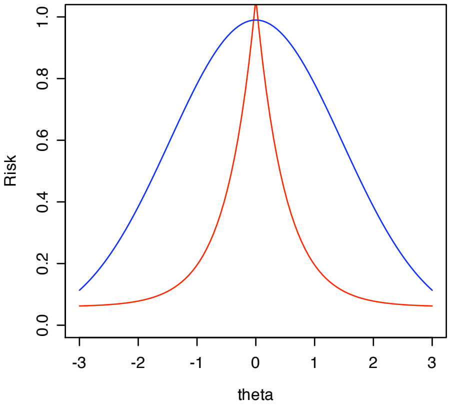
Minimax Decision Rules

- Motivated as follows: we do not know anything about θ so let us insure ourselves against the worst thing that can happen.
- Makes sense if you are in a zero-sum game: if your opponent chooses θ to maximize $\mathcal{L}$ then one should look for minimax rules. But is Nature really an opponent?
- If there is no reason to believe that Nature is trying to "do her worst", then the minimax principle is overly conservative: it places emphasis on the "bad θ".
- Minimax rules may not be unique, and may not even be admissible. A minimax rule may very well dominate another minimax rule.
- A unique minimax rule is obviously admissible.
- Minimaxity can lead to counterintuitive results. A rule may dominate another rule, except for a small region in Θ, where the other rule achieves a smaller supremum risk.

Inadmissible minimax rule



Counterintuitive minimax rule



Bayes Rules

Suppose we have some prior belief about the value of θ. How can this be incorporated in our risk-based considerations?
$\hookrightarrow$ Rather than looking at risk at every θ, concentrating on average risk.

Definition (Bayes Risk)

Let $\pi(\theta)$ be a probability density (or frequency) function on Θ and let δ be a decision rule for the experiment $(\{F_\theta\}_{\theta\in\Theta}, \mathcal{L})$. The π-Bayes risk of δ is defined as

$$r(\pi, \delta) = \int_\Theta R(\theta, \delta)\pi(\theta)d\theta = \int_\Theta \int_\mathcal{X} \mathcal{L}(\theta, \delta(\boldsymbol{x}))dF_\theta(\boldsymbol{x})\pi(\theta)d\theta.$$

The prior $\pi(\theta)$ places different emphasis for different values of θ based on our prior belief/knowledge.

Bayes Decision Rules

Bayes principle: a decision rule is *preferable* to another if it has a smaller Bayes risk (depends on the prior $\pi(\theta)$!).

Definition (Bayes Decision Rule)

Let $\mathcal{D}$ be a class of decision rules for an experiment $(\{F_\theta\}_{\theta\in\Theta}, \mathcal{L})$ and let $\pi(\cdot)$ be a probability density (or frequency) function on Θ. If $\delta \in \mathcal{D}$ is such that

$$r(\pi, \delta) \leq r(\pi, \delta') \quad \forall\ \delta' \in \mathcal{D},$$

then δ is called a *Bayes decision rule* with respect to π.

- The minimax principle aims at minimizing the maximum risk.
- The Bayes principle aims at minimizing the average risk.
- Sometimes no Bayes rule exists because the infimum may not be attained for any $\delta \in \mathcal{D}$. However in such cases $\forall \epsilon > 0\ \exists \delta_\epsilon \in \mathcal{D}$: $r(\pi, \delta_\epsilon) < \inf_{\delta\in\mathcal{D}} r(\pi, \delta) + \varepsilon$.

Admissibility of Bayes Rules

Rule of thumb: Bayes rules are nearly always admissible.

Theorem (Discrete Case Admissibility)

Assume that $\Theta = \{\theta_1, \ldots, \theta_t\}$ is a finite space and that the prior $\pi(\theta_i) > 0$, $i = 1, ..., t$. Then a Bayes rule with respect to π is admissible.

Proof.

Let δ be a Bayes rule, and suppose that κ strictly dominates δ. Then, for any j,

$$R(\theta_j, \kappa) \leq R(\theta_j, \delta),$$

and there exists $k \in \{1, \ldots, t\}$ such that $R(\theta_k, \kappa) < R(\theta_k, \delta)$. Thus, as $\pi(\theta_j) > 0$ for any j,

$$R(\theta_j, \kappa)\pi(\theta_j) \leq R(\theta_j, \delta)\pi(\theta_j) \text{ and } R(\theta_k, \kappa)\pi(\theta_k) < R(\theta_k, \delta)\pi(\theta_k),$$

which yield

$$\sum_{j=1}^{t} R(\theta_j, \kappa)\pi(\theta_j) < \sum_{j=1}^{t} R(\theta, \delta)\pi(\theta_j),$$

which contradicts the fact that δ is a Bayes rule with respect to π. $\square$

Admissibility of Bayes Rules

Theorem (Uniqueness and Admissibility)

If a Bayes rule is unique, it is admissible.

Proof.

Suppose that δ is a unique Bayes rule and assume that κ strictly dominates it. Then,

$$\int_{\Theta} R(\theta, \kappa)\pi(\theta)d\theta \leq \int_{\Theta} R(\theta, \delta)\pi(\theta)d\theta,$$

as a result of strict domination and by $\pi(\theta)$ being non-negative. If there is equality, it contradicts the uniqueness of the Bayes rule and if the inequality is strict, it contradicts the fact that δ is a Bayes rule. Either possibility contradicts our assumption. □

Admissibility of Bayes Rules

Theorem (Continuous Case Admissibility)

Let $\Theta \subset \mathbb{R}^d$. Assume that the risk functions $R(\theta, \delta)$ are continuous in θ for all decision rules $\delta \in \mathcal{D}$. Suppose that π places positive mass on any open subset of Θ. Then a Bayes rule with respect to π is admissible.

Proof.

Let κ be a decision rule that strictly dominates δ. Let Θ_0 be the set on which $R(\theta, \kappa) < R(\theta, \delta)$. Given a $\theta_0 \in \Theta_0$, we have $R(\theta_0, \kappa) < R(\theta_0, \delta)$. By continuity, there exists $\epsilon > 0$ such that $R(\theta, \kappa) < R(\theta, \delta)$ for all θ satisfying $\|\theta - \theta_0\| < \epsilon$. It follows that Θ_0 is open and hence, by our assumption, $\pi(\Theta_0) > 0$. Therefore,

$$\int_{\Theta_0} R(\theta, \kappa)\pi(\theta)d\theta < \int_{\Theta_0} R(\theta, \delta)\pi(\theta)d\theta.$$

Admissibility of Bayes Rules

(proof cont'd).

Hence, using the fact that $\int_{\Theta_0^c} R(\theta,\kappa)\pi(\theta)d\theta \leq \int_{\Theta_0^c} R(\theta,\delta)\pi(\theta)d\theta$, we obtain

$$\begin{aligned} r(\pi,\kappa) &= \int_{\Theta} R(\theta,\kappa)\pi(\theta)d\theta \\ &= \int_{\Theta_0} R(\theta,\kappa)\pi(\theta)d\theta + \int_{\Theta_0^c} R(\theta,\kappa)\pi(\theta)d\theta \\ &< \int_{\Theta_0} R(\theta,\delta)\pi(\theta)d\theta + \int_{\Theta_0^c} R(\theta,\delta)\pi(\theta)d\theta \\ &= r(\pi,\delta), \end{aligned}$$

which contradicts our assumption that δ is a Bayes rule. □

The continuity assumption and the assumption on π ensure that Θ_0 is not an isolated set and has positive measure, so that it "contributes" to the integral.

Randomized Rules

Randomized Decision Rules

Given

- decision rules $\delta_1, \ldots, \delta_k$,
- probabilities $p_i \geq 0$, $\sum_{i=1}^k p_i = 1$,

we may define a new decision rule, δ_* "$= \sum_{i=1}^k p_i \delta_i$", called a *randomized decision rule*.

Interpretation

Given data $\boldsymbol{X}$, we choose a rule δ_i with probability p_i independently of $\boldsymbol{X}$. If δ_j is the outcome $(1 \leq j \leq k)$, then we take decision/action $\delta_j(\boldsymbol{X})$.

$\rightarrow$ The risk of δ_* is the average risk: $R(\theta, \delta_*) = \sum_{i=1}^k p_i R(\theta, \delta_i)$.

- Such rules appear artificial but, often, minimax rules are randomized decision rules.
- Examples of randomized rules with $\sup_\theta R(\theta, \delta_*) < \sup_\theta R(\theta, \delta_i) \forall i$.

Summary

- Decision theory gives us tools to compare different estimators/statistical procedures inside parametric models.
- In order to use decision theory, we have to choose an appropriate loss function from which we derive a risk function.
- Comparing risk functions is hard because there is no canonical ordering on positive functions! We saw three possibilities:
  - Admissibility: corresponding to a partial order.
  - Minimax rules: ordering risk functions according to their maximum.
  - Bayes rules: corresponding to a weighting of the different θ.
- Amazingly, Bayes rules and admissible rules have a very close relationship.
- We presented randomized decision rules which might appear silly but are useful for minimaxity.

Statistical Theory (Week 9): Minimum Variance Unbiased Estimation

Erwan Koch

Ecole Polytechnique Fédérale de Lausanne (Institute of Mathematics)

EPFL

Optimality in the Decision Theory Framework

Decision Theory Framework

Saw how point estimation can be seen as a game: Nature vs Statistician.
The decision theory framework includes:

- A *family of distributions* $\mathcal{F}$, usually assumed to admit densities (frequencies) and a *parameter space* $\Theta \subseteq \mathbb{R}^p$ which parametrizes the family, i.e., $\mathcal{F} = \{F_\theta\}_{\theta \in \Theta}$.
- A *data space* $\mathcal{X}$, on which the parametric family is supported.
- An *action space* $\mathcal{A}$, which represents the space of possible *actions* available to the statistician. In point estimation, $\mathcal{A} = \Theta$.
- A *loss function* $\mathcal{L} : \Theta \times \mathcal{A} \to \mathbb{R}^+$. In point estimation, $\mathcal{L}(\theta, \alpha)$ represents the lost incurred when estimating $\theta \in \Theta$ by $\alpha \in \mathcal{A}$.
- A set $\mathcal{D}$ of *decision rules*. Any $\delta \in \mathcal{D}$ is a (measurable) function $\delta : \mathcal{X} \to \mathcal{A}$. In point estimation, decision rules are simply estimators.

The performance of decision rules has to be judged by the risk they induce:

$$R(\theta, \delta) = \mathbb{E}_\theta[\mathcal{L}(\theta, \delta(\boldsymbol{X}))], \quad \boldsymbol{\theta} \in \Theta, X \sim F_\theta, \delta \in \mathcal{D}.$$

Optimality in Point Estimation

An optimal decision rule would be one that uniformly minimizes the risk:

$$R(\theta, \delta_{\text{OPTIMAL}}) \leq R(\theta, \delta), \quad \forall \theta \in \Theta \;\&\; \forall \delta \in \mathcal{D}.$$

But such rules can very rarely be determined.
$\hookrightarrow$ Optimality becomes a *vague* concept.

$\hookrightarrow$ Can be made precise in many ways ...

Avenues to studying optimal decision rules include:

- **Restricting attention to global risk criteria rather than local**
  $\hookrightarrow$ Bayes and minimax risk.
- **Focusing on restricted classes of rules** $\mathcal{D}$
  $\hookrightarrow$ e.g., Minimum Variance Unbiased Estimation.
- **Studying the risk behaviour asymptotically** $(n \to \infty)$
  $\hookrightarrow$ e.g., Asymptotic Relative Efficiency.

Uniform Optimality in Unbiased Quadratic Estimation

Unbiased Estimators under Quadratic Loss

Focus on Point Estimation

1. Assume that F_θ is known up to the parameter θ which is unknown.
2. Let $(x_1, \dots, x_n)^\top$ be a realization of $\boldsymbol{X} \sim F_\theta$ which is available to us.
3. Estimate the value of θ that generated $\boldsymbol{X}$, given $(x_1, \dots, x_n)^\top$.

Focus on Quadratic Loss

Error incurred when estimating θ by $\hat{\theta} = \delta(\boldsymbol{X})$ is

$$\mathcal{L}(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2,$$

giving MSE as risk: $R(\theta, \hat{\theta}) = \mathbb{E}_\theta[\|\theta - \hat{\theta}\|^2] = \mathrm{Var}(\hat{\theta}) + \mathrm{bias}^2(\hat{\theta})$.

RESTRICT the class of estimators (=decision rules)

Consider ONLY *unbiased* estimators: $\mathcal{D} = \{\delta : \mathcal{X} \to \Theta | \mathbb{E}_\theta[\delta(\boldsymbol{X})] = \theta\}$.

Comments on Unbiasedness

- Unbiasedness requirement is one means of reducing the class of rules/estimators we are considering.
  ↪ Other requirements could be invariance or equivariance, e.g.,

$$\delta(\boldsymbol{X} + \boldsymbol{c}) = \delta(\boldsymbol{X}) + \boldsymbol{c}.$$

- Risk reduces to variance since bias is zero.
- Unbiased Estimators may not exist in a particular problem.
- Unbiased Estimators may be silly for a particular problem.
- Not necessarily a sensible requirement.
  ↪ e.g., violates the "likelihood principle".
- However, unbiasedness can be a reasonable/natural requirement in a wide class of point estimation problems.

Comments on Unbiasedness

Example (Unbiased Estimators Do Not Always Exist)

Let $X \sim$Binom(n, θ), with θ unknown but n known.

- We wish to find an unbiased estimator of

$$\psi = \sin\theta,$$

i.e., an estimator $\delta(X)$ such that $\mathbb{E}_\theta[\delta] = \psi = \sin\theta$. Such an estimator must satisfy

$$\sum_{x=0}^{n} \delta(x) \binom{n}{x} \theta^x (1-\theta)^{n-x} = \sin\theta,$$

but this cannot hold for all θ, since the sine function cannot be represented as a finite polynomial.

Comments on Unbiasedness

Example (Unbiased Estimators Do Not Always Exist)

- Now, we wish to find an unbiased estimator of

$$\psi = 1/\theta.$$

We need to find $\delta(0), \ldots, \delta(n)$ such that

$$\sum_{x=0}^{n} \delta(x) \binom{n}{x} \theta^x (1-\theta)^{n-x} = \frac{1}{\theta},$$

i.e.,

$$\sum_{x=0}^{n} \delta(x) \binom{n}{x} \theta^{x+1} (1-\theta)^{n-x} = \sum_{k=1}^{n+1} a(k)\theta^k = 1,$$

where $a(0), \ldots, a(n+1)$ depend on $\delta(0), \ldots, \delta(n)$. Whatever the values of $\delta(0), \ldots, \delta(n)$, the latter equation is satisfied for at most $n+1$ values of θ.

Thus, the class of unbiased estimators is empty in both cases.

Comments on Unbiased Estimators

Example (Unbiased Estimators May Be "Silly")

Let $X \sim$Poisson(λ). We wish to estimate the parameter

$$\psi = e^{-2\lambda}.$$

If $\delta(X)$ is an unbiased estimator of ψ, then we must have

$$\sum_{x=0}^{\infty} \delta(x) \frac{\lambda^x}{x!} e^{-\lambda} = e^{-2\lambda},$$

i.e.,

$$\sum_{x=0}^{\infty} \delta(x) \frac{\lambda^x}{x!} = e^{-\lambda},$$

or, equivalently,

$$\sum_{x=0}^{\infty} \delta(x) \frac{\lambda^x}{x!} = \sum_{x=0}^{\infty} (-1)^x \frac{\lambda^x}{x!}.$$

Hence $\delta(X) = (-1)^X$ is the only unbiased estimator of ψ. But as $0 < \psi < 1$ for $\lambda > 0$, this is clearly a ridiculous estimator.

Comments on Unbiased Estimators

Example (A Non-Trivial Example)

Let $X_1, \ldots, X_n$ be iid random variables with density

$$f(x;\mu) = e^{-(x-\mu)}, \quad x \geq \mu \in \mathbb{R}.$$

Two possible unbiased estimators are

$$\hat{\mu} = X_{(1)} - \frac{1}{n} \quad \& \quad \tilde{\mu} = \bar{X} - 1,$$

and, for any t, $t\hat{\mu} + (1-t)\tilde{\mu}$ is also unbiased. Simple calculations yield

$$R(\mu, \hat{\mu}) = \text{Var}(\hat{\mu}) = \frac{1}{n^2} \quad \& \quad R(\mu, \tilde{\mu}) = \text{Var}(\tilde{\mu}) = \frac{1}{n},$$

meaning that $\hat{\mu}$ strictly dominates $\tilde{\mu}$. Note that $\hat{\mu}$ depends only on the one-dimensional sufficient statistic $X_{(1)}$. Will it dominate any other unbiased estimator?

Unbiased Estimation and Sufficiency

Theorem (Rao-Blackwell Theorem)

Let $\boldsymbol{X}$ be distributed according to a distribution depending on an unknown parameter θ and let T be a sufficient statistic for θ. Let δ be a statistic such that

1. $\mathbb{E}_\theta[\delta(\boldsymbol{X})] = g(\theta)$ *for all* θ.
2. $Var_\theta(\delta(\boldsymbol{X})) < \infty$, *for all* θ.

Then $\delta^* := \mathbb{E}[\delta|T]$ *is an unbiased estimator of* $g(\theta)$ *that dominates* δ, *i.e.,*

1. $\mathbb{E}_\theta[\delta^*(\boldsymbol{X})] = g(\theta)$ *for all* θ.
2. $Var_\theta(\delta^*(\boldsymbol{X})) \leq Var_\theta(\delta(\boldsymbol{X}))$ *for all* θ.

Moreover, inequality is strict unless $\mathbb{P}_\theta[\delta^* = \delta] = 1$.

- Indicates that any candidate for the minimum variance unbiased estimator should be a function of the sufficient statistic.
- Intuitively, by conditioning on a sufficient statistic, we throw away only irrelevant information for θ, and we keep the relevant information for θ which was already contained in δ. This decreases the variance.

Proof.

Since T is sufficient for θ, $\mathbb{E}[\delta|T=t]=h(t)$ is independent of θ, and thus δ^* is a statistic (it depends only on $\boldsymbol{X}$). Then,

$$\mathbb{E}_\theta[\delta^*(\boldsymbol{X})]=\mathbb{E}_\theta[\mathbb{E}[\delta(\boldsymbol{X})|T(\boldsymbol{X})]]=\mathbb{E}_\theta[\delta(\boldsymbol{X})]=g(\theta).$$

Furthermore, we have

$$\text{Var}_\theta(\delta)=\text{Var}_\theta[\mathbb{E}(\delta|T)]+\mathbb{E}_\theta[\text{Var}(\delta|T)]\geq\text{Var}_\theta[\mathbb{E}(\delta|T)]=\text{Var}_\theta(\delta^*).$$

In addition,

$$\text{Var}(\delta|T):=\mathbb{E}[(\delta-\mathbb{E}[\delta|T])^2|T]=\mathbb{E}[(\delta-\delta^*)^2|T],$$

so that $\mathbb{E}_\theta[\text{Var}(\delta|T)]=\mathbb{E}_\theta[(\delta-\delta^*)^2]>0$ unless $\mathbb{P}_\theta(\delta^*=\delta)=1$. □

Exercise

Show that $\text{Var}(Y)=\mathbb{E}[\text{Var}(Y|\boldsymbol{X})]+\text{Var}[\mathbb{E}(Y|\boldsymbol{X})]$ when $\text{Var}(Y)<\infty$.

The role of sufficiency and "Rao-Blackwellization"

Unbiasedness and Sufficiency

- Any admissible unbiased estimator should be a function of a sufficient statistic
  - $\hookrightarrow$ If not, we can dominate it by its conditional expectation given a sufficient statistic.
- But which sufficient statistic should we choose to compute the conditional expectation? Is any function of a sufficient statistic (provided that it is unbiased) admissible?

Suppose that δ is an unbiased estimator of $g(\theta)$ and T, S are sufficient statistics for θ.

- What is the relationship between $\text{Var}_\theta(\underbrace{\mathbb{E}[\delta|T]}_{\delta_T^*}) \overset{?}{\gtreqless} \text{Var}_\theta(\underbrace{\mathbb{E}[\delta|S]}_{\delta_S^*})$?
- Intuition suggests that the statistics which carries the least irrelevant information (in addition to the relevant information) should "win".
  - $\hookrightarrow$ More formally, if $T = h(S)$ then we expect δ_T^* to dominate δ_S^*.

Unbiasedness and Sufficiency

Proposition

Let δ be an unbiased estimator of $g(\theta)$ and define

$$\delta_T^* := \mathbb{E}[\delta|T] \quad \& \quad \delta_S^* := \mathbb{E}[\delta|S],$$

where T and S are sufficient statistics for θ. Then,

$$T = h(S) \implies \mathrm{Var}_\theta(\delta_T^*) \leq \mathrm{Var}_\theta(\delta_S^*).$$

1. Essentially means that the best possible "Rao-Blackwellization" of δ is achieved by conditioning on a minimal sufficient statistic.
2. Does not necessarily imply that for T minimally sufficient and δ unbiased, $\mathbb{E}[\delta|T]$ will have the minimum variance among all unbiased estimators.
   - $\hookrightarrow$ In fact it does not even imply that $\mathbb{E}[\delta|T]$ is admissible.

Proof.

Recall the *tower property* of conditional expectation:

$$\mathbb{E}[X|g(Y)] = \mathbb{E}\{\mathbb{E}(X|Y)|g(Y)\}.$$

Thus, assuming that $T = h(S)$ we have

$$\delta_T^* = \mathbb{E}[\delta|T] = \mathbb{E}[\delta|h(S)] = \mathbb{E}[\mathbb{E}(\delta|S)|h(S)] = \mathbb{E}[\delta_S^*|T].$$

The conclusion follows from the Rao-Blackwell theorem. □

A mathematical remark

Recall that $\mathbb{E}[Z|Y]$ is the minimizer of $\mathbb{E}[(Z-\varphi(Y))^2]$ over all (measurable) functions φ of Y. Moreover, $\sqrt{\mathbb{E}[X^2]}$ defines a Hilbert norm on the space of random variables with finite variance. This yields a geometric intuition about the tower property.

The role of completeness in Uniform Optimality

Completeness, Sufficiency, Unbiasedness, and Optimality

Theorem (Lehmann-Scheffé Theorem)

Let T be a complete sufficient statistic for θ and let δ be a statistic such that $\mathbb{E}_\theta[\delta] = g(\theta)$ and $Var_\theta(\delta) < \infty$, $\forall \theta \in \Theta$. Let $\delta^* := \mathbb{E}[\delta | T]$ and V be any other unbiased estimator of $g(\theta)$. Then,

1. $Var_\theta(\delta^*) \leq Var_\theta(V)$, $\forall \theta \in \Theta$.
2. $Var_\theta(\delta^*) = Var_\theta(V) \implies \mathbb{P}_\theta[\delta^* = V] = 1$.

Thus δ^* is the unique Uniformly Minimum Variance Unbiased Estimator (UMVU estimator or UMVUE) of $g(\theta)$.

- States that if a complete sufficient statistic T exists, then the Minimum Variance Unbiased Estimator (MVUE) of $g(\theta)$ (if it exists) must be a function of T.
- Establishes that whenever there exists an UMVUE, it is unique.
- Can be used to examine whether unbiased estimators exist at all: if a complete sufficient statistic T exists, but there exists no function h with $\mathbb{E}[h(T)] = g(\theta)$, then no unbiased estimator of $g(\theta)$ exists.

Proof.

1. Let V be an arbitrary unbiased estimator of $g(\theta)$ with finite variance, and define its "Rao-Blackwellized" version $V^* := \mathbb{E}[V|T]$. Now, by unbiasedness of V and V^*, we have, for any $\theta \in \Theta$,

$$0 = \mathbb{E}_\theta[V^* - \delta^*] = \mathbb{E}_\theta[\mathbb{E}[V|T] - \mathbb{E}[\delta|T]] = \mathbb{E}_\theta[h(T)],$$

where $h(T) = \mathbb{E}[V|T] - \mathbb{E}[\delta|T]$. It follows by completeness of T that, for all θ, $\mathbb{P}_\theta[h(T) = 0] = 1$, i.e, $\mathbb{P}_\theta[V^* = \delta^*] = 1$. Now, as $\text{Var}_\theta(V^*) \leq \text{Var}_\theta(V)$ (by the Rao-Blackwell theorem), we obtain

$$\text{Var}_\theta(\delta^*) \leq \text{Var}_\theta(V).$$

2. We assume that $\text{Var}_\theta(V) = \text{Var}_\theta(\delta^*)$. From above, this implies that $\text{Var}_\theta(V) = \text{Var}_\theta(V^*)$, which, by the Rao-Blackwell theorem, yields $\mathbb{P}_\theta[V = V^*] = 1$. As $\mathbb{P}_\theta[V^* = \delta^*] = 1$, we obtain $\mathbb{P}_\theta[V = \delta^*] = 1$.

□

Completeness, Sufficiency, Unbiasedness, and Optimality

Taken together, the Rao-Blackwell and Lehmann-Scheffé theorems also suggest two approaches to finding the UMVUE when a complete sufficient statistic T exists:

1. Find a function h such that $\mathbb{E}_\theta[h(T)] = g(\theta)$. If $\text{Var}_\theta[h(T)] < \infty$ for all θ, then $\delta = h(T)$ is the unique UMVUE of $g(\theta)$.
   - $\hookrightarrow$ The function h can be found by solving the equation $\mathbb{E}_\theta[h(T)] = g(\theta)$ or by an educated guess.
2. Given an unbiased estimator δ of $g(\theta)$, we obtain the UMVUE by "Rao-Blackwellizing" it wrt the complete sufficient statistic.

Example (Bernoulli Trials)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bern}(\theta)$. What is the UMVUE of θ^2?

As already seen (see week 3), $T = X_1 + \ldots + X_n$ is sufficient and also complete. Sufficiency can easily be obtained from the Neyman factorization theorem, and completeness directly stems from the fact that the distribution of $X_1, \ldots, X_n$ belongs to a 1-parameter exponential family.

Example (Bernoulli Trials)

First suppose that $n = 2$. If a UMVUE exists, it must be of the form $h(T)$ with h satisfying

$$\theta^2 = \sum_{k=0}^{2} h(k) \binom{2}{k} \theta^k (1-\theta)^{2-k}.$$

It is easy to see that $h(0) = h(1) = 0$ while $h(2) = 1$. Thus, for $n = 2$, $h(T) = T(T-1)/2$ is the unique UMVUE of θ^2.
For $n > 2$, set $\delta = \mathbf{1}\{X_1 + X_2 = 2\}$, which is an unbiased estimator of θ^2. By the Lehmann-Scheffé theorem, $\delta^* = \mathbb{E}[\delta|T]$ is the unique UMVUE estimator of θ^2. We have

$$\begin{aligned}
\mathbb{E}[\delta|T=t] &= \mathbb{P}[X_1 + X_2 = 2|T=t] \\
&= \frac{\mathbb{P}_\theta[X_1 + X_2 = 2, X_3 + \ldots + X_n = t-2]}{\mathbb{P}_\theta[T=t]} \\
&= \left\{\begin{matrix} 0 & \text{if } t \leq 1 \\ \binom{n-2}{t-2}/\binom{n}{t} & \text{if } t \geq 2 \end{matrix}\right\} = \frac{t(t-1)}{n(n-1)}.
\end{aligned}$$

Thus, $\delta^* = T(T-1)/[n(n-1)]$ is the UMVUE of θ^2.

Lower Bounds for the Risk and Achieving them

Variance Lower Bounds for Unbiased Estimators

- Often a minimal sufficient statistic exists but is not complete. In such cases, we cannot use the Lehmann-Scheffé theorem to find an UMVUE.
- However, if we could establish a *lower bound* for the variance as a function of θ, then an estimator achieving this bound would be an UMVUE.

The Aim

For iid $X_1, \ldots, X_n$ with density (frequency) depending on θ unknown, we want to establish conditions under which

$$\mathrm{Var}_\theta[\delta] \geq \phi(\theta), \quad \forall \theta,$$

for any unbiased estimator δ. We also wish to determine $\phi(\theta)$.

Cauchy-Schwarz Bounds

Theorem (Cauchy-Schwarz Inequality)

Let U, V be random variables with finite second moment. Then,

$$Cov(U, V) \leq \sqrt{Var(U)Var(V)}.$$

It yields an immediate lower bound for the variance of an unbiased estimator δ_0:

$$\mathrm{Var}_\theta(\delta_0) \geq \frac{\mathrm{Cov}_\theta^2(\delta_0, U)}{\mathrm{Var}_\theta(U)},$$

which is valid for any random variable U with $\mathrm{Var}_\theta(U) < \infty$ for all θ.

- The bound can be made tight be choosing a suitable U.
- However this is still not very useful. The bound will be specific to δ_0, while we want a bound that holds for any unbiased estimator δ and depends merely on θ.
- Is there a smart choice of U for which $\mathrm{Cov}_\theta(\delta_0, U)$ depends on $g(\theta) = \mathbb{E}_\theta(\delta_0)$ only (and so is not specific to δ_0)?

Optimizing the Cauchy-Schwarz Bound

Let θ be a real and $f(\cdot, \theta)$ be the density of $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$. Assume that the following regularity conditions hold.

Regularity Conditions

(C1) The support of f, $\{\boldsymbol{x} \in \mathbb{R}^n : f(\boldsymbol{x}; \theta) > 0\}$, is independent of θ.

(C2) $f(\boldsymbol{x}; \theta)$ is differentiable wrt θ, $\forall \theta \in \Theta$.

(C3) $\mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} \log f(\boldsymbol{X}; \theta) \right] = 0$.

(C4) For a statistic $T = T(\boldsymbol{X})$ with $\mathbb{E}_\theta[|T|] < \infty$ and $g(\theta) = \mathbb{E}_\theta[T]$ differentiable,

$$g'(\theta) = \mathbb{E}_\theta \left[T \frac{\partial}{\partial \theta} \log f(\boldsymbol{X}; \theta) \right], \quad \forall \theta.$$

To make sense of (C3) and (C4), let us take any statistic S. Then

$$\frac{d}{d\theta} \int S(\boldsymbol{x}) f(\boldsymbol{x}; \theta) d\boldsymbol{x} \overset{!}{=} \int S(\boldsymbol{x}) \frac{f(\boldsymbol{x}; \theta)}{f(\boldsymbol{x}; \theta)} \frac{d}{d\theta} f(\boldsymbol{x}; \theta) dx = \mathbb{E}_\theta \left[S(\boldsymbol{X}) \frac{\partial}{\partial \theta} \log f(\boldsymbol{X}; \theta) \right],$$

provided integration and differentiation can be interchanged.

Theorem

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ *have joint density (frequency)* $f(\boldsymbol{x}; \theta)$ *satisfying (C1), (C2) and (C3). If the statistic* T *satisfies* $(C4)$, *then*

$$Var_\theta(T) \geq \frac{[g'(\theta)]^2}{I_n(\theta)},$$

where

$$I_n(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial\theta} \log f(\boldsymbol{X}; \theta) \right)^2 \right].$$

The Cramér-Rao Lower Bound

Proof.

By the Cauchy-Schwarz inequality with $U = \frac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)$,

$$\mathrm{Var}_\theta(T) \geq \frac{\mathrm{Cov}_\theta^2\left(T, \frac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\right)}{\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\right)}$$

Since

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\right] = 0,$$

we have

$$\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\right) = I_n(\theta),$$

and, using (C4),

$$\mathrm{Cov}_\theta\left(T, \frac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\right) = \mathbb{E}_\theta\left[T\frac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\right] = g'(\theta),$$

which completes the proof. □

The Cramér-Rao Lower Bound

When is the Cramér-Rao lower bound achieved? Note that

$$\mathrm{Var}_\theta[T] = \frac{[g'(\theta)]^2}{I_n(\theta)} \implies \mathrm{Var}_\theta[T] = \frac{\mathrm{Cov}_\theta^2\left[T, \frac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\right]}{\mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\right]}.$$

which occurs if and only if $\frac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)$ is an affine function of T with probability one (case where the correlation equals 1), i.e.,

$$\frac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta) = A(\theta)T(\boldsymbol{x}) + B(\theta).$$

Solving this differential equation yields, for all $\boldsymbol{x}$,

$$\log f(\boldsymbol{x};\theta) = A^*(\theta)T(\boldsymbol{x}) + B^*(\theta) + S(\boldsymbol{x}),$$

i.e.,

$$f(\boldsymbol{x};\theta) = \exp\{A^*(\theta)T(\boldsymbol{x}) + B^*(\theta) + S(\boldsymbol{x})\}.$$

Conclusion

Thus, $\mathrm{Var}_\theta(T)$ attains the lower bound if and only if the density (frequency) of $\boldsymbol{X}$ has a one-parameter exponential family form as above.

The Cramér-Rao bound asymptotically

If the $X_1, \ldots, X_n$ are iid, then the Fisher information is

$$I_n(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial\theta} \log f(\boldsymbol{X}; \theta) \right)^2 \right] = n\mathbb{E}_\theta \left[\left(\frac{\partial}{\partial\theta} \log f(X_1; \theta) \right)^2 \right] = nI(\theta).$$

More generally, the Fisher information of several independent observations is the sum of the Fisher informations of each one.

Definition

The *asymptotic efficiency* of a sequence of estimators $\hat{\theta}_n$ of θ based on iid observations $X_1, \ldots, X_n$ is the ratio

$$\mathrm{Var}(\hat{\theta}_n) / \left[nI(\theta) \right]^{-1}.$$

The asymptotic efficiency measures whether a given estimator asymptotically saturates the Cramér-Rao bound or falls short.

Summary

- Unbiasedness is one criteria we can follow to find a good estimator.
- "Rao-Blackwellizing" an unbiased estimator with a sufficient statistic gives a better estimator (with a lower variance).
- If there exists a complete sufficient statistic, there may exist a unique uniformly minimum variance unbiased estimator (UMVUE). But recall that, besides exponential families, a complete and sufficient statistic rarely exists!
- More generally, all estimators must obey the Cramér-Rao lower bound. If we can prove that an estimator saturates the Cramér-Rao bound, then that proves that it is optimal.

The MLE dominates

From the results presented in this lecture, we see that the MLE is a great estimator:

- It automatically depends only on a minimally sufficient statistic: its already Rao-Blackwellized!
- If there exists a complete sufficient statistic AND the MLE is unbiased, then it is the UMVUE.
- Even without completeness, the MLE is asympotically:
    - Unbiased: $\mathbb{E}(\hat{\theta}) = \theta$.
    - Gaussian with variance $1/[nI(\theta)]$. Asymptotically, it saturates the Cramér-Rao bound!

It is a great estimator **if the model is correctly specified**!

Statistical Theory (Week 10): Testing Statistical Hypotheses

Erwan Koch

Ecole Polytechnique Fédérale de Lausanne (Institute of Mathematics)

EPFL

Contrasting Theories With Experimental Evidence

Using Data to Evaluate Theories/Assertions

- Scientific theories lead to assertions/implications that are testable using empirical data.
- If the theory (or *hypothesis*) is true, then the data should be compatible with corresponding implications.
- Data may discredit the theory or not.
- Similarities with the logical/mathematical concept of necessary condition and reasoning by contradiction.
- Example: Large Hadron Collider in CERN, Genève. To gain insight about the existence of the Higgs Boson, study if particle trajectories are consistent with what theory predicts.
- Example: The theory of "luminoferous aether" in late 19th century to explain light travelling in vacuum was discredited by the Michelson-Morley's experiment.

What would be the appropriate formal statistical framework?

Hypothesis Testing Setup

Statistical Framework for Testing Hypotheses

The Problem of Hypothesis Testing

- $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ random vector with joint density/frequency $f(\boldsymbol{x}; \theta)$
- $\theta \in \Theta$ where $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \emptyset$
- We observe a realization $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ of $\boldsymbol{X} \sim f_\theta$
- Decide on the basis of $\boldsymbol{x}$ whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$

$\hookrightarrow$ Often $\dim(\Theta_0) < \dim(\Theta)$ so $\theta \in \Theta_0$ represents a *simplified model*.

Example

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$ and $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\nu, 1)$. Let $\theta = (\mu, \nu)^\top$ and

$$\Theta = \{(\mu, \nu)^\top : \mu \in \mathbb{R}, \nu \in \mathbb{R}\} = \mathbb{R}^2.$$

May be interesting to test if $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same distribution, even though they may be measurements on characteristics of different groups. In this case $\Theta_0 = \{(\mu, \nu)^\top \in \mathbb{R}^2 : \mu = \nu\}$.

Type I vs Type II Error

Decision Theory Perspective on Hypothesis Testing

- Given $\boldsymbol{X}$ we need to *decide* between two hypotheses:

  H_0: $\theta \in \Theta_0$ (the NULL HYPOTHESIS)
  H_1: $\theta \in \Theta_1$ (the ALTERNATIVE HYPOTHESIS)

- We want decision rule that allows us to choose between H_0 and H_1. We take $\delta : \mathcal{X} \to \mathcal{A} = \{0, 1\}$ and we choose H_0 if $\delta(\boldsymbol{X}) = 0$ and H_1 if $\delta(\boldsymbol{X}) = 1$.
  - In hypothesis testing δ is called a *test function*
  - Often δ depends on $\boldsymbol{X}$ only through some real-valued statistic $T = T(\boldsymbol{X})$ called a *test statistic*.

- Unlikely that a test function is perfect. Possible errors to be made?

Action / Truth	H_0	H_1
0	🙂	Type II Error
1	Type I Error	🙂

Potential asymmetry of errors in practice: false positive VS false negative (e.g., spam filters for e-mail).

Decision Theory Perspective on Hypothesis Testing

Typically the loss function is a "0–1" loss, i.e.,

$$\mathcal{L}(\theta, a) = \begin{cases} 1 & \text{if } \theta \in \Theta_0 \ \& \ a = 1 \qquad \text{(Type I Error)} \\ 1 & \text{if } \theta \in \Theta_1 \ \& \ a = 0 \qquad \text{(Type II Error)}, \\ 0 & \text{otherwise} \qquad \text{(No Error)} \end{cases}$$

i.e., we lose 1 unit whether we commit a type I or type II error. $\longrightarrow$ Leads to the risk function

$$R(\theta, \delta) = \begin{cases} \mathbb{E}_\theta[\mathbf{1}\{\delta = 1\}] = \mathbb{P}_\theta[\delta = 1] & \text{if } \theta \in \Theta_0 \quad \text{(prob of type I error)} \\ \mathbb{E}_\theta[\mathbf{1}\{\delta = 0\}] = \mathbb{P}_\theta[\delta = 0] & \text{if } \theta \in \Theta_1 \quad \text{(prob of type II error)} \end{cases}.$$

In short,

$$\begin{aligned} R(\theta, \delta) \quad &= \quad \mathbb{P}_\theta[\delta = 1]\mathbf{1}\{\theta \in \Theta_0\} + \mathbb{P}_\theta[\delta = 0]\mathbf{1}\{\theta \in \Theta_1\} \\ \text{``} = \text{''} \quad & \text{``}\mathbb{P}_\theta[\text{choose } H_1 | H_0 \text{ is true}]\text{'' or ``}\mathbb{P}_\theta[\text{choose } H_0 | H_1 \text{ is true}]\text{''}. \end{aligned}$$

Optimal Testing?

As with point estimation, we may wish to find *optimal* test functions.
$\hookrightarrow$ Test functions that uniformly minimize risk?

- Almost never exist
- In general there is a trade-off between the two error probabilities
- How to relax problem in this case? Treat each type I and type II error probabilities separately?

For example consider: $X \sim \mathcal{N}(\mu, 1)$ where $H_0 : \mu = -1$ and $H_1 : \mu = 1$. Take the parametric decision rule: $\delta_t(X) = \mathbf{1}(X \geq t)$ (it's optimal). If we increase t, probability of type I error decreases, but probability of type II error increases.

The Neyman-Pearson Setup

The Neyman-Pearson Setup

Classical approach: restrict class of test functions by "minimax reasoning"

1. We fix an $\alpha \in (0, 1)$, usually small (called the significance level)
2. We only consider test functions $\delta : \mathcal{X} \to \{0, 1\}$ such that
$$\delta \in \mathscr{D}(\Theta_0, \alpha) = \left\{\delta : \sup_{\theta \in \Theta_0} \mathbb{P}_\theta[\delta = 1] \leq \alpha\right\},$$
   i.e., rules for which probability of type I error is bounded above by α

$\hookrightarrow$ *Jargon: we fix a significance level for our test*

3. Within this restricted class of rules, we choose δ to minimize the probability of type II error uniformly on Θ_1, i.e., to minimize
$$\mathbb{P}_\theta[\delta(\boldsymbol{X}) = 0] = 1 - \mathbb{P}_\theta[\delta(\boldsymbol{X}) = 1], \quad \theta \in \Theta_1.$$
4. Equivalently, to maximize the *power* uniformly over Θ_1, i.e., maximize
$$\beta(\theta, \delta) = \mathbb{P}_\theta[\delta(\boldsymbol{X}) = 1] = \mathbb{E}_\theta[\delta(\boldsymbol{X})], \quad \theta \in \Theta_1$$

The Neyman-Pearson Setup

Intuitive rationale of the approach:

- Suppose we observe $\delta(\boldsymbol{X}) = 1$ (so we take action 1). As α is usually small and $\delta = 1$ has probability at most α under H_0, if H_0 is indeed true, we have observed something rare or unusual under H_0.
  $\Longrightarrow$ Evidence that H_0 is false (i.e., in favour of H_1)
  $\Longrightarrow$ Taking action 1 (choosing H_1) is a highly reasonable decision.

- But what if we observe $\delta(\boldsymbol{X}) = 0$ (so we take action 0)?
  - Due to the low significance level, this does not guarantee at all that our decision is the right one, i.e, that H_0 is true (a low significance level is generally associated with a low power).
  - We would be more confident in our decision if δ was such that the type II error was also low or if we had maximized the power β (given the significance level α).

The Neyman-Pearson Setup

- Neyman-Pearson setup naturally exploits any asymmetric structure
- But, if natural asymmetry absent, need judicious choice of H_0 (must depend on the goal)

Example: Obama VS Romney 2012. Pollsters gather iid sample $\boldsymbol{X}$ from Ohio with $X_i = \mathbf{1}\{\text{vote Romney}\}$. Which pair of hypotheses to test?

$$\begin{cases} H_0: & \text{Romney wins Ohio} \\ H_1: & \text{Obama wins Ohio} \end{cases} \quad \text{OR} \quad \begin{cases} H_0: & \text{Obama wins Ohio} \\ H_1: & \text{Romney wins Ohio} \end{cases}$$

- Which pair to choose to make a prediction? (confidence intervals?)
- Assume that Romney wonders whether he should spend more money to campaign in Ohio. His possible losses due to errors are:
  - (a) Spend more \$'s to campaign in Ohio even though he would win anyway: lose \$'s
  - (b) Lose Ohio to Obama because he thought he would win without any extra effort
- (b) is much worse than (a) (especially since Romney had lots of \$'s)
- Hence Romney would pick $H_0 = \{\text{Obama wins Ohio}\}$ as his null

Optimality in the Neyman-Pearson Setup

Finding Good Test Functions

We consider the simplest situation. Assume that $(X_1, \ldots, X_n)^\top \sim f(\cdot;\theta)$ with $\Theta = \{\theta_0, \theta_1\}$

The Neyman-Pearson Lemma - Continuous Case

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ have density function $f \in \{f_0, f_1\}$ and suppose we wish to test

$$H_0 : f = f_0 \qquad vs \qquad H_1 : f = f_1,$$

at the significance level $\alpha \in (0, 1)$. If $\Lambda(\boldsymbol{X}) = f_1(\boldsymbol{X})/f_0(\boldsymbol{X})$ is a continuous random variable, then there exists a $k > 0$ such that

$$\mathbb{P}_0[\Lambda \geq k] = \alpha,$$

and the test whose test function is given by

$$\delta(\boldsymbol{X}) = \mathbf{1}\{\Lambda(\boldsymbol{X}) \geq k\},$$

is a *most powerful (MP)* test of H_0 versus H_1 at significance level α.

Proof.

Use obvious notation $\mathbb{E}_0$, $\mathbb{E}_1$, $\mathbb{P}_0$, $\mathbb{P}_1$ corresponding to H_0 or H_1. Let $G_0(t) = \mathbb{P}_0[\Lambda \leq t]$. By assumption, G_0 is a continuous distribution function and thus takes values over the whole range $[0, 1]$. Consequently, the set $\mathcal{K}_{1-\alpha} = \{t : G_0(t) = 1 - \alpha\}$ is non-empty for any $\alpha \in (0, 1)$. Setting $k = \inf\{t \in \mathcal{K}_{1-\alpha}\}$, i.e., the $1 - \alpha$ quantile of the distribution G_0, we have $\mathbb{P}_0[\Lambda \geq k] = \alpha$. Thus

$$\mathbb{P}_0[\delta = 1] = \alpha \qquad (\text{since } \mathbb{P}_0[\delta = 1] = \mathbb{P}_0[\Lambda \geq k])$$

and therefore $\delta \in \mathscr{D}(\{\theta_0\}, \alpha)$ (i.e., δ indeed respects the level α).

To show that δ is also most powerful, it suffices to prove that if ψ is any function with $\psi(\boldsymbol{x}) \in \{0, 1\}$, then

$$\mathbb{E}_0[\psi(\boldsymbol{X})] \leq \underbrace{\mathbb{E}_0[\delta(\boldsymbol{X})]}_{=\alpha \text{ (by first part of proof)}} \implies \underbrace{\mathbb{E}_1[\psi(\boldsymbol{X})]}_{\beta_1(\psi)} \leq \underbrace{\mathbb{E}_1[\delta(\boldsymbol{X})]}_{\beta_1(\delta)}.$$

(recall that $\beta_1(\delta) = 1 - \mathbb{P}_1[\delta = 0] = \mathbb{P}_1[\delta = 1] = \mathbb{E}_1[\delta]$).

Since

$$f_1(\boldsymbol{x}) - k \cdot f_0(\boldsymbol{x}) \geq 0 \text{ if } \delta(\boldsymbol{x}) = 1 \quad \& \quad f_1(\boldsymbol{x}) - k \cdot f_0(\boldsymbol{x}) < 0 \text{ if } \delta(\boldsymbol{x}) = 0$$

and ψ can only take the values 0 or 1, we have

$$\begin{aligned} \psi(\boldsymbol{x})(f_1(\boldsymbol{x}) - k \cdot f_0(\boldsymbol{x})) &\leq \delta(\boldsymbol{x})(f_1(\boldsymbol{x}) - k \cdot f_0(\boldsymbol{x})), \quad \text{and thus} \\ \int_{\mathbb{R}^n} \psi(\boldsymbol{x})(f_1(\boldsymbol{x}) - k \cdot f_0(\boldsymbol{x})) d\boldsymbol{x} &\leq \int_{\mathbb{R}^n} \delta(\boldsymbol{x})(f_1(\boldsymbol{x}) - k \cdot f_0(\boldsymbol{x})) d\boldsymbol{x}. \end{aligned}$$

Rearranging the terms yields

$$\begin{aligned} \int_{\mathbb{R}^n} (\psi(\boldsymbol{x}) - \delta(\boldsymbol{x})) f_1(\boldsymbol{x}) d\boldsymbol{x} &\leq k \int_{\mathbb{R}^n} (\psi(\boldsymbol{x}) - \delta(\boldsymbol{x})) f_0(\boldsymbol{x}) d\boldsymbol{x}, \quad \text{i.e.,} \\ \mathbb{E}_1[\psi(\boldsymbol{X})] - \mathbb{E}_1[\delta(\boldsymbol{X})] &\leq k \left(\mathbb{E}_0[\psi(\boldsymbol{X})] - \mathbb{E}_0[\delta(\boldsymbol{X})]\right). \end{aligned}$$

As $k > 0$ by assumption, $\mathbb{E}_0[\psi(\boldsymbol{X})] \leq \mathbb{E}_0[\delta(\boldsymbol{X})]$ implies that the RHS is non-positive. Hence, δ is an MP test of H_0 vs H_1 at level α. $\square$

The Neyman-Pearson Lemma

- Basically we reject H_0 if the likelihood of θ_1 is at least k times higher than the likelihood of θ_0. This is called a likelihood ratio test, and Λ is the likelihood ratio statistic: *how much more plausible is the alternative than the null?*

- When Λ is a continuous random variable, the choice of k is essentially unique. That is, if k' is such that $\delta' = \mathbf{1}\{\Lambda \geq k'\} \in \mathscr{D}(\{\theta_0\}, \alpha)$, then $\delta = \delta'$ almost surely.

- The result does not guarantee uniqueness when an MP test exists.

- The existence of an MP test is guaranteed only if Λ is continuous. If Λ has a discontinuous distribution, there may exist no k for which the equation $\mathbb{P}_0[\Lambda \geq k] = \alpha$ has a solution.

- In the latter case, we need to consider *randomized decision rules* in order to guarantee the existence of a most powerful test.

The Neyman-Pearson Lemma

General version of the Neyman-Pearson lemma considers the relaxed problem:

Maximize $\mathbb{E}_1[\delta]$ subject to $\mathbb{E}_0[\delta] = \alpha$ and $0 \leq \delta(\boldsymbol{X}) \leq 1$ *a.s.*

$\rightarrow$ The solution does not need to be a test function since now $\delta : \mathcal{X} \rightarrow [0, 1]$! Interpretation? Think of relaxation≡randomization:

- We are willing to consider also <u>randomized</u> decision rules.
- How does a randomized decision rule work?
  1. If $\delta(\boldsymbol{X}) = 1$, reject.
  2. If $\delta(\boldsymbol{X}) = 0$, don't reject.
  3. If $\delta(\boldsymbol{X}) = p \in (0, 1)$, then sample an independent Bernoulli random variable Y with probability of success p.
     - (3a) If Y takes the value 1, then reject.
     - (3b) If Y takes the value 0, don't reject.

The last step is randomization: we inject randomness which is completely independent of the data.

The Neyman-Pearson Lemma

Neyman-Pearson Lemma - General Case

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ have density (frequency) function $f \in \{f_0, f_1\}$ and suppose we wish to test

$$H_0 : f = f_0 \qquad vs \qquad H_1 : f = f_1,$$

at level $\alpha \in (0,1)$. Let $\Lambda(\boldsymbol{X}) = f_1(\boldsymbol{X})/f_0(\boldsymbol{X})$. Then, there exist $k > 0$ and $p \in [0,1]$ such that the decision rule

$$\delta(\boldsymbol{X}) = \begin{cases} 1 & \text{if } \Lambda(\boldsymbol{X}) > k, \\ p & \text{if } \Lambda(\boldsymbol{X}) = k, \\ 0 & \text{if } \Lambda(\boldsymbol{X}) < k, \end{cases}$$

satisfies

$$\mathbb{E}_0[\delta(\boldsymbol{X})] = \alpha \qquad \& \qquad \mathbb{E}_1[\psi(\boldsymbol{X})] \leq \mathbb{E}_1[\delta(\boldsymbol{X})]$$

for all $\psi : \mathcal{X} \to [0,1]$ such that $\mathbb{E}_0[\psi(\boldsymbol{X})] \leq \alpha$.

Proof.

Let $G_0(t) = \mathbb{P}_0[\Lambda \leq t]$ and $k = \inf\{t : G_0(t) \geq 1 - \alpha\}$. If $G_0(k) = 1 - \alpha$, then set $p = 0$ and proceed as in the continuous version of the NP-lemma. Otherwise, if $G_0(k) > 1 - \alpha$, define $\xi := \lim_{\epsilon \to 0} G_0(k - \epsilon) < (1 - \alpha)$ and

$$p = \frac{G_0(k) - (1 - \alpha)}{G_0(k) - \xi}.$$

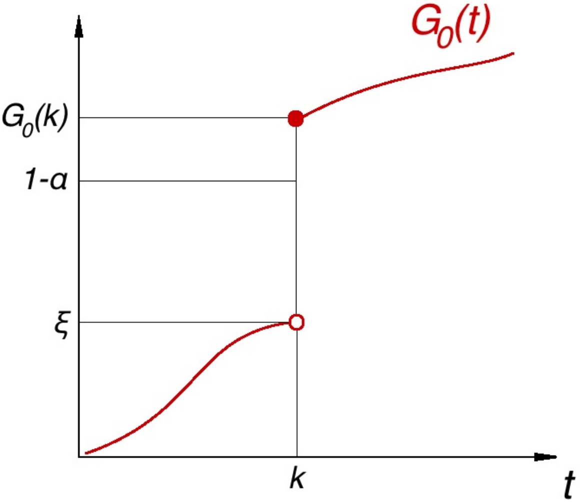
By definition of ξ, it must be that $p \in (0, 1)$. Furthermore,

$$G_0(k) - \xi = \mathbb{P}_0[\Lambda \leq k] - \lim_{\epsilon \to 0} \mathbb{P}_0[\Lambda \leq k - \epsilon] = \mathbb{P}_0[\Lambda = k]$$

($\lim_{\epsilon \to 0} \mathbb{P}_0[\Lambda \leq k - \epsilon] = \mathbb{P}_0[\Lambda < k]$ by continuity of probability measures from above), which yields

$$\begin{aligned} \mathbb{E}_0[\delta] &= 1 \times \mathbb{P}_0[\Lambda > k] + p \times \mathbb{P}_0[\Lambda = k] + 0 \times \mathbb{P}_0[\Lambda < k] \\ &= 1 - G_0(k) + \frac{G_0(k) - (1 - \alpha)}{\mathbb{P}_0[\Lambda = k]} \times \mathbb{P}_0[\Lambda = k] = \alpha. \end{aligned}$$

For the power, repeat the steps in the proof of continuous NP-lemma. □



(recall that G_0 is necessarily *càdlàg*: continue à droite, limite à gauche)

The Neyman-Pearson Setup

Example (Exponential Distribution)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\lambda)$ and $\lambda \in \{\lambda_0, \lambda_1\}$, with $\lambda_1 > \lambda_0$ (H_1 leads to small values of X_i).
We want to test

$$H_0 : \lambda = \lambda_0 \quad \text{vs} \quad H_1 : \lambda = \lambda_1$$

at the level $\alpha \in (0, 1)$. We have

$$f(\boldsymbol{X}; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i}.$$

So Neyman-Pearson Lemma says that it is optimal to base our test on the statistic

$$\Lambda = \frac{f(\boldsymbol{X}; \lambda_1)}{f(\boldsymbol{X}; \lambda_0)} = \left(\frac{\lambda_1}{\lambda_0}\right)^n \exp\left[(\lambda_0 - \lambda_1) \sum_{i=1}^{n} X_i\right]$$

and to reject the null if $\Lambda \geq k$, for k such that the level is α.

The Neyman-Pearson Setup

Example (cont'd)

Now, we note that Λ is a decreasing function of $S = \sum_{i=1}^{n} X_i$ (since $\lambda_0 < \lambda_1$), which gives that

$$\Lambda \geq k \iff S \leq K,$$

for some K, so that

$$\alpha = \mathbb{P}_{\lambda_0}[\Lambda \geq k] \iff \alpha = \mathbb{P}_{\lambda_0}[S \leq K].$$

For given values of λ_0 and α it is easy to find the appropriate K. Indeed, under the null hypothesis, S has a gamma distribution with parameters n and λ_0 and thus we reject H_0 at level α if S is below the α-quantile of the Gamma(n, λ_0) distribution.

Example (Uniform Distribution)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0, \theta)$ with $\theta \in \{\theta_0, \theta_1\}$ where $\theta_0 > \theta_1$. Consider

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_1 : \theta = \theta_1.$$

As

$$f(\boldsymbol{X}; \theta) = \frac{1}{\theta^n} \mathbf{1} \left\{ \max_{1 \le i \le n} X_i \le \theta \right\},$$

an MP test of H_0 vs H_1 can be based on the discrete test statistic

$$\Lambda = \frac{f(\boldsymbol{X}; \theta_1)}{f(\boldsymbol{X}; \theta_0)} = \left(\frac{\theta_0}{\theta_1} \right)^n \mathbf{1}\{X_{(n)} \le \theta_1\}.$$

So if the test rejects H_0 when $X_{(n)} \le \theta_1$ then it is MP for H_0 vs H_1 at

$$\alpha = \mathbb{P}_{\theta_0}[X_{(n)} \le \theta_1] = (\theta_1/\theta_0)^n$$

with power $\mathbb{P}_{\theta_1}[X_{(n)} \le \theta_1] = 1$. What about smaller values of α?

Example (cont'd)

$\hookrightarrow$ What about finding an MP test for $\alpha < (\theta_1/\theta_0)^n$?

An intuitive test statistic is the sufficient statistic $X_{(n)}$, and it would be natural to reject H_0 iff $X_{(n)} \leq k$, where k solves the equation

$$\mathbb{P}_{\theta_0}[X_{(n)} \leq k] = \left(\frac{k}{\theta_0}\right)^n = \alpha,$$

i.e., $k = \theta_0 \alpha^{1/n}$. This test has power

$$\mathbb{P}_{\theta_1}[X_{(n)} \leq \theta_0 \alpha^{1/n}] = \left(\frac{\theta_0 \alpha^{1/n}}{\theta_1}\right)^n = \alpha \left(\frac{\theta_0}{\theta_1}\right)^n.$$

Is this the MP test at level $\alpha < (\theta_1/\theta_0)^n$?

Example (cont'd)

Use general form of the Neyman-Pearson lemma to solve relaxed problem:

Maximize $\mathbb{E}_1[\delta(\boldsymbol{X})]$ subject to $\mathbb{E}_{\theta_0}[\delta(\boldsymbol{X})] = \alpha < \left(\frac{\theta_1}{\theta_0}\right)^n$ & $0 \leq \delta(\boldsymbol{x}) \leq 1$.

One solution to this problem is given by

$$\delta(\boldsymbol{X}) = \begin{cases} \alpha(\theta_0/\theta_1)^n & \text{if } X_{(n)} \leq \theta_1, \\ 0 & \text{otherwise,} \end{cases}$$

which is not a test function. However, we see that its power is

$$\mathbb{E}_{\theta_1}[\delta(\boldsymbol{X})] = \alpha\left(\frac{\theta_0}{\theta_1}\right)^n = \mathbb{P}_{\theta_1}[X_{(n)} \leq \theta_0\alpha^{1/n}],$$

which is the power of the test we proposed. Hence the test that rejects H_0 if $X_{(n)} \leq \theta_0\alpha^{1/n}$ is an MP test for all levels $\alpha < (\theta_1/\theta_0)^n$.

Summary

- Hypothesis testing is a key statistical problem.
- Key insight: the errors are not symmetric.
- Neyman-Pearson setup:
  - First, we choose a **significance level** $\alpha \in (0, 1)$.
  - We seek to maximize (if possible) the **power** of the test while maintaining the significance level.
- In a simple vs simple test, there exists an optimal test for any level α. If the likelihood ratio is a discrete random variable, this test is randomized for most values of α.
- Many statisticians strongly disagree with randomized decision rules in the context of tests.

simple vs. simple $H_0: \theta = \theta_0$, $H_1: \theta = \theta_1$

$$\Lambda(\vec{x}) = \frac{f(\vec{x}, \theta_1)}{f(\vec{x}, \theta_0)}$$

$$\delta(\vec{x}) = \mathbb{1}[\Lambda(\vec{x}) > c]$$

choose c s.t.

$$\mathbb{P}_{\theta_0}(\delta(x) = 1) = \alpha$$

<u>example</u> (exp. family)

$$f(x, \theta) = \exp\{c(\theta)T(x) - d(\theta) + S(x)\}$$

$$\Lambda(\vec{x}) = \frac{\exp\{c(\theta_1)\sum T(x_i) - d(\theta_1) + \sum_i S(x_i)\}}{\exp\{c(\theta_0)\sum_i T(x_i) - d(\theta_0) + \sum_i S(x_i)\}} > c$$

$$= \exp\{(c(\theta_1) - c(\theta_0))\sum T(x_i) - d(\theta_1) + d(\theta_0)\} > c$$

$$\Lambda(\vec{x}) = \exp\left\{ (c(\theta_1) - c(\theta_0)) \sum T(x_i) - d(\theta_1) + d(\theta_0) \right\} > c$$

$$\Updownarrow$$

$$[c(\theta_1) - c(\theta_0)] \sum T(x_i) - d(\theta_1) + d(\theta_0) > c' \quad (= \log c)$$

$$\Updownarrow$$

$$[c(\theta_1) - c(\theta_0)] \sum T(x_i) > c''$$

$$\Updownarrow$$

$$\sum_{i=1}^{n} T(x_i) \ \substack{> \\ \text{or} \\ <} \ c'''$$

depending on the sign

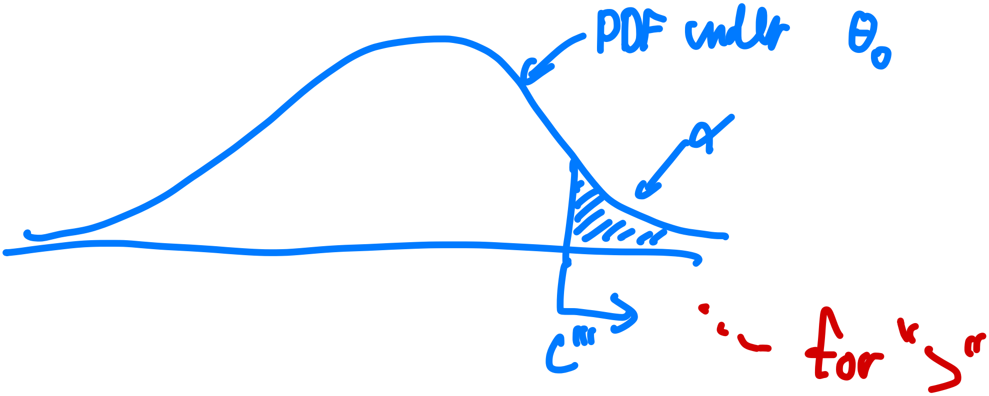
choose c''' s.t.

$$\mathbb{P}_{\theta_0}\left(\sum_i T(x_i) \gtrless c''' \right) = \alpha$$

then $\gamma(\vec{x}) = \mathbb{1}_{\{ \sum_i T(x_i) \gtrless c''' \}}$

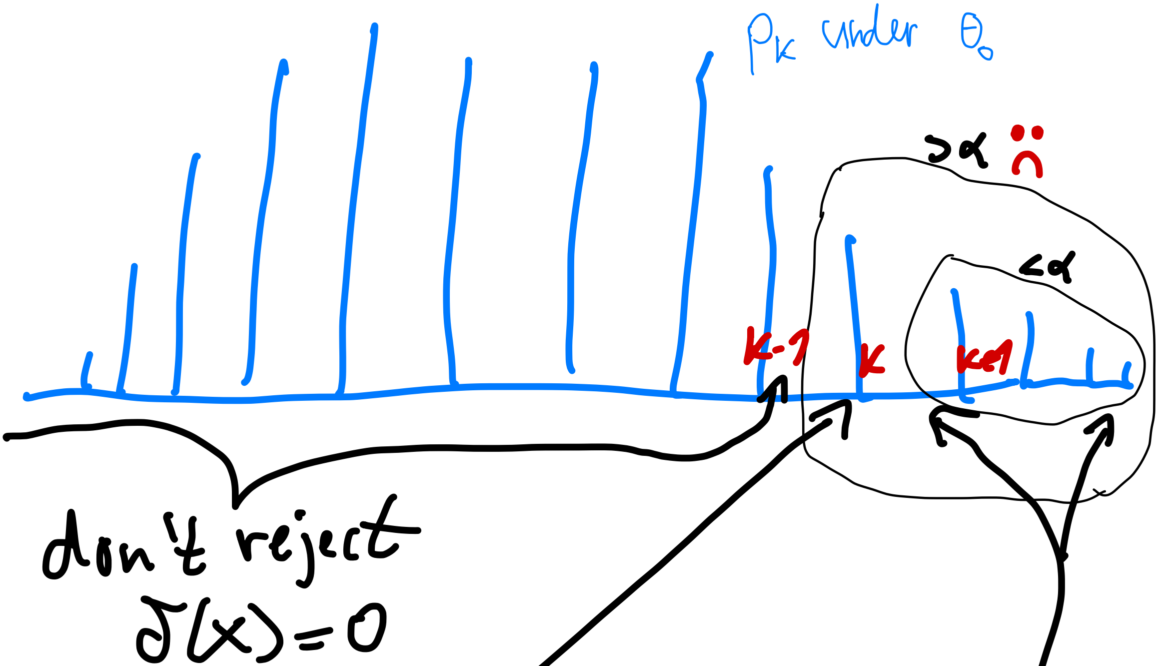
choose c''' s.t.

$$\mathbb{P}_{\theta_0}\left(\sum_i T(X_i) \gtrless c'''\right) = \alpha$$



$$\Updownarrow \quad \frac{\sum_i T(X_i) - n\,\mathbb{E}T(X_i)}{\sqrt{n\,\mathrm{Var}(X_i)}} \gtrless c''''$$

$$\dot\sim N(0,1)$$



randomize! $\delta(x) = p$

reject
$\delta(x) = 1$

$$\mathbb{E}\,\delta(x) = \underbrace{\mathbb{P}(\delta(x) > k)}_{< \alpha} + \underbrace{p}_{} \, \mathbb{P}(\delta(x) = k) + 0 \overset{!}{=} \alpha$$

choose p s.t. $\uparrow$

$p \in (0,1)$

Statistical Theory (Week 11): Testing Statistical Hypotheses II

Erwan Koch

Ecole Polytechnique Fédérale de Lausanne (Institute of Mathematics)

Uniformly Most Powerful Tests

Neyman-Pearson Framework for Testing Hypotheses

The Problem of Hypothesis Testing

- $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ random variables with joint density/frequency $f(\boldsymbol{x}; \theta)$
- $\theta \in \Theta$ where $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \emptyset$
- We observe a realization $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ of $\boldsymbol{X} \sim f_\theta$
- Decide on the basis of $\boldsymbol{x}$ whether $\theta \in \Theta_0$ (H_0) or $\theta \in \Theta_1$ (H_1)

Neyman-Pearson Framework:

1. Fix a significance level α for the test
2. Among all rules respecting the significance level, pick the one that uniformly maximizes power

When H_0/H_1 both simple$\rightarrow$ Neyman-Pearson lemma settles the problem.

$\hookrightarrow$ What about more general structure of Θ_0, Θ_1?

Uniformly Most Powerful Tests

A *uniformly most powerful (UMP) test* of $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ at level α:

1. Respects the level for all $\theta \in \Theta_0$, i.e.,

$$\delta \in \mathscr{D}(\Theta_0, \alpha) = \{\delta : \mathcal{X} \to \{0, 1\} : \mathbb{E}_\theta[\delta] \leq \alpha, \ \forall\, \theta \in \Theta_0\}$$

2. Is most powerful for all $\theta \in \Theta_1$ (for all possible simple alternatives), i.e.,

$$\mathbb{E}_\theta[\delta] \geq \mathbb{E}_\theta[\delta'] \qquad \forall \theta \in \Theta_1 \quad \& \quad \delta' \in \mathscr{D}(\Theta_0, \alpha)$$

Unfortunately UMP tests rarely exist. Why?
E.g., in the situation $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, UMP tests typically do not exist:

- A UMP test must be MP test for any $\theta_1 \neq \theta_0$.
- But the form of the MP test typically differs for $\theta_1 > \theta_0$ and $\theta_1 < \theta_0$!
  $\hookrightarrow$ e.g., recall the example with exponential distribution (week 10)

Example (No UMP test exists)

Let $X \sim$Binom(n, θ) and suppose we want to test:

$$H_0 : \theta = \theta_0 \qquad vs \qquad H_1 : \theta \neq \theta_0$$

at some level α. To this aim, consider first

$$H_0' : \theta = \theta_0 \qquad vs \qquad H_1' : \theta = \theta_1$$

Neyman-Pearson lemma states that an optimal test statistic is

$$\Lambda = \frac{f(X; \theta_1)}{f(X; \theta_0)} = \left(\frac{1-\theta_1}{1-\theta_0}\right)^n \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^X .$$

- If $\theta_1 > \theta_0$ then Λ increasing in X
  $\hookrightarrow$ MP test would reject for large values of X
- If $\theta_1 < \theta_0$ then Λ decreasing in X
  $\hookrightarrow$ MP test would reject for small values of X

Example (A UMP test exists)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\lambda)$ and suppose we wish to test

$$H_0 : \lambda \leq \lambda_0 \qquad vs \qquad H_1 : \lambda > \lambda_0$$

at some level α. To this aim, consider first the pair

$$H_0' : \lambda = \lambda_0 \qquad vs \qquad H_1' : \lambda = \lambda_1$$

with $\lambda_1 > \lambda_0$ which we saw last time to admit a MP test $\forall\ \lambda_1 > \lambda_0$:

Reject H_0' for $\sum_{i=1}^{n} X_i \leq k$, with k such that $\mathbb{P}_{\lambda_0}\left[\sum_{i=1}^{n} X_i \leq k\right] = \alpha$

But for $\lambda < \lambda_0$, $\mathbb{P}_{\lambda_0}\left[\sum_{i=1}^{n} X_i \leq k\right] = \alpha \implies \mathbb{P}_{\lambda}\left[\sum_{i=1}^{n} X_i \leq k\right] < \alpha$. So the same test respects level α for all singletons under H_0.
$\implies$ The test is UMP of H_0 vs H_1

Situations When UMP Tests Exist

When do UMP tests exist?

Previous examples give insight on which composite pairs typically admit UMP tests:

1. Hypothesis pair concerns a single real-valued parameter
2. Hypothesis pair is "one-sided"

But existence of UMP test does not only depend on hypothesis structure... $\hookrightarrow$ Also depends on the specific model considered. Sufficient condition?

Definition (Monotone Likelihood Ratio Property)

A family of density (frequency) functions $\{f(\boldsymbol{x};\theta) : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$ is said to have monotone likelihood ratio (MLR) if there exists a real-valued function $T(\boldsymbol{x})$ such that, for any $\theta_0 < \theta_1$, the function

$$f(\boldsymbol{x};\theta_1)/f(\boldsymbol{x};\theta_0)$$

is non-decreasing wrt $T(\boldsymbol{x})$ for $\boldsymbol{x}$ such that $f(\boldsymbol{x};\theta_1)/f(\boldsymbol{x};\theta_0) \in (0,\infty)$.

Such a statistic T will necessarily be sufficient for θ (Fisher-Neyman).

MLR example

Example

Let $X \sim \text{Binom}(n, \theta)$ and let $\theta_1 > \theta_0$. The likelihood ratio is

$$\frac{f(x, \theta_1)}{f(x, \theta_0)} = \left(\frac{1-\theta_1}{1-\theta_0}\right)^n \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^x,$$

and so it is an increasing function of $T(x) = x$, $x = 0, 1, \ldots, n$.

Intuition: increasing T shifts the likelihood to the right.

When do UMP tests exist?

Theorem (MLR and UMP)

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ have density (or frequency) function depending on $\theta \in \mathbb{R}$ and satisfying the monotone likelihood ratio property with respect to a statistic T. Furthermore, assume that T is a continuous random variable. Then, the test function given by

$$\delta(\boldsymbol{X}) = \begin{cases} 1 & \text{if } T(\boldsymbol{X}) \geq k \\ 0 & \text{if } T(\boldsymbol{X}) < k \end{cases} \qquad k \text{ such that } \mathbb{E}_{\theta_0}[\delta(\boldsymbol{X})] = \alpha$$

is UMP among all tests at level α for the hypothesis pair

$$\begin{cases} H_0: & \theta \leq \theta_0 \\ H_1: & \theta > \theta_0 \end{cases}$$

[The assumption of continuity of the random variable T can be removed, by considering randomized tests as well, similarly as before]

Proof.

We will show that:

1. $\delta \in \mathscr{D}(\Theta_0, \alpha)$, i.e. $\mathbb{E}_\theta[\delta] \le \alpha \, (= \mathbb{E}_{\theta_0}[\delta])$ for all $\theta \in \Theta_0 = (-\infty, \theta_0]$.
2. For any $\delta' \in \mathscr{D}(\Theta_0, \alpha)$ and all $\theta_1 \in \Theta_1$, $\mathbb{E}_{\theta_1}[\delta'] \le \mathbb{E}_{\theta_1}[\delta]$.

To show (1) it suffices to show that $\mathbb{E}_{\theta_0}[\delta] - \mathbb{E}_\theta[\delta] \ge 0$ for $\theta \le \theta_0$. Notice that δ is a non-decreasing function of T. Thus, by the MLR property, it is in fact a non-decreasing function of $f(\boldsymbol{x};\theta_0)/f(\boldsymbol{x};\theta)$ for $\theta \le \theta_0$. Call this function $q(\cdot)$. Then

$$\mathbb{E}_{\theta_0}[\delta] - \mathbb{E}_\theta[\delta] = \int_{\mathcal{X}} q\left(\frac{f(\boldsymbol{x};\theta_0)}{f(\boldsymbol{x};\theta)}\right)(f(\boldsymbol{x};\theta_0) - f(\boldsymbol{x};\theta))d\boldsymbol{x}$$

Letting $A = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x};\theta_0) > f(\boldsymbol{x};\theta)\}$, the RHS becomes

$$\int_A q\left(\frac{f(\boldsymbol{x};\theta_0)}{f(\boldsymbol{x};\theta)}\right)(f(\boldsymbol{x};\theta_0) - f(\boldsymbol{x};\theta))d\boldsymbol{x} + \int_{A^c} q\left(\frac{f(\boldsymbol{x};\theta_0)}{f(\boldsymbol{x};\theta)}\right)(f(\boldsymbol{x};\theta_0) - f(\boldsymbol{x};\theta))d\boldsymbol{x}$$

Letting $q_* = \inf_{\boldsymbol{x}\in A} q\left(\frac{f(\boldsymbol{x};\theta_0)}{f(\boldsymbol{x};\theta)}\right)$ and $q^* = \sup_{\boldsymbol{x}\in A^c} q\left(\frac{f(\boldsymbol{x};\theta_0)}{f(\boldsymbol{x};\theta)}\right)$ we may bound the last expression from below by

$$\begin{aligned}
q_* \int_A (f(\boldsymbol{x};\theta_0) - f(\boldsymbol{x};\theta))d\boldsymbol{x} + q^* \int_{A^c} (f(\boldsymbol{x};\theta_0) - f(\boldsymbol{x};\theta))d\boldsymbol{x} = \\
= q_*(\mathbb{P}_{\theta_0}[A] - \mathbb{P}_\theta[A]) + q^*(\mathbb{P}_{\theta_0}[A^c] - \mathbb{P}_\theta[A^c]) \\
= q_*(\mathbb{P}_{\theta_0}[A] - \mathbb{P}_\theta[A]) + q^*(1 - \mathbb{P}_{\theta_0}[A] - 1 + \mathbb{P}_\theta[A]) \\
= (q_* - q^*)(\mathbb{P}_{\theta_0}[A] - \mathbb{P}_\theta[A]) = (q_* - q^*)\underbrace{\int_A (f(\boldsymbol{x};\theta_0) - f(\boldsymbol{x};\theta))d\boldsymbol{x}}_{\geq 0}.
\end{aligned}$$

Part (1) will thus follow if $q_* - q^* \geq 0$. But q is nondecreasing, so

$$q\left(\frac{f(\boldsymbol{u};\theta_0)}{f(\boldsymbol{u};\theta)}\right) \geq q\left(\frac{f(\boldsymbol{v};\theta_0)}{f(\boldsymbol{v};\theta)}\right), \qquad \forall \boldsymbol{u} \in A \ \& \ \forall \boldsymbol{v} \in A^c,$$

and hence $$q_* = \inf_{\boldsymbol{u}\in A} q\left(\frac{f(\boldsymbol{u};\theta_0)}{f(\boldsymbol{u};\theta)}\right) \geq \sup_{\boldsymbol{v}\in A^c} q\left(\frac{f(\boldsymbol{v};\theta_0)}{f(\boldsymbol{v};\theta)}\right) = q^*.$$

For part (2), note that $\mathscr{D}(\Theta_0, \alpha) \subseteq \mathscr{D}(\{\theta_0\}, \alpha)$, because

$$\phi \in \mathscr{D}(\Theta_0, \alpha) \implies \sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\phi] \leq \alpha \implies \mathbb{E}_{\theta_0}[\phi] \leq \alpha \implies \phi \in \mathscr{D}(\{\theta_0\}, \alpha).$$

Thus, if we show that for any $\delta' \in \mathscr{D}(\{\theta_0\}, \alpha)$ and any $\theta_1 \in \Theta_1$, $\mathbb{E}_{\theta_1}[\delta'] \leq \mathbb{E}_{\theta_1}[\delta]$, assertion (2) will follow. For $\theta_1 \in \Theta_1$, we have $\theta_0 < \theta_1$ and thus $f(\boldsymbol{X}; \theta_1)/f(\boldsymbol{X}; \theta_0) = h(T)$ for some non-decreasing h by the MLR property of T. Let $K = h(k)$ and let

$$I_k = [k - a, k + b], \qquad a, b > 0,$$

the interval on which $h(t) = K$ (this set is an interval since h is non-decreasing; it could also be half open, or open). Define

$$\psi(\boldsymbol{X}) = \begin{cases} 1, & \text{if } f(\boldsymbol{X}; \theta_1) > Kf(\boldsymbol{X}; \theta_0) \\ \mathbb{P}[k \leq T < k + b]/\mathbb{P}[T \in I_k], & \text{if } f(\boldsymbol{X}; \theta_1) = Kf(\boldsymbol{X}; \theta_0). \\ 0, & \text{if } f(\boldsymbol{X}; \theta_1) < Kf(\boldsymbol{X}; \theta_0) \end{cases}$$

Now we note that (recall that T is continuous, so strict inequalities irrelevant)

$$\begin{aligned}\mathbb{E}_\theta[\psi] &= 0 \times \mathbb{P}_\theta[T < k - a] \\ &\quad + \frac{\mathbb{P}_\theta[k \leq T < k + b]}{\mathbb{P}_\theta[T \in I_k]} \mathbb{P}_\theta[T \in I_k] + 1 \times \mathbb{P}_\theta[T \geq k + b] \\ &= \mathbb{P}_\theta[T \geq k] \\ &= \mathbb{E}_\theta[\delta].\end{aligned}$$

Thus, $\mathbb{E}_{\theta_0}[\psi] = \mathbb{E}_{\theta_0}[\delta]$. Therefore, it follows from the generalized NP-lemma that ψ is most powerful at level $\mathbb{E}_{\theta_0}[\delta]$, i.e., $\mathbb{E}_{\theta_1}[\delta'] \leq \mathbb{E}_{\theta_1}[\psi]$ for all $\delta' \in \mathscr{D}(\{\theta_0\}, \alpha)$. As $\mathbb{E}_{\theta_1}[\psi] = \mathbb{E}_{\theta_1}[\delta]$, we obtain that $\mathbb{E}_{\theta_1}[\delta'] \leq \mathbb{E}_{\theta_1}[\delta]$ for all $\delta' \in \mathscr{D}(\{\theta_0\}, \alpha)$ and the proof is complete. $\square$

When do UMP tests exist?

Example (One-Parameter Exponential Family)

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ have a density (frequency)

$$f(\boldsymbol{x}; \theta) = \exp[c(\theta) T(\boldsymbol{x}) - b(\theta) + S(\boldsymbol{x})]$$

and assume WLOG that $c(\theta)$ is strictly increasing. For $\theta_0 < \theta_1$,

$$\frac{f(\boldsymbol{x}; \theta_1)}{f(\boldsymbol{x}; \theta_0)} = \exp\{[c(\theta_1) - c(\theta_0)] T(\boldsymbol{x}) + b(\theta_0) - b(\theta_1)\}$$

is strictly increasing in T by strict increasingness of $c(\cdot)$.

Hence the UMP test defined above of $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$ would reject H_0 iff $T(\boldsymbol{x}) \geq k$, with k such that $\alpha = \mathbb{P}_{\theta_0}[T \geq k]$.

Locally Most Powerful Tests

Locally Most Powerful Tests

$\hookrightarrow$ What if MLR property fails to be satisfied? Can optimality be "saved"?

- Consider $\theta \in \mathbb{R}$ and the test: $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$
- Intuition: if true θ far from θ_0, then any reasonable test powerful

$\star$ So focus on maximizing power in small neighbourhood of θ_0

$\rightarrow$ Consider power function $\beta(\theta) = \mathbb{E}_\theta[\delta(\boldsymbol{X})]$ of some δ

$\rightarrow$ Require $\beta(\theta_0) = \alpha$ (notice that $\theta_0 \in \Theta_0$ so $\beta(\theta_0)$ is the probability of type I error)

$\rightarrow$ Assume that $\beta(\theta)$ is differentiable, so for θ close to θ_0 and such that $\theta > \theta_0$,

$$\beta(\theta) \approx \beta(\theta_0) + \beta'(\theta_0)(\theta - \theta_0) = \alpha + \beta'(\theta_0)\underbrace{(\theta - \theta_0)}_{>0}.$$

Since $\Theta_1 = (\theta_0, \infty)$, this suggests approach for locally most powerful test

Choose δ to Maximize $\beta'(\theta_0)$ Subject to $\beta(\theta_0) = \alpha$

How do we solve this constrained optimization problem?

Supposing that $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ has density $f(\boldsymbol{x}; \theta)$, then

$$
\begin{aligned}
& \beta(\theta) = \int_{\mathbb{R}^n} \delta(\boldsymbol{x}) f(\boldsymbol{x}; \theta) d\boldsymbol{x} \\
\Longrightarrow \quad & \frac{\partial}{\partial \theta}\beta(\theta) = \int_{\mathbb{R}^n} \delta(\boldsymbol{x}) \frac{\partial}{\partial \theta} f(\boldsymbol{x}; \theta) d\boldsymbol{x} \quad \text{[provided interchange possible]} \\
= \quad & \int_{\mathbb{R}^n} \delta(\boldsymbol{x}) \frac{f(\boldsymbol{x}; \theta)}{f(\boldsymbol{x}; \theta)} \frac{\partial}{\partial \theta} f(\boldsymbol{x}; \theta) d\boldsymbol{x} \\
= \quad & \int_{\mathbb{R}^n} \delta(\boldsymbol{x}) \left[\frac{\partial}{\partial \theta} \log f(\boldsymbol{x}; \theta)\right] f(\boldsymbol{x}; \theta) d\boldsymbol{x} \\
= \quad & \mathbb{E}_\theta \left[\delta(\boldsymbol{X}) \underbrace{\frac{\partial}{\partial \theta} \log f(\boldsymbol{X}; \theta)}_{S(\boldsymbol{X}; \theta)}\right] = \text{Cov}(\delta, S(\boldsymbol{X}, \theta))
\end{aligned}
$$

The last equality follows if we can differentiate under the integral, in which case $\mathbb{E}[S(\boldsymbol{X}; \theta)] = 0$. So δ must be a "linear functional" of $S(\boldsymbol{X}; \theta)$!

Locally Most Powerful Tests

Theorem (Score Tests are Locally Most Powerful)

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ have density (frequency) $f(\boldsymbol{x};\theta)$ and define the test function

$$\delta(\boldsymbol{X}) = \begin{cases} 1 & \text{if } S(\boldsymbol{X};\theta_0) \geq k, \\ 0 & \text{otherwise} \end{cases}$$

where k is such that $\mathbb{E}_{\theta_0}[\delta(\boldsymbol{X})] = \alpha$. Then δ maximizes

$$\mathbb{E}_{\theta_0}\left[\psi(\boldsymbol{X}) S(\boldsymbol{X};\theta_0)\right]$$

over all test functions ψ satisfying the constraint $\mathbb{E}_{\theta_0}[\psi(\boldsymbol{X})] = \alpha$.

- Gives recipe for constructing LMP test
- We were concerned about power *only locally around* θ_0
- **BEWARE !** May not even give a level α test for some $\theta < \theta_0$

Proof.

Consider ψ with $\psi(\boldsymbol{x}) \in \{0,1\}\ \forall\ \boldsymbol{x}$ and $\mathbb{E}_{\theta_0}[\psi(\boldsymbol{X})] = \alpha$. Then,

$$\delta(\boldsymbol{x}) - \psi(\boldsymbol{x}) = \begin{cases} \geq 0 & \text{if } S(\boldsymbol{x};\theta_0) \geq k, \\ \leq 0 & \text{if } S(\boldsymbol{x};\theta_0) \leq k. \end{cases}$$

Therefore

$$\mathbb{E}_{\theta_0}[(\delta(\boldsymbol{X}) - \psi(\boldsymbol{X}))(S(\boldsymbol{X};\theta_0) - k)] \geq 0.$$

Expanding the product and since $\mathbb{E}_{\theta_0}[\delta(\boldsymbol{X}) - \psi(\boldsymbol{X})] = 0$, we obtain

$$\mathbb{E}_{\theta_0}\left[\delta(\boldsymbol{X})S(\boldsymbol{X};\theta_0)\right] \geq \mathbb{E}_{\theta_0}\left[\psi(\boldsymbol{X})S(\boldsymbol{X};\theta_0)\right]$$

□

How is the critical value k evaluated in practice? (obviously to give level α)

- When $X_1, \ldots, X_n$ are iid, then $S(\boldsymbol{X};\theta) = \sum_{i=1}^n \ell'(X_i;\theta)$
- Under regularity conditions, sum of iid random variables with mean zero and variance $I(\theta)$.
- Hence, for $\theta = \theta_0$ and large n, $S(\boldsymbol{X};\theta) \overset{d}{\approx} \mathcal{N}(0, nI(\theta))$

Example (Cauchy distribution)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Cauchy}(\theta)$ with density

$$f(x;\theta) = \frac{1}{\pi(1 + (x-\theta)^2)}, \quad x \in \mathbb{R},$$

and consider the hypothesis pair $\begin{cases} H_0 : & \theta \geq 0 \\ H_1 : & \theta < 0. \end{cases}$

We have

$$S(\boldsymbol{X}; 0) = \sum_{i=1}^{n} \frac{2X_i}{1 + X_i^2}$$

so that the LMP test at level α rejects the null if $S(\boldsymbol{X}; 0) \leq k$, where

$$\mathbb{P}_0[S(\boldsymbol{X}; 0) \leq k] = \alpha.$$

While the exact distribution is difficult to obtain, for large n, $S(\boldsymbol{X}; 0) \overset{d}{\approx} \mathcal{N}(0, n/2)$.

Likelihood Ratio Tests

Likelihood Ratio Tests

So far, tests for $\theta \in \mathbb{R}$ with simple vs simple or one sided vs one sided hypothesis.
$\hookrightarrow$ Extension to multiparameter case $\boldsymbol{\theta} \in \mathbb{R}^p$? General Θ_0, Θ_1?

- Unfortunately, optimality theory breaks down in higher dimensions and for more general Θ_0, Θ_1.
- General method for constructing *reasonable* tests?

$\rightarrow$ The idea: Combine Neyman-Pearson paradigm with Max Likelihood

Definition (Likelihood Ratio)

The *likelihood ratio (LR) statistic* corresponding to the pair of hypotheses $H_0 : \boldsymbol{\theta} \in \Theta_0$ vs $H_1 : \boldsymbol{\theta} \in \Theta_1$ is defined to be

$$\Lambda(\boldsymbol{X}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta_1} f(\boldsymbol{X}; \theta)}{\sup_{\boldsymbol{\theta} \in \Theta_0} f(\boldsymbol{X}; \theta)} = \frac{\sup_{\boldsymbol{\theta} \in \Theta_1} L(\theta)}{\sup_{\boldsymbol{\theta} \in \Theta_0} L(\theta)}$$

- "*Neyman-Pearson*"-esque approach: reject H_0 for large Λ.
- Intuition: choose the "most favourable" $\theta \in \Theta_0$ (in favour of H_0) and compare it against the "most favourable" $\theta \in \Theta_1$ (in favour of H_1) in a simple vs simple setting (applying NP-lemma)
- Provided the likelihood is continuous wrt θ and Θ_0 is a lower dimensional subspace of Θ, then $\sup_{\boldsymbol{\theta} \in \Theta_1} L(\theta) = \sup_{\boldsymbol{\theta} \in \Theta} L(\theta)$. In those cases, for convenience of the MLE computation, we generally take $\sup_{\boldsymbol{\theta} \in \Theta} L(\theta)$ as numerator in the above definition.

Example

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Consider

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \mu \neq \mu_0.$$

We have

$$\Lambda(\boldsymbol{X}) = \frac{\sup_{(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}^+} f(\boldsymbol{X};\mu,\sigma^2)}{\sup_{(\mu,\sigma^2)\in\{\mu_0\}\times\mathbb{R}^+} f(\boldsymbol{X};\mu,\sigma^2)} = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right)^{\frac{n}{2}} = \left(\frac{\sum_{i=1}^n (X_i-\mu_0)^2}{\sum_{i=1}^n (X_i-\bar{X})^2}\right)^{\frac{n}{2}}.$$

We reject H_0 when $\Lambda \geq k$, where k is s.t. $\mathbb{P}_0[\Lambda \geq k] = \alpha$. Distribution of Λ? By monotonicity look only at

$$\begin{aligned}\frac{\sum_{i=1}^n (X_i-\mu_0)^2}{\sum_{i=1}^n (X_i-\bar{X})^2} &= 1 + \frac{n(\bar{X}-\mu_0)^2}{\sum_{i=1}^n (X_i-\bar{X})^2} = 1 + \frac{1}{n-1}\left(\frac{n(\bar{X}-\mu_0)^2}{S^2}\right) \\ &= 1 + \frac{T^2}{n-1}.\end{aligned}$$

Denoting $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i-\bar{X})^2$, we have $T = \sqrt{n}(\bar{X}-\mu_0)/S \overset{H_0}{\sim} t_{n-1}$. So $T^2 \overset{H_0}{\sim} F_{1,n-1}$ and k may be chosen appropriately.

Example

Let $X_1, \ldots, X_m \overset{iid}{\sim} \text{Exp}(\lambda)$ and $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Exp}(\theta)$ and $\boldsymbol{X}$ indep $\boldsymbol{Y}$.

$$\text{Consider:} \quad H_0 : \theta = \lambda \qquad \text{vs} \qquad H_1 : \theta \neq \lambda.$$

$$\underset{\sup_{(\lambda,\theta)\in\mathbb{R}_+^2} f(\boldsymbol{X},\boldsymbol{Y};\lambda,\theta)}{\text{Unrestricted MLEs:}} \quad \hat{\lambda} = 1/\bar{X} \qquad \& \qquad \hat{\theta} = 1/\bar{Y}.$$

$$\underset{\sup_{(\lambda,\theta)\in\{(x,y)\in\mathbb{R}_+^2 : x=y\}} f(\boldsymbol{X},\boldsymbol{Y};\lambda,\theta)}{\text{Restricted MLEs:}} \quad \hat{\lambda}_0 = \hat{\theta}_0 = \left[\frac{m\bar{X} + n\bar{Y}}{m+n}\right]^{-1}.$$

$$\implies \Lambda = \left(\frac{m}{m+n} + \frac{n}{n+m}\frac{\bar{Y}}{\bar{X}}\right)^m \left(\frac{n}{n+m} + \frac{m}{m+n}\frac{\bar{X}}{\bar{Y}}\right)^n.$$

Depends on $T = \bar{X}/\bar{Y}$ and can make Λ large/small by varying T.

$\hookrightarrow$ But $T \overset{H_0}{\sim} F_{2m,2n}$ so given α we may find the critical value k.

Distribution of Likelihood Ratio?

More often than not, dist(Λ) intractable (and no simple dependence on a statistic T having tractable distribution).

Consider asymptotic approximations?

Setup:

- Θ open subset of $\mathbb{R}^p$
- either $\Theta_0 = \{\boldsymbol{\theta}_0\}$ or Θ_0 open subset of $\mathbb{R}^s$, where $s < p$
- $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ where the components are iid
- Initially restrict attention to $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. LR becomes:

$$\Lambda_n(\boldsymbol{X}) = \prod_{i=1}^{n} \frac{f(X_i; \hat{\boldsymbol{\theta}}_n)}{f(X_i; \boldsymbol{\theta}_0)}$$

  where $\hat{\boldsymbol{\theta}}_n$ is the MLE of $\boldsymbol{\theta}$.

- Impose regularity conditions from MLE asymptotics

Asymptotic Distribution of the Likelihood Ratio

Theorem (Wilks' Theorem, case $p = 1$)

Let $X_1, \ldots, X_n$ be iid random variables with density (frequency) depending on $\theta \in \mathbb{R}$ and satisfying conditions (A1)-(A6), with $I(\theta) = J(\theta)$. If the MLE sequence $\hat{\theta}_n$ is consistent for θ, then the likelihood ratio statistic Λ_n for $H_0 : \theta = \theta_0$ satisfies

$$2 \log \Lambda_n \xrightarrow{d} V \sim \chi_1^2$$

when H_0 is true.

- Obviously, knowing approximate distribution of $2 \log \Lambda_n$ is as good as knowing approximate distribution of Λ_n for the purposes of testing (by monotonicity and rejection method).
- Theorem extends immediately and trivially to the case of general p and for a hypothesis pair $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.
  (i.e. when null hypothesis is simple)

Asymptotic Distribution of the Likelihood Ratio

Proof.

Let $\ell(x;\theta) = \log f(x;\theta)$, $x \in \mathcal{X}$. By a Taylor series expansion around $\hat{\theta}_n$,

$$\begin{aligned}
\log \Lambda_n &= \sum_{i=1}^{n}[\ell(X_i;\hat{\theta}_n) - \ell(X_i;\theta_0)] = \sum_{i=1}^{n}[\ell(X_i;\hat{\theta}_n) - \ell(X_i;\hat{\theta}_n)] \\
&\qquad -(\theta_0 - \hat{\theta}_n)\sum_{i=1}^{n}\ell'(X_i;\hat{\theta}_n) - \frac{1}{2}(\hat{\theta}_n - \theta_0)^2\sum_{i=1}^{n}\ell''(X_i;\theta_n^*) \\
&= -\frac{1}{2}n(\hat{\theta}_n - \theta_0)^2\frac{1}{n}\sum_{i=1}^{n}\ell''(X_i;\theta_n^*)
\end{aligned}$$

where θ_n^* lies between $\hat{\theta}_n$ and θ_0.

Asymptotic Distribution of the Likelihood Ratio

If H_0 is true, then $\hat{\theta}_n \xrightarrow{p} \theta_0$ by assumption. Hence, as θ_n^* lies between $\hat{\theta}_n$ and θ_0, we have

$$\theta_n^* \xrightarrow{p} \theta_0.$$

Hence under (A1)-(A6) and if H_0 is true, a first order Taylor expansion about θ_0, Slutsky's theorem and the WLLN give

$$-\frac{1}{n}\sum_{i=1}^{n} \ell''(X_i; \theta_n^*) \xrightarrow{p} -\mathbb{E}_{\theta_0}[\ell''(X_i; \theta_0)] = I(\theta_0).$$

Now, under the conditions of the theorem and when H_0 is true,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta_0)).$$

which, by the continuous mapping theorem, yields

$$n(\hat{\theta}_n - \theta_0)^2 \xrightarrow{d} \frac{V}{I(\theta_0)}.$$

Slutsky's theorem gives the result. □

Asymptotic Distribution of the Likelihood Ratio

Theorem (Wilk's theorem, general p, general $r \leq p$)

Let $X_1, \ldots, X_n$ be iid random variables with density (frequency) depending on $\boldsymbol{\theta} \in \mathbb{R}^p$ and satisfying conditions (B1)-(B6), with $I(\boldsymbol{\theta}) = J(\boldsymbol{\theta})$. If the MLE sequence $\hat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}$, then the likelihood ratio statistic Λ_n for $H_0 : \{\theta_j = \theta_{j,0}\}_{j=1}^r$ satisfies $2 \log \Lambda_n \xrightarrow{d} V \sim \chi_r^2$ when H_0 is true.

Exercise

Prove Wilks' theorem. Note that it may potentially be that $r < p$: some of the components of $\boldsymbol{\theta}$ might be adjustable under H_0!

Hypotheses of the form $H_0 : \{g_j(\boldsymbol{\theta}) = a_j\}_{j=1}^r$, for g_j differentiable real-valued functions, can also be handled by Wilks' theorem:

- Define $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\top = g(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \ldots, g_p(\boldsymbol{\theta}))^\top$
- $g_{r+1}, \ldots, g_p$ defined so that $\boldsymbol{\theta} \mapsto g(\boldsymbol{\theta})$ is 1-1
- Apply theorem with parameter $\boldsymbol{\phi}$

Other Tests?

Many other tests possible once we "liberate" ourselves from strict optimality criteria. For example:

- Wald's test
  - ↪ For a simple null, may compare the unrestricted MLE with the MLE under the null. Large deviations indicate evidence against null hypothesis. Distributions are approximated for large n via the asymptotic normality of MLEs.
- Score Test
  - ↪ For a simple null, if the null hypothesis is false, then the loglikelihood gradient at the null should not be close to zero, at least when n reasonably large: so measure its deviations from zero. Use asymptotics for distributions (under conditions we end up with a χ^2)
- ...

Summary

- In general, UMP tests do not exist, because they would need to be MP for all pairs: $\theta_0 \in \Theta_0, \theta_1 \in \Theta_1$. However, in the case of a real-valued parameter:
  - If there is a monotone LR, one-sided vs one-sided situation has a MP test.
  - We can consider locally MP tests like the score test.
- When the parameter is a vector and/or we want to test: $\theta = \theta_0$ vs $\theta \neq \theta_0$, we need to give up on optimality.
- But we can extend the likelihood-ratio test to these situations. Wilks' theorem gives us the asymptotic sampling distribution of the likelihood-ratio under the null hypothesis.
- Other tests can also be used.

Statistical Theory (Week 12): From Hypothesis Tests to Confidence Regions

Erwan Koch

Ecole Polytechnique Fédérale de Lausanne (Institute of Mathematics)

EPFL

p-values

Beyond Neyman-Pearson?

So far we have considered the Neyman-Pearson Framework:

1. Fix a significance level α for the test
2. Consider the rules δ respecting this significance level
   $\hookrightarrow$ We choose one of those rules, δ^*, based on power considerations
3. We reject at level α if $\delta^*(\boldsymbol{x}) = 1$.

Useful for attempting to determine optimal test statistics

What if we already have a given form of test statistic in mind (e.g., LRT)?

$\hookrightarrow$ A different perspective on testing (used more in practice) says:

Rather than considering a family of test functions respecting level α...

... consider a family of test functions indexed by α

1. Fix a family $\{\delta_\alpha\}_{\alpha \in (0,1)}$ of decision rules, with δ_α having level α
   $\hookrightarrow$ for a given $\boldsymbol{x}$ some of these rules reject the null while others do not
2. Which is the smallest α for which H_0 is rejected given $\boldsymbol{x}$?

Observed Significance Level

Definition (p–Value)

Let $\{\delta_\alpha\}_{\alpha\in(0,1)}$ be a family of test functions satisfying

$$\alpha_1 < \alpha_2 \implies \{\boldsymbol{x} \in \mathcal{X} : \delta_{\alpha_1}(\boldsymbol{x}) = 1\} \subseteq \{\boldsymbol{x} \in \mathcal{X} : \delta_{\alpha_2}(\boldsymbol{x}) = 1\}.$$

The p–value (or observed significance level) of the family $\{\delta_\alpha\}$ is

$$p(\boldsymbol{x}) = \inf\{\alpha : \delta_\alpha(\boldsymbol{x}) = 1\}.$$

$\hookrightarrow$ The p–value is the smallest value of α for which the null would be rejected at level α, given $\boldsymbol{X} = \boldsymbol{x}$.

Most usual setup:

- We have $\delta_\alpha(\boldsymbol{x}) = \mathbf{1}\{T(\boldsymbol{x}) > k_\alpha\}$, where T is a single test statistic
- Then

$$p(\boldsymbol{x}) = \mathbb{P}_{H_0}[T(\boldsymbol{X}) \geq T(\boldsymbol{x})] = 1 - G(T(\boldsymbol{x})),$$

  where G is the df of T under H_0

Observed Significance Level

Notice: contrary to NP-framework we did not make explicit decision!

- We simply reported a p–value
- The p–value is used as a measure of evidence against H_0
  - $\hookrightarrow$ Small p–value provides evidence against H_0
  - $\hookrightarrow$ Large p–value provides no evidence against H_0
- How small does "small" mean?
  - $\hookrightarrow$ Depends on the specific problem...

Intuition:

- Recall that extreme values of test statistics are those that are "inconsistent" with the null (NP-framework)
- p–value = probability under the null of observing a value of the test statistic as extreme as or more extreme than the one we observed
- If this probability is small, then we have witnessed something quite unusual under the null

  $\Longrightarrow$ Gives evidence against the null hypothesis

Example (Normal Mean)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Consider

$$H_0 : \mu = 0 \qquad \text{vs} \qquad H_1 : \mu \neq 0.$$

Likelihood ratio test: reject when T^2 large, where $T = \sqrt{n}\bar{X}/S \overset{H_0}{\sim} t_{n-1}$.
Since $T^2 \overset{H_0}{\sim} F_{1,n-1}$, p–value is

$$p(\boldsymbol{x}) = \mathbb{P}_{H_0}[T^2(\boldsymbol{X}) \geq T^2(\boldsymbol{x})] = 1 - G_{F_{1,n-1}}(T^2(\boldsymbol{x})).$$

With the samples (datasets)

$$\boldsymbol{x} = (0.66, 0.28, -0.99, 0.007, -0.29, -1.88, -1.24, 0.94, 0.53, -1.2)$$

$$\boldsymbol{y} = (1.4, 0.48, 2.86, 1.02, -1.38, 1.42, 2.11, 2.77, 1.02, 1.87),$$

we obtain $p(\boldsymbol{x}) = 0.32$ while $p(\boldsymbol{y}) = 0.006$.

Significance VS Decision

- Reporting a p–value does not necessarily mean making a decision
- A small p–value can simply reflect our "confidence" in rejecting a null
  - $\hookrightarrow$ reflects how statistically significant the alternative statement is

Example

Statisticians working for Obama gather an iid sample $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ from Ohio with $X_i = \mathbf{1}\{\text{vote Obama}\}$. Obama's team wants to test

$$\begin{cases} H_0: & \text{Romney wins Ohio} \\ H_1: & \text{Obama wins Ohio} \end{cases}$$

Should statisticians decide for Obama? Perhaps better to report p–value to him and let him decide...

What if statisticians work for newspapers and not Obama?

$\hookrightarrow$ Something easier to interpret than test/p–value?

Confidence Intervals

A Glance Back at Point Estimation

- Let $X_1, ..., X_n$ be iid random variables with density (frequency) $f(\cdot;\theta)$.
- Problem with point estimation: $\mathbb{P}_\theta[\hat{\theta} = \theta]$ typically small (if not zero)
  - $\hookrightarrow$ We always attach an estimator of variability, e.g., its standard error. Interpretation?
- Hypothesis tests may provide way to interpret estimator's variability within the setup of a particular problem
  - $\hookrightarrow$ e.g., if we observe $\hat{P}[\text{obama wins}] = 0.52$, we can see what p–value we get when testing $H_0 : P[\text{obama wins}] \geq 1/2$ or $H_0 : P[\text{Obama wins}] < 1/2$.
- Something more directly interpretable?

Back to our example: What do pollsters do in newspapers?

$\hookrightarrow$ They announce their point estimate (e.g., 0.52)

$\hookrightarrow$ They give upper and lower confidence limits

What are these and how are they interpreted?

Interval Estimation

Simple underlying idea:

- Instead of estimating θ by a single value
- Present a whole range of values for θ that are consistent with the data
  $\hookrightarrow$ In the sense that they could have produced the data

Definition (Confidence Interval)

Let $\boldsymbol{X} = (X_1, ..., X_n)^\top$ be a random vector with distribution depending on $\theta \in \mathbb{R}$, $L(\boldsymbol{X})$ and $U(\boldsymbol{X})$ be two statistics with $L(\boldsymbol{X}) < U(\boldsymbol{X})$ a.s., and $\alpha \in (0, 1)$. Then, the random interval $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is called a $100(1-\alpha)\%$ confidence interval (CI) for θ if

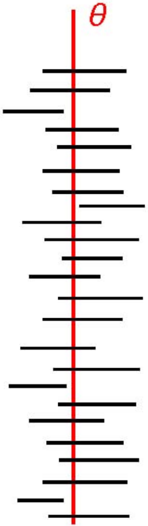
$$\mathbb{P}_\theta[L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})] \geq 1 - \alpha$$

for all $\theta \in \Theta$, with equality for at least one value of θ.

- $1 - \alpha$ is called the coverage probability or confidence level
- Beware of interpretation!

Interval Estimation: Interpretation

- Probability statement is NOT made about θ, which is constant.
- Statement is about the random interval: probability that the random interval contains the true value is at least $1 - \alpha$.
- Given any realization $\boldsymbol{X} = \boldsymbol{x}$, the interval $[L(\boldsymbol{x}), U(\boldsymbol{x})]$ will either contain or not contain θ.
- Interpretation: we expect that $100(1 - \alpha)\%$ of the time our intervals will contain the true value.



Example

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$. Then $\sqrt{n}(\bar{X} - \mu) \sim \mathcal{N}(0, 1)$, so that

$$\mathbb{P}_\mu[-1.96 \leq \sqrt{n}(\bar{X} - \mu) \leq 1.96] = 0.95.$$

Since

$$-1.96 \leq \sqrt{n}(\bar{X} - \mu) \leq 1.96 \iff \bar{X} - 1.96/\sqrt{n} \leq \mu \leq \bar{X} + 1.96/\sqrt{n}$$

we obviously have

$$\mathbb{P}_\mu\left[\bar{X} - \frac{1.96}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{1.96}{\sqrt{n}}\right] = 0.95.$$

So the random interval $[L(\boldsymbol{X}), U(\boldsymbol{X})] = \left[\bar{X} - \frac{1.96}{\sqrt{n}}, \bar{X} + \frac{1.96}{\sqrt{n}}\right]$ is a 95% confidence interval for μ.

Using the CLT, the same argument yields approximate 95% CIs when $X_1, ..., X_n$ are iid with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = 1$, regardless of their distribution.

The Pivotal Method

Pivotal Quantities

What can we learn from previous example?

Definition (Pivot)

A random function $g(\boldsymbol{X}, \theta)$ is said to be a pivotal quantity (or simply a pivot) if it is a function of both $\boldsymbol{X}$ and θ, but whose distribution does not depend on θ.

$\hookrightarrow \sqrt{n}(\bar{X} - \mu) \sim \mathcal{N}(0, 1)$ is a pivot in previous example

Why is a pivot useful?

- $\forall\ \alpha \in (0, 1)$ we can find constants $a < b$ independent of θ, such that

$$\mathbb{P}_\theta[a \leq g(\boldsymbol{X}, \theta) \leq b] = 1 - \alpha \qquad \forall\ \theta \in \Theta$$

- If $g(\boldsymbol{X}, \theta)$ can be manipulated then the above yields a CI

Example

Let $X_1, ..., X_n \overset{iid}{\sim} \text{Unif}(0, \theta)$. Recall that MLE of θ is $\hat{\theta} = X_{(n)}$, with distribution

$$\mathbb{P}_\theta\left[X_{(n)} \leq x\right] = F_{X_{(n)}}(x) = \left(\frac{x}{\theta}\right)^n, \quad x \in [0, \theta],$$

i.e.,

$$\mathbb{P}_\theta\left[\frac{X_{(n)}}{\theta} \leq y\right] = y^n, y \in [0, 1].$$

Thus $X_{(n)}/\theta$ is a pivot for θ and we can choose $a < b$ such that

$$\mathbb{P}_\theta\left[a \leq \frac{X_{(n)}}{\theta} \leq b\right] = 1 - \alpha.$$

$\rightarrow$ But there are ∞-many such choices!
$\hookrightarrow$ Idea: choose a pair (a, b) that minimizes interval's length! Solution can be seen to be $a = \alpha^{1/n}$ and $b = 1$, yielding

$$\left[X_{(n)}, \frac{X_{(n)}}{\alpha^{1/n}}\right].$$

Comments on Pivotal Quantities

Pivotal method extends to construction of CI for θ_k, when

$$\boldsymbol{\theta} = (\theta_1, ..., \theta_k, ..., \theta_p) \in \mathbb{R}^p$$

and the remaining coordinates are also unknown. $\rightarrow$ Pivotal quantity should now be function $g(\boldsymbol{X}; \theta_k)$ which

1. Depends on $\boldsymbol{X}$, θ_k, but no other parameters
2. Has a distribution independent of any of the parameters

$\hookrightarrow$ e.g.: CI for normal mean, when variance unknown

$\rightarrow$ Main difficulties with pivotal method:

- Hard to find exact pivots in general problems
- Exact distributions may be unknown or intractable

$\Longrightarrow$ We often resort to asymptotic approximations...

$\hookrightarrow$ Most classic example: $a_n(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$.

Extension to Confidence Regions

Confidence Regions

What about higher dimensional parameters?

Definition (Confidence Region)

Let $\boldsymbol{X} = (X_1, ..., X_n)^\top$ be a random vector with distribution depending on $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. A random subset $R(\boldsymbol{X})$ of Θ depending on $\boldsymbol{X}$ is called a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}$ if

$$\mathbb{P}_\theta[R(\boldsymbol{X}) \ni \boldsymbol{\theta}] \geq 1 - \alpha$$

for all $\boldsymbol{\theta} \in \Theta$, with equality for at least one value of $\boldsymbol{\theta}$.

- No restriction requiring $R(\boldsymbol{X})$ to be convex or even contiguous
  - $\hookrightarrow$ So when $p = 1$ we get more general notion than CI
- Nevertheless, many notions extend immediately to CR case
  - $\hookrightarrow$ e.g. notion of a pivotal quantity

Pivots for Confidence Regions

Let $g : \mathcal{X} \times \Theta \to \mathbb{R}$ be a function such that $\text{dist}[g(\boldsymbol{X}, \boldsymbol{\theta})]$ independent of $\boldsymbol{\theta}$
$\hookrightarrow$ Since image space is the real line, we can find $a < b$ s.t.

$$\mathbb{P}_{\boldsymbol{\theta}}[a \leq g(\boldsymbol{X}, \boldsymbol{\theta}) \leq b] = 1 - \alpha,$$

i.e.,

$$\mathbb{P}_{\boldsymbol{\theta}}[R(\boldsymbol{X}) \ni \boldsymbol{\theta}] = 1 - \alpha$$

where $R(\boldsymbol{x}) = \{\boldsymbol{\theta} \in \Theta : g(\boldsymbol{x}, \boldsymbol{\theta}) \in [a, b]\}$.

Notice that region can be "wild" since it is a random fibre of g

Example

Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \overset{iid}{\sim} \mathcal{N}_k(\boldsymbol{\mu}, \Sigma)$. Two unbiased estimators of $\boldsymbol{\mu}$ and Σ are

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i$$

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})^T$$

Example (cont'd)

Consider the random variable

$$g(\{\boldsymbol{X}_i\}_{i=1}^n, \boldsymbol{\mu}) := \frac{n(n-k)}{k(n-1)}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T \hat{\Sigma}^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}),$$

which is known to follow $F_{k,n-k}$. A pivot!
$\hookrightarrow$ If f_q is q-quantile of this distribution, then we get as $100(1-\alpha)\%$ CR for $\boldsymbol{\mu}$

$$R(\{\boldsymbol{X}_i\}_{i=1}^n) = \left\{\boldsymbol{\mu} \in \mathbb{R}^k : \frac{n(n-k)}{k(n-1)}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T \hat{\Sigma}^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \leq f_{1-\alpha}\right\}$$

- An ellipsoid in $\mathbb{R}^k$
- Ellipsoid centred at $\hat{\boldsymbol{\mu}}$
- Principle axis lengths given by eigenvalues of $\hat{\Sigma}^{-1}$
- Orientation given by eigenvectors of $\hat{\Sigma}^{-1}$

Getting Confidence Regions from Confidence Intervals

Visualisation of high-dimensional CR's can be hard

- When these are ellipsoids, spectral decomposition helps
- But more generally?

Things especially easy when dealing with rectangles - but they rarely occur!
$\hookrightarrow$ What if we construct a CR as Cartesian product of CI's?

Let $[L_i(\boldsymbol{X}), U_i(\boldsymbol{X})]$ be $100q_i\%$ CI's for θ_i, $i = 1, ..., p$, and define

$$R(\boldsymbol{X}) = [L_1(\boldsymbol{X}), U_1(\boldsymbol{X})] \times \ldots \times [L_p(\boldsymbol{X}), U_p(\boldsymbol{X})]$$

Bonferroni's inequality implies that

$$\mathbb{P}_{\boldsymbol{\theta}}[R(\boldsymbol{X}) \ni \boldsymbol{\theta}] \geq 1 - \sum_{i=1}^{p} \mathbb{P}[\theta_i \notin [L_i(\boldsymbol{X}), U_i(\boldsymbol{X})]] = 1 - \sum_{i=1}^{p}(1 - q_i)$$

$\rightarrow$ So pick q_i such that $\sum_{i=1}^{p}(1 - q_i) = \alpha$ (can be conservative...)

Inverting Hypothesis Tests

Confidence Intervals and Hypothesis Tests

- Discussion on CIs/CRs $\rightarrow$ no guidance to choose "good" regions
- But: $\exists$ close relationship between CR's and HT's! $\hookrightarrow$ can be exploited to transform good testing properties into good CR properties

From CR to HP

Suppose $R(\boldsymbol{X})$ is an exact $100(1-\alpha)\%$ CR for $\boldsymbol{\theta}$. Consider

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \qquad vs \qquad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

Define test function:

$$\delta(\boldsymbol{X}) = \begin{cases} 1 & \text{if } \boldsymbol{\theta}_0 \notin R(\boldsymbol{X}), \\ 0 & \text{if } \boldsymbol{\theta}_0 \in R(\boldsymbol{X}). \end{cases}$$

Then, $$\mathbb{E}_{\boldsymbol{\theta}_0}[\delta(\boldsymbol{X})] = 1 - \mathbb{P}_{\boldsymbol{\theta}_0}[\boldsymbol{\theta}_0 \in R(\boldsymbol{X})] \leq \alpha.$$

$\Longrightarrow$ We can use a CR to construct test with significance level α!

From HT to CR

Going the other way around, we can invert tests to get CRs. Suppose we have tests at level α for any $\boldsymbol{\theta}_0 \in \Theta$. Let $\delta(\boldsymbol{X}; \boldsymbol{\theta}_0)$ denote the appropriate test function for a given $\boldsymbol{\theta}_0$.

Define

$$R^*(\boldsymbol{X}) = \{\boldsymbol{\theta}_0 : \delta(\boldsymbol{X}; \boldsymbol{\theta}_0) = 0\}.$$

Coverage probability of $R^*(\boldsymbol{X})$ is

$$\mathbb{P}_{\boldsymbol{\theta}}[R^*(\boldsymbol{X}) \ni \boldsymbol{\theta}] = \mathbb{P}_{\boldsymbol{\theta}}[\delta(\boldsymbol{X}; \boldsymbol{\theta}) = 0] \geq 1 - \alpha.$$

$\implies$ We obtain a $100(1-\alpha)\%$ confidence region by choosing all the $\boldsymbol{\theta}$ for which the null would not be rejected given our data $\boldsymbol{X}$.

$\hookrightarrow$ If test inverted is powerful, then we get a “small” region for given $1 - \alpha$.

Summary

- p-values provide an alternative framework for hypothesis testing:
  - Strong point: more nuanced judgement on H_0.
  - Weakness: users usually forget about power.
  - Key point: in the right hands, p-values are innocuous.
    In the wrong hands though ...

- Confidence intervals provide a richer notion of estimation by returning an **interval of values of θ compatible with the data**.

- They are often constructed based on pivotal quantities.

- They have a dual relationship with hypothesis testing: an $(1-\alpha)$-CR can be turned into a family of α-tests for $\theta \stackrel{?}{=} \theta_0$ and vice-versa.

- In the rare cases in which we have UMP tests, we thus have associated Uniformly Most Accurate CIs.

Multiple testing (NOT FOR EXAM)

Multiple Testing

Modern example: looking for signals in noise

- Interested in detecting presence of a signal $\mu(x_t)$, $t = 1, \ldots, T$ over a discretized domain, $\{x_1, \ldots, x_t\}$, on the basis of noisy measurements
- This is to be detected against some known background, say 0.
- May be interested in detecting whether there is any signal over the domain or more specifically at which location x_t there is a signal

Formally:

Does there exist a $t \in \{1, \ldots, T\}$ such that $\mu(x_t) \neq 0$?

or

for which t's is $\mu(x_t) \neq 0$?

Multiple Testing

More generally:

- Observe

$$Y_t = \mu(x_t) + \varepsilon_t, \qquad t = 1, \ldots, T.$$

- Wish to test, at some significance level α:

$$\begin{cases} H_0 : \mu(x_t) = 0 & \text{for all } t \in \{1, \ldots, T\}, \\ H_A : \mu(x_t) \neq 0 & \text{for some } t \in \{1, \ldots, T\}. \end{cases}$$

- May also be interested in which specific locations signal deviates from zero
- More generally: May have T hypotheses to test simultaneously at level α (they may be related or totally unrelated)
- Suppose we have a test statistic for each individual hypothesis $H_{0,t}$ yielding a p-value p_t.

Bonferroni Method

If we test each hypothesis individually, we will not maintain the level!

Can we maintain the level α?

Idea: use Bonferroni's inequality.

Bonferroni

1. Test individual hypotheses separately at level $\alpha_t = \alpha/T$
2. Reject H_0 if at least one of the $\{H_{0,t}\}_{t=1}^{T}$ is rejected

Global level is bounded as follows:

$$\mathbb{P}[\not{H}_0 | H_0] = \mathbb{P}\left[\bigcup_{t=1}^{T} \{\not{H}_{0,t}\} \middle| H_0\right] \leq \sum_{t=1}^{T} \mathbb{P}[\not{H}_{0,t} | H_0] = T\frac{\alpha}{T} = \alpha$$

Holm-Bonferroni Method

- Advantage: Works for any (discrete domain) setup!
- Disadvantage: Too conservative when T large

Holm's modification increases average # of hypotheses rejected at level α (but does not increase power for overall rejection of $H_0 = \cap_{t \in T} H_{0,t}$)

Holm–Bonferroni's Procedure

1. We reject $H_{0,t}$ for small values of a corresponding p-value, p_t
2. Order p-values from most to least significant: $p_{(1)} \leq \ldots \leq p_{(T)}$
3. Starting from $t = 1$ and going up, reject all $H_{0,(t)}$ such that $p_{(t)}$ significant at level $\alpha/(T - t + 1)$. Stop rejecting at first insignificant $p_{(t)}$.

Genuine improvement over Bonferroni if want to detect as many signals as possible, not just existence of some signal.
Both Bonferroni and Holm–Bonferroni reject the global H_0 if and only if $\inf_t p_t$ significant at level α/T.

Taking Advantage of Structure: Independence

In the (special) case where individual test statistics are independent, one may use Sime's (in)equality,

$$\mathbb{P}\left[p_{(j)} \geq \frac{j\alpha}{T},\ \text{for all } j = 1, \ldots, T \,\middle|\, H_0\right] \geq 1 - \alpha$$

(strict equality requires continuous test statistics, otherwise $\leq \alpha$)

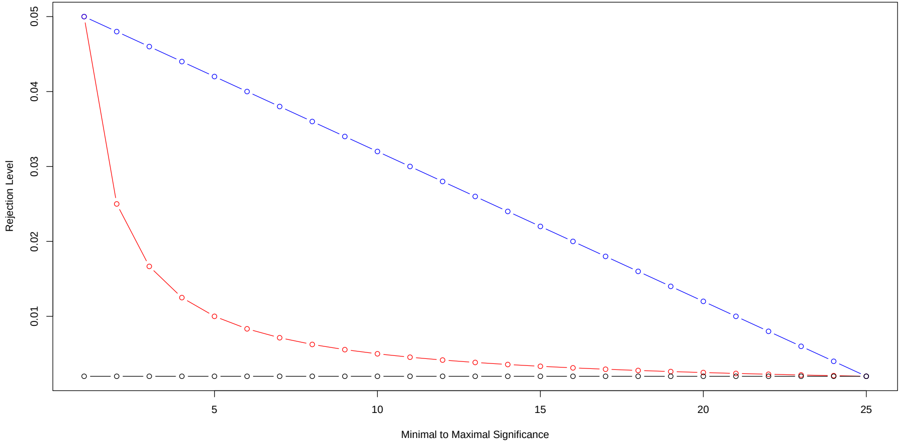
Sime's procedure (assuming independence)

1. Suppose we reject $H_{0,j}$ for small values of p_j
2. Order p-values from most to least significant: $p_{(1)} \leq \ldots \leq p_{(T)}$
3. If, for some $j = 1, \ldots, T$ the p-value $p_{(j)}$ is significant at level $\frac{j\alpha}{T}$, then reject the global H_0.

Provides a test for the global hypothesis H_0, but does not "localize" the signal at a particular x_t

Taking Advantage of Structure: Independence

Bonferroni, Hochberg, Simes



p-value

$p(x) = \inf\{\alpha \mid \text{reject on level } \alpha\}$

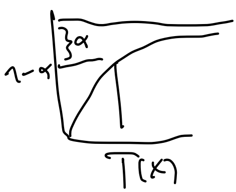
x … observed data

often $\varphi_\alpha(x) = \mathbb{1}\{T(x) > k_\alpha\}$, $\quad k_\alpha = G^{-1}(1-\alpha)$

G: cdf of $T(X)$ under θ_0

$H_0: \theta = \theta_0$

$$p(x) = \inf\{\alpha \mid T(x) > k_\alpha\}$$
$$= \inf\{\alpha \mid T(x) > G^{-1}(1-\alpha)\}$$



$T(x) = G^{-1}(1-\alpha)$ solve for α

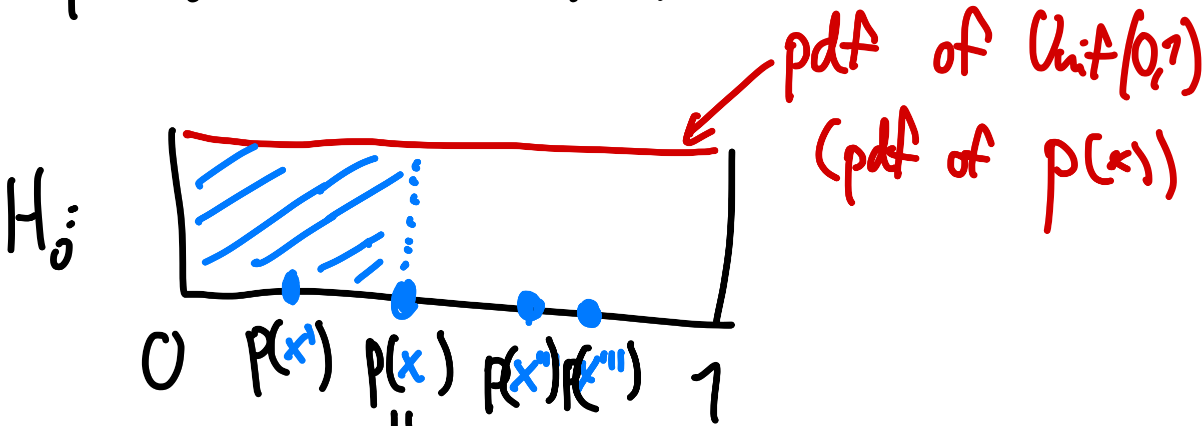
$$\alpha = 1 - G(T(x))$$

$$p(x) = \mathbb{P}_{\theta_0}(T(X) > T(x)) = 1 - G(T(x))$$

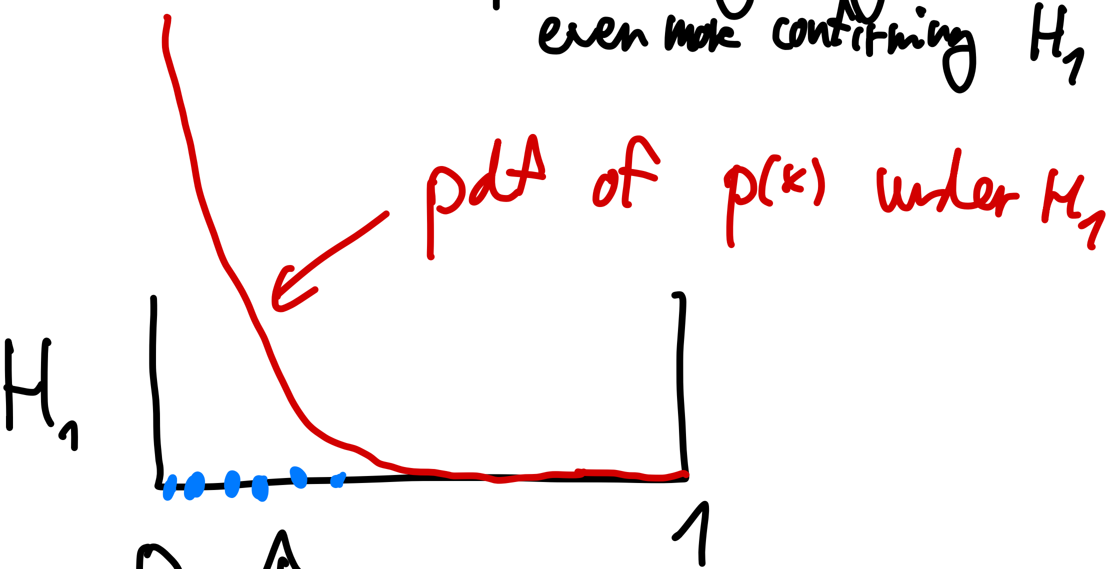
the prob. of getting a result even more extreme in favour of H_1 (even more confirming H_1)

p-value cont'd

$$p(x) \overset{H_0}{\sim} \text{Unif}(0,1)$$



p(x) = 0.4 – prob of getting a dataset even more confirming H_1



p-values tend to be small

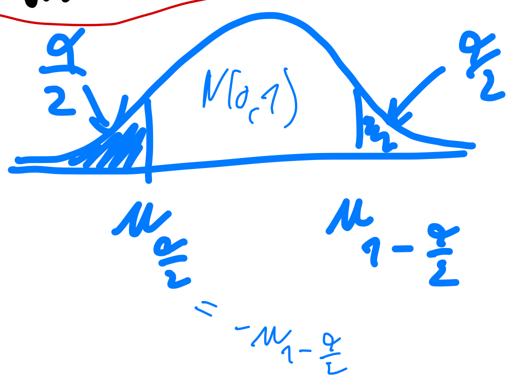
Confidence interval

$g(X, \theta)$ ... a pivot .... dist indep of θ

ex. $\dfrac{\bar{X}-\mu}{\sqrt{\frac{1}{n}}\,\hat{\sigma}}$ ... (asymptotic) pivot for the mean μ $\quad X_i \overset{iid}{\sim} N(\mu, \sigma^2)$

$\sim N(0,1)$

$$\mathbb{P}_\mu\left(-u_{1-\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{\sqrt{\frac{1}{n}}\,\hat{\sigma}} < u_{1-\frac{\alpha}{2}}\right) \doteq 1-\alpha$$



$$\frac{\bar{X}-\mu}{\sqrt{\frac{1}{n}}\,\hat{\sigma}} < u_{1-\frac{\alpha}{2}}$$

$$\Updownarrow$$

$$\mu > \bar{X} - \sqrt{\tfrac{1}{n}}\,\hat{\sigma}\, u_{1-\frac{\alpha}{2}}$$

(the other "<" analogously)

$$\downarrow$$

$$\mu \in \left(\bar{X} - \sqrt{\tfrac{1}{n}}\,\hat{\sigma}\, u_{1-\frac{\alpha}{2}},\ \bar{X} + \sqrt{\tfrac{1}{n}}\,\hat{\sigma}\, u_{1-\frac{\alpha}{2}}\right)$$

$$2\left(\ell(\hat{\theta}) - \ell(\theta)\right)$$

is also an asymptotic pivot distribution (χ^2) indep of θ under H_0

but: $\mathbb{P}(\dots < \dots < \dots) \doteq 1-\alpha$

sometimes not invertable (like on prev slide)

↳ just comment on how you would get the interval extremities

uniformly most powerful test (week 11)

$$\frac{f(x, \theta_1)}{f(x, \theta_0)} \quad \theta_1 > \theta_0$$

is non-decreasing function of $T(x)$

then the UMP test for

$$H_0: \theta \leq \theta_1, \quad H_1: \theta > \theta_0$$

is of the form

$$\partial(X) = \mathbb{1}_{[T(X) > k]}$$

choose k s.t. the level is α

Statistical Theory (Week 13): Further considerations about likelihood methods

Erwan Koch

Ecole Polytechnique Fédérale de Lausanne (Institute of Mathematics)

1. Confidence intervals based on MLE asymptotics

2. Confidence intervals based on the profile log-likelihood

3. Likelihood methods in practice

Confidence intervals based on MLE asymptotics

Reminder about Asymptotic normality of the MLE

Theorem (Asymptotic Normality of the MLE)

Let $X_1, \ldots, X_n$ be iid random variables with density (frequency) $f(x; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^d$, satisfying conditions (B1)-(B6). If $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}(X_1, \ldots, X_n)$ is a consistent sequence of MLEs, then

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N_d(\mathbf{0}, J^{-1}(\boldsymbol{\theta}) I(\boldsymbol{\theta}) J^{-1}(\boldsymbol{\theta})).$$

Generally, $I(\boldsymbol{\theta}) = J(\boldsymbol{\theta})$, so that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N_d(\mathbf{0}, I^{-1}(\boldsymbol{\theta})),$$

where $I(\boldsymbol{\theta}) = -\mathbb{E}[\nabla^2 \ell(X_1; \boldsymbol{\theta})]$ and thus has for element (i, j)

$$e_{i,j} = \mathbb{E}\left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ell(X_1; \boldsymbol{\theta})\right].$$

Denoting by $\psi_{i,j}$ the element (i, j) of $I^{-1}(\boldsymbol{\theta})$,

$$\hat{\theta}_i \sim N(\theta_i, \psi_{i,i}/n), \quad i = 1, \ldots, d.$$

CIs for individual components

Since the $\psi_{i,i}$ are usually unknown, we generally adopt one of the following solutions:

- If $I(\cdot)$ has a closed form, we can approximate $I(\boldsymbol{\theta})$ by $I(\hat{\boldsymbol{\theta}}_n)$, the so-called expected information matrix.
- We can estimate $I(\boldsymbol{\theta})$ using the so-called observed information matrix

$$I_{\mathrm{O}}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \nabla^2 \ell(X_i; \boldsymbol{\theta}),$$

  and evaluate it at $\hat{\boldsymbol{\theta}}_n$. $\implies I(\boldsymbol{\theta}) \approx I_{\mathrm{O}}(\hat{\boldsymbol{\theta}}_n)$.

Denoting by $\tilde{\psi}_{i,j}$ the element (i,j) of the inverse of the obtained estimated information matrix, we have

$$\hat{\theta}_i \sim N(\theta_i, \tilde{\psi}_{i,i}/n).$$

Thus, for $\alpha \in (0,1)$, an approximate $100(1-\alpha)\%$ confidence interval for θ_i is given by

$$\left[\hat{\theta}_i - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\tilde{\psi}_{i,i}}{n}}, \hat{\theta}_i + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\tilde{\psi}_{i,i}}{n}}\right],$$

where z_α is the α-quantile of the standard Gaussian distribution.

Use of the delta-method

Let $\hat{\boldsymbol{\theta}}_n$ be the MLE of $\boldsymbol{\theta}$. Assume that we are interested in a real-valued parameter $\phi = g(\boldsymbol{\theta})$. If

$$\hat{\boldsymbol{\theta}}_n \sim N_d(\boldsymbol{\theta}, V_{\boldsymbol{\theta}}),$$

the delta method yields

$$\hat{\phi}_n \sim N(\phi, V_\phi),$$

where

$$V_\phi = \nabla\phi^\top V_{\boldsymbol{\theta}} \nabla\phi,$$

with

$$\nabla\phi = \left(\frac{\partial\phi}{\partial\theta_1}, \ldots, \frac{\partial\phi}{\partial\theta_n}\right)^\top$$

evaluated at $\hat{\boldsymbol{\theta}}_n$. Then we can easily derive from the asymptotic normality of ϕ associated CIs.

Confidence intervals based on the profile log-likelihood

Profile log-likelihood

Alternative and usually more accurate method for making inferences on a particular component is based on **profile likelihood**.

Let $X_1, \dots, X_n \overset{iid}{\sim} f(x, \boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0 \in \mathbb{R}^d$. We denote by $\mathcal{L}$ the log-likelihood associated with $X_1, \dots, X_n$. For any $\boldsymbol{\theta} \in \mathbb{R}^d$ and $i = 1, \dots, d$, we can write (up to a reordering of the components) the vector $\boldsymbol{\theta}$ as $(\theta_i, \boldsymbol{\theta}_{-i}^\top)^\top$, where θ_i denotes the i-th component of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{-i}$ denotes all components of $\boldsymbol{\theta}$ excluding θ_i.

Definition

Let $i = 1, \dots, d$. The profile log-likelihood for θ_i is defined as

$$\mathcal{L}_p(\theta_i) = \max_{\boldsymbol{\theta}_{-i}} \mathcal{L}(\theta_i, \boldsymbol{\theta}_{-i}).$$

$\Longrightarrow$ $\mathcal{L}_p(\theta_i)$ is the profile of the log-likelihood surface viewed from the θ_i-axis.

Profile log-likelihood

- Previous definition generalizes to the situation where $\boldsymbol{\theta}$ can be partitioned into two components, $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$, where $\boldsymbol{\theta}^{(1)}$ is the r-dimensional vector of interest and $\boldsymbol{\theta}^{(2)}$ corresponds to the remaining $(d - r)$ components.

- The profile log-likelihood for $\boldsymbol{\theta}^{(1)}$ is now defined as

$$\mathcal{L}_p(\boldsymbol{\theta}^{(1)}) = \max_{\boldsymbol{\theta}^{(2)}} \mathcal{L}(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}).$$

Reminder about Wilk's theorem

Let $X_1, \ldots, X_n$ be iid random variables with density (frequency) depending on $\boldsymbol{\theta} \in \mathbb{R}^d$ and satisfying conditions (B1)-(B6), with $I(\boldsymbol{\theta}) = J(\boldsymbol{\theta})$.
Consider the likelihood ratio statistic

$$\Lambda_n(\boldsymbol{X}) = \frac{\prod_{i=1}^n f(X_i; \hat{\boldsymbol{\theta}}_n)}{\max_{\boldsymbol{\theta}^{(2)}} \prod_{i=1}^n f(X_i; \boldsymbol{\theta})}$$

where $\hat{\boldsymbol{\theta}}_n$ is the MLE of $\boldsymbol{\theta}$ and $\boldsymbol{\theta} = \left({\boldsymbol{\theta}^{(1)}}^\top, {\boldsymbol{\theta}^{(2)}}^\top \right)^\top$.
Recall Wilk's theorem.

Theorem (Wilk's theorem, general d, general $r \leq d$)

If the MLE sequence $\hat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}$, then the likelihood ratio statistic Λ_n for $H_0 : \boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}_0^{(1)}$ satisfies $2 \log \Lambda_n \xrightarrow{d} V \sim \chi_r^2$ when H_0 is true.

Link with profile log-likelihood and CIs

Assume that the true parameter is $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_0^{(1)\top}, \boldsymbol{\theta}_0^{(2)\top})^\top$. Observe that $\log \Lambda_n = \mathcal{L}(\hat{\boldsymbol{\theta}}_n) - \mathcal{L}_p(\boldsymbol{\theta}_0^{(1)})$, so that Wilk's theorem yields

$$2\left[\mathcal{L}(\hat{\boldsymbol{\theta}}_n) - \mathcal{L}_p(\boldsymbol{\theta}_0^{(1)})\right] \xrightarrow{d} V \sim \chi^2_r.$$

On top of being useful for model selection between nested models (see Week 11), valuable for making inferences about a single component. In the case where $\boldsymbol{\theta}_0 = (\theta_{0,i}, \boldsymbol{\theta}_{0,-i}^\top)^\top$, we have

$$2\left[\mathcal{L}(\hat{\boldsymbol{\theta}}_n) - \mathcal{L}_p(\theta_{0,i})\right] \xrightarrow{d} V \sim \chi^2_1.$$

Profile log-likelihood based CI

Let $\alpha \in (0,1)$ and $\chi^2_{1,1-\alpha}$ be the $(1-\alpha)$-quantile of the χ^2_1 distribution. The set

$$C_{1-\alpha} = \left\{\theta_i : 2\left[\mathcal{L}(\hat{\boldsymbol{\theta}}_n) - \mathcal{L}_p(\theta_i)\right] \leq \chi^2_{1,1-\alpha}\right\}$$

is a $100(1-\alpha)\%$ confidence interval for $\theta_{0,i}$.

Likelihood methods in practice

Likelihood methods

In this course, we have seen several methods which make heavy use of the likelihood.

1. Point Estimation: the likelihood function $L(\theta)$ represents the compatibility of each possible value of the parameter with the data. An intuitively satisfying estimator for θ is the MLE:

$$\theta_{\mathrm{MLE}} = \arg\max L(\theta).$$

2. Hypothesis testing (including model selection)/Interval estimation: the likelihood ratio statistic

$$\frac{\sup_{\theta \in \Theta_1} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)}$$

measures the relative compatibility with the data between the null and the alternative.

Likelihood methods

Likelihood methods follow the likelihood principle:

Likelihood principle

1. The likelihood function contains all the relevant information present in a dataset.
2. Statistical analyses should only take into account the likelihood and no other aspect of the data.

The likelihood principle is probably too extreme, but good to have principles.

Likelihood methods are superior

Throughout the course, we have seen many arguments in favour of the likelihood principle:

1. The normalized likelihood is a minimally sufficient statistic: It holds as much information as the data with as little ancillary information as possible. As such, any statistic computed from the likelihood is already Rao-Blackwellized = can't be improved further in this way.
2. Furthermore, asymptotically, the MLE is unbiased, Gaussian, and saturates the Cramér-Rao bound: It is maximally efficient (among regular estimators).
3. When there exist optimal tests of a null hypothesis H_0 vs H_1, they are
    - the likelihood ratio test (simple vs simple).
    - directly deduced from the likelihood (MLR property).

Limits of optimality

It is prominent to remember the restrictions which we had to impose in order to reach these optimality results:

1. Optimality in point estimation:
   - Only among unbiased estimators or asymptotically.
   - The MLE might very-well be dominated.

2. Optimality in testing (including model selection)/interval estimation:
   - Optimal tests only rarely exist.
   - The LRT is intuitively satisfying and respects the likelihood principle. This is all we can say given the content of this course; generally it is not UMP.

Asymptotics

In the course, we have seen two main asymptotic results:

1. Asymptotically, the MLE is generally a Gaussian unbiased estimator of the true parameter value. But beware that it can be biased for finite n! Consistency issues are also possible.
2. Asymptotically, the Likelihood Ratio Statistic follows a χ^2 distribution under the null hypothesis for nested models.

These two results are **crucial for inference**. Especially, enable the construction of CIs from the MLE or the LR Statistic (in link with profile likelihood) and the choice of an appropriate threshold for the LRT.

Misspecification

- A key limit of likelihood methods is **misspecification**.
- Misspecification **almost always occurs**.
  - You might be the greatest statistician on earth, but you will never be able to guess correctly the true model that generated the data.
  - A statistical model is always a simplification of reality.
- Misspecification implies that some good properties of likelihood methods are modified or vanish. E.g, pertaining to asymptotics, misspecification changes the covariance of the MLE and kills the LR Statistic result.
- Importantly, misspecification **doesn't make likelihood methods meaningless**! For example, for point estimation, we have seen that the MLE tries to estimate the best approximation to the truth within the assumed parametric class.

Statistics in practice

My personal opinion is that likelihood methods constitute the best way to do statistics. Two steps:

1. You choose a good model. It is very hard but mild misspecification is completely fine. E.g., using a Gaussian model instead of a Student t with 50 degrees of freedom is no problem at all!
2. You figure out how to compute the MLE or the LRT.

Two crucial advantages:

- No step in which you have to guess a good estimator that you then have to analyze. $\Longrightarrow$ Being a "likelihoodist" entails never having to deal with this annoying side of statistics.
- Method is guaranteed to be (almost) optimal as long as your model is almost correctly specified.

Computational aspects and optimization

- Statistics is, at its heart, a computational discipline. If your method has great theoretical properties but can't be performed by a computer, it is useless.
- Finding the MLE or the LRT are intrinsically **optimization problems**.
- Essential to understand optimization to be a good independent statistician.
- Some optimization methods: gradient descent and its variants, BFGS, Nelder–Mead . . .

Summary

In this course we focused inter alia on three important topics:

- Providing a general framework for statistical inference: likelihood methods.
- Analyzing the behaviour of statistical methods when the number of data points tends to ∞: asymptotic results.
- Analyzing the efficiency of various approaches to statistics: is there an optimal way to do statistics (estimation, hypothesis testing, ...)?

Important aspects we did not really have time to tackle:

- Computational issues.
- How to choose a good model?

Statistical Theory (Week 14): The Stein Phenomenon and Superefficiency

Erwan Koch

Ecole Polytechnique Fédérale de Lausanne (Institute of Mathematics)

Motivation: Is Likelihood Always Sensible?

Likelihood Reminder

We've seen that the likelihood possesses several appealing properties:

- When there exists a complete sufficient statistic, the MLE is a function of this statistic
  - $\hookrightarrow$ Hence an unbiased MLE in an exponential family is UMVUE
- Asymptotically, the MLE is unbiased and has variance that approximates the Cramér-Rao bound.

Though the likelihood is not always unbiased, it generally produces estimators with sensible mean squared error.

- For example, it was long believed that, except for pathological situations, the MLE would always be admissible.
- Fisher's position was that likelihood was always the way to go.
  - (arguing that the cases where it was shown to not perform well were artificial and monstrous constructions).

Enter Charles Stein

In the late 50's, Charles Stein presented a paper in the Berkeley Probability/Statistics Symposium that **shocked** the statistical community:

- He produced a non-artificial example of another estimator that dominates the MLE.
- As a matter of fact, the likelihood was inadmissible in his example.
- Most shockingly, the example was about **estimating the mean of a Gaussian!**
  - Perhaps the most natural of estimation problems!

Let's see the precise setting.

Gaussian Estimation Under Quadratic Loss

Stein's Setup

Gaussian Estimation Under Quadratic Loss

1. Let $X_1, ..., X_n$ be independent random variables.
2. Assume that $X_i \sim \mathcal{N}(\mu_i, \sigma^2)$.
   - Notice that each X_i has a different mean but same variance.
3. Suppose that σ^2 is known, say $\sigma^2 = 1$ (wlog)
4. Unknown parameter to estimate: $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T \in \mathbb{R}^n$
5. Consider quadratic loss, $\mathcal{L}(\delta, \boldsymbol{\mu}) = \|\delta - \boldsymbol{\mu}\|^2$
6. Hence risk is mean squared error, as usual.

$\hookrightarrow$ Looks like the usual setup, but notice the subtlety: the dimension of the parameter dim($\boldsymbol{\mu}$)=n grows along with the dimension of the sample size.

Is this an artificiality? No: Modern problems have # parameters comparable to # observations (high dimensional statistics).

The MLE in Stein's Setup

By independence, the loglikelihood is

$$\ell(\boldsymbol{\mu}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu_i)^2$$

and by differentiation and convexity,

$$\hat{\boldsymbol{\mu}} = (X_1, ..., X_n)^\top = \boldsymbol{X}$$

is the unique MLE of $\boldsymbol{\mu}$.

- Intuition: we essentially have n Gaussian mean separate problems, each of sample size 1.
- Hence separately estimate each of these means by corresponding sample mean
  (which is X_i since there is only 1 observation in each sample)

MLE Risk

MLE Risk in Stein's Setup

Let $\hat{\boldsymbol{\mu}}$ be the MLE in Stein's setup. Then

$$R(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = n, \qquad \forall \boldsymbol{\mu} \in \mathbb{R}^n.$$

Proof.

$R(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = \mathbb{E}\left[\sum_{i=1}^n (X_i - \mu_i)^2\right] = n\sigma^2 = n.$ □

Contrary to the usual setup, the risk increases with n (since the number of parameters increases in n).

Now let's see what estimator Stein defined...

The James-Stein Estimator

The James-Stein Estimator

Theorem (James-Stein)

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ be such that $\boldsymbol{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, I)$, $\boldsymbol{\mu} \in \mathbb{R}^n$ (Stein's setup). Let δ_a be an estimator defined as

$$\delta_a(\boldsymbol{X}) = \left(1 - \frac{a}{\|\boldsymbol{X}\|^2}\right)\boldsymbol{X}.$$

Then, under a quadratic loss function, and if $n \geq 3$,

1. For all $a \in (0, 2n-4)$, $R(\delta_a, \boldsymbol{\mu}) \leq R(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$.
2. For $a = n-2$, $2 = R(\delta_{n-2}, \mathbf{0}) < R(\hat{\boldsymbol{\mu}}, \mathbf{0}) = n$.
3. $R(\delta_{n-2}, \boldsymbol{\mu}) \leq R(\delta_a, \boldsymbol{\mu})$, for all $\boldsymbol{\mu} \in \mathbb{R}^n$ and all $a \in (0, 2n-4)$.

Corollary

The MLE is inadmissible in Stein's setup for $n \geq 3$

The James-Stein Estimator

The result is surprising, not just because the MLE is inadmissible

- The JS estimator takes the MLE and shrinks it towards zero.
- The amount of shrinkage depends on $\|\boldsymbol{X}\|$
- That is, we take into account the estimate of μ_i in order to estimate μ_j $(i \neq j)$, even though in principle these are unrelated!
- (for example, we are violating the sufficiency principle)

Notice also that the performance of the MLE as compared to the JS estimator becomes worse and worse as n grows.

- The proof is surprisingly elementary
  (once one knows what to look for!)

The James-Stein Estimator

Lemma

Let $Y \sim \mathcal{N}(\theta, \sigma^2)$ and $h : \mathbb{R} \to \mathbb{R}$ be differentiable. If

1. $\mathbb{E}|h(Y)| < \infty$,
2. $\lim_{y \to \pm\infty} \left\{ h(y) \exp\left[-\frac{1}{2\sigma^2}(y-\theta)^2 \right] \right\} = 0$,

then

$$\mathbb{E}[h(Y)(Y-\theta)] = \sigma^2 \mathbb{E}\left[h'(Y)\right].$$

Proof.

By definition, $\mathbb{E}[h(Y)(Y-\theta)] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} h(y)(y-\theta) e^{-\frac{1}{2\sigma^2}(y-\theta)^2} dy$.

Integration by parts transforms the right hand side into

$$\underbrace{-\frac{\sigma^2}{\sigma\sqrt{2\pi}} \left(h(y) e^{-\frac{1}{2\sigma^2}(y-\theta)^2} \right) \Big|_{-\infty}^{+\infty}}_{=0} + \underbrace{\frac{\sigma^2}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} h'(y) e^{-\frac{1}{2\sigma^2}(y-\theta)^2} dy}_{=\sigma^2 \mathbb{E}[h'(Y)]}$$

□

Proof of the James-Stein Theorem.

$$
\begin{aligned}
R(\delta_a, \boldsymbol{\mu}) &= \mathbb{E}\left\|\left(1-\frac{a}{\|\boldsymbol{X}\|^2}\right)\boldsymbol{X}-\boldsymbol{\mu}\right\|^2 = \mathbb{E}\left\|\boldsymbol{X}-\boldsymbol{\mu}-\frac{a\boldsymbol{X}}{\|\boldsymbol{X}\|^2}\right\|^2 \\
&= \mathbb{E}\,\|\boldsymbol{X}-\boldsymbol{\mu}\|^2 - 2\mathbb{E}\left(\frac{a\boldsymbol{X}^\top(\boldsymbol{X}-\boldsymbol{\mu})}{\|\boldsymbol{X}\|^2}\right) + \mathbb{E}\left[\frac{a^2\|\boldsymbol{X}\|^2}{\|\boldsymbol{X}\|^4}\right] \\
&= n - 2a\sum_{i=1}^{n}\mathbb{E}\left[\frac{X_i(X_i-\mu_i)}{\sum_{j=1}^{n}X_j^2}\right] + a^2\mathbb{E}\left[\frac{1}{\|\boldsymbol{X}\|^2}\right].
\end{aligned}
$$

Now define n differentiable functions $h_i : \mathbb{R} \to \mathbb{R}$ by

$$h_i(x) = \frac{x}{x^2 + \sum_{j\neq i}^{n} X_j^2}$$

and observe that, for all $i \in \{1, ..., n\}$,

$$\lim_{x_i \to \pm\infty}\left\{h(x_i)\exp\left[-\frac{1}{2\sigma^2}(x_i-\mu_i)^2\right]\right\} = 0$$

(proof ct'd)

We now use the tower property and apply our lemma to obtain

$$\mathbb{E}\left[\frac{X_i(X_i-\mu_i)}{\sum_{j=1}^n X_j^2}\right] = \mathbb{E}\left[h_i(X_i)(X_i-\mu_i)\right] = \mathbb{E}\big\{\mathbb{E}\left[h_i(X_i)(X_i-\mu_i)|\,\{X_j\}_{j\neq i}\right]\big\}$$

$$= \mathbb{E}\big\{\mathbb{E}\left[h_i'(X_i)|\{X_j\}_{j\neq i}\right]\big\} = \mathbb{E}\left[h_i'(X_i)\right] = \mathbb{E}\left[\frac{\sum_{j=1}^n X_j^2 - 2X_i^2}{\left(\sum_{j=1}^n X_j^2\right)^2}\right].$$

It follows that the risk can be written as

$$\begin{aligned} R(\delta_a, \boldsymbol{\mu}) &= n - 2a\mathbb{E}\left[\frac{n\|\boldsymbol{X}\|^2 - 2\|\boldsymbol{X}\|^2}{\|\boldsymbol{X}\|^4}\right] + a^2\mathbb{E}\left[\frac{1}{\|\boldsymbol{X}\|^2}\right] \\ &= n + [a^2 - 2a(n-2)]\underbrace{\mathbb{E}\left[\frac{1}{\|\boldsymbol{X}\|^2}\right]}_{>0}. \end{aligned}$$

(proof ct'd).

Now, the polynomial

$$p(a) = a^2 - 2a(n-2) = a[a - 2(n-2)]$$

is strictly negative in the range $(0, 2n-4)$. Therefore, we have proven part (1).

Furthermore, on the same range, $p(a)$ has a unique minimum at $a = n-2$, which proves part (3).

For part (2), note that if $\boldsymbol{\mu} = \mathbf{0}$, $\|\boldsymbol{X}\|^2 \sim \chi_n^2$, so $\mathbb{E}[1/\|\boldsymbol{X}\|^2] = 1/(n-2)$ (recall that $n \geq 3$). Consequently, $R(\delta_{n-2}, \mathbf{0}) = 2$. □

Summary on JSE vs MLE

- The MLE has constant risk $R_{MLE} = n$.
- Around $\boldsymbol{\mu} = \mathbf{0}$, the JSE dominates the MLE by a mile! $R_{JSE} = 2 \ll n$.
- For every other value of $\boldsymbol{\mu}$, the JSE dominates the MLE (possibly by a hair).
- The Stein setup can be extended to the case where we have $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent p-dimensional random vectors. The same phenomenon appears wrt p for $p \geq 3$. In this setting, we see that the domination region shrinks when the sample size n grows.
- The Stein setup is written for the Gaussian model, but the same phenomenon occurs **asymptotically** for any MLE: $\hat{\boldsymbol{\theta}} \dot{\sim} N(\boldsymbol{\theta}_0, \Sigma/n)$.
- We could construct a JSE biased towards any point of space instead of $\boldsymbol{\mu} = \mathbf{0}$: this just shifts the domination zone. We can also have multiple shrinkages.

Summary on JSE vs MLE

- Critically, the domination only occurs in a small region around $\boldsymbol{\mu} = \mathbf{0}$. As soon as $\|\boldsymbol{\mu}\|^2 \gg p/n$, their risks are approximately equal. Furthermore, if you have been able to choose the shrinkage region correctly, you have been able to **locate a priori** the true parameter value at the same precision as the data. **That's a miracle: go play the lottery instead of doing stats**.
  $\Longrightarrow$ the domination of the JSE is mostly theoretical: I don't think I have ever seen it used in practice.

- However, Stein's example demonstrates the huge benefits of bias in high-dimensions: a small bias can result in a huge reduction in variance.
  Canonically, we induce bias through the addition of an L_2 loss on top of the log-likelihood. The L_1 loss can also be used to induce sparsity in the estimator. The relative size of the additional loss is chosen through a validation set or cross-validation.

Asymptotic Optimality and Superefficiency

What about asymptotic optimality?

An optimal decision rule would is one that uniformly minimizes risk:

$$R(\theta, \delta_{\text{OPTIMAL}}) \leq R(\theta, \delta), \quad \forall \theta \in \Theta \ \& \ \forall \delta \in \mathcal{D}.$$

Such rules can very rarely be determined.

Some avenues to studying optimal decision rules include:

- **Restricting attention to global risk criteria rather than local**
  $\hookrightarrow$ Bayes and minimax risk.
- **Focusing on restricted classes of rules** $\mathcal{D}$
  $\hookrightarrow$ e.g. Minimum Variance Unbiased Estimation.
- **Studying risk behaviour asymptotically** $(n \to \infty)$
  $\hookrightarrow$ e.g. Asymptotic Relative Efficiency.

Asymptotically Gaussian Estimators

Comparing Asymptotically Gaussian Estimators

- Have two possible estimators $\hat{\theta}$ and $\tilde{\theta}$ of θ based on $X_1, .., X_n$.
- Risk comparisons may be intractable (including minimax/Bayes)
- Idea: Compare as $n \to \infty$

Definition

Let $\{X_i\}_{i=1}^n$ be a sequence of random variables and suppose that $\hat{\theta}_n$ and $\tilde{\theta}_n$ are estimators of θ based on $X_1, ..., X_n$ satisfying

$$\frac{\hat{\theta} - \theta}{\sigma_{1n}(\theta)} \xrightarrow{d} \mathcal{N}(0,1) \quad \& \quad \frac{\tilde{\theta} - \theta}{\sigma_{2n}(\theta)} \xrightarrow{d} \mathcal{N}(0,1)$$

for some sequences $\{\sigma_{2n}\}$ and $\{\sigma_{1n}\}$. We define the *asymptotic relative efficiency* of $\hat{\theta}$ to $\tilde{\theta}$ to be

$$ARE_\theta(\hat{\theta}, \tilde{\theta}) = \lim_{n\to\infty} \left(\sigma_{2n}^2 / \sigma_{1n}^2\right)$$

provided that the limit exists.

Comparing Asymptotically Gaussian Estimators

Interpretation of asymptotic relative efficiency?

In many examples (e.g. if $X_1, ..., X_n$ are iid) we have

$$\sigma_{1n} = \frac{\sigma_1(\theta)}{\sqrt{n}} \quad \& \quad \sigma_{2n} = \frac{\sigma_2(\theta)}{\sqrt{n}} \quad \text{so that} \quad ARE_\theta(\hat{\theta}, \tilde{\theta}) = \frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}.$$

Suppose that we have a choice between $\hat{\theta}_n$ and $\tilde{\theta}_m$ as estimators of θ
$\hookrightarrow$ Notice that we allow for different sample sizes n and m
Suppose we choose n and m so that

$$\mathbb{P}_\theta[|\hat{\theta}_n - \theta| < \Delta] \approx \mathbb{P}_\theta[|\tilde{\theta}_m - \theta| < \Delta].$$

If n, m are sufficiently large, this is equivalent to

$$\mathbb{P}[|Z| < \Delta\sqrt{n}/\sigma_1(\theta)] \approx \mathbb{P}[|Z| < \Delta\sqrt{m}/\sigma_2(\theta)]$$

for $Z \sim \mathcal{N}(0, 1)$.

Comparing Asymptotically Gaussian Estimators

We conclude that $\frac{\sqrt{n}}{\sigma_1(\theta)} \approx \frac{\sqrt{m}}{\sigma_2(\theta)}$ or, equivalently, $\frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)} \approx \frac{m}{n}$

- The ratio of sample sizes needed to achieve the same accuracy is approximately equal to ARE
- e.g. if $ARE_\theta(\hat{\theta}, \tilde{\theta}) = 2$ we need double the amount of data to achieve $\hat{\theta}$'s precision when using $\tilde{\theta}$
- Warning: interpretation valid for large sample sizes and ARE may change for different values θ of the true parameter.

Example

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. We have

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \quad \& \quad \sqrt{n}(\text{med}(X_1, \ldots, X_n) - \mu) \xrightarrow{d} \mathcal{N}(0, \pi\sigma^2/2)$$

Hence $ARE(\bar{X}, \text{med}(X_1, \ldots, X_n)) = \pi/2 \approx 1.571$.

Example

Let $X_1, ..., X_n \overset{iid}{\sim}$ Poisson(λ). Suppose we want to estimate $\exp(-\lambda) = \mathbb{P}_\lambda(X_i = 0)$. Consider the estimators

$$\hat{\theta}_n = \exp(-\bar{X}_n) \quad \& \quad \tilde{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i = 0\}.$$

Using the CLT and the Delta method we have

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{d}{=} \mathcal{N}(0, \lambda \exp(-2\lambda))$$
$$\sqrt{n}(\tilde{\theta}_n - \theta) \overset{d}{=} \mathcal{N}(0, \exp(-\lambda) - \exp(-2\lambda))$$

yielding

$$ARE_\lambda(\hat{\theta}, \tilde{\theta}) = \frac{\exp(\lambda) - 1}{\lambda}$$

Using a McLaurin expansion, it is easy to see that this expression is greater than 1 for all λ, but close to 1 for small values of λ.

Example

Let $X_1, ..., X_n \overset{iid}{\sim} \text{Exp}(\lambda)$. When discussing MoM estimators, we derived a family of estimators of λ through the equation $\mathbb{E}_\lambda[X_i^r] = \frac{\Gamma(r+1)}{\lambda^r}$,

$$\hat{\lambda}_n^{(r)} := \left(\frac{1}{n\Gamma(r+1)} \sum_{i=1}^{n} X_i^r \right)^{-\frac{1}{r}}.$$

Since $\text{Var}_\lambda(X_i^r) = (\Gamma(2r+1) - \Gamma^2(r+1))/\lambda^{2r})$, we may apply the CLT followed by the Delta Method and obtain

$$\sqrt{n}(\hat{\lambda}_n^{(r)} - \lambda) \xrightarrow{d} \mathcal{N}\left(0, \frac{\lambda^2}{r^2}\left[\frac{\Gamma(2r+1)}{\Gamma^2(r+1)} - 1\right]\right).$$

The variance term turns out to be minimized for $r = 1$, so that $1/\bar{X}$ is (asymptotically) the most efficient estimator within this family.

Asymptotic Efficiency

Asymptotic Normality and the Cramér-Rao Bound

We have seen that, under regularity conditions, the MLE $\hat{\theta}$ of θ satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta))$$

where $I(\theta) = \text{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1;\theta)\right]$. In other words, for sufficiently large n,

$$\mathbb{E}_\theta[\hat{\theta}_n] \approx \theta \quad \& \quad \text{Var}_\theta(\hat{\theta}_n) \approx \frac{1}{nI(\theta)}.$$

On the other hand, the Cramér-Rao bound informs us that for any unbiased estimator T, based on $X_1, ..., X_n$ it must be that

$$\text{Var}_\theta[T] \geq n^{-1}I^{-1}(\theta)$$

Raises question:

- If $\tilde{\theta}_n$ is such that $\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$ then is $\sigma^2(\theta) \geq I^{-1}(\theta)$ $\forall\theta \in \Theta$?
- In other words, is the MLE *asymptotically optimal* among consistent estimators that asymptotically have a Gaussian distribution?

Hodges' Superefficient Estimator

Hodges' Counterexample

The answer to our question is NO in general

$\hookrightarrow$ Hodges' example of a *superefficient* estimator

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, 1)$. Observe that, for this model,

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X_i;\theta)\right)^2\right] = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} - \frac{1}{2}(X_i - \theta)^2\right)^2\right] = \mathrm{Var}(X_i) = 1$$

Define an estimator

$$\tilde{\theta}_n := \begin{cases} \bar{X}_n & \text{if } |\bar{X}_n| \geq n^{-1/4}, \\ \alpha\bar{X}_n & \text{otherwise.} \end{cases}$$

where α is some fixed constant with $|\alpha| < 1$.

Let's study the asymptotics of this estimator...

Hodges' Counterexample

Taking note that $\sqrt{n}(\bar{X}_n - \theta) \stackrel{d}{=} Z \sim \mathcal{N}(0,1)$, for all $n \geq 1$,

$$\begin{aligned}
\sqrt{n}(\tilde{\theta} - \theta) &= \sqrt{n}(\bar{X}_n - \theta)\mathbf{1}\{|\bar{X}_n| \geq n^{-\frac{1}{4}}\} + \sqrt{n}(\alpha \bar{X}_n - \theta)\mathbf{1}\{|\bar{X}_n| < n^{-\frac{1}{4}}\} \\
&= \sqrt{n}(\bar{X}_n - \theta)\mathbf{1}\{\sqrt{n}|\bar{X}_n - \theta + \theta| \geq n^{\frac{1}{4}}\} + \\
&\quad + \sqrt{n}(\alpha \bar{X}_n - \alpha\theta + \alpha\theta - \theta)\mathbf{1}\{\sqrt{n}|\bar{X}_n - \theta + \theta| < n^{\frac{1}{4}}\} \\
&\stackrel{d}{=} Z\mathbf{1}\{|Z + \sqrt{n}\theta| \geq n^{\frac{1}{4}}\} + \\
&\quad + [\alpha Z + \sqrt{n}\theta(\alpha - 1)]1\{|Z + \sqrt{n}\theta| < n^{\frac{1}{4}}\}
\end{aligned}$$

Observe that $Z + \sqrt{n}\theta \sim \mathcal{N}(\sqrt{n}\theta, 1)$ so that

$$\mathbf{1}\{|Z + \sqrt{n}\theta| \geq n^{\frac{1}{4}}\} \xrightarrow{p} \begin{cases} 0 & \text{if } \theta = 0, \\ 1 & \text{if } \theta \neq 0. \end{cases}$$

which implies that

$$Z\mathbf{1}\{|Z + \sqrt{n}\theta| \geq n^{\frac{1}{4}}\} \xrightarrow{p} \begin{cases} 0 & \text{if } \theta = 0, \\ Z & \text{if } \theta \neq 0. \end{cases}$$

Hodges' Counterexample

Similarly, the fact that

$$\mathbf{1}\{|Z+\sqrt{n}\theta| < n^{\frac{1}{4}}\} \xrightarrow{p} \begin{cases} 1 & \text{if } \theta = 0, \\ 0 & \text{if } \theta \neq 0. \end{cases}$$

yields

$$[\alpha Z + \sqrt{n}\theta(\alpha - 1)]1\{|Z+\sqrt{n}\theta| < n^{\frac{1}{4}}\} \xrightarrow{p} \begin{cases} \alpha Z & \text{if } \theta = 0, \\ 0 & \text{if } \theta \neq 0. \end{cases}.$$

Combining our findings, we conlcude that

$$\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} \begin{cases} \alpha Z & \text{if } \theta = 0, \\ Z & \text{if } \theta \neq 0. \end{cases}.$$

It follows that $\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$ with

$$I^{-1}(\theta) = 1 \geq \sigma^2(\theta) = 1 \cdot \mathbf{1}\{\theta \neq 0\} + \alpha^2 \cdot \mathbf{1}\{\theta = 0\}$$

- Observe that in the example, $\sigma^2(\theta) \leq I^{-1}(\theta)$ and not just

$$\exists\, \theta : \quad \sigma^2(\theta) < I^{-1}(\theta).$$

- Such estimators are called *superefficient*, as they asymptotically dominate estimators that asymptotically achieve the CR-bound.
- What causes this phenomenon?. It turns out that if $\sigma^2(\theta)$ is continuous then $\sigma^2(\theta) \geq I^{-1}(\theta)$ always
  - $\hookrightarrow$ In the presence of continuity the answer to our question on MLE asymptotic optimality is YES.

Subject to weak regularity conditions,

$$\{\theta : \sigma^2(\theta) < I^{-1}(\theta)\} \quad \text{is at most a countable set}$$

Crucial notion behind superefficiency?

Optimality of the MLE

- One critical feature of the Hodges estimator is the fact that $\sigma^2(\theta)$ has a discontinuity at $\theta = 0$ where the superefficiency is achieved.
- We can define **regular** estimators which are such that such discontinuities are forbidden.
- It turns out that, among regular estimators, it is true that $\sigma^2(\theta) \geq I(\theta)$ everywhere. Thus, **the MLE maximizes efficiency for regular estimators**.

This is one possible way to defend the MLE against Hodge super-efficiency ...

Going on the offensive

However, it is much better to observe that Hodges-style super-efficient estimators are actually terrible:

- We pay for efficiency around $\theta = 0$ in other positions.
- Furthermore, the Hodges estimator is also biased.
- Finally, the Hodges estimator is very non-Gaussian for $\theta \approx n^{-1/4}$.

Going to the limit $n \to \infty$ hides these properties of the Hodge estimator. Be wary of limits (Jayne Probability Theory, the logic of science).

Summary

In a correctly specified model, the MLE is a great estimator because it is asymptotically unbiased and saturates the Cramér-Rao bound.
Today, we saw two results that challenge this view on the MLE:

- - The JSE is a biased estimator that dominates the MLE everywhere.
  - Very general and interesting result.
  - However, the zone where this domination is significant is very small: $\|\boldsymbol{\mu}\| \ll p/n$.
  - The JSE example tells us about the **strength of bias** in high-dimensional inference.
  - the JSE is a super-efficient estimator, but not Hodges-style.
- - The Hodges superefficient estimator has superior Asymptotic Efficiency compared to the MLE.
  - This is a (fairly boring) case of the danger of limits
  - For finite n the Hodges estimator is better at $\theta = 0$ and worse everywhere else.
  - We can exclude the Hodges estimator by focusing on **regular estimators**.

The MLE is a great estimator. Regularized MLEs are also great estimators.

Regular Sequences of Estimators

Hodge's Counterexample, Superefficiency and Regularity

Definition (Hájek Regularity)

A sequence of estimators $\{\hat{\theta}_n\}$ is *regular* at θ if, for $\theta_n = \theta + c/\sqrt{n}$,

$$\lim_{n\to\infty} \mathbb{P}_{\theta_n}\left[\sqrt{n}(\hat{\theta}_n - \theta_n) \leq x\right] = G_\theta(x)$$

where G_θ may depend on θ but not on c.

- Intuition: limit theorem is stable to $n^{-1/2}$ perturbations of the true parameter (limit theorem is continuous at θ at scale $n^{-1/2}$).
- Hodges' estimator is not regular, MLE is regular

Example (Normal Distribution)

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\theta, 1)$ and $\hat{\theta}_n = \bar{X}_n$. Under the parameter $\theta_n = \theta + c/\sqrt{n}$, we have $\hat{\theta}_n \sim \mathcal{N}(\theta_n, \frac{1}{n})$.
Hence $\sqrt{n}(\hat{\theta}_n - \theta_n) \sim \mathcal{N}(0, 1)$ $\forall n$ and $\hat{\theta}_n$ is regular.

Hájek Regularity

Regularity and Superefficincy

Example (Exponential Distribution)

Let $X_1, ..., X_n \overset{iid}{\sim} \text{Exp}(\lambda)$ and define $\hat{\lambda}_n := 1/\bar{X}_n$ and $\lambda_n = \lambda + c/\sqrt{n}$. By the Lyapunov CLT:

$$\mathbb{P}_{\lambda_n}\left[\sqrt{n}\left(\bar{X}_n - \frac{1}{\lambda_n}\right) \leq x\right] \to \Phi(\lambda x)$$

where Φ is the standard Gaussian distribution function. A "Delta-Method"-type argument yields

$$\mathbb{P}_{\lambda_n}\left[\sqrt{n}\left(\hat{\lambda}_n - \lambda_n\right) \leq x\right] \to \Phi(x/\lambda)$$

and so $\{\hat{\lambda}_n\}$ is a regular sequence of estimators.

So why care about regularity?

Regularity and Asymptotic Efficiency

Theorem

Let $X_1, \dots, X_n$ be iid random variables with density (frequency) $f(x;\theta)$ and suppose that $\{\hat{\theta}_n\}$ is a regular sequence of estimators for θ. If

$$\sum_{i=1}^{n}\left[\log f\left(X_i;\theta+\frac{c}{\sqrt{n}}\right)-\log f(X_i;\theta)\right]=cS_n(\theta)-\frac{1}{2}c^2I(\theta)+R_n(c,\theta)$$

where $S_n(\theta)\xrightarrow{d}\mathcal{N}(0,I(\theta))$ and $R_n(c,\theta)\xrightarrow{p}0$ for all c, then

$$\sqrt{n}(\hat{\theta}_n-\theta)\xrightarrow{d}Z_1+Z_2$$

where $Z_1\sim\mathcal{N}(0,I^{-1}(\theta))$ and Z_2 is independent of Z_1.

- Gives an asymptotic representation of regular sequences.
- Can be thought of as an asymptotic version of the Cramér-Rao bound.
- Condition is quadratic expansion of likelihood in neighbourhood of θ

Regularity and Asymptotic Efficiency

- In most cases $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$ (i.e. Z_2 also Gaussian)
- When $\sigma^2(\theta) = I^{-1}(\theta)$ then $\hat{\theta}_n$ is said to be *asymptotically efficient*.

Asymptotic Efficiency of MLEs

Under the assumptions of the theorem, the MLE $\hat{\theta}_n$ typically satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta))$$

which establishes the MLE as the *most efficient of all regular estimators*.

$\hookrightarrow$ However, there may exist other regular estimators with *the same asymptotic properties* and *superior finite sample properties*

- Theorem extends to vector parameter case $\boldsymbol{\theta} \in \mathbb{R}^p$, in which case Z_1 is distributed as $\mathcal{N}_p(\mathbf{0}, I^{-1}(\boldsymbol{\theta}))$.

Sketch of proof. (rigorous proof quite technical)

Make stronger assumptions

$$\mathbb{E}_\theta\left[\exp\left(t_1\sqrt{n}(\hat\theta_n-\theta)+t_2S_n(\theta)\right)\right]\overset{n\to\infty}{\longrightarrow} m(t_1,t_2)$$

$$\mathbb{E}_{\theta_n}\left[\exp\left(t_1\sqrt{n}(\hat\theta_n-\theta_n)\right)\right]\overset{n\to\infty}{\longrightarrow} m(t_1,0)$$

for $\theta_n=\theta+c/\sqrt{n}$ and $|t_1|,|t_2|\leq b$, some $b>0$. We need to show that $m(t,0)$ is the product of two moment generating functions, one of which is that of a $\mathcal{N}(0,I^{-1}(\theta))$. Now, note that

$$\begin{aligned}\mathbb{E}_{\theta_n}\left[\exp\left(t_1\sqrt{n}(\hat\theta_n-\theta)\right)\right] &= \exp(t_1c)\mathbb{E}_{\theta_n}\left[\exp\left(t_1\sqrt{n}(\hat\theta_n-\theta_n)\right)\right]\\ &\overset{n\to\infty}{\longrightarrow} \exp(t_1c)m(t_1,0)\end{aligned}$$

Set

$$W_n(\theta,c)=\sum_{i=1}^n\left[\log f(X_i,\theta+c/\sqrt{n})-\log f(X_i;\theta)\right]$$

Regularity and Asymptotic Efficiency

Moreover, it is not too difficult to see that

$$\begin{aligned}\mathbb{E}_{\theta_n}\left[\exp\left(t_1\sqrt{n}(\hat{\theta}_n-\theta)\right)\right] &= \mathbb{E}_{\theta}\left[\exp\left(t_1\sqrt{n}(\hat{\theta}_n-\theta)+W_n(\theta,c)\right)\right] \\ &\stackrel{n\to\infty}{\longrightarrow} m(t_1,c)\exp(-\frac{1}{2}c^2 I(\theta))\end{aligned}$$

since we may substitute the approximately quadratic function for $W_n(\theta, c)$. Equating the two limits,

$$m(t_1,0)=m(t_1,c)\exp\left(-t_1c-\frac{1}{2}c^2I(\theta)\right).$$

Now set $c=-t_1/I(\theta)$ to obtain

$$m(t_1,0)=m\left(t_1,-\frac{t_1}{I(\theta)}\right)\exp\left(\frac{t_1^2}{2I(\theta)}\right)$$

Regularity and Asymptotic Efficiency

It is easy to see that $m(t_1, -t_1/I(\theta))$ is an mgf , and, of course, $\exp\left(\frac{t_1^2}{2I(\theta)}\right)$ is the mgf of a $\mathcal{N}(0, I^{-1}(\theta))$.

- Rigorous proof very similar, but uses cf's and takes care of may technical issues (and of course the points we took as assumptions).

The question that naturally arises then is how to establish regularity?
$\hookrightarrow$ Usually a tedious process.

$\rightarrow$ Hájek regularity assumption may be replaced by *Tierney regularity*:

$$\lim_{n\to\infty} \mathbb{P}_\theta \left[\sqrt{n}(\hat{\theta}_n - \theta) \leq x\right] = G_\theta(x)$$

where G_θ has the property that $\int_{-\infty}^{+\infty} h(x) G_\theta(dx)$ is continuous w.r.t. θ for all bounded h.
$\rightarrow$ If $G_\theta = \mathcal{N}(0, \sigma^2(\theta))$ and $\sigma^2(\theta)$ continuous, then Tierney regularity satisfied.