Statistical Theory: Introduction

Statistical Theory

Victor Panaretos
Ecole polytechnique fédérale de Lausanne

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

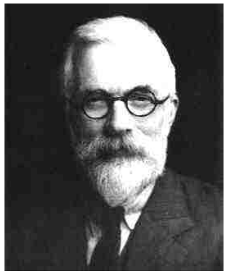
What is This Course About

Statistics $\longrightarrow$ Extracting Information from Data

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

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

What is This Course About?



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

Ronald A. Fisher

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

Jacques Hadamard

What is This Course About?

Statistical Theory: What and How?

What? The rigorous study of the procedure of extracting information from data using the formalism and machinery of mathematics.

How? Thinking of data as outcomes of probability experiments

- Probability offers a natural language to describe uncertainty or partial knowledge
- Deep connections between probability and formal logic
- 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 (uknown) probability model $\mathbb{P}$ that generated the outcome.

(*something in addition to what I knew before observing the outcome A)

Such questions can be:

1. Are the data consistent with a certain model?
2. Given a family of models, can we determine which model generated the data?

These give birth to more questions: how can we answer 1,2? is there a best way? how much "off" is our answer?

A Probabilist and a Statistician Flip a Coin

Example

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

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

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

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

Probabilist Asks:

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

A Probabilist and a Statistician Flip a Coin

Example (cont'd)

Statistician Asks:

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

The Basic Setup

Elements of a Statistical Model:

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

Goal

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

The Basic Setup: An Ilustration

Example (Coin Tossing)

Consider the following probability space:

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

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

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

Describing Families of Distributions: Parametric Models

Definition (Parametrization)

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

Definition (Parametric Model)

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

Example (IID Normal Model)

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

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

Parametric Models

Example (Geometric Distribution)

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

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

Example (Poisson Distribution)

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

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

Identifiability

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

Definition (Identifiability)

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

When a parametrization is not identifiable:

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

Definition (Parameter)

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

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

Identifiability

Example (Binomial Thinning)

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

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

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

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

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

Parametric Inference for Regular Models

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

Regular Models

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

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

Will be considering the mathematical aspects of problems such as:

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

Examples

Example (Five Examples)

- Sampling Inspection (Hypergeometric Distribution)
- Problem of Location (Location-Scale Families)
- Regression Models (Non-identically distributed data)
- Autoregressive Measurement Error Model (Dependent data)
- Random Projections of Triangles (Shape Theory)

Overview of Stochastic Convergence

Statistical Theory

Victor Panaretos
Ecole Polytechnique Fédérale de Lausanne

Functions of Random Variables

Let $X_1, ..., X_n$ be i.i.d. with $\mathbb{E}X_i = \mu$ and $\text{var}[X_i] = \sigma^2$. Consider:

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

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

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 $\text{dist}[\bar{X}_n]$ 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"

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

More generally $\longrightarrow$ Want to understand distribution of $Y = g(X_1, ..., X_n)$ for some general g:

- Often intractable
- Resort to asymptotic approximations to understand behaviour of Y

Warning: While lots known about asymptotics, often they are misused (n small!)

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 "Y_n behaves like Y" for large n means
- $\text{dist}[g(X_1, ..., X_n)] \overset{n\to\infty}{\approx}$?

$\hookrightarrow$ Need appropriate notions of convergence for random variables

Recall: random variables are *functions* between *measurable spaces*

$\implies$ 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)

Each of these is qualitatively different - Some notions 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 with high probability $X_n \approx X$ for large n.

Example

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{U}[0,1]$, 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 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 $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{U}[0,1]$, $M_n = \max\{X_1, ..., X_n\}$, 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 $Q \sim exp(1)$.

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

- Convergence in probability implies convergence in distribution.
- Convergence in distribution does NOT imply convergence in probability
  - $\hookrightarrow$ Consider $X \sim \mathcal{N}(0,1)$, $-X + \frac{1}{n} \xrightarrow{d} X$ but $-X + \frac{1}{n} \xrightarrow{p} -X$.
- "$\xrightarrow{d}$" relates *distribution functions*
  - $\hookrightarrow$ Can use to approximate distributions (approximation error?).
- Both notions of convergence are *metrizable*
  - $\hookrightarrow$ i.e. there exist metrics on the space of random variables and distribution functions that are compatible with the notion of convergence.
  - $\hookrightarrow$ Hence can use things such as the triangle inequality etc.
- "$\xrightarrow{d}$" is also known as "weak convergence" (will see why).

Equivalent Def: $X \xrightarrow{d} X \iff \mathbb{E}f(X_n) \to \mathbb{E}f(X)\ \forall$ cts and bounded f

Some Basic Results

Theorem

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

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

Proof

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

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

since $\{X \le x + \epsilon\}$ contains $\{X_n \le x, |X_n - X| \le \epsilon\}$. Similarly,

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

(proof cont'd).

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

Combining the two inequalities and "sandwitching" yields the result.

(b) Let F be the distribution function of a constant r.v. c,

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

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

Since $X_n \xrightarrow{d} c$. □

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)

Let $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c \in \mathbb{R}$. 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 may assume $c = 0$. Let x be a continuity point of F_X. We have

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

Similarly, $\mathbb{P}[X_n \le x - \epsilon] \le \mathbb{P}[X_n + Y_n \le x] + \mathbb{P}[|Y_n| > \epsilon]$

Therefore,

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

Taking $n \to \infty$, and then $\epsilon \to 0$ proves (a).

(b) By (a) we may assume that $c = 0$ (check). Let $\epsilon, M > 0$:

$$\begin{aligned}\mathbb{P}[|X_n Y_n| > \epsilon] &\le \mathbb{P}[|X_n Y_n| > \epsilon, |Y_n| \le 1/M] + \mathbb{P}[|Y_n| \ge 1/M] \\ &\le \mathbb{P}[|X_n| > \epsilon M] + \mathbb{P}[|Y_n| \ge 1/M] \\ &\xrightarrow{n\to\infty} \mathbb{P}[|X| > \epsilon M] + 0\end{aligned}$$

The first term can be made arbitrarily small by letting $M \to \infty$. □

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, $g(X_n, Y_n) \to g(X, c)$ as $n \to \infty$.

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

- The continuous mapping theorem would be applicable if (X_n, Y_n) weakly converged jointly (i.e. their joint distribution) 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.
- Notice that even the special cases (addition and multiplication) of Slutsky's theorem fail of 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

Taylor expanding around θ gives:

$$g(X_n) = g(\theta) + g'(\theta_n^*)(X_n - \theta), \quad \theta_n^* \text{ between } X_n, \theta.$$

Thus $|\theta_n^* - \theta| < |X_n - \theta| = a_n^{-1} \cdot |a_n(X_n - \theta)| = a_n^{-1} Z_n \xrightarrow{p} 0$ [by Slutsky]
Therefore, $\theta_n^* \xrightarrow{p} \theta$. By the continuous mapping theorem $g'(\theta_n^*) \xrightarrow{p} g'(\theta)$.

$$\text{Thus} \quad a_n(g(X_n) - g(\theta)) = a_n(g(\theta) + g'(\theta_n^*)(X_n - \theta) - g(\theta)) = g'(\theta_n^*) a_n(X - \theta) \xrightarrow{d} g'(\theta)Z.$$

The delta method actually applies even when $g'(\theta)$ is not continuous (proof uses Skorokhod representation).

Exercise: Give a counterexample to show that neither of $X_n \xrightarrow{p} X$ or $X_n \xrightarrow{d} X$ ensures that $\mathbb{E}X_n \to \mathbb{E}X$ as $n \to \infty$.

Theorem (Convergence of Expecations)

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 $\forall\ n$. Then,

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

since $M < \infty$ and the integration domain is bounded. $\square$

Exercise: Generalise the proof to arbitrary random variables.

Remarks on Weak Convergence

- Often difficult to establish weak convergence directly (from definition)
- Indeed, if F_n known, establishing weak convergence is "useless"
- Need other more "handy" sufficient conditions

Scheffé's Theorem

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

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

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

Continuity Theorem

Let X_n and X have characteristic functions $\varphi_n(t) = \mathbb{E}[e^{itX_n}]$, and $\varphi(t) = \mathbb{E}[e^{itX}]$, respectively. Then,
(a) $X_n \xrightarrow{d} X \Leftrightarrow \phi_n \to \phi$ pointwise
(b) If $\phi_n(t)$ converges pointwise to some limit function $\psi(t)$ that is continuous at zero, then:
(i) $\exists$ a measure ν with c.f. ψ
(ii) $F_{X_n} \xrightarrow{w} \nu$.

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 $\mathbf{X}_n = (X_n^{(1)}, ..., X_n^{(d)})^\top$ and $\mathbf{X} = (X^{(1)}, ..., X^{(d)})^\top$. Define the distribution functions $F_{\mathbf{X}_n}(\mathbf{x}) = \mathbb{P}[X_n^{(1)} \leq x^{(1)}, ..., X_n^{(d)} \leq x^{(d)}]$ and $F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}[X^{(1)} \leq x^{(1)}, ..., X^{(d)} \leq x^{(d)}]$, for $\mathbf{x} = (x^{(1)}, ..., x^{(d)})^\top \in \mathbb{R}^d$. We say that $\mathbf{X}_n$ converges in distribution to $\mathbf{X}$ as $n \to \infty$ (and write $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$) if for every continuity point of $F_{\mathbf{X}}$ we have

$$F_{\mathbf{X}_n}(\mathbf{X}) \overset{n\to\infty}{\longrightarrow} F_{\mathbf{X}}(\mathbf{x}).$$

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,

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

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 $A := \{\omega \in \Omega : X_n(\omega) \overset{n\to\infty}{\longrightarrow} X(\omega)\}$. 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 $X_n \xrightarrow{a.s.} X$ 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 $X_n \xrightarrow{L^p} X$) if

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

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 $X_n \xrightarrow{d} X$. 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.

Recalling two basic Theorems

Multivariate Random Variables $\to$ "$\xrightarrow{d}$" defined coordinatewise

Theorem (Strong Law of Large Numbers)

Let $\{X_n\}$ be pairwise 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 requires $\mathbb{E}X_k^2 < \infty$ instead of $\mathbb{E}|X_k| < \infty$ and gives "$\xrightarrow{p}$" instead of "$\xrightarrow{a.s.}$"

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,

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

Convergence Rates

Often convergence not enough $\longrightarrow$ How fast?
$\hookrightarrow$[quality of approximation]

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

Theorem (Berry-Essen)

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be iid random vectors taking values in $\mathbb{R}^d$ and such that $\mathbb{E}[\mathbf{X}_i] = 0$, $cov[\mathbf{X}_i] = I_d$. Define,

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

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

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

Principles of Data Reduction

Statistical Theory

Victor Panaretos
Ecole Polytechnique Fédérale de Lausanne

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Statistical Models and The Problem of Inference

Recall our setup:

- Collection of r.v.'s (a random vector) $\mathbf{X} = (X_1, ..., X_n)$
- $\mathbf{X} \sim F_\theta \in \mathcal{F}$
- $\mathcal{F}$ a parametric class with parameter $\theta \in \Theta \subseteq \mathbb{R}^d$

The Problem of Point Estimation

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

The only guide (apart from knowledge of $\mathcal{F}$) at hand is the data:

$\hookrightarrow$ Anything we "do" will be a function of the data $g(x_1, ..., x_n)$

$\hookrightarrow$ Need to study properties of such functions and information loss incurred (any function of $(x_1, .., x_n)$ will carry at most the same information but usually less)

Statistics

Definition (Statistic)

Let $\mathbf{X}$ be a random sample from F_θ. A *statistic* is a (measurable) function T that maps $\mathbf{X}$ into $\mathbb{R}^d$ and does not depend on θ.

$\hookrightarrow$ Intuitively, any function of the sample alone is a statistic.

$\hookrightarrow$ Any statistics is itself a r.v. with its own distribution.

Example

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

Example

$T(\mathbf{X}) = (X_{(1)}, \ldots, X_{(n)})$ where $X_{(1)} \leq X_{(2)} \leq \ldots X_{(n)}$ are the order statistics of $\mathbf{X}$. Since T depends only on the values of $\mathbf{X}$, T is a statistic.

Example

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

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 $T(\mathbf{X})$ that is not "1-1" carries less information about θ than $\mathbf{X}$
- 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 $\forall\ \theta \in \Theta$.

Example

Suppose that $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ (where μ unknown but σ^2 known). Let $T(X_1, ..., X_n) = X_1 - X_2$; then T has a Normal distribution with mean 0 and variance $2\sigma^2$. Thus T is ancillary for the unknown parameter μ. 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 dist(T) must depend explicitly on θ.
- Intuitively, the amount of information T gives on θ increases as the dependence of dist(T) on θ increases

Example

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{U}[0, \theta]$, $S = \min(X_1, \ldots, X_n)$ and $T = \max(X_1, \ldots, X_n)$.

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

$\hookrightarrow$ 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 while

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

Statistics and Information about θ

- $\mathbf{X} = (X_1, \ldots, X_n) \overset{iid}{\sim} F_\theta$ and $T(\mathbf{X})$ a statistic.
- The *fibres* or *level sets* or *contours* of T are the sets

$$A_t = \{\mathbf{x} \in \mathbb{R}^n : T(\mathbf{x}) = t\}.$$

$\hookrightarrow$ T is constant when restricted to an fibre.

- Any realization of $\mathbf{X}$ that falls in a given fibre is equivalent as far as T is concerned

Statistics and Information about θ

- Look at the dist($\mathbf{X}$) on an fibre A_t: $f_{\mathbf{X}|T=t}(\mathbf{x})$
- Suppose $f_{\mathbf{X}|T=t}$ is independent of θ

$\Longrightarrow$ Then $\mathbf{X}$ contains no information about θ on the set A_t

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

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

Definition (Sufficient Statistic)

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

Sufficient Statistics

Example (Bernoulli Trials)

Let $X_1, ..., X_n \stackrel{iid}{\sim}$ Bernoulli(θ) and $T(\mathbf{X}) = \sum_{i=1}^n X_i$. Given $\mathbf{x} \in \{0,1\}^n$,

$$\begin{aligned}
\mathbb{P}[\mathbf{X}=\mathbf{x}|T=t] &= \frac{\mathbb{P}[\mathbf{X}=\mathbf{x}, T=t]}{\mathbb{P}[T=t]} = \frac{\mathbb{P}[\mathbf{X}=\mathbf{x}]}{\mathbb{P}[T=t]}\mathbf{1}\{\Sigma_{i=1}^n x_i = t\} \\
&= \frac{\theta^{\Sigma_{i=1}^n x_i}(1-\theta)^{n-\Sigma_{i=1}^n x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}}\mathbf{1}\{\Sigma_{i=1}^n x_i = t\} \\
&= \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1}.
\end{aligned}$$

- T is sufficient for $\theta \to$ Given the number of tosses that came heads, knowing which tosses exactly came heads is irrelevant in deciding if the coin is fair:

0 0 1 1 1 0 1 VS 1 0 0 0 1 1 1 VS 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 $\mathbf{X} = (X_1, \dots, X_n)$ has a joint density or frequency function $f(\mathbf{x};\theta)$, $\theta \in \Theta$. A statistic $T = T(\mathbf{X})$ is sufficient for θ if and only if

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

Example

Let $X_1, ..., X_n \stackrel{iid}{\sim} \mathcal{U}[0,\theta]$ with pdf $f(x;\theta) = \mathbf{1}\{x \in [0,\theta]\}/\theta$. Then,

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

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

Sufficient Statistics

Proof of Neyman-Fisher Theorem - Discrete Case.

Suppose first that T is sufficient. Then

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

Since T is sufficient, $\mathbb{P}[\mathbf{X}=\mathbf{x}|T=T(\mathbf{x})]$ is independent of θ and so $f(x;\theta) = g(T(\mathbf{x});\theta)h(\mathbf{x})$. Now suppose that $f(x;\theta) = g(T(\mathbf{x});\theta)h(\mathbf{x})$. Then if $T(\mathbf{x}) = t$,

$$\begin{aligned}
\mathbb{P}[\mathbf{X}=\mathbf{x}|T=t] &= \frac{\mathbb{P}[\mathbf{X}=\mathbf{x}, T=t]}{\mathbb{P}[T=t]} = \frac{\mathbb{P}[\mathbf{X}=\mathbf{x}]}{\mathbb{P}[T=t]}\mathbf{1}\{T(\mathbf{x})=t\} \\
&= \frac{g(T(\mathbf{x});\theta)h(\mathbf{x})\mathbf{1}\{T(\mathbf{x})=t\}}{\sum_{\mathbf{y}:T(\mathbf{y})=t} g(T(\mathbf{y});\theta)h(\mathbf{y})} = \frac{h(\mathbf{x})\mathbf{1}\{T(\mathbf{x})=t\}}{\sum_{T(\mathbf{y})=t} h(\mathbf{y})}.
\end{aligned}$$

which does not depend on θ. $\square$

Minimally Sufficient Statistics

- Saw that sufficient statistic keeps what is important and leaves out irrelevant information.
- How much info can we through away? Is there a "necessary" statistic?

Definition (Minimally Sufficient Statistic)

A statistic $T = T(\mathbf{X})$ is said to be *minimally sufficient* for the parameter θ if for any sufficient statistic $S = S(\mathbf{X})$ there exists a function $g(\cdot)$ with

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

Lemma

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

Theorem

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

Proof.

Assume for simplicity that $f(\mathbf{x}; \theta) > 0$ for all $\mathbf{x} \in \mathbb{R}^n$ and $\theta \in \Theta$.
[sufficiency part] Let $\mathcal{T} = \{T(\mathbf{y}) : y \in \mathbb{R}^n\}$ be the image of $\mathbb{R}^n$ under T and let A_t be the level sets of T. For each t, choose an element $\mathbf{y}_t \in A_t$. Notice that for any $\mathbf{x}$, $\mathbf{y}_{T(\mathbf{x})}$ is in the same level set as $\mathbf{x}$, so that

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

does not depend on θ by assumption. Let $g(t, \theta) := f(\mathbf{y}_t; \theta)$ and notice

$$f(\mathbf{x}; \theta) = \frac{f(\mathbf{x}_{T(\mathbf{x})}; \theta) f(\mathbf{x}; \theta)}{f(\mathbf{x}_{T(\mathbf{x})}; \theta)} = g(T(\mathbf{x}), \theta) h(\mathbf{x})$$

and the claim follows from the factorization theorem.

[minimality part] Suppose that T' is another sufficient statistic. By the factorization thm: $\exists g', h' : \quad f(\mathbf{x}; \theta) = g'(T'(\mathbf{x}); \theta) h'(\mathbf{x})$. Let $\mathbf{x}, \mathbf{y}$ be such that $T'(\mathbf{x}) = T'(\mathbf{y})$. Then

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

Since ratio does not depend on θ, we have by assumption $T(\mathbf{x}) = T(\mathbf{y})$. Hence T is a function of T'; so is minimal by arbitrary choice of T'. □

Example (Bernoulli Trials)

Let $X_1, ..., X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$. Let $\mathbf{x}, \mathbf{y} \in \{0, 1\}^n$ be two possible outcomes. Then

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

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

Complete Statistics

- Ancillary Statistic $\rightarrow$ Contains no info on θ
- Minimally Sufficient Statistic $\rightarrow$ Contains all relevant info and as little irrelevant as possible.
- Should they be mutually independent?

Definition (Complete Statistic)

Let $\{g(t; \theta) : \theta \in \Theta\}$ be a family of densities of frequencies corresponding to a statistic $T(\mathbf{X})$. The statistic T is called *complete* if

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

Example

If $\hat{\theta} = T(\mathbf{X})$ is an unbiased estimator of θ (i.e. $\mathbb{E}\hat{\theta} = \theta$) which can be written as a function of a complete sufficient statistic, then it is the unique such estimator.

Complete Statistics

Theorem (Basu's Theorem)

A complete sufficient statistic is independent of every ancillary statistic.

Proof.

We consider the discrete case only. It suffices to show that,

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

$$\text{Define: } h(t) = \mathbb{P}[S(\mathbf{X}) = s | T(\mathbf{X}) = t] - \mathbb{P}[S(\mathbf{X}) = s]$$

and observe that:

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

and so h does not depend on θ.

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

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

But T is complete so it follows that $h(t) = 0$ for all t. QED. □

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

- No need to determine their joint distribution
- Needs showing completeness (usually hard analytical problem)
- Will see models in which completeness is easy to check

Completeness and Minimal Sufficiency

Theorem (Lehmann-Scheffé)

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

Proof.

First of all we show that a minimally sufficient statistic exists. Define an equivalence relation as $\mathbf{x} \equiv \mathbf{x}'$ if and only if $f(\mathbf{x}; \theta)/f(\mathbf{x}'; \theta)$ is independent of θ. If S is any function such that $S = c$ on these equivalent classes, then S is a minimally sufficient, establishing existence (rigorous proof by Lehmann-Scheffé (1950)).
Therefore, it must be the case that $S = g_1(T)$, for some g_1. Let $g_2(S) = \mathbb{E}[T|S]$ (does not depend on θ since S sufficient). Consider:

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

Write $\mathbb{E}[g(T)] = \mathbb{E}[T] - \mathbb{E}\{\mathbb{E}[T|S]\} = \mathbb{E}T - \mathbb{E}T = 0$ for all θ.

(proof cont'd).

By completeness of T, it follows that $g_2(S) = T$ a.s. In fact, g_2 has to be injective, or otherwise we would contradict minimal sufficiency of S. But then T is 1-1 a function of S and S is a $1-1$ function of T. Invoking our previous lemma proves that T is minimally sufficient. □

One can also prove:

Theorem

If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.

Special Families of Models

Statistical Theory

Victor Panaretos
Ecole Polytechnique Fédérale de Lausanne

Focus on Parametric Families

Recall our setup:

- Collection of r.v.'s (a random vector) $\mathbf{X} = (X_1, ..., X_n)$
- $\mathbf{X} \sim F_\theta \in \mathcal{F}$
- $\mathcal{F}$ a parametric class with parameter $\theta \in \Theta \subseteq \mathbb{R}^d$

The Problem of Point Estimation

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

The only guide (apart from knowledge of $\mathcal{F}$) at hand is the data:

$\hookrightarrow$ Anything we "do" will be a function of the data $g(x_1, ..., x_n)$

So far have concentrated on aspects of data: approximate distributions + data reduction...... 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 $g : \Theta \to \mathcal{F}$ an onto mapping. The pair (Θ, g) is called a *parametrization* of $\mathcal{F}$.

$\hookrightarrow$ 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 have seen a number of examples of distributions...
...have worked out certain properties individually

Question

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

Definition (Exponential Family)

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

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

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

- k need not be equal to p, although they often coincide.
- The value of k may be reduced if c or T satisfy linear constraints.
- We will assume that the representation above is minimal.
  - $\hookrightarrow$ Can re-parametrize via $\phi_i = c_i(\theta)$, the *natural parameter*.

Consider the following variational problem:

Determine the probability distribution f supported on $\mathcal{X}$ with maximum entropy

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

subject to the linear constraints

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

Philosophy: How to choose a probability model for a given situation?

Maximum entropy approach:

- In any given situation, choose the distribution that gives *highest uncertainty* while satisfying situation–specific required constraints.

Proposition.

The unique solution of the constrained optimisation problem has the form

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

Proof.

Let $g(\mathbf{x})$ be a density also satisfying the constraints. Then,

$$\begin{aligned} H(g) &= -\int_{\mathcal{X}} g(\mathbf{x})\log g(\mathbf{x})d\mathbf{x} = -\int_{\mathcal{X}} g(\mathbf{x})\log\left[\frac{g(\mathbf{x})}{f(\mathbf{x})}f(\mathbf{x})\right]d\mathbf{x} \\ &= -KL(g\|f) - \int_{\mathcal{X}} g(\mathbf{x})\log f(\mathbf{x})d\mathbf{x} \\ &\le -\log Q\underbrace{\int_{\mathcal{X}} g(\mathbf{x})dx}_{=1} - \int_{\mathcal{X}} g(\mathbf{x})\left(\sum_{i=1}^k \lambda_i T_i(\mathbf{x})\right)d\mathbf{x} \end{aligned}$$

But g also satisfies the moment constraints, so the last term is

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

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

- The λ's are the Lagrange multipliers derived by the Lagrange form of the optimisation 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 $X \sim$Binomial(n,θ) with n known. Then

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

Example (Gamma Distribution)

Let $X_1,...,X_n \overset{iid}{\sim}$Gamma with unknown shape parameter α and unknown scale parameter λ. Then,

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

Example (Heteroskedastic Gaussian Distribution)

Let $X_1,...,X_n \overset{iid}{\sim} \mathcal{N}(\theta,\theta^2)$. Then,

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

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 $X\sim\mathcal{U}[0,\theta]$. Then, $f_X(x;\theta)=\frac{\mathbf{1}\{x\in[0,\theta]\}}{\theta}$. Since the support of f, $\mathcal{X}$, depends on θ, we do *not* have an exponential family.

The Exponential Family of Distributions

Proposition

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

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

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

(a) the parameter space Θ is open,

(b) $c(\theta)$ is a one-to-one function on Θ,

(c) $c(\theta), c^{-1}(\theta), d(\theta)$ are twice differentiable functions on Θ.

Then,

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

Proof.

Define $\phi=c(\theta)$ the *natural parameter* of the exponential family. Let $d_0(\phi)=d(c^{-1}(\phi))$, where c^{-1} is well-defined since c is 1-1. Since c is a homeomorphism, $\Phi=c(\Theta)$ is open. Choose s sufficiently small so that $\phi+s\in\Phi$, and observe that the m.g.f. of T is

$$\begin{aligned}\mathbb{E}\exp[sT(\mathbf{X})] &= \int e^{sT(\mathbf{x})}e^{\phi T(\mathbf{x})-d_0(\phi)+S(\mathbf{x})}d\mathbf{x} \\ &= e^{d_0(\phi+s)-d_0(\phi)}\underbrace{\int e^{(\phi+s)T(\mathbf{x})-d_0(\phi+s)+S(\mathbf{x})}d\mathbf{x}}_{=1} \\ &= \exp[d_0(\phi+s)-d_0(\phi)],\end{aligned}$$

By our assumptions we may differentiate w.r.t. s, and, setting $s=0$, we get $\mathbb{E}[T(\mathbf{X})]=d_0'(\phi)$ and $\mathrm{Var}[T(\mathbf{X})]=d_0''(\phi)$. But

$$d_0'(\phi)=d'(\theta)/c'(\theta) \text{ and } d_0''(\phi)=[d''(\theta)c'(\theta)-d'(\theta)c''(\theta)]/[c'(\theta)]^3$$

and so the conclusion follows. □

Exponential Families and Sufficiency

Exercise

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

Lemma

Suppose that $\mathbf{X} = (X_1, \ldots, X_n)$ has a k-parameter exponential family distribution with density or frequency function

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

for $x \in \mathcal{X}$. Then, the statistic $(T_1(\mathbf{x}), \ldots, T_k(\mathbf{x}))$ is sufficient for θ

Proof.

Set $g(\mathbf{T}(\mathbf{x}); \boldsymbol{\theta}) = \exp\{\sum_i T_i(\mathbf{x}) c_i(\boldsymbol{\theta}) + d(\theta)\}$ and $h(\mathbf{x}) = e^{S(\mathbf{x})}\mathbf{1}\{\mathbf{x} \in \mathcal{X}\}$, and apply the factorization theorem. □

Exponential Families and Completeness

Theorem

Suppose that $\mathbf{X} = (X_1, \ldots, X_n)$ has a k-parameter exponential family distribution with density or frequency function

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

for $x \in \mathcal{X}$. Define $C = \{(c_1(\theta), \ldots, c_k(\theta)) : \theta \in \Theta\}$. If the set C contains an open set (rectangle) of the form $(a_1, b_1) \times \ldots \times (a_k, b_k)$ then the statistic $(T_1(\mathbf{X}), \ldots, T_k(\mathbf{X}))$ is complete for θ, and so minimally sufficient.

- The result is essentially a consequence of the uniqueness of characteristic functions.
- Intuitively, result says that a k-dimensional sufficient statistic in a k-parameter exponential family will also be complete provided that the effective dimension of the natural parameter space is k.

Sampling Exponential Families

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

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

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

The Natural Statistics

Lemma

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

Proof. (discrete case).

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

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

□

The Natural Statistics

Lemma

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

Proof. (discrete case).

Let $\mathcal{T}_i = \tau_i(\mathbf{X})$. Have $\mathbb{P}[\mathcal{T} = \mathbf{y}] = \delta(\theta)\mathcal{S}(\mathbf{y}) \exp\left[\sum_{i=1}^k c_i(\theta) y_i\right]$, so

$$\mathbb{P}[\mathcal{T}_A = \mathbf{y}_A | \mathcal{T}_{A^c} = \mathbf{y}_{A^c}] = \frac{\mathbb{P}[\mathcal{T}_A = \mathbf{y}_A, \mathcal{T}_{A^c} = \mathbf{y}_{A^c}]}{\sum_{\mathbf{w} \in \mathbb{R}^l} \mathbb{P}[\mathcal{T}_A = \mathbf{w}, \mathcal{T}_{A^c} = \mathbf{y}_{A^c}]}$$

$$= \frac{\delta(\theta)\mathcal{S}((\mathbf{y}_A, \mathbf{y}_{A^c})) \exp\left[\sum_{i \in A} c_i(\theta) y_i\right] \exp\left[\sum_{i \in A^c} c_i(\theta) y_i\right]}{\delta(\theta) \exp\left[\sum_{i \in A^c} c_i(\theta) y_i\right] \sum_{\mathbf{w} \in \mathbb{R}^l} \mathcal{S}((\mathbf{w}, \mathbf{y}_{A^c})) \exp\left[\sum_{i \in A} c_i(\theta) w_i\right]}$$

$$= \Delta(\{c_i(\theta) : i \in A\}) h(\mathbf{y}_A) \exp\left[\sum_{i \in A} c_i(\theta) y_i\right] \qquad \square$$

The Natural Statistics and Sufficiency

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

- Already know that $\boldsymbol{\tau}$ is sufficient for $\phi = c(\theta)$.
- But result tells us something even stronger:

> that each τ_i is sufficient for $\phi_i = c_i(\theta)$

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

Groups Acting on the Sample Space

Basic Idea

Often 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 Sample Space

- Have a group of transformations G, 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 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*.

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

Invariance and Equivalence

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

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

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 $T(x) = T(gx)\ \forall g \in G$.

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 Equivalence

- 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.

Invariance and Equivalence

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. $\square$

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 $g_\theta x = g_{(\eta,\tau)}x = \eta + \tau x$ so

$$g_{(\eta,\tau)} \circ g_{(\mu,\sigma)}x = \eta + \tau\mu + \tau\sigma x = g_{(\eta+\tau\mu,\tau\sigma)}x$$

- set of transformations is closed under composition
- $g_{(0,1)} \circ g_{(\eta,\tau)} = g_{\eta,\tau} \circ g_{(0,1)} = g_{(\eta,\tau)}$ (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 $G = \{g_\theta : \theta \in \mathbb{R} \times \mathbb{R}_+\}$ is a group under $\circ$.
- Action of G on random sample $\mathbf{X} = \{X_i\}_{i=1}^n$ is $g_{(\eta,\tau)}\mathbf{X} = \eta\mathbf{1}_n + \tau\mathbf{X}$.
- 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 $S(\mathbf{X}) = (\bar{X}, V^{1/2})$ where $V = \frac{1}{n-1}\sum(X_j - \bar{X})^2$:

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

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

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

Example (The Multivariate Gaussian Distribution)

- Let $\mathbf{Z} \sim \mathcal{N}_d(0, I)$ and consider $\mathbf{X} = \boldsymbol{\mu} + \Omega\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \Omega\Omega^{\mathsf{T}})$
- Parameter is $(\boldsymbol{\mu}, \Omega) \in \mathbb{R}^d \times \mathrm{GL}(d)$
- 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 $G = \{g_\theta : \theta \in \mathbb{R} \times \mathbb{R}_+\}$ is a group under $\circ$ (affine group).
- Action of G on $\mathbf{X}$ is $g_{(\boldsymbol{\mu},\Omega)}\mathbf{X} = \boldsymbol{\mu} + \Omega\mathbf{X}$.
- Induced group action on Θ is $(\boldsymbol{\mu}, \Omega) * (\boldsymbol{\nu}, \Psi) = (\boldsymbol{\nu} + \Psi\mu, \Psi\Omega)$.

Basic Principles of Point Estimation

Statistical Theory

Victor Panaretos
Ecole Polytechnique Fédérale de Lausanne

Point Estimation for Parametric Families

- Collection of r.v.'s (a random vector) $\mathbf{X} = (X_1, ..., X_n)$
- $\mathbf{X} \sim F_\theta \in \mathcal{F}$
- $\mathcal{F}$ a parametric class with parameter $\theta \in \Theta \subseteq \mathbb{R}^d$

The Problem of Point Estimation

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

So far considered aspects related to point estimation:

- Considered approximate distributions of $g(X_1, ..., X_n)$ as $n \uparrow \infty$
- Studied the information carried by $g(X_1, .., X_n)$ w.r.t θ
- Examined general parametric models

Today: How do we estimate θ in general? 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 $\mathbf{X} = (X_1, ..., X_n) \sim F_{\theta_0}$ 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!

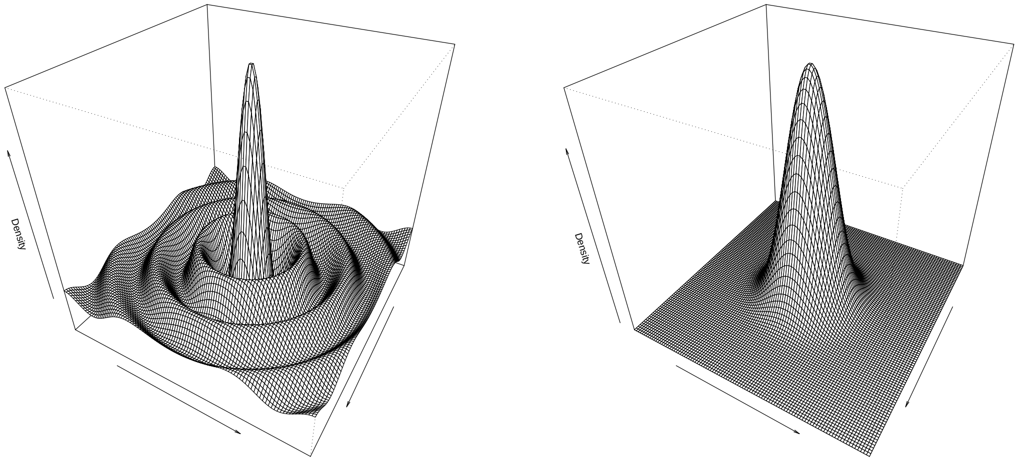
$\hookrightarrow$ Harder to answer what a *good* estimator is!

- 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 ∞-dimensional description of quality.
- Look at some simpler measures of quality?

Concentration around a Parameter



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're 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. $\text{bias}(\hat{\theta}) = 0$.

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

Definition (Mean Squared Error)

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

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

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 $\text{dist}(\hat{\theta})$ relative to θ (somehow assumes mean is good measure of location)
- MSE gives us a measure of spread/dispersion of $\text{dist}(\hat{\theta})$ around θ
- If $\hat{\theta}$ is unbiased for $\theta \in \mathbb{R}$ then $\text{Var}(\hat{\theta}) = MSE(\hat{\theta})$
- for $\Theta \subseteq \mathbb{R}^d$ have $MSE(\hat{\theta}) := \mathbb{E}\|\hat{\theta} - \theta\|^2$.

Example

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

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

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

The Bias-Variance Decomposition of MSE

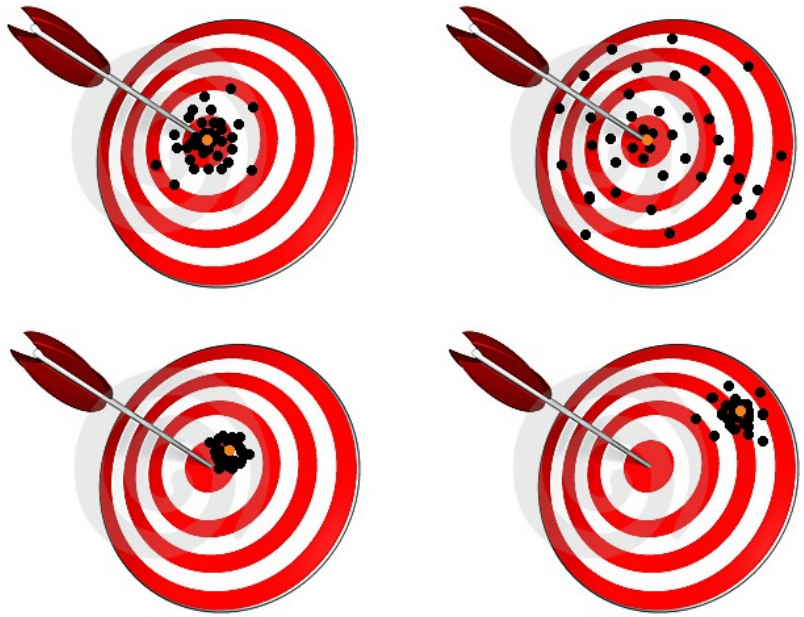
$$\begin{aligned}
\mathbb{E}[\hat{\theta} - \theta]^2 &= \mathbb{E}[\hat{\theta} - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \theta]^2 \\
&= \mathbb{E}\left\{(\hat{\theta} - \mathbb{E}\hat{\theta})^2 + (\mathbb{E}\hat{\theta} - \theta)^2 + 2(\hat{\theta} - \mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta}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

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

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

- 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 have bias/variance tradeoff (e.g. nonparametric regression)

Bias–Variance Tradeoff



Consistency

Can also consider quality of an estimator not for given sample size, but also as sample size increases.

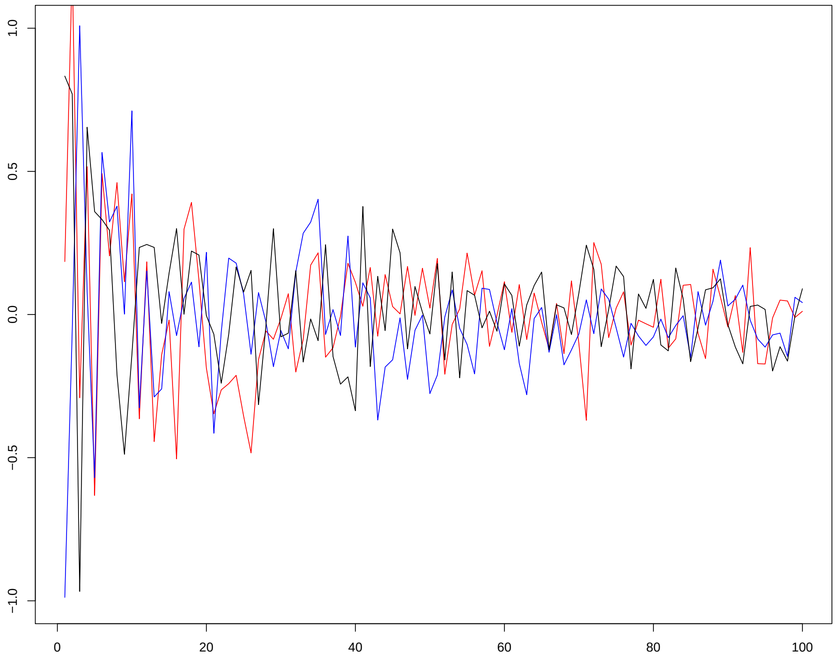
Consistency

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 sample size grows (usually have $\hat{\theta}_n$ being an estimator based on n iid values).
- 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: $X_1, ..., X_n \sim \mathcal{N}(0,1)$, plot $\bar{X}_n$ for $n = 1, 2, ...$



Plug-In Estimators

Want to find general procedures for constructing estimators.
$\hookrightarrow$ An idea: $\theta \mapsto F_\theta$ is bijection under identifiability.

- Recall that more generally, a parameter is a function $\nu : \mathcal{F} \to \mathcal{N}$
- Under identifiability $\nu(F_\theta) = q(\theta)$, some q.

The Plug-In Principle

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

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

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: $\theta(F) := \frac{d}{dx} F(x)\Big|_{x=x_0}$

The Empirical Distribution Function

Plug-in Principle

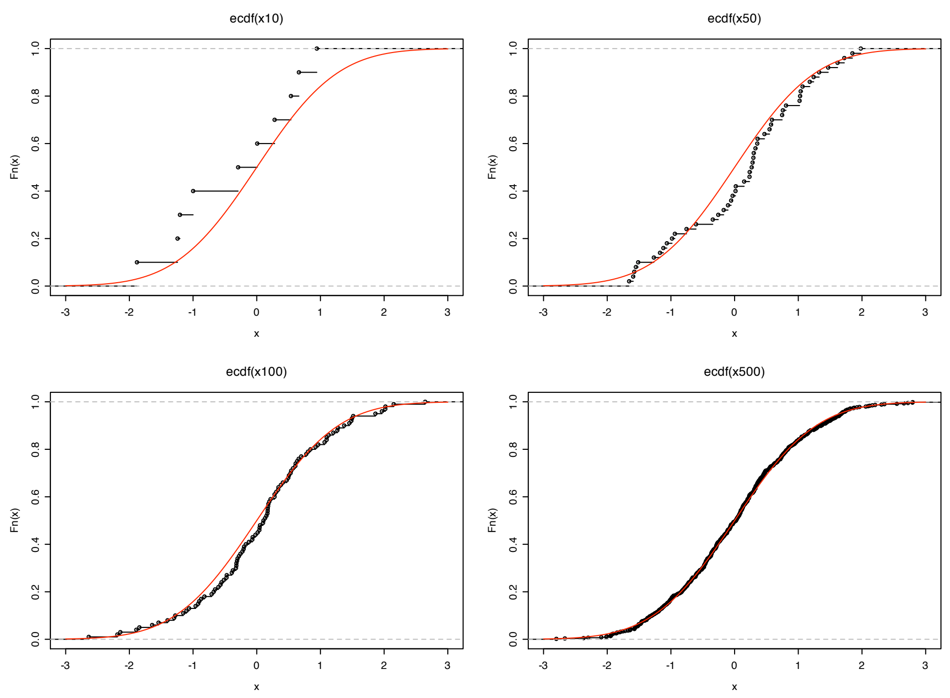
Converts problem of estimating θ into problem of estimating F. But how?

Consider the case when $\mathbf{X} = (X_1, .., X_n)$ has iid coordinates. We may define the empirical version of the distribution function $F_{X_i}(\cdot)$ as

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

- Places mass $1/n$ on each observation
- SLLN $\implies \hat{F}_n(y) \xrightarrow{a.s.} F(y)\ \forall y \in \mathbb{R}$
  - $\hookrightarrow$ since $\mathbf{1}\{X_i \leq y\}$ are iid Bernoulli$(F(y))$ random variables

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



The Empirical Distribution Function

Seems that we're actually doing better than just pointwise convergence...

Theorem (Glivenko-Cantelli)

Let $X_1, .., X_n$ be independent random variables, distributed according to F. Then, $\hat{F}_n(y) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq y\}$ 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 $F(y) = y\mathbf{1}\{y \in [0, 1]\}$. Fix a regular finite partition $0 = x_1 \leq x_2 \leq \ldots \leq x_m = 1$ of [0,1] (so $x_{k+1} - x_k = (m-1)^{-1}$). By monotonicity of $F, \hat{F}_n$

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

Adding and subtracting $F(x_k)$ within each term we can bound above by

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

by an application of the triangle inequality to each term. Letting $n \uparrow \infty$, the SSLN implies that the first term vanishes almost surely. Since m is arbitrary we have proven that, given 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 the cdf F is uniform.

For a general cdf F, we let $U_1, U_2, \ldots \overset{iid}{\sim} \mathcal{U}[0,1]$ and define

$$W_i := F^{-1}(U_i) = \inf\{x : F(x) \geq U_i\}.$$

Observe that

$$W_i \leq x \iff U_i \leq F(x)$$

so that $W_i \overset{d}{=} X_i$. By Skorokhod's representation theorem, we may thus assume that

$$W_i = X_i \qquad \text{a.s.}$$

Letting $\hat{G}_n$ be the edf of $(U_1, \ldots, U_n)$ we note that

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

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

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

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

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

Example (Mean of a function)

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

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

Example (Variance)

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

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

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,

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

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 G concentrated at 0.

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

The Method of Moments

Perhaps the oldest estimation method (Karl Pearson, late 1800's).

Method of Moments

Let $X_1, ..., X_n$ be an iid sample from F_θ, $\theta \in \mathbb{R}^p$. The *method of moments* estimator $\hat{\theta}$ of θ is the solution w.r.t 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 $\theta(F)$ unavailble
- While the method was introduced by equating moments, it may be generalized to equating p theoretical functionals to their empirical analogues.
  $\hookrightarrow$ Choice of equations 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,

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

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

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

Example (Exponential Distribution)

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

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

Tune value of r so as 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 estimates $\hat{\alpha} = \bar{X}^2/\hat{\sigma}^2$ and $\hat{\lambda} = \bar{X}/\hat{\sigma^2}$.

Example (Discrete Uniform Distribution)

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{U}\{1, 2, ..., \theta\}$, 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 $\mathbf{X} = (X_1, ..., X_n)$ 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

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

$\rightarrow$ Important here that T_k is a reasonable estimator of $\mathbb{E}T_k$. (motivation)

Comments on Plug-In and MoM Estimators

- Usually easy to compute and can be valuable as preliminary estimates for algorithms that attempt to compute more efficient (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, so they are likely to be close to the true parameter value for large sample size.
  - ↪ Use empirical process theory for plug-ins
  - ↪ Estimating equation theory for MoM's
- Can lead to biased estimators, or even completely ridiculous estimators (will 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, both n and p unknown).
- Will later discuss optimality in estimation, and appropriateness (or inappropriateness) will become clearer.
- Observation: many of these estimators do not depend solely on sufficient statistics
  - ↪ Sufficiency seems to play an important role in optimality – and it does (more later)
- Will now see a method where estimator depends *only* on a sufficient statistic, when such a statistic exists.

The Likelihood Function

A central theme in statistics. Introduced by Ronald Fisher.

Definition (The Likelihood Function)

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

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

↪ Notice that we consider L as a function of θ NOT of $\mathbf{X}$.

Interpretation: Most easily interpreted in the discrete case → How likely does the value θ make what we observed?

(can extend interpretation to continuous case by thinking of $L(\theta)$ as how likely θ makes something in a small neighbourhood of what we observed)

When $\mathbf{X}$ has iid coordinates with density $f(\cdot; \theta)$, then likelihood is:

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

Maximum Likelihood Estimators

Definition (Maximum Likelihood Estimators)

Let $\mathbf{X} = (X_1, ..., X_n)$ 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 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 that is 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 MLE's

Saw that MoMs and Plug-Ins often do not depend only on sufficient statistics.

$\hookrightarrow$ i.e. they also use "irrelevant" information

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

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

i.e. any MLE depends on data ONLY through the sufficient statistic

- MLE's 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 MLE's

- When the support of a distribution depends on a parameter, maximization is usually carried out 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:

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

- Notice that 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 does not occur on the boundary of Θ.

Example (Uniform Distribution)

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{U}[0, \theta]$. 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 $\theta \leq X_{(n)}$ the likelihood is zero. In the domain $[X_{(n)}, \infty)$, the likelihood is a decreasing function of θ. Hence $\hat{\theta} = X_{(n)}$.

Example (Poisson Distribution)

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

$$L(\lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \right\} \implies \log L(\lambda) = -n\lambda + \log \lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log(x_i!)$$

Setting $\nabla_\theta \log L(\theta) = -n + \lambda^{-1} \sum x_i = 0$ we obtain $\hat{\lambda} = \bar{x}$ since $\nabla^2_\theta \log L(\theta) = -\lambda^{-2} \sum x_i < 0$.

Maximum Likelihood Estimation

Statistical Theory

Victor Panaretos
Ecole Polytechnique Fédérale de Lausanne

Point Estimation for Parametric Families

- Collection of r.v.'s (a random vector) $\mathbf{X} = (X_1, ..., X_n)$
- $\mathbf{X} \sim F_\theta \in \mathcal{F}$
- $\mathcal{F}$ a parametric class with parameter $\theta \in \Theta \subseteq \mathbb{R}^d$

The Problem of Point Estimation

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

Last week, we saw three estimation methods:

- the plug-in method,
- the method of moments,
- maximum likelihood.

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

Maximum Likelihood Estimators

Recall our definition of a maximum likelihood estimator:

Definition (Maximum Likelihood Estimators)

Let $\mathbf{X} = (X_1, ..., X_n)$ 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 of* θ.

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

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

$\rightarrow$ $\hat{\theta}$ makes what we observed *most probable, 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 (frequency) functions on $\mathbb{R}$. The *Kullback-Leibler divergence*, of q with respect to p is defined as:

$$KL(q\|p) := \int_{-\infty}^{+\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx.$$

- Have $KL(p\|p) = \int_{-\infty}^{+\infty} p(x)\log(1)dx = 0$.
- By Jensen's inequality, for $X \sim p(\cdot)$ we have

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

  since q integrates to 1.
- $p \neq q$ implies that $KL(q\|p) > 0$.
- KL is, in a sense, a distance between probability distributions
- KL is not a metric: no symmetry and no triangle inequality!

Likelihood through KL-divergence

Lemma (Maximum Likelihood as Minimum KL-Divergence)

An estimator $\hat{\theta}$ based on an iid sample $X_1, ..., X_n$ is a maximum likelihood estimator if and only if $KL(F(x;\hat{\theta})\|\hat{F}_n(x)) \leq KL(F(x;\theta)\|\hat{F}_n(x))\ \forall\theta \in \Theta$.

Proof (discrete case).

We recall that $\int h(x)d\hat{F}_n(x) = n^{-1}\sum h(X_i)$ so that

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

Likelihood through KL-divergence

<u>Intuition</u>:

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

Final comments 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).

Asymptotics for MLE's

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

Often, when MLE coincides with an MoM estimator, this can be seen directly.

Example (Geometric distribution)

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

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

MLE of θ is

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

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

Example (Geometric distribution)

Now apply the delta method with $g(x) = 1/(1+x)$, so that $g'(x) = -1/(1+x)^2$:

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

Example (Uniform distribution)

Suppose that $X_1, ..., X_n \stackrel{iid}{\sim} \mathcal{U}[0, \theta]$. MLE of θ is

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

with distribution function $\mathbb{P}[\hat{\theta}_n \le x] = (x/\theta)^n \mathbf{1}\{x \in [0, \theta]\}$. Thus for $\epsilon > 0$,

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

so that the MLE is a consistent estimator.

Example (Uniform distribution)

To determine the asymptotic concentration of $\text{dist}(\hat{\theta}_n)$ around θ,

$$\begin{aligned} \mathbb{P}[n(\theta - \hat{\theta}_n) \le x] &= \mathbb{P}\left[\hat{\theta}_n \ge \theta - \frac{x}{n}\right] \\ &= 1 - \left(1 - \frac{x}{\theta n}\right)^n \\ &\xrightarrow{n\to\infty} 1 - \exp(-x/\theta) \end{aligned}$$

so that $n(\theta - \hat{\theta}_n)$ weakly converges to an exponential random variable. Thus we understand the concentration of $\text{dist}(\hat{\theta}_n - \theta)$ around zero for large n as that of an exponential distribution with variance $\frac{1}{\theta^2 n^2}$.

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

- $\ell(x;\theta) = \log f(x;\theta)$
- $\ell'(x;\theta)$, $\ell''(x;\theta)$ and $\ell'''(x;\theta)$ are partial derivatives w.r.t θ.

Asymptotics for 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 w.r.t. θ for all $x \in \text{supp}f$.

(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

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

Let's take a closer look at these conditions...

If Θ is open, then for θ_0 the true parameter, it always makes sense for an estimator $\hat{\theta}$ to have a symmetric distribution around θ_0 (e.g. Gaussian).

Asymptotics for the MLE

Under condition (A2) we have $\frac{d}{d\theta}\int_{\text{supp}\,f} f(x;\theta)dx = 0$ for all $\theta \in \Theta$ so that, if we can interchange integration and differentiation,

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

so that in the presence of (A2), (A4) is essentially a condition that enables differentiation under the integral and asks that the r.v. ℓ' have a finite second moment for all θ. Similarly, (A5) requires that ℓ'' have a first moment for all θ.

Conditions (A2) and (A6) are smoothness conditions that will allow us to "linearize" the problem, while the other conditions will allow us to "control" the random linearization.

Furthermore, if we can differentiate twice under the integral sign

$$0 = \int \frac{d}{d\theta}[\ell'(x;\theta)f(x;\theta)]dx = \int \ell''(x;\theta)f(x;\theta)dx + \int (\ell'(x;\theta))^2 f(x;\theta)dx$$

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

Example (Exponential Family)

Let $X_1, \dots, X_n$ 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,

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

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

Example (Exponential Family)

Furthermore,

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

and

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

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

Asymptotic Normality of the MLE

Theorem (Asymptotic Distribution of the MLE)

Let $X_1, \dots, X_n$ be iid random variables with density (frequency) $f(x;\theta)$ and satisfying conditions (A1)-(A6). Suppose that the sequence of MLE's $\hat{\theta}_n$ satisfies $\hat{\theta}_n \xrightarrow{p} \theta$ where

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

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 $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I(\theta)}\right)$.

Proof.

Under conditions (A1)-(A3), if $\hat{\theta}_n$ maximizes the likelihood, we have

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

Expanding this equation in a Taylor series, we get

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

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

Dividing accross by $\sqrt{n}$ yields

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

which suggests 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 the central limit theorem and condition (A4), it follows that

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

Next, the weak law of large numbers along with condition (A5) implies

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

Now we turn to show that the remainder vanishes in probability,

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

We have that for any $\epsilon > 0$

$$\mathbb{P}[|R_n| > \epsilon] = \mathbb{P}[|R_n| > \epsilon, |\hat{\theta}_n - \theta| > \delta] + \mathbb{P}[|R_n| > \epsilon, |\hat{\theta}_n - \theta| \leq \delta]$$

and

$$\mathbb{P}[|R_n| > \epsilon, |\hat{\theta}_n - \theta| > \delta] \leq \mathbb{P}[|\hat{\theta}_n - \theta| > \delta] \xrightarrow{p} 0.$$

If $|\hat{\theta}_n - \theta| < \delta$, (A6) gives $|R_n| \leq \frac{\delta}{2n}\sum_{i=1}^{n}M(X_i)$.

Since the weak law of large numbers implies that

$$\frac{1}{n}\sum_{i=1}^{n}M(X_i) \xrightarrow{p} \mathbb{E}[M(X_1)] < \infty,$$

the quantity $\mathbb{P}[|R_n| > \epsilon, |\hat{\theta}_n - \theta| \leq \delta]$ can be made arbitrarily small (for large n) by taking δ sufficiently small. Thus, $R_n \xrightarrow{p} 0$ and applying Slutsky's theorem we may conclude that

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

□

- Notice that for our proof, we assumed that the sequence of MLE's was consistent.
- Proving consistency of an MLE can be subtle

Consistency of the MLE

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.

- The latter is minimized when $t = \theta$ and so $\phi(t)$ is maximized at $t = \theta$.
- 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 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 $M > 0$, $\sup_{|t|\le M} |\phi_n(t) - \phi(t)| \xrightarrow{p} 0$*
2. *T_n maximizes $\phi_n(t)$ and T_0 is the unique maximizer of $\phi(t)$*
3. *$\forall \epsilon > 0$, there exists M_ϵ such that $\mathbb{P}[|T_n| > M_\epsilon] < \epsilon\ \forall n$*

Then, $T_n \xrightarrow{p} T_0$

If ϕ_n are concave, can 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 may define $u = c(t)$ and consider

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

with $d_0(u) = d(c^{-1}(u))$. It follows that ϕ_n^* is a concave function, since its second derivative is, $(\phi_n^*)''(u) = -d_0''(u)$, which is negative ($d_0''(u) = \text{Var}\, T(X_i)$).

Example (Exponential Families)

Now, by the weak law of large numbers, for each u, we have

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

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

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

we must have that ϕ^* is maximized when

$$d_0'(u) = d_0'(c(\theta))$$

The condition holds is if we set $u = c(\theta)$, so $c(\theta)$ is a maximizer of ϕ^*. By concavity, it is the unique maximizer.
It follows from our theorem that if $\hat{u}_n = c(\hat{\theta}_n)$ then $\hat{u}_n = c(\hat{\theta}_n) \xrightarrow{p} c(\theta)$.
But c is 1-1 and continuous, so the continuous mapping theorem implues

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

More on Maximum Likelihood Estimation

Statistical Theory

Victor Panaretos
Ecole Polytechnique Fédérale de Lausanne

Maximum Likelihood Estimators

Recall our definition of a maximum likelihood estimator:

Definition (Maximum Likelihood Estimators)

Let $\mathbf{X} = (X_1, ..., X_n)$ 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 of* θ.

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 when this is real.

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

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,

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

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.
- Hints that extrema of $L(\theta;\mathbf{X})$ 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.

Notice that

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

By the WLLN,

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

But we have seen that the KL-divergence is zero only at θ_0 and positive 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 (frequencies) and let $X_1,...,X_n$ be iid random variables each having density $f(x;\theta_0)$. Assume that the 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, given any $\epsilon > 0$, with probability tending to 1 as $n\to\infty$, the likelihood equation

$$\frac{\partial}{\partial\theta}\ell(\theta;X_1,...,X_n) = 0$$

has a root $\hat{\theta}_n(X_1,...,X_n)$ such that $|\hat{\theta}_n(X_1,...,X_n)-\theta_0| < \epsilon$.

- Does not tell us *which* root to choose, so not useful in practice
- Actually the consistent sequence is *essentially* unique

Corollary (Consistency of Unique Solutions)

Under the assumptions of the previous theorem, if the likelihood equation has a unique root δ_n for each n and all $\mathbf{x}$, then δ_n is a consistent sequence of estimators 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 $n\to\infty$.
- Notice that 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

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

Lemma

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

- 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 require that the estimator used is consistent
  - $\hookrightarrow$ Often the case with Plug-In or MoM estimators
- Very often, the roots will not be available in closed form. In these cases, an iterative approach will be required to approximate the roots

The Newton-Raphson Algorithm

We wish to solve the equation

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

Supposing that $\tilde{\theta}$ is close to a root (perhaps is a consistent estimator),

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

By using a second-order Taylor expansion. This suggests

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

The procedure can then be iterated by replacing $\tilde{\theta}$ by the right hand side of the above relation.
$\rightarrow$ Many issues regarding convergence, speed of convergence, etc... (numerical analysis course)

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. Then, the sequence of estimators

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

satisfies

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

- Therefore, with a single Newton-Raphson step, we may obtain an estimator that, asymptotically, behaves like a consistent MLE.
  $\hookrightarrow$ Provided that we have a $\sqrt{n}$—consistent estimator!
- The "one-step" estimator does not necessarily behave like an MLE for finite n!

Proof.

We Taylor expand around the true value, θ_0,

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

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

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

Exercise.

Use the central limit theorem and the law of large numbers to complete the proof. □

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:

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

The Multiparameter Case

Regularity Conditions

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

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

(B3) All mixed partial derivatives of ℓ w.r.t. θ up to degree 3 exist and are continuous.

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

(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 $1 \leq j, k, l \leq p$,

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

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

- The interpretation of the conditions is the same as with the one-dimensional case

The Multiparameter Case

Theorem (Asymptotic Normality of the MLE)

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

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

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

Misspecification of Models

- Statistical models are typically mere 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 $X_1, ..., X_n$ be iid Exponential(λ) r.v.'s but we have modelled them as having the following two parameter density

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

with α and θ positive unknown parameters to be estimated.

Example (cont'd)

- Notice that the exponential distribution is not a member of this parametric family.
- However, letting $\alpha, \theta \to \infty$ at rates 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 $\hat{\lambda}$ will yield a density "close" to the true density.

Example

Let $X_1, ..., X_n$ 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

- $X_1, ..., X_n$ are iid r.v.'s with distribution F
- We have assumed that the X_i admit a density in $\{f(x;\theta)\}_{\theta\in\Theta}$.
- The true distribution F 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 w.r.t. $f(\cdot|\theta)$.

- 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 equations $\sum_{i=1}^{n} \ell'(X_i; \theta) = 0$ for θ in the open set Θ. If

(a) *ℓ' is a strictly monotone function on Θ for each x*

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

$$\hat{\theta}_n - \theta(F) > \epsilon \implies \hat{\theta}_n > \theta(F) + \epsilon \implies \frac{1}{n}\sum_{i=1}^{n} \ell'(X_i; \theta(F) + \epsilon) > 0$$

because $\hat{\theta}_n$ is the solution to the equation $\frac{1}{n}\sum_{i=1}^{n} \ell'(X_i; \theta) = 0$. Similarly we also obtain

$$\theta(F) - \hat{\theta}_n > \epsilon \implies \theta(F) - \epsilon > \hat{\theta}_n \implies \frac{1}{n}\sum_{i=1}^{n} \ell'(X_i; \theta(F) - \epsilon) < 0.$$

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

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \ell'(X_i;\theta(F)+\epsilon) > 0\right] = \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \ell'(X_i;\theta(F)+\epsilon) - \int_{-\infty}^{\infty}\ell'(x;\theta(F)+\epsilon)dF(x) > -\int_{-\infty}^{\infty}\ell'(x;\theta(F)+\epsilon)dF(x)\right].$$

This converges to zero because the monotonicity assumption implies that $-\int_{-\infty}^{\infty}\ell'(x;\theta(F)+\epsilon)dF(x) > 0$ and the law of large numbers implies that

$$\frac{1}{n}\sum_{i=1}^n \ell'(X_i;\theta(F)+\epsilon) \xrightarrow{p} \int_{-\infty}^{\infty}\ell'(x;\theta(F))dF(x).$$

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

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

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
- Indicated the subtleties that are involved when proving convergence for indirectly defined estimators

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:

$$\int_{-\infty}^{+\infty}\frac{d}{d\theta}\log f(x;\theta)dF(x) = 0 \iff \frac{d}{d\theta}\int_{-\infty}^{+\infty}\log f(x;\theta)dF(x) = 0$$

$$\iff \frac{d}{d\theta}\left[\int_{-\infty}^{+\infty}\log f(x;\theta)dF(x) - \int_{-\infty}^{+\infty}\log g(x)dF(x)\right] = 0$$

$$\iff \frac{d}{d\theta}KL(f(x;\theta)\|g(x)) = 0$$

- When the equation is assumed to have a unique solution, then this is to be though as a minimum of the KL-distance
- Hence we may intuitively think of the $\theta(F)$ as the element of Θ for which f_θ is "closest" to F in the KL-sense.

The Decision Theory Framework

Statistical Theory

Victor Panaretos
Ecole Polytechnique Fédérale de Lausanne

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 (frequencies). This is the variant of the game we decide to play.
- A *parameter space* $\Theta \subseteq \mathbb{R}^p$ which parametrizes the family $\mathcal{F} = \{F_\theta\}_{\theta\in\Theta}$. 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 $\delta : \mathcal{X} \to \mathcal{A}$. These represent the possible strategies available to the statistician.

Statistics as a Random Game?

How the game is played:

- First we agree on the rules:
  1. Fix a parametric family $\{F_\theta\}_{\theta\in\Theta}$
  2. Fix an action space $\mathcal{A}$
  3. Fix a loss function $\mathcal{L}$
- Then we play:
  1. Nature selects (plays) $\theta_0 \in \Theta$.
  2. The statistician observes $\mathbf{X} \sim F_{\theta_0}$
  3. The statistician plays $\alpha \in \mathcal{A}$ in response.
  4. The statistician has to pay nature $\mathcal{L}(\theta_0, \alpha)$.

Framework proposed by A. Wald in 1939. Encompasses three basic statistical problems:

- Point estimation
- Hypothesis testing
- Interval estimation

Point Estimation as a Game

In the problem of point estimation we have:

1. Fixed parametric family $\{F_\theta\}_{\theta\in\Theta}$
2. Fixed an action space $\mathcal{A} = \Theta$
3. 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 $\mathbf{X} \sim F_{\theta_0}$
3. The statistician plays $\delta(\mathbf{X}) \in \mathcal{A} = \Theta$
4. The statistician loses $\mathcal{L}(\theta_0, \delta(\mathbf{X}))$

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

Statistician would like to pick strategy δ so as to minimize his losses.
But losses are random, as they depend on $\mathbf{X}$.

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: $R(\theta, \delta) = \mathbb{E}_\theta [\mathcal{L}(\theta, \delta(\mathbf{X}))]$.

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

$$MSE(\delta(\mathbf{X})) = \mathbb{E}_\theta[\|\theta - \delta(\mathbf{X})\|^2]$$

is the risk corresponding to a squared error loss function.

Coin Tossing Revisited

Consider the "coin tossing game" with quadratic loss:

- Nature picks $\theta \in [0, 1]$
- We observe n variables $X_i \overset{iid}{\sim}$ Bernoulli(θ).
- Action space is $\mathcal{A} = [0, 1]$
- Loss function is $\mathcal{L}(\theta, \alpha) = (\theta - \alpha)^2$.

Consider 3 different decision procedures $\{\delta_j\}_{j=1}^3$:

1. $\delta_1(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n X_i$
2. $\delta_2(\mathbf{X}) = X_1$
3. $\delta_3(\mathbf{X}) = \frac{1}{2}$

Let us compare these using their associated risks as benchmarks.

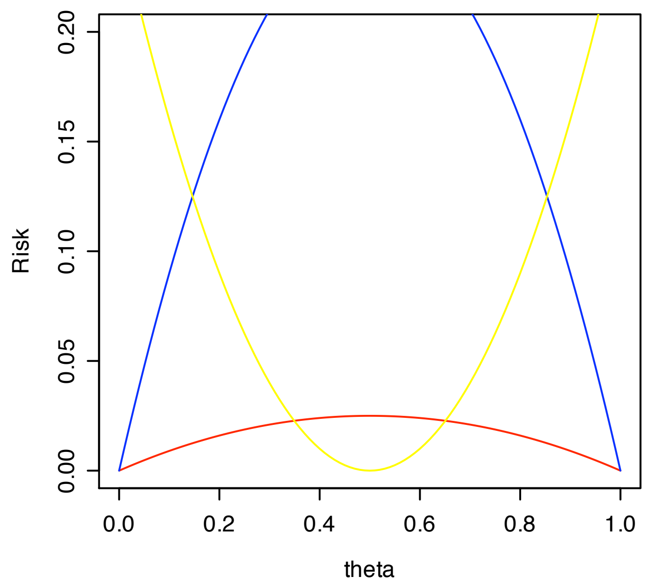
Coin Tossing Revisited

Risks associated with different decision rules:

$$R_j(\theta) = R(\theta, \delta_j(\mathbf{X})) = \mathbb{E}_\theta[(\theta - \delta_j(\mathbf{X}))^2]$$

- $R_1(\theta) = \frac{1}{n}\theta(1-\theta)$
- $R_2(\theta) = \theta(1-\theta)$
- $R_3(\theta) = \left(\theta - \frac{1}{2}\right)^2$

Coin Tossing Revisited



$R_1(\theta)$, $R_2(\theta)$, $R_3(\theta)$

Risk of a Decision Rule

Saw that decision rule may strictly *dominate* another rule ($R_2(\theta) > R_1(\theta)$).

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.

$$R(\theta,\delta^*) \leq R(\theta,\delta),\ \forall\theta\in\Theta \quad \& \quad \exists\ \theta'\in\Theta : R(\theta',\delta^*) < R(\theta',\delta),$$

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 or different parameter space) then domination may break down.

Risk of a Decision Rule

Example (Exponential Distribution)

Let $X_1, ..., X_n \stackrel{iid}{\sim}$ Exponential(λ), $n \geq 2$. The MLE of λ is

$$\hat{\lambda} = \frac{1}{\bar{X}}$$

with $\bar{X}$ the empirical mean. Observe that

$$\mathbb{E}_\lambda[\hat{\lambda}] = \frac{n\lambda}{n-1}.$$

It follows that $\tilde{\lambda} = (n-1)\hat{\lambda}/n$ is an unbiased estimator of λ. Observe now that

$$MSE_\lambda(\tilde{\lambda}) < MSE_\lambda(\hat{\lambda})$$

since $\tilde{\lambda}$ is unbiased and $\mathrm{Var}_\lambda(\tilde{\lambda}) < \mathrm{Var}_\lambda(\hat{\lambda})$. Hence the MLE is an inadmissible rule for quadratic loss.

Risk of a Decision Rule

Example (Exponential Distribution)

Notice that the parameter space in this example is $(0,\infty)$. In such cases, quadratic loss tends to penalize over-estimation more heavily than under-estimation (the maximum possible under-estimation is bounded!). Considering a different loss function gives the opposite result! Let

$$\mathcal{L}(a,b) = \frac{a}{b} - 1 - \log(a/b)$$

where, for each fixed a, $lim_{b\to 0}\mathcal{L}(a,b) = lim_{b\to\infty}\mathcal{L}(a,b) = \infty$. Now,

$$\begin{aligned} R(\lambda,\tilde{\lambda}) &= \mathbb{E}_\lambda\left[\frac{n\lambda\bar{X}}{n-1} - 1 - \log\left(\frac{n\lambda\bar{X}}{n-1}\right)\right] \\ &= \mathbb{E}_\lambda\left[\lambda\bar{X} - 1 - \log(\lambda\bar{X})\right] + \frac{\mathbb{E}_\lambda(\lambda\bar{X})}{n-1} - \log\left(\frac{n}{n-1}\right) \\ &> \mathbb{E}_\lambda\left[\lambda\bar{X} - 1 - \log(\lambda\bar{X})\right] = R(\lambda,\hat{\lambda}). \end{aligned}$$

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.
- 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 will be no decision rule that makes risk uniformly smallest over θ
- Narrow down class of possible decision rules by unbiasedness/symmetry/... considerations, and try to find *uniformly dominating* rules of all other rules (next week!).

Minimax Decision Rules

- Another approach to good procedures is to use global rather than local criteria (with respect to θ).

Rather than look at risk at every θ $\leftrightarrow$ 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.

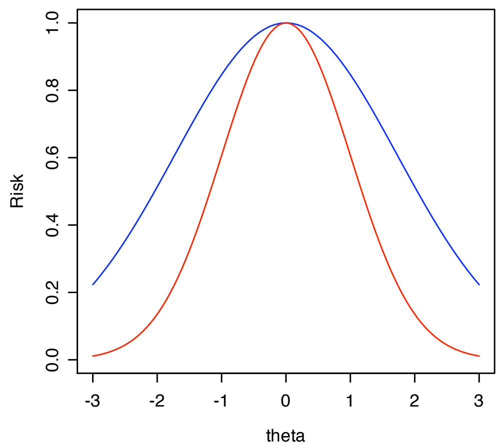
Minimax Decision Rules

A few comments on minimaxity:

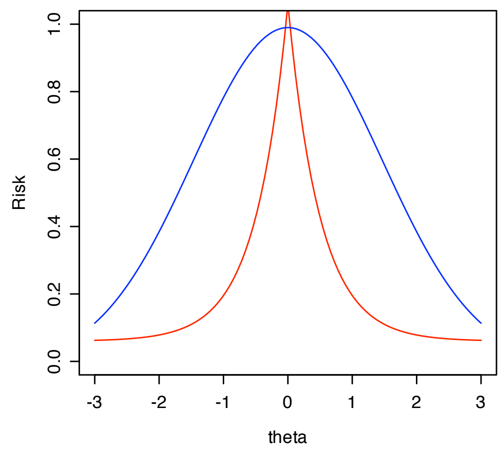
- Motivated as follows: we do not know anything about θ so let us insure ourselves against the worst thing that can happen.
- Makes sense if you are in a zero-sum game: if your opponent chooses θ to maximize $\mathcal{L}$ then one should look for minimax rules. But is nature really an opponent?
- If 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.

Minimax Decision Rules

Inadmissible minimax rule



Counterintuitive minimax rule



Bayes Decision Rules

- Wanted to compare decision procedures using global rather than local criteria (with respect to θ).
- We arrived at the minimax principle by assuming we have no idea about the true value of θ.
- Suppose we have some prior belief about the value of θ. How can this be factored in our risk-based considerations?

Rather than look at risk at every θ $\leftrightarrow$ Concentrate on average risk

Definition (Bayes Risk)

Let $\pi(\theta)$ be a probability density (frequency) 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

$$r(\pi,\delta) = \int_\Theta R(\theta,\delta)\pi(\theta)d\theta = \int_\Theta \int_\mathcal{X} \mathcal{L}(\theta,\delta(\mathbf{x}))F_\theta[d\mathbf{x}]\pi(\theta)d\theta$$

The prior $\pi(\theta)$ places different emphasis for different values of θ based on our prior belief/knowedge.

Bayes Decision Rules

- Bayes principle: a decision rule is *preferable* to another if it has 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 (frequency) on Θ. If $\delta \in \mathcal{D}$ is such that

$$r(\pi,\delta) \le r(\pi,\delta') \quad \forall\ \delta' \in \mathcal{D},$$

then δ is called a *Bayes decision rule* with respect to π.

- The minimax principle aims to minimize the maximum risk.
- The Bayes principle aims to minimize the average risk
- Sometime no Bayes rule exist becaise 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 $\Theta = \{\theta_1, ..., \theta_t\}$ 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

$$\begin{aligned} R(\theta_j,\kappa) &\le R(\theta_j,\delta), \quad \forall j \\ R(\theta_j,\kappa)\pi(\theta_j) &\le R(\theta_j,\delta)\pi(\theta_j), \quad \forall \theta\in\Theta \\ \sum_j R(\theta_j,\kappa)\pi(\theta_j) &< \sum_j R(\theta,\delta)\pi(\theta_j) \end{aligned}$$

which is a contradiction (strict inequality follows by strict domination and the fact that $\pi(\theta_j)$ is always positive). □

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,

$$\int_\Theta R(\theta,\kappa)\pi(\theta)d\theta \le \int_\Theta R(\theta,\delta)\pi(\theta)d\theta.$$

as a result of strict domination and by $\pi(\theta)$ being non-negative. This implies that κ either improves upon δ, or κ is a Bayes rule. Either possibility contradicts our assumption. □

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 must exist an $\epsilon > 0$ such that $R(\theta, \kappa) < R(\theta, \delta)$ for all theta satisfying $\|\theta - \theta_0\| < \epsilon$. It follows that Θ_0 is open and hence, by our assumption, $\pi[\Theta_0] > 0$. Therefore, it must be that

$$\int_{\Theta_0} R(\theta, \kappa)\pi(\theta)d\theta < \int_{\Theta_0} R(\theta, \delta)\pi(\theta)d\theta$$

Observe now that

$$\begin{aligned} r(\pi, \kappa) &= \int_{\Theta} R(\theta, \kappa)\pi(\theta)d\theta \\ &= \int_{\Theta_0} R(\theta, \kappa)\pi(\theta)d\theta + \int_{\Theta_0^c} R(\theta, \kappa)\pi(\theta)d\theta \\ &< \int_{\Theta_0} R(\theta, \delta)\pi(\theta)d\theta + \int_{\Theta_0^c} R(\theta, \delta)\pi(\theta)d\theta \\ &= r(\pi, \delta), \end{aligned}$$

since $\int_{\Theta_0^c} R(\theta, \kappa)\pi(\theta)d\theta \le \int_{\Theta_0^c} R(\theta, \delta)\pi(\theta)d\theta$, while we have strict inequality on Θ_0, contradicting 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.

Randomised Decision Rules

Given

- decision rules $\delta_1, ..., \delta_k$
- probabilities $\pi_i \ge 0$, $\sum_{i=1}^k p_i = 1$

we may define a new decision rule

$$\delta_* = \sum_{i=1}^k p_i \delta_i$$

called a *randomised decision rule*. Interpretation:

Given data $\mathbf{X}$, choose a δ_i randomly according to p but independent of $\mathbf{X}$. If δ_j is the outcome $(1 \le j \le k)$, then take action $\delta_j(\mathbf{X})$.

$\rightarrow$ Risk of δ_* is average risk: $R(\theta, \delta_*) = \sum_{i=1}^k p_i R(\theta, \delta_i)$

- Appears artificial but often minimax rules are randomised
- Examples of randomised rules with $\sup_\theta R(\theta, \delta_*) < \sup_\theta R(\theta, \delta_i) \forall i$

Minimum Variance Unbiased Estimation

Statistical Theory

Victor Panaretos
Ecole Polytechnique Fédérale de Lausanne

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 $\mathcal{F} = \{F_\theta\}_{\theta\in\Theta}$.
- 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 $\mathcal{A} \equiv \Theta$
- A *loss function* $\mathcal{L} : \Theta \times \mathcal{A} \to \mathbb{R}^+$. This 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.

Performance of decision rules was to be judged by the risk they induce:

$$R(\theta, \delta) = \mathbb{E}_\theta[\mathcal{L}(\theta, \delta(\mathbf{X}))], \quad \boldsymbol{\theta} \in \Theta, X \sim F_\theta, \delta \in \mathcal{D}$$

Optimality in Point Estimation

An optimal decision rule would be one that uniformly minimizes 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 risk behaviour asymptotically** $(n \to \infty)$
  $\hookrightarrow$ e.g. Asymptotic Relative Efficiency.

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 $(x_1, \dots, x_n)$ be a realization of $\mathbf{X} \sim F_\theta$ which is available to us
  3. Estimate the value of θ that generated the sample given $(x_1, \dots, x_n)$

- **Focus on Quadratic Loss**

  Error incurred when estimating θ by $\hat{\theta} = \delta(\mathbf{X})$ is

  $$\mathcal{L}(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$$

  giving MSE as risk $R(\theta, \hat{\theta}) = \mathbb{E}_\theta\|\theta - \hat{\theta}\|^2 = \text{Variance} + \text{Bias}^2$.

- **RESTRICT class of estimators (=decision rules)**

  Consider ONLY *ubiased* estimators: $\mathcal{D} := \{\delta : \mathcal{X} \to \Theta | \mathbb{E}_\theta[\delta(\mathbf{X})] = \theta\}$.

Comments on Unbiasedness

- Unbiasedness requirement is one means of reducing the class of rules/estimators we are considering
  - $\hookrightarrow$ Other requirements could be invariance or equivariance, e.g.

  $$\delta(\mathbf{X} + \mathbf{c}) = \delta(\mathbf{X}) + \mathbf{c}$$

- Risk reduces to variance since bias is zero.
- Not necessarily a sensible requirement
  - $\hookrightarrow$ e.g. violates "likelihood principle"
- Unbiased Estimators may not exist in a particular problem
- Unbiased Estimators may be silly for a particular problem
- However unbiasedness can be a reasonable/natural requirement in a wide class of point estimation problems.
- Unbiasedness can be defined for more general loss functions, but not as conceptually clear (and with tractable theory) as for quadratic loss.
  $\hookrightarrow$ δ is unbiased under $\mathcal{L}$ if $\mathbb{E}_\theta[\mathcal{L}(\theta', \delta)] \geq \mathbb{E}_\theta[\mathcal{L}(\theta, \delta)] \quad \forall\, \theta, \theta' \in \Theta$.

Comments on Unbiasedness

Example (Unbiased Estimators Need not Exist)

Let $X \sim$Binomial(n, θ), with θ unknown but n known. We wish to estimate

$$\psi = \sin\theta$$

We require that our estimator $\delta(X)$ be unbiased, $\mathbb{E}_\theta[\delta] = \psi = \sin\theta$. Such an estimator satisfies

$$\sum_{x=0}^{n} \delta(x) \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad = \quad \sin\theta$$

but this cannot hold for all θ, since the sine function cannot be represented as a finite polynomial.

The class of unbiased estimators in this case is empty.

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

$$\sum_{x=0}^{\infty} \delta(x) \frac{\lambda^x}{x!} e^{-\lambda} = e^{-2\lambda} \implies \sum_{x=0}^{\infty} \delta(x) \frac{\lambda^x}{x!} \quad = \quad e^{-\lambda}$$

$$\implies \sum_{x=0}^{\infty} \delta(x) \frac{\lambda^x}{x!} \quad = \quad \sum_{x=0}^{\infty} (-1)^x \frac{\lambda^x}{x!}$$

so that $\delta(X) = (-1)^X$ is the only unbiased estimator of ψ.

But $0 < \psi < 1$ for $\lambda > 0$, so this is clearly a silly estimator

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

$$\hat{\mu} = X_{(1)} - \frac{1}{n} \quad \& \quad \tilde{\mu} = \bar{X} - 1.$$

In fact, $t\hat{\mu} + (1-t)\tilde{\mu}$ is unbiased for any t. Simple calculations reveal

$$R(\mu,\hat{\mu}) = \mathrm{Var}(\hat{\mu}) = \frac{1}{n^2} \quad \& \quad R(\mu,\tilde{\mu}) = \mathrm{Var}(\tilde{\mu}) = \frac{1}{n}$$

so that $\hat{\mu}$ dominates $\tilde{\mu}$. Will it dominate any other unbiased estimator? (note that $\hat{\mu}$ depends only on the one-dimensional sufficient statistic $X_{(1)}$)

Theorem (Rao-Blackwell Theorem)

Let $\mathbf{X}$ be distributed according to a distribution depending on an unknown parameter θ and let T be a sufficient statistic for θ. Let δ be decision rule such that

1. $\mathbb{E}_\theta[\delta(\mathbf{X})] = g(\theta)$ *for all* θ
2. $Var_\theta(\delta(\mathbf{X})) < \infty$, *for all* θ.

Then $\delta^* := \mathbb{E}[\delta|T]$ is an unbiased estimator of $g(\theta)$ that dominates δ, i.e.

1. $\mathbb{E}_\theta[\delta^*(\mathbf{X})] = g(\theta)$ *for all* θ.
2. $Var_\theta(\delta^*(\mathbf{X})) \leq Var_\theta(\delta(\mathbf{X}))$ *for all* θ.

Moreover, inequality is replaced by equality if and only if $\mathbb{P}_\theta[\delta^* = \delta] = 1$.

- The theorem indicates that any candidate minimum variance unbiased estimator should be a functions of the sufficient statistic.
- Intuitively, an estimator that takes into account aspects of the sample that are irrelevant with respect to θ, can always be improved

Proof.

Since T is sufficient for θ, $\mathbb{E}[\delta|T=t] = h(t)$ is independent of θ, so that δ^* is well-defined as a statistic (depends only on $\mathbf{X}$). Then,

$$\mathbb{E}_\theta[\delta^*(\mathbf{X})] = \mathbb{E}_\theta[\mathbb{E}[\delta(\mathbf{X})|T(\mathbf{X})]] = \mathbb{E}_\theta[\delta(\mathbf{X})] = g(\theta).$$

Furthermore, we have

$$\begin{aligned}\mathrm{Var}_\theta(\delta) = \mathrm{Var}_\theta[\mathbb{E}(\delta|T)] + \mathbb{E}_\theta[\mathrm{Var}(\delta|T)] &\geq \mathrm{Var}_\theta[\mathbb{E}(\delta|T)] \\ &= \mathrm{Var}_\theta(\delta^*)\end{aligned}$$

In addition, note that

$$\mathrm{Var}(\delta|T) := \mathbb{E}[(\delta - \mathbb{E}[\delta|T])^2|T] = \mathbb{E}[(\delta - \delta^*)^2|T]$$

so that $\mathbb{E}_\theta[\mathrm{Var}(\delta|T)] = \mathbb{E}_\theta(\delta - \delta^*)^2 > 0$ unless if $\mathbb{P}_\theta(\delta^* = \delta) = 1$. $\square$

Exercise

Show that $\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|\mathbf{X})] + \mathrm{Var}[\mathbb{E}(Y|\mathbf{X})]$ when $\mathrm{Var}(Y) < \infty$.

- 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 is any function of a sufficient statistic admissible? (provided that it is unbiased)

Suppose that δ is an unbiased estimator of $g(\theta)$ and T, S are θ-sufficient.

- What is the relationship between $\mathrm{Var}_\theta(\underbrace{\mathbb{E}[\delta|T]}_{\delta_T^*}) \overset{?}{\gtreqless} \mathrm{Var}_\theta(\underbrace{\mathbb{E}[\delta|S]}_{\delta_S^*})$
- Intuition suggests that whichever of T, S carries the least irrelevant information (in addition to the relevant information) should "win"
  - $\hookrightarrow$ More formally, if $T = h(S)$ then we should expect that δ_T^* dominate δ_S^*.

Unbiasedness and Sufficiency

Proposition

Let δ be an unbiased estimator of $g(\theta)$ and for T, S two θ-sufficient statistics define

$$\delta_T^* := \mathbb{E}[\delta|T] \quad \& \quad \delta_S^* := \mathbb{E}[\delta|S].$$

Then, the following implication holds

$$T = h(S) \implies \mathrm{Var}_\theta(\delta_T^*) \leq \mathrm{Var}_\theta(\delta_S^*)$$

1. Essentially this means that the best possible "Rao-Blackwellization" is achieved by conditioning on a minimal sufficient statistic.
2. This does not necessarily imply that for T minimally sufficient and δ unbiased, $\mathbb{E}[\delta|T]$ has minimum variance.
   - $\hookrightarrow$ In fact it does not even imply that $\mathbb{E}[\delta|T]$ is admissible.

Proof.

Recall the *tower property* of conditional expectation: if $Y = f(X)$, then

$$\mathbb{E}[Z|Y] = \mathbb{E}\{\mathbb{E}(Z|X)|Y\}.$$

Since $T = f(S)$ we have

$$\begin{aligned} \delta_T^* &= \mathbb{E}[\delta|T] \\ &= \mathbb{E}[\mathbb{E}(\delta|S)|T] \\ &= \mathbb{E}[\delta_S^*|T] \end{aligned}$$

The conclusion now follows from the Rao-Blackwell theorem. □

A mathematical remark

To better understand the tower property intuitively, recall that $\mathbb{E}[Z|Y]$ is the minimizer of $\mathbb{E}[(Z-\varphi(Y))^2]$ over all (measurable) functions φ of Y. You can combine that with the fact that $\sqrt{\mathbb{E}[(X-Y)^2]}$ defines a Hilbert norm on random variables with finite variance to get geometric intuition.

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 $Var_\theta(\delta) < \infty$, $\forall\theta \in \Theta$. If $\delta^* := \mathbb{E}[\delta|T]$ and V is any other unbiased estimator of $g(\theta)$, then

1. $Var_\theta(\delta^*) \leq Var_\theta(V)$, $\forall\theta \in \Theta$
2. $Var_\theta(\delta^*) = Var_\theta(V) \implies \mathbb{P}_\theta[\delta^* = V] = 1$.

That is, $\delta^* := \mathbb{E}[\delta|T]$ is the unique Uniformly Minimum Variance Unbiased Estimator of $g(\theta)$.

- The theorem says that if a complete sufficient statistic T exists, then the MVUE of $g(\theta)$ (if it exists) must be a function of T.
- Moreover it establishes that whenever $\exists$ 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.

To prove (1) we go through the following steps:

- Take V to be any unbiased estimator with finite variance.
- Define its "Rao-Blackwellized" version $V^* := \mathbb{E}[V|T]$
- By unbiasedness of both estimators,

$$0 = \mathbb{E}_\theta[V^* - \delta^*] = \mathbb{E}_\theta[\mathbb{E}[V|T] - \mathbb{E}[\delta|T]] = \mathbb{E}_\theta[h(T)], \quad \forall\theta \in \Theta.$$

- By completeness of T we conclude $\mathbb{P}_\theta[h(T) = 0] = 1$ for all θ.
- In other words, $\mathbb{P}_\theta[V^* = \delta^*] = 1$ for all θ.
- But V^* dominates V by the Rao-Blackwell theorem.
- Hence $\mathrm{Var}_\theta(\delta^*) = \mathrm{Var}_\theta(V^*) \leq \mathrm{Var}_\theta(V)$.

For part (2) (the uniqueness part) notice that from our reasoning above

- $\mathrm{Var}_\theta(V) = \mathrm{Var}_\theta(\delta^*) \implies \mathrm{Var}_\theta(V) = \mathrm{Var}_\theta(V^*)$
- But Rao-Blackwell theorem says $\mathrm{Var}_\theta(V) = \mathrm{Var}_\theta(V^*) \iff \mathbb{P}_\theta[V = V^*] = 1$.

□

Completeness, Sufficiency, Unbiasedness, and Optimality

Taken together, the Rao-Blackwell and Lehmann-Scheffé theorems also suggest approaches to finding UMVUE estimators 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 $g(\theta)$.
   - ↪ 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 may obtain the UMVUE by "Rao-Blackwellizing" with respect to the complete sufficient statistic:

Example (Bernoulli Trials)

Let $X_1, ..., X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$. What is the UMVUE of θ^2?

- By the Neyman factorization theorem $T = X_1 + \ldots + X_n$ is sufficient,
- Since the distribution of $(X_1, ..., X_n)$ is a 1-parameter exponential family, T is also complete.

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 $\delta = \mathbf{1}\{X_1 + X_2 = 2\}$ and note that this 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

$$\begin{aligned}
\mathbb{E}[S|T=t] &= \mathbb{P}[X_1 + X_2 = 2|T=t] \\
&= \frac{\mathbb{P}_\theta[X_1 + X_2 = 2, X_3 + \ldots + X_n = t-2]}{\mathbb{P}_\theta[T=t]} \\
&= \left\{\begin{matrix} 0 & \text{if } t \le 1 \\ \binom{n-2}{t-2}/\binom{n}{t} & \text{if } t \ge 2 \end{matrix}\right\} = \frac{t(t-1)}{n(n-1)}.
\end{aligned}$$

Variance Lower Bounds for Unbiased Estimators

- Often → minimal sufficient statistic exists but is not complete.
  - ↪ Cannot appeal to the Lehmann-Scheffé theorem in search of a UMVUE.
- However, if we could establish a *lower bound* for the variance as a function of θ, than an estimator achieving this bound will be the unique UMVUE.

The Aim

For iid $X_1, ..., X_n$ with density (frequency) depending on θ unknown, we want to establish conditions under which

$$Var_\theta[\delta] \ge \phi(\theta), \quad \forall\theta$$

for any unbiased estimator δ. We also wish to determine $\phi(\theta)$.

Let's take a closer look at this...

Cuachy-Schwarz Bounds

Theorem (Cauchy-Schwarz Inequality)

Let U, V be random variables with finite variance. Then,

$$Cov(U,V) \le \sqrt{Var(U)Var(V)}$$

The theorems yields an immediate lower bound for the variance of an unbiased estimator δ_0:

$$\text{Var}_\theta(\delta_0) \ge \frac{\text{Cov}^2_\theta(\delta_0, U)}{\text{Var}_\theta(U)}$$

which is valid for any random variable U with $\text{Var}_\theta(U) < \infty$ for all θ.

- The bound can be made tight be choosing a suitable U.
- However this is still not very useful as it falls short of our aim
  - The bound will be specific to δ_0, while we want a bound that holds for any unbiased estimator δ.
- Is there a smart choice of U for which $\text{Cov}_\theta(\delta_0, U)$ depends on $g(\theta) = \mathbb{E}_\theta(\delta_0)$ only? (and so is not specific to δ_0)

Optimizing the Cauchy-Schwartz Bound

Assume that θ is real and the following regularity conditions hold

Regularity Conditions

(C1) The support of $A := \{\mathbf{x} : f(\mathbf{x};\theta) > 0\}$ is independent of θ

(C2) $f(\mathbf{x};\theta)$ is differentiable w.r.t. θ, $\forall \theta \in \Theta$

(C3) $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right] = 0$

(C4) For a statistic $T = T(\mathbf{X})$ with $\mathbb{E}_\theta|T| < \infty$ and $g(\theta) = \mathbb{E}_\theta T$ differentiable,

$$g'(\theta) = \mathbb{E}_\theta\left[T\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right], \quad \forall\theta$$

To make sense of (C3) and (C4), suppose that $f(\cdot;\theta)$ is a density. Then

$$\frac{d}{d\theta}\int S(\mathbf{x})f(\mathbf{x};\theta)d\mathbf{x} \overset{!}{=} \int S(\mathbf{x})\frac{f(x;\theta)}{f(x;\theta)}\frac{d}{d\theta}f(\mathbf{x};\theta)dx = \int S(\mathbf{x})f(\mathbf{x};\theta)\frac{d}{d\theta}\log f(\mathbf{x};$$

provided integration/differentiation can be interchanged.

The Cramér-Rao Lower Bound

Theorem

Let $\mathbf{X} = (X_1, \ldots, X_n)$ have joint density (frequency) $f(\mathbf{x};\theta)$ satisfying conditions (C1), (C2) and (C3). If the statistic T satisfies condition (C4), then

$$Var_\theta(T) \geq \frac{[g'(\theta)]^2}{I(\theta)}$$

with $I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right)^2\right]$

Proof.

By the Cauchy-Schartz inequality with $U = \frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)$,

$$\mathrm{Var}_\theta(T) \geq \frac{\mathrm{Cov}^2_\theta\left(T, \frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right)}{\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right)}$$

Since $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right] = 0$ we have $\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right) = I(\theta)$.

The Cramér-Rao Lower Bound

Also, observe that

$$\begin{aligned}
\mathrm{Cov}_\theta\left(T, \frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right) &= \mathbb{E}_\theta\left[T\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right] \\
&\quad -\mathbb{E}_\theta[T]\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right] \\
&= \mathbb{E}_\theta\left[T\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right] \\
&= \frac{d}{d\theta}\mathbb{E}_\theta[T] \\
&= g'(\theta)
\end{aligned}$$

which completes the proof. □

The Cramér-Rao Lower Bound

When is the Cramér-Rao lower bound achieved?

$$\text{if} \quad \mathrm{Var}_\theta[T] = \frac{[g'(\theta)]^2}{I(\theta)}$$

$$\text{then} \quad \mathrm{Var}_\theta[T] = \frac{\mathrm{Cov}^2_\theta\left[T, \frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right]}{\mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right]}$$

which occurs if and only if $\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)$ is a linear function of T (correlation 1). That is, w.p.1:

$$\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta) = A(\theta)T(\mathbf{x}) + B(\theta)$$

Solving this differential equation yields, for all $\mathbf{x}$,

$$\log f(\mathbf{x};\theta) = A^*T(\mathbf{x}) + B^*(\theta) + S(\mathbf{x})$$

so that $\mathrm{Var}_\theta(T)$ attains the lower bound if and only if the density (frequency) of $\mathbf{X}$ is a one-parameter exponential family as above

Testing Statistical Hypotheses

Statistical Theory

Victor Panaretos
Ecole Polytechnique Fédérale de Lausanne

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Using Data to Evaluate Theories/Assertions

- Scientific theories lead to assertions that are testable using empirical data.
- Data may discredit the theory (call it a *hypothesis*) or not
  ↪ i.e. are empirical findings reasonable under hypothesis?
- Example: Large Hadron Collider in CERN, Genève. Does the Higgs Boson exist? Study if particle trajectories are consistent with what theory predicts.
- Example: Theory of "luminoferous aether" in late 19th century to explain light travelling in vacuum. Discredited by Michelson-Morley experiment.
- Similarities with the logical/mathematical concept of a necessary condition

Formal statistical framework?

Statistical Framework for Testing Hypotheses

The Problem of Hypothesis Testing

- $\mathbf{X} = (X_1, ..., X_n)$ random variables with joint density/frequency $f(\mathbf{x}; \theta)$
- $\theta \in \Theta$ where $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \emptyset$ (or $\Lambda(\Theta_0 \cap \Theta_1) = 0$)
- Observe realization $\mathbf{x} = (x_1, ..., x_n)$ of $\mathbf{X} \sim f_\theta$
- Decide on the basis of $\mathbf{x}$ whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$

↪ Often $\dim(\Theta_0) < \dim(\Theta)$ so $\theta \in \Theta_0$ represents a *simplified model*.

Example

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$ and $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{N}(\nu, 1)$. Have $\theta = (\mu, \nu)$ and

$$\Theta = \{(\mu, \nu) : \mu \in \mathbb{R}, \nu \in \mathbb{R}\} = \mathbb{R}^2$$

May be interested to see if $\mathbf{X}$ and $\mathbf{Y}$ have same distribution, even though they may be measurements on characteristics of different groups. In this case $\Theta_0 = \{(\mu, \nu) \in \mathbb{R}^2 : \mu = \nu\}$

Decision Theory Perspective on Hypothesis Testing

Given $\mathbf{X}$ 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)

$\to$ Want decision rule $\delta : \mathcal{X} \to \mathcal{A} = \{0, 1\}$ (chooses between H_0 and H_1)

- In hypothesis testing δ is called a *test function*
- Often δ depends on $\mathbf{X}$ only through some real-valued statistic $T = T(\mathbf{X})$ called a *test statistic*.

Unlikely that a test function is perfect. Possible errors to be made?

Action / Truth	H_0	H_1
0	$\ddot\smile$	Type II Error
1	Type I Error	$\ddot\smile$

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 loss function is "0–1" loss, i.e.

$$\mathcal{L}(\theta, a) = \begin{cases} 1 & \text{if } \theta \in \Theta_0 \ \& \ a = 1 & \text{(Type I Error)} \\ 1 & \text{if } \theta \in \Theta_1 \ \& \ a = 0 & \text{(Type II Error)} \\ 0 & \text{otherwise} & \text{(No Error)} \end{cases}$$

i.e. way lose 1 unit whenever committing a type I or type II error.
$\longrightarrow$ Leads to the following risk function:

$$R(\theta, \delta) = \begin{cases} \mathbb{E}_\theta[\mathbf{1}\{\delta = 1\}] = \mathbb{P}_\theta[\delta = 1] & \text{if } \theta \in \Theta_0 & \text{(prob of type I error)} \\ \mathbb{E}_\theta[\mathbf{1}\{\delta = 0\}] = \mathbb{P}_\theta[\delta = 0] & \text{if } \theta \in \Theta_1 & \text{(prob of type II error)} \end{cases}$$

In short,

$$\begin{aligned} R(\theta, \delta) &= \mathbb{P}_\theta[\delta = 1]\mathbf{1}\{\theta \in \Theta_0\} + \mathbb{P}_\theta[\delta = 0]\mathbf{1}\{\theta \in \Theta_1\} \\ \text{“}=\text{”} &\ \text{“}\mathbb{P}_\theta[\text{choose } H_1 | H_0 \text{ is true}]\text{” OR “}\mathbb{P}_\theta[\text{choose } H_0 | H_1 \text{ is true}]\text{”} \end{aligned}$$

Optimal Testing?

As with point estimation, we may wish to find *optimal* test functions
$\hookrightarrow$ Find test functions that uniformly minimize risk?

- Possible to do in some problems, but in intractable in general
  - $\hookrightarrow$ As in point estimation
- How to relax problem in this case? Minimize each type I and type II error probabilities separately?
- In general there is a trade-off between the two error probabilities

For example, consider two test functions δ_1 and δ_2 and let

$$R_1 = \{\mathbf{x} : \delta_1(\mathbf{x}) = 1\} \ \& \ R_2 = \{\mathbf{x} : \delta_2(\mathbf{x}) = 1\}$$

Assume that $R_1 \subset R_2$. Then, for all $\theta \in \Theta$,

$$\begin{aligned} \mathbb{P}_\theta[\delta_1(\mathbf{X}) = 1] &\leq \mathbb{P}_\theta[\delta_2(\mathbf{X}) = 1] \\ \mathbb{P}_\theta[\delta = 0] = 1 - \mathbb{P}_\theta[\delta_1(\mathbf{X}) = 1] &\geq 1 - \mathbb{P}_\theta[\delta_2(\mathbf{X}) = 1] = \mathbb{P}_\theta[\delta_2(\mathbf{X}) = 0] \end{aligned}$$

so by attempting to reduce the probability of error when $\theta \in \Theta_0$ we may increase the the probability of error when $\theta \in \Theta_1$!

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 declare that we will only consider test functions δ such that

$$\mathbb{P}_\theta[\delta = 1] \leq \alpha \qquad \forall \theta \in \Theta_0 \qquad \left(\text{i.e. } \sup_{\theta \in \Theta_0} \mathbb{P}_\theta[\delta = 1] \leq \alpha\right)$$

   i.e. rules for which prob 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, choose δ to minimize prob of type II error uniformly on Θ_1:

$$\mathbb{P}_\theta[\delta(\mathbf{X}) = 0] = 1 - \mathbb{P}_\theta[\delta(\mathbf{X}) = 1]$$

4. Equivalently, maximize the *power* uniformly over Θ_1

$$\beta(\theta, \delta) = \mathbb{P}_\theta[\delta(\mathbf{X}) = 1] = \mathbb{E}_\theta[\mathbf{1}\{\delta(\mathbf{X}) = 1\}] = \mathbb{E}_\theta[\delta(\mathbf{X})], \quad \theta \in \Theta_1$$

(since $\delta = 1 \iff \mathbf{1}\{\delta = 1\} = 1$ and $\delta = 0 \iff \mathbf{1}\{\delta = 1\} = 0$)

The Neyman-Pearson Setup

Intuitive rationale of the approach:

- Want to test H_0 against H_1 at significance level α
- Suppose we observe $\delta(\mathbf{X}) = 1$ (so we take action 1)
- α is usually small, so that if H_0 is indeed true, we have observed something rare or unusual
  - $\hookrightarrow$ since $\delta = 1$ has probability at most α under H_0
- Evidence that H_0 is false (i.e. in favour of H_1)
- So taking action 1 is a highly reasonable decision

But what if we observe $\delta(\mathbf{X}) = 0$? (so we take action 0)

- Our significance level does not guarantee that our decision is necessarily reasonable
- Our decision would have been reasonable if δ was such that the type II error was also low (given the significance level).
- If we had maximized power β at level α though, then we would be reassured of our decision.

The Neyman-Pearson Setup

- Neyman-Pearson setup naturally exploits any asymmetric structure
- But, if natural asymmetry absent, need judicious choice of H_0

Example: Obama VS McCain 2008. Pollsters gather iid sample $\mathbf{X}$ from Ohio with $X_i = \mathbf{1}\{\text{vote Obama}\}$. Which pair of hypotheses to test?

$$\begin{cases} H_0: & \text{Obama wins Ohio} \\ H_1: & \text{McCain wins Ohio} \end{cases} \qquad \text{OR} \qquad \begin{cases} H_0: & \text{McCain wins Ohio} \\ H_1: & \text{Obama wins Ohio} \end{cases}$$

- Which pair to choose to make a prediction? (confidence intervals?)
- If Obama is conducting poll to decide whether he'll spend more money to campaign in Ohio, then 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 McCain because he thought he would win without any extra effort.
- (b) is much worse than (a) (especially since Obama had lots of $'s)
- Hence Obama would pick $H_0 = \{\text{McCain wins Ohio}\}$ as his null

Finding Good Test Functions

Consider simplest situation:

- Have $(X_1, ..., X_n) \sim f(\cdot; \theta)$ with $\Theta = \{\theta_0, \theta_1\}$
- Want to test $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$

The Neyman-Pearson Lemma

Let $\mathbf{X} = (X_1, ..., X_n)$ have joint density (frquency) function $f \in \{f_0, f_1\}$ and suppose we wish to test

$$H_0 : f = f_0 \qquad VS \qquad H_1 : f = f_1.$$

Then, the test whose test function is given by

$$\delta(\mathbf{X}) = \begin{cases} 1 & \text{if } f_1(\mathbf{X}) \geq k \cdot f_0(\mathbf{X}), \\ 0 & \text{otherwise} \end{cases}$$

for some $k \in (0, \infty)$, is a *most powerful (MP)* test of H_0 versus H_1 at significance level $\alpha = \mathbb{P}_0[\delta(\mathbf{X}) = 1](= \mathbb{E}_0[\delta(\mathbf{X})] = \mathbb{P}_0[f_1(\mathbf{X}) \geq k \cdot f_0(\mathbf{X})])$.

Proof.

Use obvious notation $\mathbb{E}_0$, $\mathbb{E}_1$, $\mathbb{P}_0$, $\mathbb{P}_1$ corresponding to H_0 or H_1.
It suffices to prove that if ψ is any function with $\psi(\mathbf{x}) \in \{0, 1\}$, then

$$\mathbb{E}_0[\psi(\mathbf{X})] \leq \underbrace{\mathbb{E}_0[\delta(\mathbf{X})]}_{=\alpha(\text{by definition})} \implies \underbrace{\mathbb{E}_1[\psi(\mathbf{X})]}_{\beta_1(\psi)} \leq \underbrace{\mathbb{E}_1[\delta(\mathbf{X})]}_{\beta_1(\delta)}.$$

(recall that $\beta_1(\delta) = 1 - \mathbb{P}_1[\delta = 0] = \mathbb{P}_1[\delta = 1] = \mathbb{E}_1[\delta]$).
WLOG assume that f_0 and f_1 are density functions. Note that

$$f_1(\mathbf{x}) - k \cdot f_0(\mathbf{x}) \geq 0 \text{ if } \delta(\mathbf{x}) = 1 \quad \& \quad f_1(\mathbf{x}) - k \cdot f_0(\mathbf{x}) < 0 \text{ if } \delta(\mathbf{x}) = 0.$$

Since ψ can only take the values 0 or 1,

$$\begin{aligned} \psi(\mathbf{x})(f_1(\mathbf{x}) - k \cdot f_0(\mathbf{x})) &\leq \delta(\mathbf{x})(f_1(\mathbf{x}) - k \cdot f_0(\mathbf{x})) \\ \int_{\mathbb{R}^n} \psi(\mathbf{x})(f_1(\mathbf{x}) - k \cdot f_0(\mathbf{x}))d\mathbf{x} &\leq \int_{\mathbb{R}^n} \delta(\mathbf{x})(f_1(\mathbf{x}) - k \cdot f_0(\mathbf{x}))d\mathbf{x} \end{aligned}$$

Rearranging the terms yields

$$\int_{\mathbb{R}^n} (\psi(\mathbf{x}) - \delta(\mathbf{x})) f_1(\mathbf{x}) d\mathbf{x} \leq k \int_{\mathbb{R}^n} (\psi(\mathbf{x}) - \delta(\mathbf{x})) f_0(\mathbf{x}) d\mathbf{x}$$
$$\Longrightarrow \mathbb{E}_1[\psi(\mathbf{X})] - \mathbb{E}_1[\delta(\mathbf{X})] \leq k\,(\mathbb{E}_0[\psi(\mathbf{X})] - \mathbb{E}_0[\delta(\mathbf{X})])$$

So when $\mathbb{E}_0[\psi(\mathbf{X})] \leq \mathbb{E}_0[\delta(\mathbf{X})]$ the RHS is negative, i.e. δ is an MP test of H_0 vs H_1 at level α. □

- Essentially the result says that the optimal test statistic for simple hypotheses vs simple alternatives is $T(\mathbf{X}) = f_1(\mathbf{X})/f_0(\mathbf{X})$
- The optimal test function would then reject the null whenever $T \geq k$
- k is chosen so that the test has desirable level α
- The result does not guarantee existence of an MP test
- The result does not guarantee uniqueness when MP test exists

The Neyman-Pearson Setup

General version of Neyman-Pearson lemma considers relaxed problem:

maximize $\mathbb{E}_1[\delta]$ subject to $\mathbb{E}_0[\delta] = \alpha$ & $0 \leq \delta(\mathbf{X}) \leq 1$ *a.s.*

It is then proven that an optimal δ exists and is given by

$$\delta(\mathbf{X}) = \begin{cases} 1 & \text{if } f_1(\mathbf{X}) > kf_0(\mathbf{X}), \\ c & \text{if } f_1(\mathbf{X}) = kf_0(\mathbf{X}) \\ 0 & \text{if } f_1(\mathbf{X}) < kf_0(\mathbf{X}) \end{cases}$$

where k and $c \in [0,1]$ are such that the conditions are satisfied
$\rightarrow$ The optimum need not be a test function (relaxation≡randomization)
$\hookrightarrow$ But when the test statistic $T = f_1/f_0$ is a continuous RV, then δ can be taken to have range $\{0,1\}$, i.e. be a test function
$\hookrightarrow$ In this case an MP test function of $H_0 : f = f_0$ against $H_1 : f = f_1$ exists for any significance level $\alpha > 0$.
$\rightarrow$ When T is discrete then the optimum need not be a test function for certain levels α, unless we consider *randomized tests* as well.

The Neyman-Pearson Setup

Example (Exponential Distribution)

Let $X_1, ..., X_n \stackrel{iid}{\sim} \text{Exp}(\lambda)$ and $\lambda \in \{\lambda_1, \lambda_2\}$, with $\lambda_1 > \lambda_0$ (say). Consider

$$\begin{cases} H_0 : & \lambda = \lambda_0 \\ H_1 : & \lambda = \lambda_1 \end{cases}$$

Have

$$f(\mathbf{x}; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$$

So Neyman-Pearson say we must base our test on the statistic

$$T = \frac{f(\mathbf{X}; \lambda_1)}{f(\mathbf{X}; \lambda_0)} = \left(\frac{\lambda_1}{\lambda_0}\right)^n \exp\left[(\lambda_0 - \lambda_1)\sum_{i=1}^n X_i\right]$$

rejecting the null if $T \geq k$, for k such that the level is α.

The Neyman-Pearson Setup

Example (cont'd)

To determine k we note that T is a decreasing function of $S = \sum_{i=1}^n X_1$ (since $\lambda_0 < \lambda_1$). Therefore

$$T \geq k \iff S \leq K$$

for some K, so that

$$\alpha = \mathbb{P}_{\lambda_0}[T \geq k] \iff \alpha = \mathbb{P}_{\lambda_0}\left[\sum_{i=1}^n X_i \leq K\right]$$

For given values of λ_0 and α it is entirely feasible to find the appropriate K: under the null hypothesis, S has a gamma distribution with parameters n and λ_0. Hence we reject H_0 at level α if S exceeds that α-quantile of a gamma(n, λ_0) distribution.

Example (Uniform Distribution)

Let $X_1, \ldots X_n \overset{iid}{\sim} \mathcal{U}[0,\theta]$ with $\theta \in \{\theta_0, \theta_1\}$ where $\theta_0 > \theta_1$. Consider

$$\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta = \theta_1 \end{cases}$$

Recall that

$$f(\mathbf{x};\theta) = \frac{1}{\theta^n}\mathbf{1}\left\{\max_{1\le i\le n} X_i \le \theta\right\}$$

so an MP test of H_0 vs H_1 should be based on the discrete test statistic

$$T = \frac{f(\mathbf{X};\theta_1)}{f(\mathbf{X};\theta_0)} = \left(\frac{\theta_0}{\theta_1}\right)^n \mathbf{1}\{X_{(n)} \le \theta_1\}.$$

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 $X_{(n)}$, giving the test

$$\text{reject } H_0 \quad \text{iff} \quad X_{(n)} \le k$$

with k solving the equation:

$$\mathbb{P}_{\theta_0}[X_{(n)} \le k] = \left(\frac{k}{\theta_0}\right)^n = \alpha,$$

i.e. with $k = \theta_0\alpha^{1/n}$, with power

$$\mathbb{P}_{\theta_1}[X_{(n)} \le \theta_0\alpha^{1/n}] = \left(\frac{\theta_0\alpha^{1/n}}{\theta_1}\right)^n = \alpha\left(\frac{\theta_0}{\theta_1}\right)^n.$$

Is this the MP test at level $\alpha < (\theta_1/\theta_0)^n$ though?

Example (cont'd)

Use general form of the Neyman-Pearson lemma to solve relaxed problem:

maximize $\mathbb{E}_1[\delta(\mathbf{X})]$ subject to $\mathbb{E}_{\theta_0}[\delta(\mathbf{X})] = \alpha < \left(\frac{\theta_1}{\theta_0}\right)^n$ & $0 \le \delta(\mathbf{x}) \le 1$.

One solution to this problem is given by

$$\delta(\mathbf{X}) = \begin{cases} \alpha(\theta_0/\theta_1)^n & \text{if } X_{(n)} \le \theta_1, \\ 0 & \text{otherwise.} \end{cases}$$

which is not a test function. However, we see that its power is

$$\mathbb{E}_{\theta_1}[\delta(\mathbf{X})] = \alpha\left(\frac{\theta_0}{\theta_1}\right)^n = \mathbb{P}_{\theta_1}[X_{(n)} \le \theta_0\alpha^{1/n}]$$

which is the power of the test we proposed.

Hence the test that rejects H_0 if $X_{(n)} \le \theta_0\alpha^{1/n}$ is an MP test for all levels $\alpha < (\theta_1/\theta_0)^n$.

Testing Statistical Hypotheses II

Statistical Theory

Victor Panaretos
Ecole Polytechnique Fédérale de Lausanne

Neyman-Pearson Framework for Testing Hypotheses

The Problem of Hypothesis Testing

- $\mathbf{X} = (X_1, ..., X_n)$ random variables with joint density/frequency $f(\mathbf{x}; \theta)$
- $\theta \in \Theta$ where $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \emptyset$
- Observe realization $\mathbf{x} = (x_1, ..., x_n)$ of $\mathbf{X} \sim f_\theta$
- Decide on the basis of $\mathbf{x}$ 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 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. for all possible simple nulls),

$$\mathbb{E}_\theta[\delta] \leq \alpha \qquad \forall \theta \in \Theta_0$$

2. Is most powerful for all $\theta \in \Theta_1$ (i.e. for all possible simple alternatives),

$$\mathbb{E}_\theta[\delta] \geq \mathbb{E}_\theta[\delta'] \qquad \forall \theta \in \Theta_1 \quad \& \quad \delta' \text{ respecting level } \alpha$$

Unfortunately UMP tests rarely exist. Why?

$\hookrightarrow$ Consider $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$

- A UMP test must be MP test for any $\theta \neq \theta_1$.
- But by the form of the MP test typically differs for $\theta_1 > \theta_0$ and $\theta_1 < \theta_0$!

Example (No UMP test exists)

Let $X \sim \text{Binom}(n, \theta)$ 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 gives test statistics

$$T = \frac{f(X;\theta_1)}{f(X;\theta_0)} = \left(\frac{1-\theta_0}{1-\theta_1}\right)^n \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^X$$

- If $\theta_1 > \theta_0$ then T increasing in X
  $\hookrightarrow$ MP test would reject for large values of X
- If $\theta_1 < \theta_0$ then T 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 $\forall\ \lambda_1 > \lambda_0$:

$$\text{Reject } H_0 \quad \text{for} \quad \sum_{i=1}^n X_i \leq k, \qquad \text{with } k \text{ such that } \mathbb{P}_{\lambda_0}\left[\sum_{i=1}^n X_i \leq k\right] = \alpha$$

But for $\lambda < \lambda_0$, $\mathbb{P}_{\lambda_0}[\sum_{i=1}^n X_i \leq k] = \alpha \implies \mathbb{P}_{\lambda}[\sum_{i=1}^n X_i \leq k] < \alpha$.

So the same test respects level α for all singletons under the the null.
$\hookrightarrow$ The test is UMP of H_0 vs H_1

When do UMP tests exist?

Examples: insight on which composite pairs typically admit UMP tests:

1. Hypothesis pair concerns a single real-valued parameter
2. Hypothesis pair is "one-sided"

However existence of UMP test does not only depend on hypothesis structure, as was the case with simple vs simple...

$\hookrightarrow$ Also depends on specific model. Sufficient condition?

Definition (Monotone Likelihood Ratio Property)

A family of density (frequency) functions $\{f(\mathbf{x};\theta) : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$ is said to have monotone likelihood ratio if there exists a real-valued function $T(\mathbf{x})$ such that, for any $\theta_1 < \theta_2$, the function

$$\frac{f(\mathbf{x};\theta_2)}{f(\mathbf{x};\theta_1)}$$

is a non-decreasing function of $T(\mathbf{x})$.

When do UMP tests exist?

Proposition

Let $\mathbf{X} = (X_1, \ldots, X_n)$ have joint distribution of monotone likelihood ratio with respect to a statistic T, depending on $\theta \in \mathbb{R}$. Further assume that T is a continuous random variable. Then, the test function given by

$$\delta(\mathbf{X}) = \begin{cases} 1 & \text{if } T(\mathbf{X}) > k, \\ 0 & \text{if } T(\mathbf{X}) \leq k \end{cases}$$

is UMP among all tests with type one error bounded above by $\mathbb{E}_{\theta_0}[\delta(\mathbf{X})]$ for the hypothesis pair

$$\begin{cases} H_0 : & \theta \leq \theta_0 \\ H_1 : & \theta > \theta_1 \end{cases}$$

[The assumption of continuity of the random variable T can be removed, by considering randomized tests as well, similarly as before]

When do UMP tests exist?

- T yielding monotone likelihood ratio necessarily a sufficient statistic

Example (One-Parameter Exponential Family)

Let $\mathbf{X} = (X_1, ..., X_n)$ have a joint density (frequency)

$$f(\mathbf{x};\theta) = \exp[c(\theta)T(\mathbf{x}) - b(\theta) + S(\mathbf{x})]$$

and assume WLOG that $c(\theta)$ is strictly increasing. For $\theta_1 < \theta_2$,

$$\frac{f(\mathbf{x};\theta_2)}{f(\mathbf{x};\theta_1)} = \exp\{[c(\theta_2) - c(\theta_1)]T(\mathbf{x}) + b(\theta_1 - b(\theta_2))\}$$

is increasing in T by monotonicity of $c(\cdot)$.

Hence the UMP test of $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$ would reject iff $T(\mathbf{x}) \geq k$, with $\alpha = \mathbb{P}_{\theta_0}[T \geq k]$.

Locally Most Powerful Tests

$\hookrightarrow$ What if MLR property fails to be satisfied? Can optimality be "saved"?

- Consider $\theta \in \mathbb{R}$ and wish to test: $H_0 : \theta \leq \theta_0$ vs $H_0 : \theta > \theta_0$
- Intuition: if true θ far from θ_0 any reasonable test powerful
- ⋆ So focus on maximizing power in small neighbourhood of θ_0

$\rightarrow$ Consider power function $\beta(\theta) = \mathbb{E}_\theta[\delta(\mathbf{X})]$ of some δ.
$\rightarrow$ Require $\beta(\theta_0) = \alpha$ (*a boundary condition*, similar with MLR setup)
$\rightarrow$ Assume that $\beta(\theta)$ is differentiable, so for θ close to θ_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 $\beta(\theta_0) = \alpha$

$\boxed{\text{How do we solve this constrained optimization problem?}}$

$\hookrightarrow$ Solution similar to Neyman-Pearson lemma?

Supposing that $\mathbf{X} = (X_1, ...X_n)$ has density $f(\mathbf{x};\theta)$, then

$$\begin{aligned}
\beta(\theta) &= \int_{\mathbb{R}^n} \delta(\mathbf{x}) f(\mathbf{x};\theta) d\mathbf{x} \\
\frac{\partial}{\partial\theta}\beta(\theta) &= \int_{\mathbb{R}^n} \delta(\mathbf{x}) \frac{\partial}{\partial\theta} f(\mathbf{x};\theta) d\mathbf{x} \quad \text{[provided interchange possible]} \\
&= \int_{\mathbb{R}^n} \delta(\mathbf{x}) \frac{f(\mathbf{x};\theta)}{f(\mathbf{x};\theta)} \frac{\partial}{\partial\theta} f(\mathbf{x};\theta) d\mathbf{x} \\
&= \int_{\mathbb{R}^n} \delta(\mathbf{x}) \left[\frac{\partial}{\partial\theta} \log f(\mathbf{x};\theta)\right] f(\mathbf{x};\theta) d\mathbf{x} \\
&= \mathbb{E}_\theta\left[\delta(\mathbf{X}) \underbrace{\frac{\partial}{\partial\theta} \log f(\mathbf{x};\theta)}_{S(\mathbf{X};\theta)}\right]
\end{aligned}$$

Locally Most Powerful Tests

Theorem

Let $\mathbf{X} = (X_1, ..., X_n)$ have joint density (frequency) $f(\mathbf{x};\theta)$ and define the test function

$$\delta(\mathbf{X}) = \begin{cases} 1 & \text{if } S(\mathbf{X};\theta_0) \geq k, \\ 0 & \text{otherwise} \end{cases}$$

where k is such that $\mathbb{E}_{\theta_0}[\delta(\mathbf{X})] = \alpha$. Then δ maximizes

$$\mathbb{E}_{\theta_0}[\psi(\mathbf{X})S(\mathbf{X};\theta_0)]$$

over all test functions ψ satisfying the constraint $\mathbb{E}_{\theta_0}[\psi(\mathbf{X})] = \alpha$.

- Gives recipe for constructing LMP test
- We were concerned about power *only locally around* θ_0
- May not even give a level α test for some $\theta < \theta_0$

Proof.

Consider ψ with $\psi(\mathbf{x}) \in \{0,1\}\ \forall\ \mathbf{x}$ and $\mathbb{E}_{\theta_0}[\psi(\mathbf{X})] = \alpha$. Then,

$$\delta(\mathbf{x}) - \psi(\mathbf{x}) = \begin{cases} \geq 0 & \text{if } S(\mathbf{x};\theta_0) \geq k, \\ \leq 0 & \text{if } S(\mathbf{x};\theta_0) \leq k \end{cases}$$

Therefore

$$\mathbb{E}_{\theta_0}[(\delta(\mathbf{X}) - \psi(\mathbf{X}))(S(\mathbf{X};\theta_0) - k)] \geq 0$$

Since $\mathbb{E}_{\theta_0}[\delta(\mathbf{X}) - \psi(\mathbf{X})] = 0$ it must be that

$$\mathbb{E}_{\theta_0}\left[\delta(\mathbf{X})S(\mathbf{X};\theta_0)\right] \geq \mathbb{E}_{\theta_0}\left[\psi(\mathbf{X})S(\mathbf{X};\theta_0)\right]$$

$\square$

How is the critical value k evaluated in practice? (obviously to give level α)

- When $\{X_i\}$ are iid then $S(\mathbf{X};\theta) = \sum_{i=1}^n \ell'(X_i;\theta)$
- Under regularity conditions: sum of iid rv's mean zero variance $I(\theta)$
- So, for $\theta = \theta_0$ and large n, $S(\mathbf{X};\theta) \stackrel{d}{\approx} \mathcal{N}(0, nI(\theta))$

Example (Cauchy distribution)

Let $X_1, ..., X_n \stackrel{iid}{\sim} \text{Cauchy}(\theta)$, with density,

$$f(x;\theta) = \frac{1}{\pi(1 + (x-\theta)^2)}$$

and consider the hypothesis pair $\begin{cases} H_0: & \theta \geq 0 \\ H_1: & \theta \leq 0 \end{cases}$

We have

$$S(\mathbf{X};0) = \sum_{i=1}^n \frac{2X_i}{1+X_i^2}$$

so that the LMP test at level α rejects the null if $S(\mathbf{X};0) \leq k$, where

$$\mathbb{P}_0[S(\mathbf{X};0) \leq k] = \alpha$$

While the exact distribution is difficult to obtain, for large n, $S(\mathbf{X};0) \stackrel{d}{\approx} \mathcal{N}(0, n/2)$.

Likelihood Ratio Tests

$\boxed{\text{So far seen} \rightarrow \text{Tests for } \Theta = \mathbb{R}\text{, simple vs simple, one sided vs one sided}}$

$\hookrightarrow$ Extension to multiparameter case $\boldsymbol{\theta} \in \mathbb{R}^p$? General Θ_0, Θ_1?

- Unfortunately, optimality theory breaks down in higher dimensions
- General method for constructive *reasonable* tests?

$\rightarrow$ The idea: Combine Neyman-Pearson paradigm with Max Likelihood

Definition (Likelihood Ratio)

The *likelihood ratio 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

$$\Lambda(\mathbf{X}) = \frac{\sup_{\boldsymbol{\theta}\in\Theta} f(\mathbf{X};\theta)}{\sup_{\boldsymbol{\theta}\in\Theta_0} f(\mathbf{X};\theta)} = \frac{\sup_{\boldsymbol{\theta}\in\Theta} L(\theta)}{\sup_{\boldsymbol{\theta}\in\Theta_0} L(\theta)}$$

- "*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)

Example

Let $X_1, ..., X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Consider:

$$H_0: \mu = \mu_0 \qquad \text{vs} \qquad H_1: \mu \neq \mu_0$$

$$\Lambda(\mathbf{X}) = \frac{\sup_{(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}^+} f(\mathbf{X};\mu,\sigma^2)}{\sup_{(\mu,\sigma^2)\in\{\mu_0\}\times\mathbb{R}^+} f(\mathbf{X};\mu,\sigma^2)} = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right)^{\frac{n}{2}} = \left(\frac{\sum_{i=1}^n (X_i-\mu_0)^2}{\sum_{i=1}^n (X_i-\bar{X})^2}\right)^{\frac{n}{2}}$$

So reject when $\Lambda \geq k$, where k is s.t. $\mathbb{P}_0[\Lambda \geq k] = \alpha$. Distribution of Λ?

By monotonicity look only at

$$\begin{aligned}\frac{\sum_{i=1}^n (X_i-\mu_0)^2}{\sum_{i=1}^n (X_i-\bar{X})^2} &= 1 + \frac{n(\bar{X}-\mu_0)^2}{\sum_{i=1}^n (X_i-\bar{X})^2} = 1 + \frac{1}{n-1}\left(\frac{n(\bar{X}-\mu_0)^2}{S^2}\right) \\ &= 1 + \frac{T^2}{n-1}\end{aligned}$$

With $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i-\bar{X})^2$ and $T = \sqrt{n}(\bar{X}-\mu_0)/S \stackrel{H_0}{\sim} t_{n-1}$.

So $T^2 \stackrel{H_0}{\sim} F_{1,n-1}$ and k may be chosen appropriately.

Example

Let $X_1, ..., X_m \stackrel{iid}{\sim} \text{Exp}(\lambda)$ and $Y_1, ..., Y_n \stackrel{iid}{\sim} \text{Exp}(\theta)$. Assume $\mathbf{X}$ indep $\mathbf{Y}$.

$$\text{Consider:} \quad H_0 : \theta = \lambda \quad \text{vs} \quad H_1 : \theta \neq \lambda$$

$$\underset{\sup_{(\lambda,\theta)\in\mathbb{R}^2_+} f(\mathbf{X},\mathbf{Y};\lambda,\theta)}{\text{Unrestricted MLEs:}} \quad \hat{\lambda} = 1/\bar{X} \quad \& \quad \hat{\theta} = 1/\bar{Y}$$

$$\underset{\sup_{(\lambda,\theta)\in\{(x,y)\in\mathbb{R}^2_+ : x=y\}} f(\mathbf{X},\mathbf{Y};\lambda,\theta)}{\text{Restricted MLEs:}} \quad \hat{\lambda}_0 = \hat{\theta}_0 = \left[\frac{m\bar{X} + n\bar{Y}}{m+n}\right]^{-1}$$

$$\implies \Lambda = \left(\frac{m}{m+n} + \frac{n}{n+m}\frac{\bar{Y}}{\bar{X}}\right)^m \left(\frac{n}{n+m} + \frac{m}{m+n}\frac{\bar{X}}{\bar{Y}}\right)^n$$

Depends on $T = \bar{X}/\bar{Y}$ and can make Λ large/small by varying T.
$\hookrightarrow$ But $T \stackrel{H_0}{\sim} F_{2m,2n}$ so given α we may find the critical value k.

Distribution of Likelihood Ratio?

More often than not, dist(Λ) intractable
$\hookrightarrow$(and no simple dependence on T with tractable distribution either)

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 $s < p$
- Concentrate on $\mathbf{X} = (X_1, ..., X_n)$ has iid components.
- Initially restrict attention to $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. LR becomes:

$$\Lambda_n(\mathbf{X}) = \prod_{i=1}^{n} \frac{f(X_i; \hat{\boldsymbol{\theta}}_n)}{f(X_i; \boldsymbol{\theta}_0)}$$

  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 $X_1, ..., X_n$ 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

$$2 \log \Lambda_n \xrightarrow{d} V \sim \chi^2_1$$

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.

Under the conditions of the theorem and when H_0 is true,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta))$$

Now take logarithms and expand in a Taylor series

$$\begin{aligned}
\log \Lambda_n &= \sum_{i=1}^{n} [\ell(X_i; \hat{\theta}_n) - \ell(X_i; \theta_0)] \\
&= (\theta_0 - \hat{\theta}_n) \sum_{i=1}^{n} \ell'(X_i; \hat{\theta}_n) - \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 \sum_{i=1}^{n} \ell''(X_i; \theta^*_n) \\
&= -\frac{1}{2} n(\hat{\theta}_n - \theta_0)^2 \frac{1}{n} \sum_{i=1}^{n} \ell''(X_i; \theta^*_n)
\end{aligned}$$

where θ^*_n lies between $\hat{\theta}_n$ and θ_0.

But under assumptions (A1)-(A6), it follows that when H_0 is true,

$$\frac{1}{n}\sum_{i=1}^{n}\ell''(X_i;\theta_n^*) \xrightarrow{p} -\mathbb{E}_{\theta_0}[\ell''(X_i;\theta_0)] = I(\theta_0)$$

On the other hand, by the continuous mapping theorem,

$$n(\hat{\theta}_n - \theta_0)^2 \xrightarrow{d} \frac{V}{I(\theta_0)}$$

Applying Slutsky's theorem now yields the result. □

Theorem (Wilk's theorem, general p, general $s \leq p$)

Let $X_1, ..., X_n$ 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 $H_0 : \{\theta_j = \theta_{j,0}\}_{j=1}^s$ satisfies $2\log\Lambda_n \xrightarrow{d} V \sim \chi_s^2$ when H_0 is true.

Exercise

Prove Wilks' theorem. Note that it may potentially be that $s < p$.

Hypotheses of the form $H_0 : \{g_j(\boldsymbol{\theta}) = a_j\}_{j=1}^s$, for g_j differentiable real functions, can also be handled by Wilks' theorem:

- Define $(\phi_1, ..., \phi_p) = g(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), ..., g_p(\boldsymbol{\theta}))$
- $g_{s+1}, ..., g_p$ defined so that $\boldsymbol{\theta} \mapsto g(\boldsymbol{\theta})$ is 1-1
- Apply theorem with parameter ϕ

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 form zero. Use asymptotics for distributions (under conditions we end up with a χ^2)
- ...

From Hypothesis Tests to Confidence Regions

Statistical Theory

Victor Panaretos
Ecole Polytechnique Fédérale de Lausanne

Beyond Neyman-Pearson?

So far restricted to Neyman-Pearson Framework:

1. Fix a significance level α for the test
2. Consider 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 $\delta^*(\mathbf{x}) = 1$.

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 then consider a family of test functions respecting level α...

... consider 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 $\mathbf{x}$ some of these rules reject the null, while others do not
2. Which is the smallest α for which H_0 is rejected given $\mathbf{x}$?

Observed Significance Level

Definition (p–Value)

Let $\{\delta_\alpha\}_{\alpha\in(0,1)}$ be a family of test functions satisfying

$$\alpha_1 < \alpha_2 \implies \{\mathbf{x}\in\mathcal{X} : \delta_{\alpha_1}(\mathbf{x}) = 1\} \subseteq \{\mathbf{x}\in\mathcal{X} : \delta_{\alpha_2}(\mathbf{x}) = 1\}.$$

The p–value (or observed significance level) of the family $\{\delta_\alpha\}$ is

$$p(\mathbf{x}) = \inf\{\alpha : \delta_\alpha(\mathbf{x}) = 1\}$$

$\hookrightarrow$ The p–value is the smallest value of α for which the null would be rejected at level α, given $\mathbf{X} = \mathbf{x}$.

Most usual setup:

- Have a single test statistic T
- Construct family $\delta_\alpha(\mathbf{x}) = \mathbf{1}\{T(\mathbf{x}) > k_\alpha\}$
- If $\mathbb{P}_{H_0}[T \le t] = G(t)$ then $p(\mathbf{x}) = \mathbb{P}_{H_0}[T(\mathbf{X}) \ge T(\mathbf{x})] = 1 - G(T(\mathbf{x}))$

Observed Significance Level

Notice: contrary to NP-framework did not make explicit decision!

- We simply reported a p–value
- The p–value is used as a measure of evidence against H_0
  - ↪ Small p–value provides evidence against H_0
  - ↪ Large p–value provides no evidence against H_0
- How small does "small" mean?
  - ↪ Depends on the specific problem...

Intuition:

- Recall that extreme values of test statistics are those that are "inconsistent" with null (NP-framework)
- p–value is probability of observing a value of the test statistic as extreme as or more extreme than the one we observed, under the null
- If this probability is small, the we have witnessed something quite unusual under the null hypothesis
- Gives evidence against the null hypothesis

Example (Normal Mean)

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ 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, $T = \sqrt{n}\bar{X}/S \overset{H_0}{\sim} t_{n-1}$.

Since $T^2 \overset{H_0}{\sim} F_{1,n-1}$, p–value is

$$p(\mathbf{x}) = \mathbb{P}_{H_0}[T^2(\mathbf{X}) \geq T(\mathbf{x})] = 1 - G_{F_{1,n-1}}(T^2(\mathbf{x}))$$

Consider two samples (datasets),

$$\mathbf{x} = (0.66, 0.28, -0.99, 0.007, -0.29, -1.88, -1.24, 0.94, 0.53, -1.2)$$

$$\mathbf{y} = (1.4, 0.48, 2.86, 1.02, -1.38, 1.42, 2.11, 2.77, 1.02, 1.87)$$

Obtain $p(\mathbf{x}) = 0.32$ while $p(\mathbf{y}) = 0.006$.

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
  - ↪ reflects how statistically significant the alternative statement is

Recall example: Statisticians working for Obama gather iid sample $\mathbf{X}$ from Ohio with $X_i = \mathbf{1}\{\text{vote Obama}\}$. Obama team want to test

$$\begin{cases} H_0 : & \text{McCain wins Ohio} \\ H_1 : & \text{Obama wins Ohio} \end{cases}$$

- Will statisticians decide for Obama?
- Perhaps better to report p–value to him and let him decide...

What if statisticians working for newspaper, not Obama?

- Something easier to interpret than test/p–value?

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)
  - ↪ always attach an estimator of variability, e.g. standard error
  - ↪ interpretation?
- Hypothesis tests may provide way to interpret estimator's variability within the setup of a particular problem
  - ↪ e.g. if observe $\hat{P}[\text{obama wins}] = 0.52$ can actually see what p–value we get when testing $H_0 : P[\text{obama wins}] \geq 1/2$.
- Something more directly interpretable?

Back to our example: What do pollsters do in newspapers?

↪ They announce their point estimate (e.g. 0.52)

↪ 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 $\mathbf{X} = (X_1, ..., X_n)$ be random variables with joint distribution depending on $\theta \in \mathbb{R}$ and let $L(\mathbf{X})$ and $U(\mathbf{X})$ be two statistics with $L(\mathbf{X}) < U(\mathbf{X})$ a.s. Then, the random interval $[L(\mathbf{X}), U(\mathbf{X})]$ is called a $100(1-\alpha)\%$ confidence interval for θ if

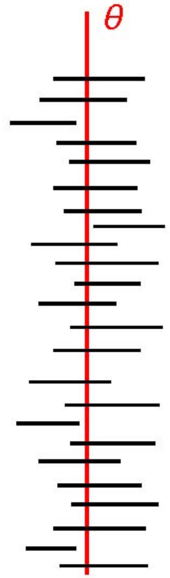
$$\mathbb{P}_\theta[L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})] \geq 1 - \alpha$$

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 interval: probability that the interval contains the true value is at least $1-\alpha$.
- Given any realization $\mathbf{X} = \mathbf{x}$, the interval $[L(\mathbf{x}), U(\mathbf{x})]$ will either contain or not contain θ.
- Interpretation: if we construct intervals with this method, then we expect that $100(1-\alpha)\%$ of the time our intervals will engulf 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

$$\mathbb{P}_\mu[-1.96 \leq \sqrt{n}(\bar{X} - \mu) \leq 1.96] = 0.95$$

and 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 that the random interval $[L(\mathbf{X}), U(\mathbf{X})] = \left[\bar{X} - \frac{1.96}{\sqrt{n}}, \bar{X} + \frac{1.96}{\sqrt{n}}\right]$ is a 95% confidence interval for μ.
Central Limit Theorem: same argument can yield approximate 95% CI when $X_1, ..., X_n$ are iid, $\mathbb{E}X_i = \mu$ and $\mathrm{Var}(X_i) = 1$, regardless of their distribution.

Pivotal Quantities

What can we learn from previous example?

Definition (Pivot)

A random function $g(\mathbf{X}, \theta)$ is said to be a pivotal quantity (or simply a pivot) if it is a function both of $\mathbf{X}$ and θ 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

$$\mathbb{P}_\theta[a \leq g(\mathbf{X}, \theta) \leq b] = 1 - \alpha \qquad \forall\ \theta \in \Theta$$

- If $g(\mathbf{X}, \theta)$ can be manipulated then the above yields a CI

Example

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{U}[0, \theta]$. Recall that MLE $\hat{\theta}$ is $\hat{\theta} = X_{(n)}$, with distribution

$$\mathbb{P}_\theta \left[X_{(n)} \leq x\right] = F_{X_{(n)}}(x) = \left(\frac{x}{\theta}\right)^n \implies \mathbb{P}_\theta \left[\frac{X_{(n)}}{\theta} \leq y\right] = y^n$$

$\rightarrow$ Hence $X_{(n)}/\theta$ is a pivot for θ. Can now 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 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 $g(\mathbf{X}; \theta_k)$ which

1. Depends on $\mathbf{X}$, 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 intractable

Resort to asymptotic approximations...
$\hookrightarrow$ Most classic example when have $a_n(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$.

Confidence Regions

What about higher dimensional parameters?

Definition (Confidence Region)

Let $\mathbf{X} = (X_1, ..., X_n)$ be random variables with joint distribution depending on $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. A random subset $R(\mathbf{X})$ of Θ depending on $\mathbf{X}$ is called a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}$ if

$$\mathbb{P}_\theta[R(\mathbf{X}) \ni \boldsymbol{\theta}] \geq 1 - \alpha$$

for all $\boldsymbol{\theta} \in \Theta$, with equality for at least one value of $\boldsymbol{\theta}$.

- No restriction requiring $R(\mathbf{X})$ to be convex or even connected
  - $\hookrightarrow$ So when $p = 1$ 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 dist$[g(\mathbf{X}, \boldsymbol{\theta})]$ independent of $\boldsymbol{\theta}$
$\hookrightarrow$ Since image space is the real line, can find $a < b$ s.t.

$$\mathbb{P}_{\boldsymbol{\theta}}[a \leq g(\mathbf{X}, \boldsymbol{\theta}) \leq b] = 1 - \alpha$$
$$\implies \mathbb{P}_{\boldsymbol{\theta}}[R(\mathbf{X}) \ni \boldsymbol{\theta}] = 1 - \alpha$$

where $R(\mathbf{x}) = \{\boldsymbol{\theta} \in \Theta : g(\mathbf{x}, \boldsymbol{\theta}) \in [a, b]\}$

Notice that region can be "wild" since it is a random fibre of g

Example

Let $\mathbf{X}_1, ..., \mathbf{X}_n \overset{iid}{\sim} \mathcal{N}_k(\boldsymbol{\mu}, \Sigma)$. Two unbiased estimators of $\boldsymbol{\mu}$ and Σ are

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i$$

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^n (\mathbf{X}_i - \hat{\boldsymbol{\mu}})(\mathbf{X}_i - \hat{\boldsymbol{\mu}})^T$$

Example (cont'd)

Consider the random variable

$$g(\{\mathbf{X}\}_{i=1}^n, \boldsymbol{\mu}) := \frac{n(n-k)}{k(n-1)}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^T\hat{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) \sim F\text{-dist with k and n-k d.f.}$$

A pivot!
↪ If f_q is q-quantile of this distribution, then get $100q\%$ CR as

$$R(\{\mathbf{X}\}_{i=1}^n) = \left\{\boldsymbol{\theta} \in \mathbb{R}^n : \frac{n(n-k)}{k(n-1)}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^T\hat{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) \le f_q\right\}$$

- An ellipsoid in $\mathbb{R}^n$
- 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!
↪ What if we construct a CR as Cartesian product of CI's?

Let $[L_i(\mathbf{X}), U_i(\mathbf{X})]$ be $100q_i\%$ CI's for θ_i, $i = 1, .., p$, and define

$$R(\mathbf{X}) = [L_1(\mathbf{X}), U_1(\mathbf{X})] \times \ldots \times [L_p(\mathbf{X}), U_p(\mathbf{X})]$$

Bonferroni's inequality implies that

$$\mathbb{P}_{\boldsymbol{\theta}}[R(\mathbf{X}) \ni \boldsymbol{\theta}] \ge 1 - \sum_{i=1}^{p} \mathbb{P}[\theta_i \notin [L_i(\mathbf{X}), U_i(\mathbf{X})]] = 1 - \sum_{i=1}^{p}(1-q_i)$$

→ So pick q_i such that $\sum_{i=1}^p (1-q_i) = \alpha$ (can be conservative...)

Confidence Intervals and Hypothesis Tests

Discussion on CR's → no guidance to choosing "good" regions

But: ∃ close relationship between CR's and HT's!
↪ exploit to transform good testing properties into good CR properties

Suppose $R(\mathbf{X})$ is an exact $100q\%$=$100(1-\alpha)\%$ CR for $\boldsymbol{\theta}$. Consider

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \qquad vs \qquad H_1 : \theta \neq \theta_0$$

Define test function:

$$\delta(\mathbf{X}) = \begin{cases} 1 & \text{if } \boldsymbol{\theta}_0 \notin R(\mathbf{X}), \\ 0 & \text{if } \boldsymbol{\theta}_0 \in R(\mathbf{X}). \end{cases}$$

Then, $$\mathbb{E}_{\boldsymbol{\theta}_0}[\delta(\mathbf{X})] = 1 - \mathbb{P}_{\boldsymbol{\theta}_0}[\boldsymbol{\theta}_0 \in R(\mathbf{X})] \le \alpha$$

Can use a CR to construct test with significance level α!

Confidence Intervals and Hypothesis Tests

Going the other way around, can invert tests to get CR's:

Suppose we have tests at level α for any choice of simple null, $\boldsymbol{\theta}_0 \in \Theta$.
↪ Say that $\delta(\mathbf{X}; \boldsymbol{\theta}_0)$ is appropriate test function for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$

Define $$R^*(\mathbf{X}) = \{\boldsymbol{\theta}_0 : \delta(\mathbf{X}; \boldsymbol{\theta}_0) = 0\}$$

Coverage probability of $R^*(\mathbf{X})$ is

$$\mathbb{P}_{\boldsymbol{\theta}}[\boldsymbol{\theta} \in R^*(\mathbf{X})] = \mathbb{P}_{\boldsymbol{\theta}}[\delta(\mathbf{X}; \boldsymbol{\theta}) = 0] \ge 1 - \alpha$$

Obtain a $100(1-\alpha)\%$ confidence region by choosing all the $\boldsymbol{\theta}$ for which the null would not be rejected given our data $\mathbf{X}$.
↪ If test inverted is powerful, then get "small" region for given $1-\alpha$.

Multiple Testing

Modern example: looking for signals in noise

- Interested in detecting presence of a signal $\mu(x_t)$, $t = 1, \dots, T$ over a discretised domain, $\{x_1, \dots, x_t\}$, on the basis of noisy measurements
- This is to be detected against some known background, say 0.
- May or may not be specifically interested in detecting the presence of the signal in some particular location x_t, but in detecting whether the a signal is present anywhere in the domain.

Formally:

Does there exist a $t \in \{1, \dots, T\}$ such that $\mu(x_t) \neq 0$?

or

for which t's is $\mu(x_t) \neq 0$?

Multiple Testing

More generally:

- Observe
$$Y_t = \mu(x_t) + \varepsilon_t, \qquad t = 1, \dots, T.$$
- Wish to test, at some significance level α:
$$\begin{cases} H_0 : \mu(x_t) = 0 & \text{for all } t \in \{1, \dots, T\}, \\ H_A : \mu(x_t) \neq 0 & \text{for some } t \in \{1, \dots, T\}. \end{cases}$$
- 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 the same trick as for confidence regions!

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:

$$\mathbb{P}[\cancel{H_0} | H_0] = \mathbb{P}\left[\bigcup_{t=1}^{T} \{\cancel{H_{0,t}}\} \middle| H_0\right] \leq \sum_{t=1}^{T} \mathbb{P}[\cancel{H_{0,t}} | H_0] = T\frac{\alpha}{T} = \alpha$$

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's Procedure

1. We reject $H_{0,t}$ for large values of a corresponding p-value, p_t
2. Order p-values from most to least significant: $p_{(1)} \leq \dots \leq p_{(T)}$
3. Starting from $t = 1$ and going up, reject all $H_{0,(t)}$ such that $p_{(t)}$ significant at level α / t. 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 Holm and 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,

$$\mathbb{P}\left[p_{(j)} \geq \frac{j\alpha}{T}, \text{ for all } j = 1, ..., T \,\middle|\, H_0\right] \geq 1 - \alpha$$

(strict equality requires continuous test statistics, otherwise $\leq \alpha$)

Yields 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: $p_{(1)} \leq \dots \leq p_{(T)}$
3. If, for some $j = 1, \dots, T$ 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 "localise" the signal at a particular x_t

Taking Advantage of Structure: Independence

One can, however, devise a sequential procedure to "localise" Sime's procedure, at the expense of lower power for the global hypothesis H_0:

Hochberg's procedure (assuming independence)

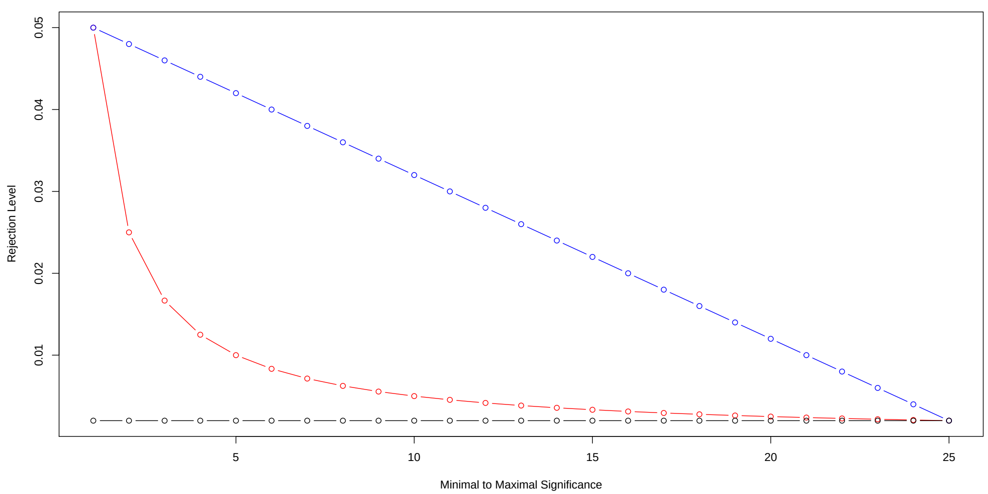
1. Suppose we reject $H_{0,j}$ for small values of p_j
2. Order p-values from most to least significant: $p_{(1)} \leq \dots \leq p_{(T)}$
3. Starting from $j = T, T-1, \dots$ and down, accept all $H_{0,(j)}$ such that $p_{(j)}$ insignificant at level α/j.
4. Stop accepting for the first j such that $p_{(j)}$ is significant at level α/j, and reject all the remaining ordered hypotheses past that j going down.

Genuine improvement over Holm-Bonferroni both overall (H_0) and in terms of signal localisation:

1. Rejects "more" individual hypotheses than Holm-Bonferroni
2. Power for overall H_0 "weaker" than Sime's (for $T > 2$), much "stronger" than Holm (for $T > 1$).

Taking Advantage of Structure: Independence

Bonferroni, Hochberg, Simes



Some Elements of Bayesian Inference

Statistical Theory

Victor Panaretos
Ecole Polytechnique Fédérale de Lausanne

1. θ as a random variable
2. Using Bayes' Rule to Update a Prior
3. Empirical Bayes
4. Choice of Prior: Informative, Conjugate, Uninformative
5. Inference in the Bayesian Framework

Back to Basics!

Classical Perspective

1. $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ is unknown fixed
2. data are realization of random $\mathbf{X} \sim f(\mathbf{x}; \boldsymbol{\theta})$
3. use data to learn about value of $\boldsymbol{\theta}$ (estimation, testing,...)

- Somehow assumes complete ignorance about $\boldsymbol{\theta}$
- What if have some prior knowledge/belief about plausible values?

$\hookrightarrow$ We did this when defining Bayes risk

$\hookrightarrow$ Placed a *prior* $\pi(\cdot)$ on θ

Bayesian Perspective

1. $\boldsymbol{\theta}$ is RANDOM VARIABLE with prior distribution $\pi(\cdot)$
2. data are realization of random $\mathbf{X}$ such that $\mathbf{X}|\boldsymbol{\theta} \sim f(\mathbf{x}|\boldsymbol{\theta})$
3. use data $\{\mathbf{X} = \mathbf{x}\}$ to UPDATE the distribution of $\boldsymbol{\theta}$

Updating a Prior

1. Have knowledge/belief about $\boldsymbol{\theta}$: expressed via $\pi(\boldsymbol{\theta})$
2. Observe data $\mathbf{X} = \mathbf{x}$

How do we readjust our belief incorporating this new evidence?

$\hookrightarrow$ Answer: use Bayes' rule to obtain a posterior for $\boldsymbol{\theta}$

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

since denominator is constant,

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto \underbrace{f(\mathbf{x}|\boldsymbol{\theta})}_{\text{Likelihood}}\underbrace{\pi(\boldsymbol{\theta})}_{\text{Prior}}$$

Bayesian Principle: Everything we want to learn about $\boldsymbol{\theta}$ from the data contained in the posterior

A Few Remarks

Some strengths of the Bayesian approach:
(provided we accept viewing $\boldsymbol{\theta}$ as random)

- Inferences on $\boldsymbol{\theta}$ take into account only observed data and prior
  - ↪ Contrast to classical approach which worries about all samples that could have but did not occur (risk)
- Provides unified approach for solving (almost) *any* inference problem
- Allows natural way to incorporate prior information in inference
- Posterior can become your prior for next exepriment!

(updating process: can unify a series of experiments naturally)

But... one basic weaknesses:

- Choice of prior? Objective priors typically NOT available...

Example (Coin Flips)

Let $(X_1, \ldots, X_n)|\theta \stackrel{iid}{\sim}$ Bernoulli(θ) and consider a Beta prior density on θ:

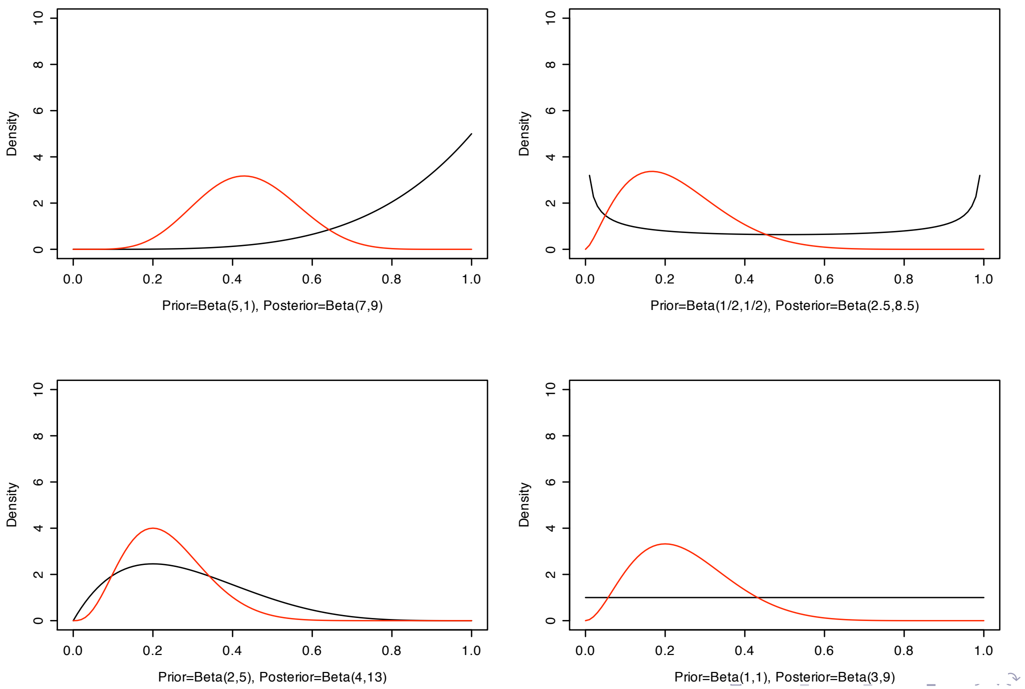
$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad \theta \in (0,1)$$

Given $X_1 = x_1, \ldots, X_n = x_n$ with $y = x_1 + \ldots + x_n$, have

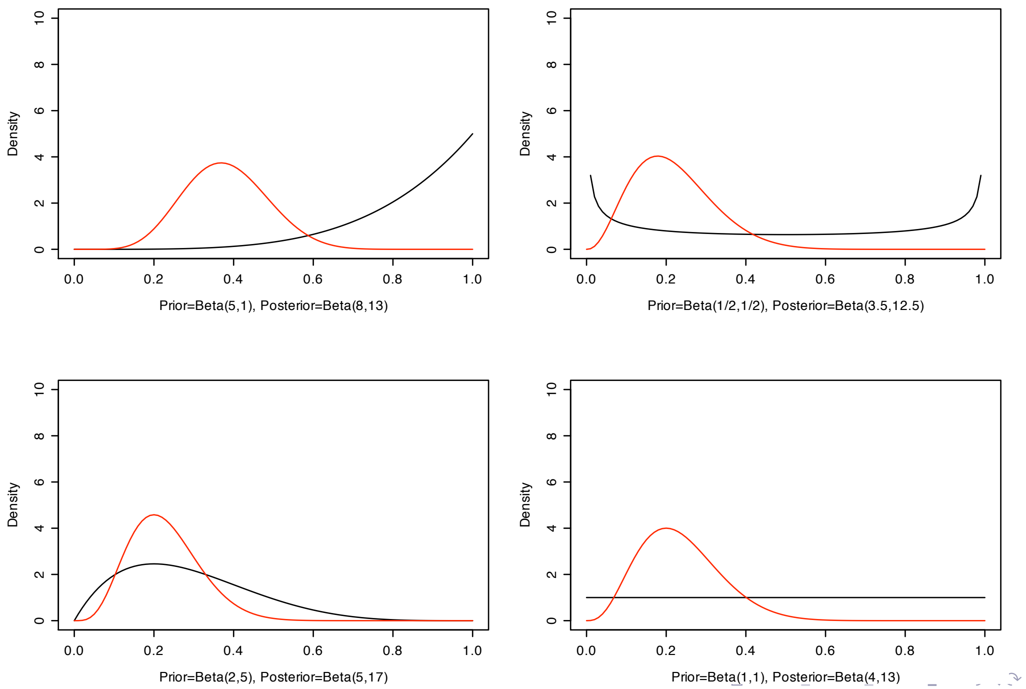
$$\begin{aligned}\pi(\theta|\mathbf{x}) &= \frac{f(x_1,\ldots,x_n|\theta)\pi(\theta)}{\int_0^1 f(x_1,\ldots,x_n|\theta)\pi(\theta)d\theta} \\ &= \frac{\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}}{\int_0^1 \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}d\theta} \\ &= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+y)\Gamma(\beta+n-y)}\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}\end{aligned}$$

Recall that $\Gamma(k) = (k-1)!$ for $k \in \mathbb{Z}^+$
↪ our choice of prior includes great variety of densities, including uniform distribution

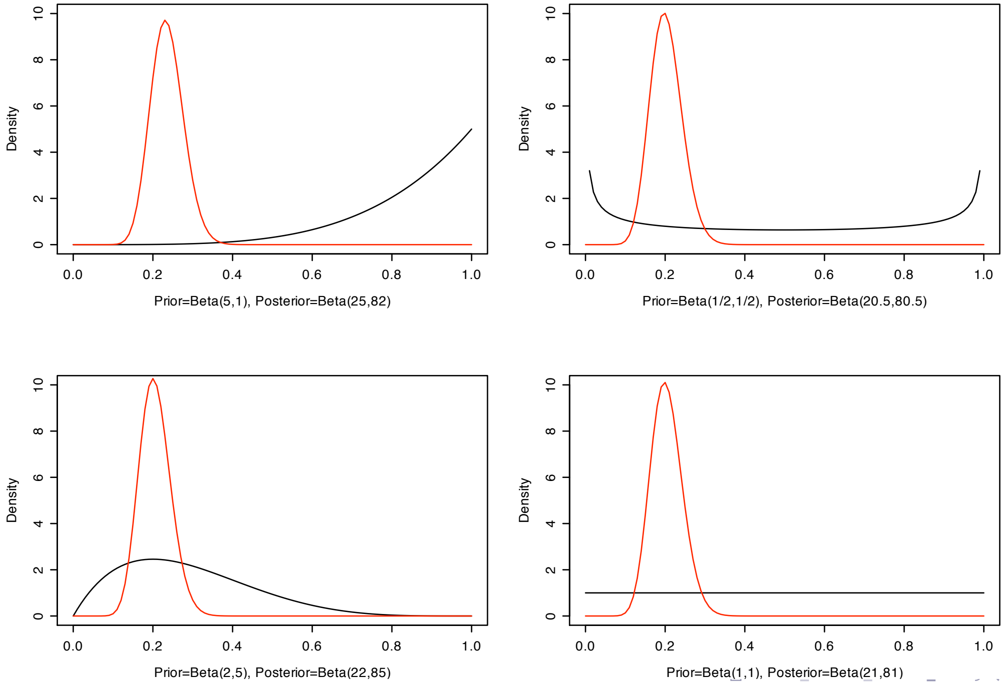
Coin Flipping Example: $n = 10$, $\sum x_i = 2$, prior vs posterior



Coin Flipping Example: $n = 15$, $\sum x_i = 3$, prior vs posterior



Coin Flipping Example: $n = 100$, $\sum x_i = 20$, prior vs posterior



Hyperparameters and Empirical Bayes

Typically $\rightarrow$ Prior depends itself on parameters (=*hyperparameters*)
$\hookrightarrow$ They are tuned to reflect prior knowledge/belief

- "Orthodox" Bayesians:
  1. Accept the role of subjective probability / a priori beliefs
  2. Hyperparameters should be specified independent of the data
- Empirical Bayes Approach:
  1. Not willing to a prior specify hyperparameters
  2. Tune prior do observed data (estimate hyperparameters from data)

(essentially a non-Bayesian approach since prior is tuned to data)

Empirical Bayes

Prior $\pi(\boldsymbol{\theta}; \boldsymbol{\alpha})$ depending on hyperparameter $\boldsymbol{\alpha}$. Write

$$\underbrace{g(\mathbf{x}; \boldsymbol{\alpha})}_{\text{marginal of } \mathbf{x}} = \int_{\Theta} \underbrace{f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}; \boldsymbol{\alpha})}_{\text{joint density of } \boldsymbol{\theta} \text{ and } \mathbf{x}} d\boldsymbol{\theta}$$

Marginal of $\mathbf{x}$ depends on α
$\hookrightarrow$ classical point estimation setting - estimate α!

- Maximum likelihood
- Plug in principle
- ...

$\hookrightarrow$ Then plug $\hat{\boldsymbol{\alpha}}$ into prior, and obtain posterior

$$\pi(\boldsymbol{\theta}|\mathbf{x}; \hat{\boldsymbol{\alpha}}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}; \hat{\boldsymbol{\alpha}})}{\int_{\Theta} f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}; \hat{\boldsymbol{\alpha}})d\boldsymbol{\theta}} = \frac{f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}; \hat{\boldsymbol{\alpha}})}{g(\mathbf{x}; \hat{\boldsymbol{\alpha}})}$$

Example

Let $X_1, ..., X_n|\lambda \overset{iid}{\sim} \text{Poisson}(\lambda)$ with a gamma prior on λ

$$\pi(\lambda; \alpha, \beta) = \frac{\alpha^\beta}{\Gamma(\beta)} \lambda^{\beta-1} \exp(-\alpha\lambda)$$

Letting $y = x_1 + \ldots + x_n$, observe that the marginal for $\mathbf{x}$ is

$$\begin{aligned} g(\mathbf{x}; \alpha, \beta) &= \int_0^\infty \frac{\exp(-n\lambda)\lambda^y}{x_1! \ldots x_n!} \pi(\lambda; \alpha, \beta) d\lambda \\ &= \left(\frac{\alpha}{n+\alpha}\right)^\beta \frac{\Gamma(y+\beta)}{\Gamma(\alpha)x_1! \ldots x_n!} \left(\frac{1}{n+\alpha}\right)^y \end{aligned}$$

So use ML estimation with $L(\alpha, \beta) = g(\mathbf{x}; \alpha, \beta)$. Alternatively, MoM:

$$\begin{aligned} \mathbb{E}[X_i] &= \mathbb{E}[\mathbb{E}(X_i|\lambda)] = \int_0^\infty \lambda\pi(\lambda; \alpha; \beta)d\lambda = \frac{\alpha}{\beta} \\ Var[X_i] &= \mathbb{E}[Var(X_i|\lambda)] + Var[\mathbb{E}(X_i|\lambda)] = \frac{\alpha}{\beta} + \frac{\alpha}{\beta^2} \end{aligned}$$

Informative, Conjugate and Ignorance Priors

The "catch" with Bayesian inference is picking a prior. Can be done:

1. By expressing prior knowledge/opinion (informative)
2. For convenience (conjugate)
3. As objectively as possible (ignorance)

Focus on (2) and (3).

The most convenient priors to work with are *conjugate families*

Definition (Conjugate Family)

A parametric family $\mathcal{P} = \{\pi(\cdot;\boldsymbol{\alpha})\}_{\boldsymbol{\alpha}\in\mathbf{A}}$ on $\boldsymbol{\Theta}$ is called a *conjugate family* for a family of distributions $\mathcal{F} = \{f(\cdot;\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}$ on $\mathcal{X}$ if,

$$\frac{f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta};\boldsymbol{\alpha})}{\int_{\boldsymbol{\Theta}} f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta};\boldsymbol{\alpha})d\boldsymbol{\theta}} = \pi(\boldsymbol{\theta}|\mathbf{x};\boldsymbol{\alpha}) \in \mathcal{P}, \qquad \forall\ \boldsymbol{\alpha}\in\mathbf{A}\ \&\ \mathbf{x}\in\mathcal{X}.$$

Informative, Conjugate and Ignorance Priors

Conjugate families: posterior immediately available
$\hookrightarrow$ great simplification to Bayesian inference

Example (Exponential Family)

Let $(X_1,\dots,X_n)|\theta$ follow a one-parameter exponential family,

$$f(\mathbf{x}|\theta) = \exp[c(\theta)T(\mathbf{x}) - d(\theta) + S(\mathbf{x})]$$

Consider prior: $\quad \pi(\theta) = K(\alpha,\beta)\exp[\alpha c(\theta) - \beta d(\theta)]$

$$\text{Posterior:}\quad \pi(\theta|\mathbf{x}) \propto \pi(\theta)f(\mathbf{x};\theta) \propto \exp[(T(\mathbf{x})+\alpha)c(\theta) - (\beta+1)d(\theta)]$$

So obtain a posterior in the same family as the prior

$$\pi(\theta|\mathbf{x}) = K(\underbrace{T(\mathbf{x})+\alpha}_{\alpha'}, \underbrace{\beta+1}_{\beta'})\exp[(\underbrace{T(\mathbf{x})+\alpha}_{\alpha'})c(\theta) - (\underbrace{\beta+1}_{\beta'})d(\theta)]$$

Informative, Conjugate and Ignorance Priors

Bayesian inference ofetn perceived as not objective.

$\downarrow$

Can we find priors that express an indifference on values of $\boldsymbol{\theta}$?

- If $\boldsymbol{\Theta}$ finite: easy, place mass $1/|\boldsymbol{\Theta}|$ on each point
- Infinite case?

Consider initially $\Theta = [a,b]$. Natural uninformative prior $\mathcal{U}[a,b]$
$\hookrightarrow$ Uninformative for θ. But what about $g(\theta)$?
$\hookrightarrow$ If $g(\cdot)$ non-linear, we are being informative about $g(\theta)$

What if $\boldsymbol{\Theta}$ not bounded? No uniform probability distribution:

$$\int_{\boldsymbol{\Theta}} k d\theta = \infty \quad \forall\ k > 0$$

Some "improper" priors (i.e. infinite mass) yield valid posterior densities

Example (Lebesgue Prior for Normal Distribution)

Let $X_1,\dots,X_n|\mu \sim \mathcal{N}(\mu,1)$. Assume prior π is "uniform" on $\mathbb{R}$,

$$\pi[a,b] = b - a = \int_a^b dx, \quad \forall\ a<b$$

(density 1 with respect to Lebesgue measure). Obtain posterior

$$\pi(\mu|\mathbf{x}) = k(\mathbf{x})\exp\left[-\frac{1}{2}\sum_{i=1}^n (x_i-\mu)^2\right]$$

$$k(\mathbf{x}) = \left(\int_{-\infty}^{+\infty}\exp\left[-\frac{1}{2}\sum_{i=1}^n(x_i-\tau)^2\right]d\tau\right)^{-1} = \sqrt{\frac{n}{2\pi}}\exp\left[\frac{1}{2}\sum_{i=1}^n(x_i-\bar{x})^2\right]$$

so the posterior is $\mathcal{N}(\bar{x},1/n)$.

Jeffreys' Prior

Invariance problem remains even with improper priors.
$\hookrightarrow$ Jeffreys (1961) proposed the following approach. Assume $\mathbf{X}|\theta \sim f(\cdot|\theta)$

1. Let g be monotone on Θ, define $\nu = g(\theta)$
2. Define $\pi(\theta) \propto |I(\theta)|^{1/2}$ (Fisher information)
3. Fisher information for ν:

$$
\begin{aligned}
I(\nu) &= \mathrm{Var}_\nu\left[\frac{d}{d\nu}\log f(\mathbf{X}; g^{-1}(\nu))\right] \\
&= \left(\frac{d}{d\nu}g^{-1}(\nu)\right)^2 \mathrm{Var}_\theta\left[\frac{d}{d\theta}\log f(\mathbf{X};\theta)\right] = \left|\frac{d\theta}{d\nu}\right|^2 \times I(\theta)
\end{aligned}
$$

4. Thus $|I(\theta)|^{1/2}d\theta = |I(\nu)|^{1/2}d\nu$

Gives widely-accepted solution to some standard problems.

Note that this prior distribution can be improper.

Computational Issues

When prior is conjugate: easy to obtain posterior

However for a general prior: posterior not easy to obtain

$\hookrightarrow$ Problems especially with evaluation of $\int_{\boldsymbol{\Theta}} f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$
Explicit intergration infeasible - alternatives

- Numerical Integration
- Monte Carlo methods
  - $\hookrightarrow$ Monte Carlo Integration
  - $\hookrightarrow$ Gibbs sampling
  - $\hookrightarrow$ Markov Chain Monte Carlo

Inference in the Bayesian Framework?

In principle: posterior contains everything you want to know

$\hookrightarrow$ So any inference really is just a descriptive measure of posterior
$\hookrightarrow$ Any descriptive measure contains less information than posterior

- Point estimators: posterior mean, mode, median,...
  - $\hookrightarrow$ Relate to Bayes decision rules
- Highest Posterior Density Regions (vs Confidence Regions)
  - $\hookrightarrow$ Different Interpretation from CRs!
- Hypothesis Testing: Bayes Factors