Econ 508B: Lecture 6

Statistical Theory

Hongyi Liu

Washington University in St. Louis

August 3, 2017

Outline

Outline

Motivation

In a statistical problem, there is not just one probability measure in question, but a whole family of measures P_θ, indexed by a parameter $\theta \in \Theta$.

- Note that P_θ is a notation different from the induced measure one, such as P_X.
- a classical example, ***exponential family*** is a broad class of distributions. That is, given dominating measure μ, an exponential family has density function (Radon-Nikodym derivative with respect to μ) of the form

$$p_\theta(x) = e^{\eta(\theta)\cdot T(x)+A(\theta)}h(x),$$

where $\eta(\theta), T(x), A(\theta)$ and $h(x)$ are some functions. There are a number of common distributions as exponential family distributions with natural parameters, such as binomial distribution, normal distribution, beta distribution, etc.

Question: given a family of probability measures, Could we induce other probability measures based on such a base measure?

Question: given a family of probability measures, Could we induce other probability measures based on such a base measure?

The common way of solving this problem is to incorporate a group of transportations.

Example 1.1

Let P_0 be a probability measure with symmetric density p_0 with respect to Lebesgue measure on $\mathbb{R}$. Symmetry implies that the median is 0; if the expected value exists, then it equals 0, too. For $X \sim P_0$, define $X' = X + \theta$ for some real number θ. Then the distribution of X' is $P_\theta(A) \equiv P_0(X + \theta \in A)$. Doing this for all θ generates the family $\{P_\theta : \theta \in \mathbb{R}\}$.

The normal family $N(\theta, 1)$ is a special case.

The family of distributions in the example above are generated by a single distribution, centered at 0, and a collection of 'location shifts'.

location shifts

There are four key properties of these shifts:

- Shifting by zero doesn't change anything.
- The result of any two consecutive location shifts can be achieved by a single location shift.
- The order in which location shifts are made is irrelevant.
- For any given location, there is a shift that takes the location back to 0.

It turns out that these properties characterize what is called a *group* of transformations.

Group transformation

To generalize the location shift example, start with a fixed probability measure P on $(\mathbb{X}, \mathcal{A})$.

- Now, introduce a group $\mathcal{G}$ of transformations on $\mathbb{X}$ and take $P_e = P$; the subscript 'e' refers to the the group identity e.
- Then define the family $\{P_g : g \in \mathcal{G}\}$ as

$$P_g(A) = P_e(g^{-1}A), \ A \in \mathcal{A}.$$

  That is, $P_g(A)$ is the probability, under $X \sim P_e$, that $g(X)$ lands in A.
- In this case where P_e has a density p_e with respect to Lebesgue measure, we have

$$p_g(x) = p_e(g^{-1}x) \left| \frac{dg^{-1}x}{dx} \right|$$

  the right hand size term is just the usual change of variable formula or is referred to as *Jacobian transformation matrix*.

Outline

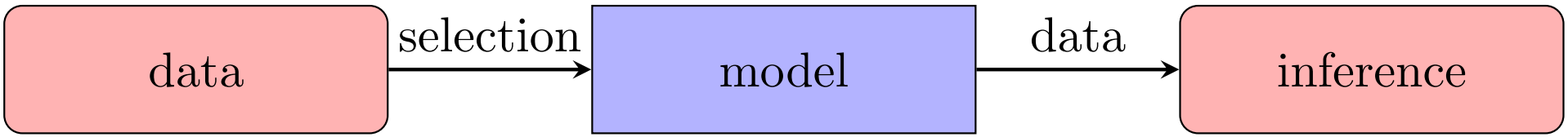
Classical Inference v.s. Post-Selection Inference

Classical Inference

Classical Inference v.s. Post-Selection Inference

Classical Inference

Post-Selection Inference



- Statisticians are tasked with turning the large amount of data generated by experiments and observations into inferences about the world
- Econometricians are tasked to "give empirical content to economic relations based on the concurrent development of theory and observation, related by appropriate methods of inference" (Samuelson, Koopmans, and Stone (1954), *Econometrica*).

Decision Theory Framework

Decision theory, developed by Abraham Wald (mid-20th century) furnishes us with a framework of statistical inference to address our core questions.

Hereafter, we view our data as a realization of a random variable X taking values in a sample space, say $\mathcal{X}$.

- Often X will be a factor $(X_1, ..., X_n)$ with i.i.d. components.
- In the decision theory framework, a decision problem will consist of three elements:
  (1) statistical model;
  (2) decision procedures;
  (3) loss function.

Statistical Model & Decision Procedure

Proposition 2.1 (Statistical Model)

A statistical model is a family of distributions $\mathcal{P}$, indexed by a parameter θ, denoted as $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$.

(1) θ is the parameter, Ω is the parameter space, and each P_θ is a distribution.

(2) Assume that the data X come from some $P_\theta \in \mathcal{P}$ but that the true θ is unknown.

Statistical Model & Decision Procedure

Proposition 2.1 (Statistical Model)

A statistical model is a family of distributions $\mathcal{P}$, indexed by a parameter θ, denoted as $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$.

(1) θ is the parameter, Ω is the parameter space, and each P_θ is a distribution.

(2) Assume that the data X come from some $P_\theta \in \mathcal{P}$ but that the true θ is unknown.

Proposition 2.2

Decision Procedure A decision procedure δ is a map from $\mathcal{X}$ (the sample space) to the decision space $\mathcal{D}$.

Loss Function & Risk Function

A *loss function* is a mapping $L : \Theta \times \mathcal{D} \to \mathbb{R}^+$.

- $L(\theta, d)$ represents the penalty for making the decsion d when θ is true.

Example 2.1

A common loss function is the squared error loss, namely $L(\theta, d) = (\theta - d)^2$.

- Especially in OLS, the loss function $(y_i - x_i\beta)^2$.
- On the dual space of loss function, we often consider the 'loss function' as a utility function.

Loss Function & Risk Function

A *loss function* is a mapping $L : \Theta \times \mathcal{D} \to \mathbb{R}^+$.

- $L(\theta, d)$ represents the penalty for making the decsion d when θ is true.

Example 2.1

A common loss function is the squared error loss, namely $L(\theta, d) = (\theta - d)^2$.

- Especially in OLS, the loss function $(y_i - x_i\beta)^2$.
- On the dual space of loss function, we often consider the 'loss function' as a utility function.

Indeed, the three components of a decision problem together give rise to our primary basis for evaluation, the **risk function**

$$R(\theta, \delta) = E_\theta(L(\theta, \delta(X))).$$

The risk $R(\theta, \delta)$ is the average loss incurred when the decision procedure δ is used and θ is true.

Risk Function

The risk function gives us a way to compare and rule out procedures.

- We say a procedure δ is inadmissible if another procedure never has greater risk than δ but sometimes has strictly lower risk.
- Decision theory rules out inadmissible procedures δ in favor of dominating procedures δ'.

Therefore, minimizing the risk function is a way of providing us the *admissible* or *optimal* procedure if it exists the minimization of loss function. For instance, this is how we define the linear projection coefficient in OLS regression model

$$\beta = \arg\min_{\beta \in \mathbb{R}^k} \mathbb{E}(y_i - x_i\beta)^2$$

Optimal Inference

However, searching for the uniformly best risk estimation directly cannot be done effectively in most of time.

- ***Constrain*** the set of decision procedures under consideration, by requiring our procedures to satisfy criteria like unbiasedness or invariance.
  - (i) Unbiasedness for estimating $g(\theta)$: $E_\theta(\delta(X)) = g(\theta)$, such as **budget constraint**.
  - (ii) Equivariance or invariance: enforce symmetries. For example, location invariance requires an estimator to satisfy $\delta(X + c) = \delta(X) + c$, such as **quasilinear utility**.
- ***Collapse*** the risk function into a single numerical summary, and minimize this overall summary of risk instead of requiring
  - Bayes procedures: minimize the average risk $\int R(\theta, \delta) d\pi(\theta)$ where $\pi(\theta)$ is a *prior distribution* on Ω.
  - Minimax procedures: minimize the worst-case risk $\sup_{\theta \in \Omega} R(\theta, \delta)$.

Outline

Data Reduction

Before constraining or collapsing, we first attend to a more basic consideration that will aid us in the design of optimal procedures: **Not all data is relevant**.

Definition 3.1 (Statistic)

Statistical theory defines a **statistic** as a function of a sample, i.e. $T : \mathcal{X} \to \mathcal{T}$.

Definition 3.2 (Sufficient Statistic)

A statistic is **sufficient** for a model $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ if for all t, the conditional distribution or density of sample given the statistic is free of the parameter θ, i.e., $X|T(x) = t$.

Example

Example 3.1 (A sufficient statistic for Bernoulli)

Consider an i.i.d. random sample $X_1, X_2, ..., X_n$ of size n, from the Bernoulli distribution with

$$P[X_i = 0] = 1 - p = 1 - p[X_i = 1]$$

where p is the population parameter. $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic.

Proof

The joint distribution of the random sample is

$$f(x_1, ..., x_n; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{(\sum_{i=1}^{n} x_i)}(1-p)^{(n-\sum_{i=1}^{n} x_i)}$$

and the joint probability distribution of $X_1 = x_1, ..., X_n = x_n$ and $T = \sum_{i=1}^{n} X_i = t$ is

$$f(x_1, ..., x_n, t; p) = p^t(1-p)^{(n-t)} \text{ where } t = \sum_{i=1}^{n} x_i$$

Note that $T = \sum_{i=1}^{n} X_i$ has the binomial distribution

$$P(T = t; p) = \binom{n}{t} p^t(1-p)^{(n-t)} \text{for } t = 0, 1, ..., n$$

Then the conditional distribution of the random sample, given $T = t$ is

$$f(x_1, ..., x_n; T = t; p) = \frac{f(x_1, ..., x_n, t; p)}{P(T = t; p)} = \frac{p^t(1-p)^{(n-t)}}{\binom{n}{t} p^t(1-p)^{(n-t)}} = \frac{1}{\binom{n}{t}}$$

which does not depend on p. Then $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic.

Other examples

Example 3.2

- *Maximum of Uniform:* Let $X_1, X_2, ..., X_n$ be i.i.d. uniform, $U[0, \theta]$. Then $T(x) = \max(X_1, ..., X_n)$ is sufficient.
- *Order Statistics:* Let $X_1, ..., X_n$ be i.i.d. with any model. The *order statistics* $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ are sufficient.

Why data reduction?

Theorem 3.1

If $X \sim P_\theta \in \mathcal{P}$ and T is sufficient for $\mathcal{P}$, then any decision procedure δ has a corresponding $\hat{\delta}(T(X))$ with equal risk.

- To see that, sample X' from conditional distribution $P(X|T=t)$ and define $\hat{\delta}(T(X)) = \delta(X')$. It suffices since $X|T=t$ is free of θ.

- Data reduction via sufficient statistics can't hurt, plus (we will see that) irrelevant attributes can lead to increased risk.
- Data reduction can decrease computational burden.
- Data reduction can increase interpretability.

Neyman-Fisher Factorization Criterion

Although the definition of sufficiency is often difficult to work with directly, there is a much simpler characterization of sufficiency whenever model distributions admits densities w.r.t. a common σ-finite measure.

Theorem 3.2 (Neyman-Fisher Factorization Criterion(NFFC))

Suppose each $P_\theta \in \mathcal{P}$ has density $p(x;\theta)$ w.r.t. a common σ-finite measure μ, i.e. $\frac{dP_\theta}{d\mu} = p(x;\theta)$. Then $T(X)$ is sufficient if and only if $p(x;\theta) = g_\theta(T(x))h(x)$ for some g_θ, h.

Example

Example 3.3 (Normal distribution)

Consider i.i.d. $X_i \sim N(\mu, \sigma^2), i = 1, ..., n$ and $\theta = (\mu, \sigma^2)$. The joint distribution is

$$\begin{aligned} p(x;\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \\ &= (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\{\frac{1}{2\sigma^2}(-\sum_{i=1}^{n} x_i^2 + 2\mu\sum_{i=1}^{n} x_i - n\mu^2)\} \\ &= g_\theta[T(x)]. \end{aligned}$$

where $T(x) = (\sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} x_i)$ is sufficient and is a 2 dimensional vector, including sufficient information of θ rather than $(X_1, ..., X_n)$.