STAT 611-600

Theory of Inference

Lecture 5: Point Estimation

Tiandong Wang

Point Estimation

What is a statistical model? A Probability Model is a triple

$$
\begin{aligned}
S &= \text{sample space} \\
\mathcal{A} &= [\text{ events; subsets of } S\ ] \\
\mathbb{P} &= \text{rule for assigning probabilities to events} \\
X, Y, Z \ldots &\text{rv's; ie. functions on } S.
\end{aligned}
$$

A Statistical Model is a family of probability models:

$$
\begin{aligned}
S &= \text{sample space} \\
\mathcal{A} &= [\text{ events }] \\
\{\mathbb{P}_\theta, \theta \in \Theta\} &= \text{parametric family of probabilties} \\
X, Y, Z \ldots &\text{rv's; ie. functions on } S.
\end{aligned}
$$

Note θ could be multidimensional.

Examples:

- Suppose we have an experiment where we want the outcome exponentially distributed:

$$\begin{aligned} S =& [0, \infty) \\ \mathcal{A} =& \text{ subsets of positive numbers} \\ \{\mathbb{P}_\lambda, \lambda > 0\} =& \text{ probabilitiy assignments.} \end{aligned}$$

Here

$$\mathbb{P}_\lambda([0, x]) = 1 - e^{-\lambda x}, \quad x > 0,$$

and

$$\theta = \lambda, \qquad \Theta = [0, \infty).$$

So Θ is a one-dimensional set.

- Suppose we have a normally distributed experiment with

$$S = \mathbb{R}, \qquad \mathbb{P}_{\mu,\sigma}$$

where

$$\mathbb{P}_{\mu,\sigma}(-\infty, x] = N(x; \mu, \sigma^2).$$

Then $\theta = (\mu, \sigma)$ and

$$\Theta = (-\infty, \infty) \times [0, \infty) = \{(u, v) : -\infty < u < \infty, v \geq 0\}.$$

So Θ is a 2-dimensional set.

- Build a statistical model for the experiment: Randomly sample a normal variable n times.

$$S = \mathbb{R}^n = \{(x_1, \ldots, x_n) : x_i \in \mathbb{R},\ i = 1, \ldots, n\},$$
$$\mathcal{A} = \text{ events built from multi-dimensional rectangles,}$$
$$\{\mathbb{P}_{\mu,\sigma};\ (\mu, \sigma) \in \mathbb{R} \times [0, \infty)\} = \text{ assigment of probabilities.}$$

Here we define

$$\mathbb{P}_{\mu,\sigma}\Big(\{(u_1, \ldots, u_n) : u_i \leq x_i;\ i = 1, \ldots, n\}\Big) = \prod_{i=1}^{n} N(x_i; \mu, \sigma).$$

Goal in estimation

The Statistics Fairy Tale

- Once upon a time there was a **true model**.
  - This **true model** gave us data $x_1, \ldots, x_n$ representing observations from random variables $X_1, \ldots, X_n$.
  - We want to find this **true model**. We want to identify the distribution of $X_1, \ldots, X_n$.
- Often, we assume
  - the data is from $X_1, \ldots, X_n$ where $X_1, \ldots, X_n$ is a random sample; ie, iid from $\mathbb{P}_\theta$ for some $\theta \in \Theta$. (This requires replication to make sense.)
  - a linear model (later).
  - standard time series model.
- We consult histograms, boxplots, probability plots, acf plots and crystal balls to help decide the correct model class.
- Use the data to estimate θ which is tantamount to making a guess what the true model is; that is, the model which produced the data.

Point estimate for θ:

Use a known function of the data–a statistic $g(x_1, \ldots, x_n)$–that can reasonably represent the unknown parameter.

Examples of statistics:

- mean
- median
- sample acf
- IQR

Point estimate for θ:
Use a known function of the data–a statistic $g(x_1, \ldots, x_n)$–that can reasonably represent the unknown parameter.

Examples of statistics:

- mean
- median
- sample acf
- IQR

An **estimator** of θ is a function $g(X_1, \ldots, X_n)$; that is a statistic. Frequently this random variable is denoted

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n).$$

An **estimate** of θ is the value of the function

$$g(x_1, \ldots, x_n) = \hat{\theta}(x_1, \ldots, x_n)$$

when we observe

$$X_1 = x_1, \ldots, X_n = x_n.$$

Note:

- The estimator is the prescription that we have before we carry out the experiment and observe $X_1 = x_1, \ldots, X_n = x_n$.
- The estimate is what we get after the observations are collected.

Note:

- The estimator is the prescription that we have before we carry out the experiment and observe $X_1 = x_1, \ldots, X_n = x_n$.
- The estimate is what we get after the observations are collected.

Example: Republican (0) or Democrat (1)?
A random sample of size 400 from the population of eligible voters is taken.

> *(Note: in practice we sample without replacement and neglect the difference between sampling with and without replacement because 400 is so small relative to the size of the entire population.)*

Suppose 202 claim they will vote for Democrat. Estimate p the proportion of the population that will vote Democrat.

The class of models is binomial with $0 < p < 1$. We observe $X_1, \ldots, X_{400}$ where each X_i is 0 (vote for Republican) or 1 (vote for Democrat). We have a class of models indexed by p where $\Theta = (0, 1)$. Then

$$X = X_1 + \cdots + X_{400} \sim Bin(n, p), \quad p \in \Theta = (0, 1).$$

Before we sample, an estimator of p is

$$\hat{p} = \hat{p}(X_1, \ldots, X_n) = X/n = \bar{X}.$$

After we sample, the estimate is

$$\hat{p} = 202/400 = 0.505.$$

(How do we quantify the error in the estimate?)

Comments.

1. There are typically a variety of estimators for a parameter. For example, if we estimate the normal mean, could use
   - Sample mean $\bar{X}$.
   - Sample median $\tilde{X}$.
   - 20% trimmed mean: Take the data, trim off the 20% largest and 20% smallest observations and average the rest.

   How does one choose between competing estimators?
2. Sometimes one wants to estimate *population characteristics* rather than θ directly. By a population characteristic, we mean a function of the density or pmf (which typically leads to a function of θ). Here is the setup:
   Given $X_1, \ldots, X_n$ is iid with common density or pmf $\{f_\theta(x), \theta \in \Theta\}$.
   Some population characteristics are
   - Population (theoretical) mean

$$\mathbb{E}_\theta(X_1) = \int x f_\theta(x) dx \text{ (in the continuous case).}$$

For example if we have an exponential model

$$f_\lambda(x) = \lambda e^{-\lambda x} 1_{[0,\infty)}(x), \quad \lambda \in \Theta = (0, \infty)$$

then

$$\mathbb{E}_\lambda(X_1) = \int_0^\infty x\lambda e^{-\lambda x} 1_{[0,\infty)}(x) dx = \frac{1}{\lambda}.$$

- Population (theoretical) variance:

$$\text{Var}_\theta(X_1) = \int (x - \mathbb{E}_\theta(X_1))^2 f_\theta(x) dx.$$

  For the exponential model, the population variance is

$$\frac{1}{\lambda^2}.$$

  For the normal model the population variance is σ^2.
- Population percentiles: Solution of

$$F_\theta(\eta(p)) = p = \int_{-\infty}^{\eta(p)} f_\theta(x) dx.$$

Possible estimators of population characteristics:

population characteristic	possible estimator
mean	$\bar{X}$ sample median 20% trimmed mean
population variance	$\mathcal{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
population median	sample median
population perentile	sample percentile (order statistic)

Imagine now the probability specifications are given by the distribution function (not the pmf or pdf). In general if we want to estimate some function of $F_\theta(x)$ (mean, variance, percentile, higher moments) we can always try to estimate the characteristic by replacing $F_\theta(x)$ with the empirical distribution

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

and computing the characteristic with respect to $\hat{F}_n(x)$. The idea is that since $\hat{F}_n(x)$ is close to the true distribution, the function of $\hat{F}_n(x)$ should be close to the desired population characteristic.

Imagine now the probability specifications are given by the distribution function (not the pmf or pdf). In general if we want to estimate some function of $F_\theta(x)$ (mean, variance, percentile, higher moments) we can always try to estimate the characteristic by replacing $F_\theta(x)$ with the empirical distribution

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{[X_i \le x]},$$

and computing the characteristic with respect to $\hat{F}_n(x)$. The idea is that since $\hat{F}_n(x)$ is close to the true distribution, the function of $\hat{F}_n(x)$ should be close to the desired population characteristic.

Note $\hat{F}_n(x)$ is the discrete distribution function placing probability mass $1/n$ at each possible value X_i. The pmf corresponding to $\hat{F}_n(x)$ is

possible values	X_1	X_2	...	X_n
probability	1/n	1/n	...	1/n

After we observe the data, the capital letters $\{X_i, 1 \le i \le n\}$ get replaced by the observed values $\{x_i, 1 \le i \le n\}$.

So for example:

- Estimate
$$E_\theta(X_1) = \int x f_\theta(x) dx = \int x \, dF_\theta(x)$$
by
$$\int_{-\infty}^{\infty} x d\hat{F}_n(x) = \sum_{i=1}^{n} \frac{1}{n} X_i = \bar{X},$$
since the mean of the discrete distribution is the sum of the possible values of the function times the probability of obtaining the possible value.
- Estimate
$$\mathrm{Var}_\theta(X_1) = \int (x - \mathbb{E}_\theta(X_1))^2 f_\theta(x) dx = \int (x - \mathbb{E}_\theta(X_1))^2 dF_\theta(x),$$
by
$$\int_{-\infty}^{\infty} (x - \bar{X})^2 d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$
which is the small variant of the sample variance.
- Estimate the median $F_\theta^{\leftarrow}(1/2)$ by $\hat{F}_n^{\leftarrow}(1/2)$= sample median.

Desirable Properties of Estimators

- There may be more than one estimator for a parameter or population characteristic.
  - Should we estimate a normal mean with sample mean, sample median, ...
- Need a list of desirable properties of estimators which may allow us to prefer one estimator over another.
- This may not always lead inexorably to a choice, but helps. Guidelines help.

Desirable Properties of Estimators

- There may be more than one estimator for a parameter or population characteristic.
  - Should we estimate a normal mean with sample mean, sample median, ...
- Need a list of desirable properties of estimators which may allow us to prefer one estimator over another.
- This may not always lead inexorably to a choice, but helps. Guidelines help.

Examples:

1. For the population variance, should we use
   - Sum of squares divided by $n-1$:

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

or $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ as suggested in the previous section?

2. Consider the parametric family: uniform on $[0, \theta]$:

$$f_\theta(x) = \frac{1}{\theta} 1_{[0,\theta]}(x), \qquad \theta > 0.$$

Should we estimate θ with

$$\hat{\theta}_1(X_1, \ldots, X_n) = \max\{X_1, \ldots, X_n\},$$

or

$$\hat{\theta}_2(X_1, \ldots, X_n) = \max\{X_1, \ldots, X_n\} + \frac{1}{n},$$

on the theory that $\hat{\theta}_1$ surely under-estimates the true θ?

2. Consider the parametric family: uniform on $[0, \theta]$:

$$f_\theta(x) = \frac{1}{\theta} 1_{[0,\theta]}(x), \qquad \theta > 0.$$

Should we estimate θ with

$$\hat{\theta}_1(X_1, \ldots, X_n) = \max\{X_1, \ldots, X_n\},$$

or

$$\hat{\theta}_2(X_1, \ldots, X_n) = \max\{X_1, \ldots, X_n\} + \frac{1}{n},$$

on the theory that $\hat{\theta}_1$ surely under-estimates the true θ?

List of nice properties.

- Unbiasedness
- Small variance
- Desirable large sample behavior.
- Desirable precision: narrow confidence interval or small standard error.

How to get point estimators?

How do we generate sensible estimators for model parameters. Two simple and common methods are:

- Method of moments estimators (MME's); replace $F(x)$ by $\hat{F}_n(x)$.
- Maximum likelihood estimators (MLE's).

MME.

General philosophy: If $\boldsymbol{\theta}$ is d-dimensional, equate d different population characteristics with the corresponding sample characteristics. Solve for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$.

For example:

- If $d = 1$, equate

$$\mathbb{E}_\theta(X_1) = \underbrace{\int x f_\theta(x) dx}_{\substack{\text{population mean} \\ =\text{theoretical mean}}} = \underbrace{\int x d\hat{F}_n(x)}_{\substack{\text{sample mean} \\ =\text{empirical mean}}} = \bar{X},$$

  and solve for θ.

- For general d: equate for $r = 1, \ldots, d$,

$$\mathbb{E}_\theta(X_1^r) = \underbrace{\int x^r f_\theta(x) dx}_{\text{pop } r\text{th moment}} = \underbrace{\int x^r d\hat{F}_n(x)}_{\text{sample } r\text{th moment}} = \frac{1}{n}\sum_{i=1}^{n} X_i^r;$$

  and solve the d equations for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$.

Features:

- This is a recipe which provides a method. This is a way to proceed if other methods are complex or computationally expensive.
- Resulting estimators may or may not be optimal
- The method is often simple.
- Occasionally the method gives dumb estimators–but this is true of most methods.
- Ambiguity: If, for example, $d = 1$, why not solve

$$\mathbb{E}_\theta(X_1^{17}) = \frac{1}{n}\sum_{i=1}^{n} X_i^{17}$$

for θ. This makes equally good sense as the stated recipe.

Examples.

1. **Poisson.** Suppose $Z_1, \ldots, Z_n$ are a random sample from the Poisson mass function

$$p(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \qquad k \geq 0.$$

Then the theoretical = population mean is

$$\mathbb{E}(Z_1) = \lambda$$

and equating this with the sample characteristic = sample mean = $\bar{Z}$ yields the equation

$$\lambda = \bar{Z} = \frac{1}{n} \sum_{i=1}^{n} Z_i.$$

The solution is $\hat{\lambda}_{MME}$:

$$\hat{\lambda}_{MME} = \bar{Z}.$$

2. $N(\mu, \sigma^2)$. Suppose $X_1, \ldots, X_1$ are a random sample from $N(\mu, \sigma^2)$. Now there are 2 parameters

$$\boldsymbol{\theta} = (\mu, \sigma^2)$$

and there must be 2 equations:

$$\mu = \mathbb{E}_{\mu,\sigma^2}(X_1) = \bar{X}, \qquad \text{(1st equation)}$$

$$\mu^2 + \sigma^2 = \mathbb{E}_{\mu,\sigma^2}(X_1^2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \qquad \text{(2nd equation).}$$

This leads to $\hat{\mu}_{MME} = \bar{X}$ and

$$\hat{\sigma}^2_{\text{MME}} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Note $\hat{\sigma}^2_{\text{MME}}$ is the biased estimate of variance.