CE 211
Mathematics for Engineers

Lecture 10

Statistical Estimation

Previously on Mathematics for Engineers

Definition (Convergence in Distribution)

A sequence of random variables $\{X_n\}$ converges to X in distribution and is denoted as $X_n \xrightarrow{d} X$ or $X_n \xrightarrow{\mathcal{D}} X$ if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for all x where $F_X(x)$ is continuous.

Definition (Convergence in Probability)

A sequence of random variables $\{X_n\}$ converges to X in probability and is denoted as $X_n \xrightarrow{p} X$ if

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0$$

for all $\epsilon > 0$.

Previously on Mathematics for Engineers

Claim (WLLN)

Let $\{X_n\}$ be a sequence of iid random variables with finite mean μ.

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \xrightarrow{p} \mu$$

Claim (CLT)

Let $\{X_n\}$ be a sequence of iid random variables with expected value $\mu < \infty$ and variance $\sigma^2 < \infty$ and also suppose. $Z_n = X_1 + X_2 + \ldots + X_n$ Then,

$$\frac{Z_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

Lecture Outline

1. The Problem
2. Point Estimation
3. Interval Estimation
4. A Solution

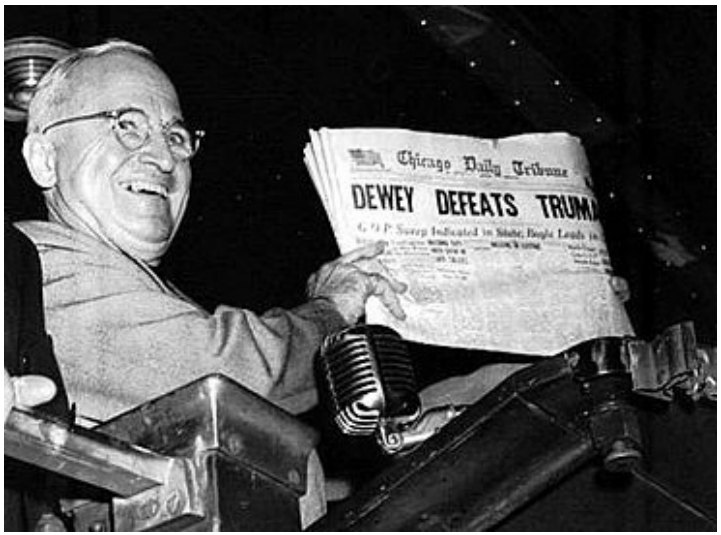
Lecture Outline

The Problem

The Problem

In the 1948 US presidential election, republican candidate Dewey was predicted to win against the incumbent democratic candidate Truman.

Sampled polls from several places motivated the Chicago Daily Tribune to go to press with a headline 'Dewey Defeats Truman' the night before the results were announced.



However, Truman scored a landslide victory with 303 electoral votes against Dewey's 189. What could have gone wrong with the sampled polls?

Lecture Outline

Point Estimation

Point Estimation

Introduction

So far, we were able to analytically analyze the odds with which a certain outcome or set of outcomes could occur.

We made assumptions of the knowledge of the parameters of the problem such as the success probability or the arrival rate etc.

Statistics involves the study of the inverse problem. If we see a certain set of outcome(s), what can we say about the random process or what can we infer about the experiment.

This process requires results from repeated trails and each trial can be treated as a random variable $X_1, X_2, \ldots$.

Point Estimation

Introduction

Suppose the random variables $X_1, X_2, \ldots$ follow the same distribution. Imagine that we sample from that distribution to get realizations of these random variables.

Point estimation procedures can be used to answer questions of the following type:

- What is the mean and variance of the random variable?
- What are the parameters of the random variable?

Parameters here are scalars used in the PDF and CDF. For example,

- Success probability of a Binomial random variable
- Mean and variance of a normal and Poisson distribution (They are parameters too)
- α and β of Weibull distribution

We will use θ to denote a generic parameter which could represent the mean and variance or the parameters of the distribution.

Let us first estimate the mean and variance of a random variable. Suppose we have a random sample $X_1, X_2, \ldots, X_n$, define another random variable $\hat{\Theta}_n$ as a function of the random variables $X_1, X_2, \ldots, X_n$.

Point Estimation

Introduction

Imagine the travel time on a street was lognormal and we are interested in its mean and variance. Suppose, we make the following measurements 5,7,3,4, and 7.

To estimate the mean of the random variable μ, we will define $\hat{\Theta}_n$ as

$$\hat{\Theta}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$\hat{\Theta}_n$ is a random variable since it is a function of random variables. Every set of measurements of Xs we make will give us a realization of $\hat{\Theta}_n$.

Likewise, to estimate the variance σ^2, we define

$$\hat{\Theta}_n = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \frac{X_1 + X_2 + \ldots + X_n}{n} \right)^2$$

Point Estimation

Properties

The specific choice of the estimator $\hat{\Theta}$ for a parameter θ is governed by a few desirable properties.

Definition (Bias)

An estimator $\hat{\Theta}_n$ of θ is unbiased if the bias $B(\hat{\Theta}_n) = \mathbb{E}(\hat{\Theta}_n) - \theta$ is 0.

Definition (MSE)

Given an estimator $\hat{\Theta}_n$ of θ, the mean squared error of the estimator is defined as $MSE(\hat{\Theta}_n) = \mathbb{E}\Big((\hat{\Theta}_n - \theta)^2\Big)$

Definition (Consistency)

An estimator $\hat{\Theta}_n$ of θ, is said to be consistent if $\hat{\Theta}_n \xrightarrow{p} \theta$

A **consistent** estimator with **zero bias** and **lower MSE** is always preferred.

Point Estimation

Estimator for Mean

In the example earlier, we can think of several candidate estimators of the mean μ. For example,

$$\hat{\Theta}_n = X_1$$

$$\hat{\Theta}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Are the above estimators consistent and unbiased? Which of the two estimators have lower MSE?

Point Estimation

Estimator for Standard Deviation

To estimate the variance σ^2 we proposed the following estimator

$$\hat{\Theta}_n = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \frac{X_1 + X_2 + \ldots + X_n}{n} \right)^2$$

Is this estimator unbiased? That is, is $\mathbb{E}(\hat{\Theta}_n) = \sigma^2$?

So far, we have looked at estimators for mean and variance. What if we want to estimate the parameters of the random variable instead.

The maximum likelihood estimator selects parameters such that the probability of realizing the observed data is maximized.

For example, consider the Binomial random variable in which we toss a coin n times. We do not know if our coin is unbiased and suppose we want to determine p.

If k out of the n trials result in H, it is natural to suppose k/n is the probability of H.

Let's now derive this result using the idea of maximizing the likelihood. Suppose p is unknown and we observed k Hs from n trials.

The probability of observing k heads is given by

$$\mathcal{L}(p) = \binom{n}{k} p^k (1-p)^{n-k}$$

We now maximize this probability assuming p is the unknown or the decision variable.

$$\frac{d\mathcal{L}}{dp} = \binom{n}{k} [k(1-p)^{n-k} p^{k-1} - p^k (n-k)(1-p)^{n-k-1}] = 0$$

Solving this we get $p = k/n$

Consider another example of estimating the parameter of a Poisson random variable. Suppose that in different one-hour intervals there are 3, 5, 1, 2, and 8 arrivals at an ATM. What is the average rate of arrivals λ?

Given a λ, the probability of observing 3 arrivals is

$$\frac{\lambda^3 e^{-\lambda}}{3!}$$

What is the probability of observing the data, i.e., observing 3, 5, 1, 2, and 8 arrivals.

$$\mathcal{L}(\lambda) = \frac{\lambda^3 e^{-\lambda}}{3!} \frac{\lambda^5 e^{-\lambda}}{5!} \frac{\lambda^1 e^{-\lambda}}{1!} \frac{\lambda^2 e^{-\lambda}}{2!} \frac{\lambda^8 e^{-\lambda}}{8!}$$

We now select a λ which maximizes the above probability.

Maximum likelihood estimation always involves maximization of probabilities. For a larger data set this would lead to more complex expressions for the objective.

Instead, we take the logarithm of it and maximize the log-likelihood. Such a transformation does not change the optimal solution since the log function is strictly increasing.

$$\begin{aligned}\mathcal{LL}(\lambda) &= \ln \frac{\lambda^3 e^{-\lambda}}{3!} \frac{\lambda^5 e^{-\lambda}}{5!} \frac{\lambda^1 e^{-\lambda}}{1!} \frac{\lambda^2 e^{-\lambda}}{2!} \frac{\lambda^8 e^{-\lambda}}{8!} \\ &= 19 \ln \lambda - 5\lambda - \text{Constant}\end{aligned}$$

Hence, the optimal λ is $19/5$. This is also the point estimate of the mean!

For continuous random variables, we use the PDF function at each data point and multiply them to derive the likelihood function.

As an example, suppose we want to estimate the parameters μ and σ of a lognormal distribution. Assume that we observe realizations $x_1, x_2, \ldots, x_n$.

The likelihood objective can then be written as

$$\mathcal{L}(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{x_i \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln x_i - \mu)^2}{2\sigma^2} \right)$$

The log-likelihood takes the form

$$\mathcal{LL}(\mu, \sigma) = \sum_{i=1}^{n} \ln\left(\frac{1}{x_i \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln x_i - \mu)^2}{2\sigma^2} \right) \right)$$

The objective is to thus maximize $\mathcal{LL}(\mu, \sigma)$ by changing both μ and σ subject to $\sigma > 0$.

Interval Estimation

Interval Estimation

Introduction

We have discussed methods to estimate the mean of a random variable using samples by constructing a realization of an estimator.

If we selected another set of samples, it is very likely that our estimate of the mean is going to be different.

Using sample data is also possible to provide an interval which might contain the mean with a certain degree of confidence.

Interval Estimation

Introduction

Consider the following example for reference. Assume that we wish to estimate the mean income of individuals in Bangalore.

The exact answer involves collecting data from everyone in the city, which is clearly prohibitive.

Instead, suppose we sample individuals and ask them their income and take its average. How close or far is this from the true population mean?

Interval Estimation

Confidence Interval

For the simple case of finding a confidence interval for the mean, using CLT, we know that

$$\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

converges to the standard normal in distribution. Hence, from standard normal tables,

$$\mathbb{P}\Big(-1.96 \leq \frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \leq 1.96\Big) = 0.95$$

$$\mathbb{P}\Big(\hat{\Theta}_n - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\Theta}_n - 1.96\frac{\sigma}{\sqrt{n}}\Big) = 0.95$$

The RHS is written as $1-\alpha$, where $\alpha = 0.05$ and the interval $[\hat{\Theta}_n - 1.96\frac{\sigma}{\sqrt{n}}, \hat{\Theta}_n + 1.96\frac{\sigma}{\sqrt{n}}]$ is described as the $100(1-\alpha)\%$ **Confidence Interval** for the mean.

Interval Estimation

Example

To estimate the confidence interval, we need to know the variance of the random variable which may not be available.

For large n (typically greater than 40), the variance can be replaced with its estimator which is the sample standard deviation as shown earlier.

Suppose, the sample average of income from 400 samples is ₹20,000 and the standard deviation is ₹5,000. Calculate a 95% confidence interval for the true population mean.

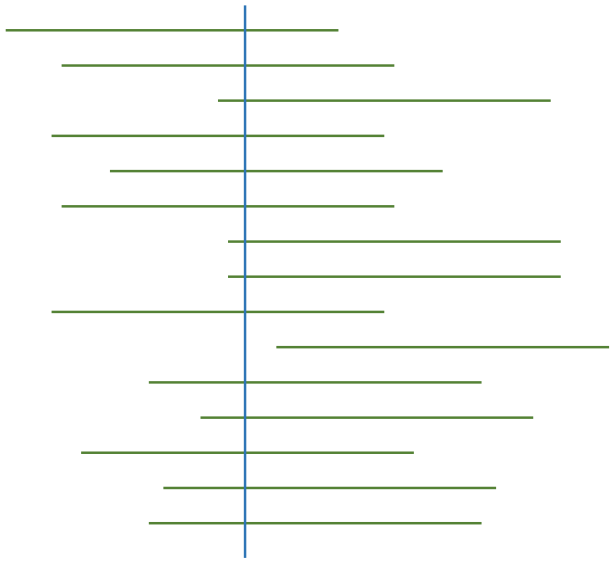
Interval Estimation

Interpretation

From the following equation, confidence intervals are often interpreted that the true population means lies in the CI with 95% probability.

$$\mathbb{P}\left(\hat{\Theta}_n - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\Theta}_n - 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

This interpretation is wrong since the true population mean is not a random variable. The right way to look at a CI is that if we do this experiment by considering another sample and construct another CI and keep repeating it, 95% of such intervals will contain the true population mean.



A Solution

A Solution

Sampling Bias

Most statistical procedures face an issue called sampling bias. A sample that is a good representation of the population is hard to find.

Thus, the process of selecting a sample for the estimator can skew the parameter estimates.

The election polls were conducted mainly using telephones as it was convenient to get large samples. However, in 1948 republicans were more likely to have telephones than democrats!

Your Moment of Zen

"Mister Jackson! You know how I feel about sampling."