Lecture 10: Estimation

- Introduction to Estimation
- Point Estimation and Unbiasedness
- Minimum Variance Unbiased Estimators
- Robust Estimators

Introduction

– Today we will begin our introduction to statistics/statistical inference.

– Usually in practice, we observe data first without knowing the true distribution. Then the question of interest becomes: given the observed data, what can we say about the underlying population?

– To do that we will first learn about estimators and desirable properties of estimators.

– We will also learn about unbiased estimators, robust estimators and minimum variance unbiased estimators.

Examples of Estimation

- **Biased coin:** We have a biased coin and we want to estimate the probability of getting heads by gathering data from n coin flips. The assumed model is Binomial(n, p), where p is unknown and to be estimated from the data
- **Blood pressure:** We want to estimate the average blood pressure among Duke undergraduates. We can model the blood pressure as a Normal(μ, σ^2), where μ is to be estimated from the data
- **Call center:** Suppose we have a call center and we want to estimate the probability that we receive more than 10 calls in a day. Given our experience, we believe that the Poisson(λ) is a reasonable probability model for the data, but we don't know λ. Note that we can estimate the probability of interest if we estimate λ from the data
- **Policy:** We want to estimate the number of Duke undergraduates that are in favor of a policy using a sample of n individuals. A reasonable model for the data is the Hypergeometric(N, M, n), where N (total number of undergraduates) and n (sample size) are known, but M (undergraduates in favor of the policy) is unknown.

Statistical Reasoning

In this course will always following this sort of logic (roughly):

1. We want to learn something we don't know from data
2. We come up with a "reasonable" probability model for the data that has some unspecified parameters
3. Use the assumed structure of the probability model to answer our question

One should always make sure that the assumptions of the probability model are reasonable before and after data are gathered. We won't be doing it much in this course, but that doesn't mean it isn't important.

Terminology

- Our starting point will be a **random sample** of n individuals X_1, $X_2, \ldots, X_n$, which we will assume to be **i.i.d.** (independent and identically distributed) from some **model**
- We want to estimate an unknown characteristic θ of the model from which we assume the data come from. The usual name that we will given to θ is **parameter**. Some examples of parameter could be the mean, median, or the variance of the distribution
- We will find an **estimator** to estimate the parameter, which we will typically denote $\widehat{\theta}$. The estimator has to be something we can compute using the data, i.e., a function $g(X_1, X_2, \ldots, X_n)$. Functions that depend on unknowns (θ in particular) don't count as estimators

Examples of Estimation

- **Biased coin:** The random sample is the number of heads in n coin flips. The parameter θ is the probability that we get heads, and the model is the Binomial(n, p). An intuitive estimator is the proportion of times that we get heads in the sample. Using mathematical notation, Y is the number of times we get heads in the n trials. The estimator is $\widehat{p} = Y/n$, which is clearly something we can compute using the data
- **Blood pressure:** The random sample is a collection of n blood pressure levels, which we model as $X_i \sim \text{Normal}(\mu, \sigma^2)$. In this example, the parameter is μ. An intuitive estimator for μ is the sample mean $\overline{X}_n = \sum_{i=1}^{n} X_i/n$
- **Call center:** The random sample is the number of calls in n hours, which we model as $C_i \sim \text{Poisson}(\lambda)$. λ can be estimated from the sample average $\overline{C_n} = \sum_{i=1}^{n} C_i/n$
- **Policy:** The random sample is the total number of undergraduates in the sample that are in favor of the policy, which we model as $X \sim \text{Hypergeometric}(N, M, n)$. The parameter is M and a reasonable estimator is $\widehat{M} = NX/n$. (do you see why?)

Some Desirable Characteristics

The ideal estimator $\widehat{\theta}$ for θ is one that equals θ with probability 1. But since θ is something we don't know, finding a "perfect" estimator is (almost always) impossible. Therefore, we have to think about desirable properties that good estimator should have, and come up with some sensible criteria to compare between estimators.

1. **Unbiasedness**: An estimator is said to be unbiased if $E(\widehat{\theta}) = \theta$. We have already seen that the sample average is an unbiased estimator of the population mean, i.e., $E(\overline{X}_n) = \mu$
2. **Low variance**: an unbiased estimator that is highly variable is not very useful
3. **Consistency**: An estimator is consistent if $\widehat{\theta} \to \theta$ as $n \to \infty$. For instance, the sample mean is a consistent estimator of the population mean by LLN

Unbiased Estimates

A point estimate $\hat{\theta} = h(X_1, \ldots, X_n)$ is said to be an **unbiased estimator** θ for a population parameter θ if $\mathbb{E}[\hat{\theta}] = \theta$.

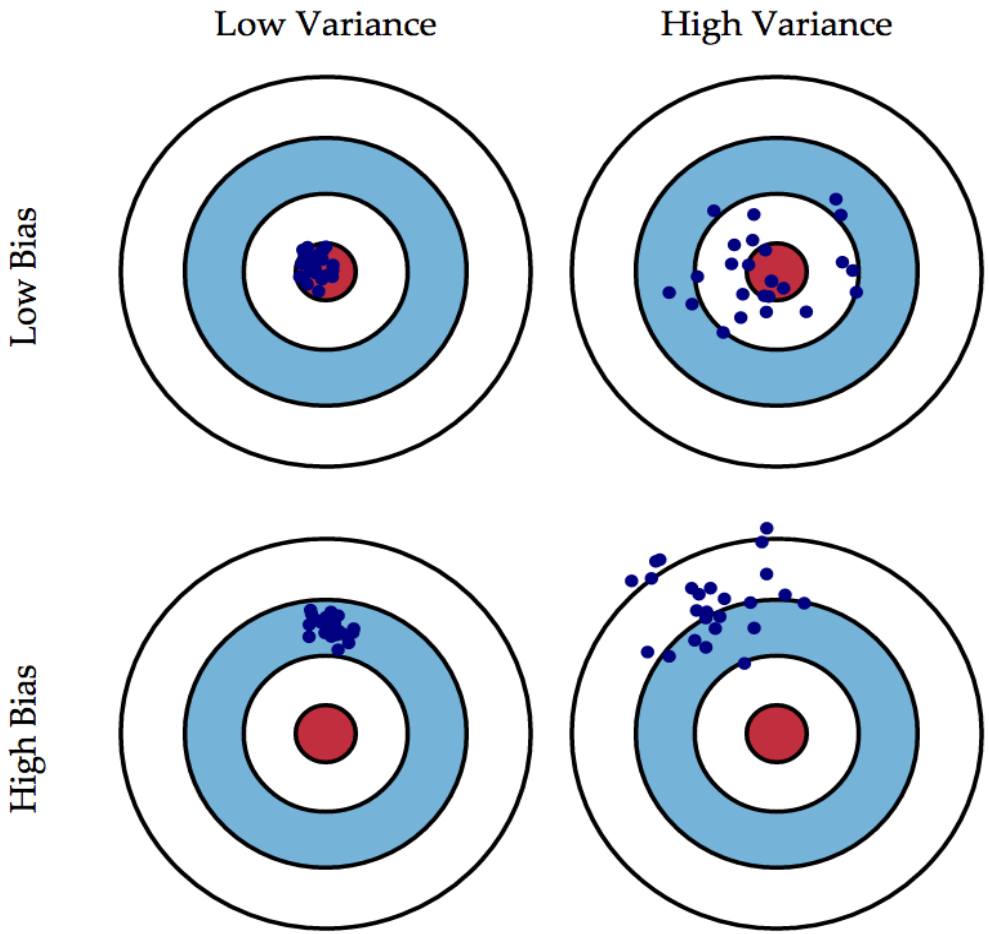
The **bias** in a point estimate is $\mathbb{E}[\hat{\theta}] - \theta = \text{bias}(\hat{\theta})$. For unbiased estimates, this is clearly zero.

Bias-Variance Decomposition: the **mean squared error (MSE)** of a point estimate is

$$\mathbb{E}[\hat{\theta} - \theta]^2 = \mathbb{V}[\hat{\theta}] + [\text{bias}(\hat{\theta})]^2.$$

The MSE has many attractive features. In particular, it is sometimes possible to trade-off a small bias for a large reduction in variance, and this leads to better accuracy.

Bias-Variance Tradeoff



Unbiasedness isn't everything: low variance (aka high precision) is also very important.

Computing Time

All the estimators that we will find in this course will be simple and easy to compute, but in harder problems this becomes a very serious issue. We might have an estimator that has great theoretical properties, but it might take days/years to compute (because the data is huge and the method is complicated).

On the other hand, we might have another estimator that might be theoretically worse, but is way faster to compute. If we care about computing time (for example, think that our application is internet ads and the parameter is "best ads to show"), we will probably end up reporting the latter.

Point Estimates

Statisticians often provide two things:

- a point estimate of some quantity of interest, and
- a statement of the uncertainty in that estimate.

Usually, other disciplines only provide the point estimate.

A **parameter** is some property of a distribution (or density function), such as the mean, median, standard deviation, and so forth.

A **point estimate** for a parameter is some statistic $h(X_1, \ldots, X_n)$ which, when evaluated for a random sample, gives a sensible approximation to the parameter.

Common Examples

– **Sample mean**: As we have seen, if $X_1, X_2, \ldots, X_n$ are a random sample such that $E(X_i) = \mu$ and $V(X_i) = \sigma^2$, $\overline{X}_n$ is an unbiased estimator of μ, with variance $V(X_i) = \sigma^2/n$ and MSE equal to σ^2/n

– **Sample variance**: Suppose that $X_1, X_2, \ldots, X_n$ are random sample such that $E(X_i) = \mu$ and $V(X_i) = \sigma^2$, and now we want to estimate σ^2 from the data. The sample variance is defined as

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2,$$

which is an unbiased estimator of σ^2:

$$\begin{aligned} E(s_n^2) &= \frac{1}{n-1}\left[E\left(\sum_{i=1}^{n} X_i^2\right) - nE\left(\overline{X}_n^2\right)\right] \\ &= \frac{1}{n-1}\left[n(\mu^2+\sigma^2) - n(\mu^2+\sigma^2/n)\right] = \sigma^2 \end{aligned}$$

Finding the variance of s_n^2 is more complicated, especially if we don't assume anything about the distribution of the X_i,
$(V(s_n^2) = \mu_4/n - \sigma^4(n-3)/[n(n-1)]$, where $\mu_4 = E[(X-\mu)^4])$

Common Examples

– **Estimating proportions**: Let Y be Binomial(n, p), where n is known and p is unknown. Our intuition tells us that the sample proportion $\widehat{p} = Y/n$ should be a reasonable estimator of p. The sample proportion is unbiased $E(\widehat{p}) = E(Y/n) = p$ and its variance is $V(\widehat{p}) = p(1-p)/n$. Note that $\widehat{p}$ is just a sample average (a sum of Bernoulli(p) over n), so we could have used what we know about sample averages directly.

– **Hypergeometric**: Let's go back to the policy example. The number of individuals in the sample that support the policy is $X \sim$ Hypergeometric(N, M, n), and we argued that $\widehat{M} = NX/n$ was an intuitively reasonable estimator for M. Well, it turns out that it is an unbiased estimator because $E(\widehat{M}) = NE(X)/n = M$.

Common Examples

– Clinical trial, sample size determination: We are in charge of designing a clinical trail whose goal is to estimate the proportion of patients that will recover from a disease after taking a new treatment. We know we will model our data as $Y \sim \text{Binomial}(n, p)$, but now we want to determine a sample size that would allow us to estimate p sufficiently well. The variance of the sample proportion is $p(1-p)/n$ which depends on p. However, $p(1-p)$ is maximized at $p = 0.5$, so we know that the variance of our estimator will be smaller than (or equal to) $0.25/n$. Therefore, if we want to design an experiment such that the variance of the estimator is less than or equal to 0.005 (standard deviation of approximately 0.071), we should select a sample size of at least 50 individuals.

Common Examples

– **The average squared deviation**,

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

for the population variance.

– **The 10% trimmed sample mean** for the population mean; this is the average of the sample after removing the largest 5% of the values and the smallest 5% of the values.

Notice that a point estimator has to be random since it is a function of a random sample from some distribution, but the true parameter itself is constant.

Examples

Example 1: Let X be uniformly distributed on the interval $[0, \theta]$ where $f(x) = 1/\theta$ so that θ is the unknown parameter. You have a random sample $X_1, \ldots X_n$ and use the statistic $\hat{\theta}_1 = Z = \max\{X_1, \ldots, X_n\}$ as an estimate of θ. Then

$$F(x) = \mathbb{P}[X \leq x] = \int_0^x 1/\theta \, dt = x/\theta.$$

and from the Lecture 8, $G(z) = \mathbb{P}[Z \leq z] = \mathbb{P}[\max\{X_1, \ldots, X_n\} \leq z]$, and

$$\begin{aligned} \mathbb{P}[\max\{X_1, \ldots, X_n\} \leq z] &= \mathbb{P}[X_1 \leq z \text{ and } \cdots \text{ and } X_n \leq z] \\ &= \prod_{i=1}^{n} \mathbb{P}[X_i \leq z] \\ &= \prod_{i=1}^{n} \frac{z}{\theta} = \left(\frac{z}{\theta}\right)^n. \end{aligned}$$

Examples

So the distribution of the sample maximum is $G(z) = (z/\theta)^n$ for $0 \leq z \leq \theta$ and thus the probability density function of the maximum is $g(z) = n(1/\theta)^n z^{n-1}$ on $0 \leq z \leq \theta$.

Since we know the density, we can find the expected value of Z, where Z is the sample maximum and check if it is unbiased:

$$\begin{aligned} \mathbb{E}[\hat{\theta}_1] = \mathbb{E}[Z] &= \int_0^\theta z * \frac{n}{\theta^n} z^{n-1}\, dz = \frac{n}{\theta^n} \frac{1}{n+1} z^{n+1} \Big|_0^\theta \\ &= \frac{n}{n+1}\theta. \end{aligned}$$

Examples

So the estimator $\hat{\theta}_1$ of θ has a small bias:

$$\frac{n}{n+1}\theta - \theta = -\frac{\theta}{n+1}.$$

One can make $\hat{\theta}_1$ into an unbiased estimator by using the new estimator

$$\hat{\theta}_2 = \frac{(n+1)\hat{\theta}_1}{n} = (1+\frac{1}{n})\hat{\theta}_1 = \frac{(n+1)}{n}\max\{X_1, \ldots, X_n\}.$$

But note that there is a price we have to pay, since

$$V(\hat{\theta}_2) = (1+\frac{1}{n})^2 V(\hat{\theta}_1) \geq V(\hat{\theta}_1)$$

Minimum Variance Unbiased Estimators (MVUE)

Usually, a first requirement for a good estimator of a parameter is that it be unbiased. When there are several unbiased estimators, one should use the one that has smallest variance.

This is not the only way to frame the problem of selecting an estimator. For example, one might want the estimator which:

- minimizes the the mean squared error,
- has the largest probability of being within some fixed distance from the true value,
- is unbiased and minimizes something more practical than the variance.

Examples

Consider again the case of a random sample from the Unif$(0, \theta)$ distribution. Then for any $X \sim \text{Unif}(0, \theta)$, $\mathbb{E}[X] = \frac{\theta}{2}$. Clearly, $\mathbb{E}[\bar{X}] = \frac{\theta}{2}$ so $\hat{\theta}_3 = 2\bar{X}$ is an unbiased estimator of θ.

We now have two candidate estimators:

$$\hat{\theta}_2 = \frac{(n+1)}{n} \max\{X_1, \ldots, X_n\} \text{ and } \hat{\theta}_3 = 2\bar{X}.$$

Which has the smaller variance?

Since $\hat{\theta}_3$ is a linear combination, its variance is $\frac{4\sigma^2}{n}$ where σ^2 is the variance of the Unif$(0, \theta)$ distribution. You can check that the variance of the Unif$(0, \theta)$ distribution is $\theta^2/12$. Thus

$$\mathbb{V}[\hat{\theta}_3] = \frac{\theta^2}{3n}.$$

Examples

To find the variance of $\hat{\theta}_2$ we first find

$$\mathbb{E}[Z^2] = \int_0^\theta z^2 g(z)\, dz = \int_0^\theta z^2 * \frac{n}{\theta^n} z^{n-1}\, dz = \frac{n}{n+2}\theta^2.$$

Since $\mathbb{V}[Z] = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2$, we have

$$\mathbb{V}[Z] = \frac{n}{n+2}\theta^2 - \left[\frac{n}{n+1}\theta\right]^2 = \left[\frac{n}{(n+2)(n+1)^2}\right]\theta^2.$$

Since $\hat{\theta}_2 = \frac{(n+1)}{n} Z$, then

$$\mathbb{V}[\hat{\theta}_2] = \left(\frac{n+1}{n}\right)^2 \left[\frac{n}{(n+2)(n+1)^2}\right]\theta^2 = \frac{1}{n(n+2)}\theta^2.$$

A little algebra shows that $n(n+2) > 3n$ for all $n > 1$, so $\hat{\theta}_2$ is better than $\hat{\theta}_3$.

Robust Estimators

Previously, we claimed to like estimators that are unbiased, have minimum variance, and/or have minimum mean squared error. Typically, one cannot achieve all of these properties with the same estimator.

An estimator may have good properties for one distribution, but not for another. We saw that $\frac{n}{n-1}Z$, for Z the sample maximum, was excellent in estimating θ for a Unif$(0, \theta)$ distribution. But it would not be excellent for estimating θ every pdf supported on $[0, \theta]$.

A **robust estimator** is one that works well across many families of distributions. In particular, it works well when there may be outliers in the data.

Robust Estimators

The **10% trimmed mean** is a robust estimator of the population mean. It discards the 5% largest and 5% smallest observations, and averages the rest. Obviously, one could trim by some fraction other than 10%, but this is a commonly-used value.

Surveyors distinguish errors from blunders. Errors are measurement jitter attributable to chance effects, and are approximately Gaussian. Blunders occur when the guy with the theodolite is standing on the wrong hill.

A trimmed mean throws out the blunders and averages the good data. If all the data are good, one has lost some sample size. But in exchange, you are protected from the corrosive effect of outliers.

Recap

Today we covered:

- Introduction to Estimation
- Point Estimation and Unbiasedness
- Unbiasedness
- Minimum Variance Unbiased Estimators
- Robust Estimators

Suggested reading:

- D.S. Sec. 7.1, 7.9