Lecture 9: Estimation

Donglei Du  
(ddu@unb.edu)

Faculty of Business Administration, University of New Brunswick, NB Canada Fredericton  
E3B 9Y2
Table of contents
What is Statistical inference?

- Statistical inference/inferential statistics is the process of drawing conclusions about population from samples that are subject to random variation.
- We will learn three basic Statistical inference techniques in the rest of the course
  - Estimation
  - Hypothesis testing
  - Linear regression analysis
Estimation theory is a branch of statistics that deals with estimating the values of parameters based on measured/empirical data that has a random component.

An **estimate** is a single value that is calculated based on samples and used to estimate a population value.

An **estimator** is a function that maps the sample space to a set of estimates.

The entire purpose of estimation theory is to arrive at an estimator, which takes the sample as input and produces an estimate of the parameters with the corresponding accuracy.
Two Types of Estimator

- There are two types of estimators
  - Point estimator
  - Interval estimator
A **point estimator** is a statistic (that is, a function of the data) that is used to infer the value of an unknown parameter in a statistical model.

A **point estimate** is one of the possible values a point estimator can assume.

Mathematically, suppose there is a fixed parameter $\theta$ that needs to be estimated and $X$ is a random variable corresponding to the observed data. Then an estimator of $\theta$, usually denoted by the symbol $\hat{\theta}$, is a function of the random variable $X$, and hence itself a random variable $\hat{\theta}(X)$.

A point estimate for a particular observed dataset (i.e. for $X = x$) is then $\hat{\theta}(x)$, which is a fixed value.
A point estimator can be evaluated based on:

- **Unbiasedness (mean):** whether the mean of this estimator is close to the actual parameter?
- **Efficiency (variance):** whether the standard deviation of this estimator is close to the actual parameter.
- **Consistency (size):** whether the probability distribution of the estimator becomes concentrated on the parameter as the sample sizes increases.
Sampling Error, Bias and Mean squared error

- **Sampling Error**: The error of the estimator $\hat{\theta}(X)$ for the parameter $\theta$ is defined as:
  \[
e(\hat{\theta}(X)) := \hat{\theta}(X) - \theta
  \]

- The **bias** of the estimator $\hat{\theta}$ is defined as the expected value of the error:
  \[
  B(\hat{\theta}(X)) := E[\hat{\theta}(X)] - \theta = E[\hat{\theta}(X) - \theta]
  \]

- The **mean squared error** of $\hat{\theta}(X)$ is defined as the expected value (probability-weighted average, over all samples) of the squared errors; namely,
  \[
  \text{MSE}(\hat{\theta}(X)) = E[(e(\hat{\theta}(X)))^2] = E[(\hat{\theta}(X) - \theta)^2].
  \]

- The MSE, variance, and bias, are related:
  \[
  \text{MSE}(\hat{\theta}) = \text{var}(\hat{\theta}) + (B(\hat{\theta}))^2
  \]
A point estimator $\hat{\theta}$ for the parameter $\theta$ is **unbiasedness** if its bias is zero; namely

$$E(\hat{\theta}(X)) - \theta = E(\hat{\theta}(X) - \theta) = 0$$

Sample mean $\bar{X}$ is an unbiasedness point estimator for population mean $\mu$; namely

$$E(\bar{X}) - \mu = E(\bar{X} - \mu) = 0$$

Sample variance $S^2$ is an unbiasedness point estimator for population variance $\sigma^2$; namely

$$E(S^2) - \sigma^2 = E(S^2 - \sigma^2) = 0$$
Efficient Point Estimator

- For the same parameter $\theta$, an unbiased point estimator $\hat{\theta}_1$ is more **efficient** than another unbiased point estimator $\hat{\theta}_2$ if
  
  \[
  \text{Var}(\hat{\theta}_1(X)) < \text{Var}(\hat{\theta}_2(X))
  \]

- The minimum-variance unbiased estimator (MVUE) is an unbiased estimator that has lower variance than any other unbiased estimator for all possible values of the parameter.

- For a normal distribution with unknown mean and variance:
  - Sample mean $\bar{X}$ is the MVUE for population mean $\mu$; namely
    \[
    E(\bar{X}) - \mu = E(\bar{X} - \mu) = 0
    \]
  - Sample variance $S^2$ is the MVUE for population variance $\sigma^2$; namely
    \[
    E(S^2) - \sigma^2 = E(S^2 - \sigma^2) = 0
    \]

- For other distributions the sample mean and sample variance are not in general MVUEs.
Interval Estimator

- An interval estimator of a population parameter under random sampling consists of two random variables, which are called the upper and lower limits of the interval estimator, whose values decide intervals which expect to contain the parameter estimated.
- Interval estimates are the all the ranges an interval estimator can assume. Each interval estimate states an range within which a population parameter probably lies.
- An interval estimator can be assessed through
  - Accuracy (confidence level)
  - Precision (margin of error)
Accuracy

- Accuracy is also called the **confidence level** of the estimator, and it is defined to be the probability that an interval estimator obtained will contain the value of the population parameter.
- Any possible outcome of an interval estimator is called an interval estimate.
- An interval estimate with confidence level \((1 - \alpha)\) is called an \((1 - \alpha)\)-confidence interval.
- The upper and lower limits of a confidence interval are called the upper and lower limits, respectively.
Precision is also called the **margin of error**.

It is measured by the half of width of the interval estimates, i.e., the difference of the upper and lower limits of the confidence interval.
Balancing Accuracy with Precision

- Accuracy and precision are two opposite driving forces, they are inversely related.
- Our objective is to design an interval estimator such that
  - the confidence level sufficiently high
  - the margin of error sufficiently small.
An interval estimator is usually designed in the following way:

(i) take an unbiased point estimator, and
(ii) define an interval of reasonable width around it.

We will use the above technique to design an interval estimator for population mean.
An interval Estimator for Population Mean

(i) Sample mean \( \bar{X} \) is an unbiased point estimator.

(ii) Define an interval estimator around it with margin of error equal to \( w \):

\[
[\bar{X} - w, \bar{X} + w]
\]
A notation

For any given $\alpha$, denote $z_{\alpha/2}$ to be the $z$-score such that

$$P \left( Z \geq z_{\alpha/2} \right) = \frac{\alpha}{2}$$

For example:

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$z_{\alpha/2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1.65</td>
</tr>
<tr>
<td>0.05</td>
<td>1.96</td>
</tr>
<tr>
<td>0.02</td>
<td>2.33</td>
</tr>
<tr>
<td>0.01</td>
<td>2.58</td>
</tr>
</tbody>
</table>
A notation

\[ 1 - \alpha \]

\[ -z_{\alpha/2} \quad \frac{\alpha}{2} \]

\[ z_{\alpha/2} \quad \frac{\alpha}{2} \]
(1 − α)-confidence interval for Population Mean

- Given a the confidence level \((1 − \alpha)\), an \((1 − \alpha)\)-confidence interval for population mean \(\mu\) is:

\[
\left[ \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] := \bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{1}
\]

- In order to apply the above formula, one of the following assumptions must be satisfied:
  
  (i) If \(n \geq 30\) and \(\sigma\) is known.
  (ii) Or, if \(n < 30\), \(\sigma\) is known, and the population is normally distributed.

- If \(n \geq 30\) and \(\sigma\) is unknown, the standard deviation \(s\) of the sample is used to approximate the population standard deviation \(\sigma\):

\[
\left[ \bar{X} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right] := \bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} \tag{2}
\]
(1 − α)-confidence interval for Population Mean

- If \( n < 30 \), \( \sigma \) is unknown, and the population is normally distributed, then we should use the Student \( t \) distribution:

\[
\left[ \bar{X} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2} \frac{s}{\sqrt{n}} \right] := \bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}
\]  

(3)

- If \( n < 30 \), \( \sigma \) is unknown, and the population is not normally distributed, then we should use non-parametric statistics, which will be discussed in Advanced Statistics, such as ADM 3628.
For the weight example, find the 95 percent confidence interval for the population mean.

Let us check the assumption first:

(i) \( n = 62 \geq 30 \)
(ii) \( \sigma \) is unknown

So we should use (2). \( \bar{x} = 155.8548, s \approx 12.0964, \) and \( z_{\alpha/2} = z_{0.95/2} = 1.96 \) from the normal table. The 95%-CI for \( \mu \) is:

\[
\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} \approx 155.8548 \pm 1.96 \frac{12.0964}{\sqrt{62}} = [152.8438, 158.8658]
\]
R code: \( z \)-test

```r
> weight <- read.csv("weight.csv")
> x<-weight$Weight.01A.2013Fall
> library(BSDA)
> z.test(x, sigma.x=sd(x), conf.level = 0.95)

One-sample z-Test

data:  x
z = 101.4519, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  152.8439 158.8658
sample estimates:
mean of x
  155.8548
```
Interpreting confidence intervals correctly

- Incorrect: A 95%-CI for population mean $\mu$ does not mean:
  The probability that this interval contains the mean with probability 95%.

- Correct (well, almost): Instead it means
  If we randomly construct 100 such confidence intervals, 95 of them will contain the mean!
Selecting the Sample Size

Given the confidence level \((1 - \alpha)\) and the margin of error \(w\), what should be the minimum number of sample size?

Depending on the assumptions, \(n\) should be chosen such that

\[
\begin{align*}
n &\geq \left( \frac{z_{\alpha/2}\sigma}{w} \right)^2 \quad \text{(4)} \\
n &\geq \left( \frac{z_{\alpha/2}s}{w} \right)^2 \quad \text{(5)} \\
n &\geq \left( \frac{t_{\alpha/2}s}{w} \right)^2 \quad \text{(6)}
\end{align*}
\]
Example

Example: A consumer group would like to estimate the mean monthly electricity charge for a single family house in July (within $5) using a 99 percent level of confidence. Based on similar studies the standard deviation is estimated to be $20.00.

Problem: How large a sample is required

Solution: Let us check the assumption first:

(i) \( n \) is unknown and we want to find it out. But assume it is at least 30!
(ii) \( \sigma \) is known

So we could use (4): \( z_{\alpha/2} = z_{0.01/2} = 2.58 \) from the normal table, \( \sigma = 20 \), and \( w = 5 \). So

\[
n \geq \left( \frac{z_{\alpha/2}\sigma}{w} \right)^2 = \left( \frac{2.58(20)}{5} \right)^2 \approx (10.32)^2 \approx 106.5.
\]

A minimum of 107 homes must be sampled.