BNAD 276
Lecture 8
Statistical Inference II
Interval Estimation

Phuong Ho

May 30, 2017

Outline

Outline

Interval Estimation

- Though $\bar{X}$ is really close to $E(X) = \mu$ and provides good estimate for μ, it is not exactly equal to $\mu = E(X)$.
- Furthermore, $\bar{X}$ varies as we choose different samples.
- Hence, it is maybe better to state that "we are 95% sure that μ is contained in an interval".
- This motivates interval estimation.
- We will estimate an interval that highly likely contains the true mean μ or true variance σ^2.

Interval Estimation

- The statement we want to make is that "we are 95% sure that μ is contained in the interval," or "95 % likely that the interval contains μ."
- Then, our job is to figure out an interval that allows us to state the above statement.
- We want to find two numbers, a and b, that specifies an **interval** satisfying the following equation.

$$P[a \leq \mu \leq b] = 0.95$$

- That is, the probability that the interval $[a, b]$ contains μ is 0.95.

Interval Estimation

$$P[a \leq \mu \leq b] = 0.95$$

- Once we find two numbers, a (lower bound) and b (upper bound) that make the above equation hold, we can finally say that 'we are 95% sure that μ is placed between a and b."
- Finding a and b is the same as finding an interval $[a, b]$.
- Thus, this job (finding $[a, b]$) is called interval estimation.
- We will start with the simplest case where $Var(X) = \sigma^2$ is known.

Outline

Interval Estimation: σ known

- Suppose that we have population random variable X that follows a population distribution.
- Of course, we absolutely have no idea about the values of $\mu = E(X)$ and $\sigma = \sqrt{Var(X)}$.
- However, we will assume that σ is somehow magically known to us to introduce the new idea, interval estimation.
- Given that σ is known to us, we will construct an interval that contains (unknown) μ with probability 0.95.
- We will use the sampling distribution $\bar{X}$ to construct the interval.

Example: Construction of an interval

- Suppose that we have a random variable X that follows a unknown population distribution.
- Fortunately, $\sqrt{Var(X)} = \sigma$ is known to us but μ is unknown.
- Suppose that we have a sample with sample size N. Then, we can calculate $\bar{X}$.
- We also know that, by CLT,

$$\bar{X} \overset{approx}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$$

- Let's construct an interval that contains μ with probability 0.95.

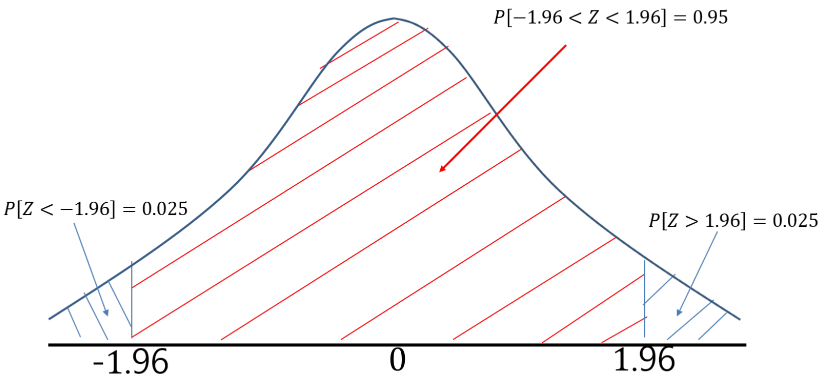
Example cont'd

- Since we can't make any statement with probability with this general normal distribution, let's convert it to the standard normal.

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{N}} \overset{approx}{\sim} \mathcal{N}(0, 1)$$

Example cont'd

- Recall that we are interested in an interval related to probability 0.95. If we find such an interval with the standard normal distribution, that will be the interval from -1.96 to 1.96 as in the following graph.



Example cont'd

- Thus, we can write the following equation.

$$P[-1.96 < Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{N}} < 1.96] = P[-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{N}} < 1.96] = 0.95$$

- What we want are two numbers, a and b, satisfying the following.

$$P[a < \mu < b] = 0.95$$

- Let's manipulate $P[-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{N}} < 1.96]$ to have the expression in the above.

Example cont'd

$$
\begin{aligned}
0.95 &= P[-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{N}} < 1.96] \\
&= P[-1.96\frac{\sigma}{\sqrt{N}} < \bar{X}-\mu < 1.96\frac{\sigma}{\sqrt{N}}] \\
&= P[1.96\frac{\sigma}{\sqrt{N}} > -\bar{X}+\mu > -1.96\frac{\sigma}{\sqrt{N}}] \\
&= P[-1.96\frac{\sigma}{\sqrt{N}} < \mu-\bar{X} < 1.96\frac{\sigma}{\sqrt{N}}] \\
&= P\left[\bar{X}-1.96\frac{\sigma}{\sqrt{N}} < \mu < \bar{X}+1.96\frac{\sigma}{\sqrt{N}}\right] = 0.95
\end{aligned}
$$

- Finally we have the interval we wanted, the interval such that $P[a < \mu < b] = 0.95$.
- And, a is $\bar{X} - 1.96\frac{\sigma}{\sqrt{N}}$ and b is $\bar{X} + 1.96\frac{\sigma}{\sqrt{N}}$.

Example cont'd

- Using the sampling distribution with manipulation, what we found is

$$P\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{N}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{N}}\right] = 0.95.$$

- With the above result, we can say that
the interval from $\bar{X} - 1.96\frac{\sigma}{\sqrt{N}}$ to $\bar{X} + 1.96\frac{\sigma}{\sqrt{N}}$ contains μ with probability 0.95

Confidence Intervals

- The interval from $\bar{X} - 1.96\frac{\sigma}{\sqrt{N}}$ to $\bar{X} + 1.96\frac{\sigma}{\sqrt{N}}$ is called **95% confidence interval** for μ since we are 95% confident that μ is contained in this interval.
- The value 95% is referred to as the **confidence level** of an confidence interval.
- The value 0.95 is referred to as the **confidence coefficient** of an confidence interval.
- There is another term which is called the **the level of significance** which is denoted by α.

$$\alpha = \text{Level of Significance} = 1 - \text{Confidence Coefficient}$$

Example 1

Suppose that a store selects a simple random sample of 100 customers in order to learn about the amount spent per shopping. X represents the amount spent per shopping. $\bar{X}$ is calculated and it is equal to 82. This store already somehow knows σ is equal to 20.

- We have $N = 100$, $\bar{X} = 82$, and $\sigma = 20$.
- What is the 95% confidence interval for μ?
- We know that

$$\frac{\bar{X} - \mu}{20/\sqrt{100}} \overset{approx}{\sim} \mathcal{N}(0, 1)$$

Example 1 cont'd

- With the manipulation we did before, we also know

$$P\left[82 - 1.96\frac{20}{\sqrt{100}} < \mu < 82 + 1.96\frac{20}{\sqrt{100}}\right] = 0.95$$

- Thus, the 95% confidence interval is the interval from 78.08 to 85.92.
- And we say that μ will be contained between 78.08 to 85.92 with probability 0.95.

Interpretation

- What does it mean by saying that μ will be contained in the interval from $\bar{X} - 1.96\frac{\sigma}{\sqrt{N}}$ to $\bar{X} + 1.96\frac{\sigma}{\sqrt{N}}$ with probability 0.95?
- Recall that $\bar{X}$ is a random variable and the value of it changes as we re-sample data.
- Thus, if we do sampling 100 times, we will have 100 confidence intervals.
- The statement that μ will be contained in the interval $\bar{X} - 1.96\frac{\sigma}{\sqrt{N}}$ to $\bar{X} + 1.96\frac{\sigma}{\sqrt{N}}$ with probability 0.95 means that 95 numbers of confidence intervals out of 100 will contain μ.
- That is, the probability that our confidence interval is one of 95 confidence intervals containing μ is 0.95.

Confidence Intervals

- We may have different confidence intervals with different confidence levels.
- We have the general form of an interval estimate of μ for the case that σ is known as follows.

Interval Estimate of a population mean: σ known

$$\bar{X} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{N}},$$

where $(1-\alpha)$ is the confidence coefficient and $z_{\alpha/2}$ is the value such that $P[Z \geq z] = \frac{\alpha}{2}$.

Confidence Intervals

- Though we may have many different confidence intervals as confidence level varies, we are typically interested in the following 3 confidence intervals: 90%, 95%, and 99% confidence intervals.
- The following table summarizes the values of α and $z_{\alpha/2}$ for 3 confidence intervals in the above.

Confidence Level	α	$\alpha/2$	$z_{\alpha/2}$
90%	.10	.05	1.645
95%	.05	.025	1.96
99%	.01	.005	2.575

Example 1 again

- We know $\bar{X} = 82$, $N = 100$, and $\sigma = 20$.
- What is 90% confidence interval?
- What is 99% confidence interval?
- What is 85% confidence interval?

Issue with the sample size N

- All works we've done is under the assumption that N is large enough.
- If N is not large enough (practically, if $N < 30$), CLT will not work well.
- Thus, we need to involve more assumptions on the population distribution.
- Remember that what we've done is valid only if we have a large sample size N.
- If we have small sample, we need to **assume that the population distribution is a normal distribution.**

Outline

Interval Estimation: σ unknown

- In the previous section, we assume that σ is known to us.
- However, in practice, σ is not known.
- Recall that our confidence intervals in the previous section take the following form.

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{N}}$$

- Hence, if we don't know σ, the above expression will not be a number.
- Thus, we need to estimate σ with s.

Interval Estimation: σ unknown

- Recall that s^2 is a consistent estimator for σ^2.
- Thus, s will be a consistent estimator for σ too.
- Hence, we may replace σ with s when we have data and don't know σ.
- However, s is not exactly same as σ.
- Because of this, we will have some issues to do interval estimation.

Interval Estimation: Unknown σ

- When σ is unknown, we need to involve one more assumption.
- Previously, we didn't make an assumption on the population distribution when N is large.
  - However, if N is small, we have to assume that the population distribution is a normal distribution.
- When σ is unknown, we also need to assume that the population distribution is a normal distribution.
- This is one more assumption we need to construct a confidence interval.

Interval Estimation: Unknown σ

- Assume that the population distribution is a normal distribution.
- We don't know either μ or σ.
- At first, we need to **estimate σ with s.**
- When we have data we can calculate s^2. Then, $\sqrt{s^2} = s$ is a consistent estimator for σ.
- Now, we **can use s instead of unknown σ.**

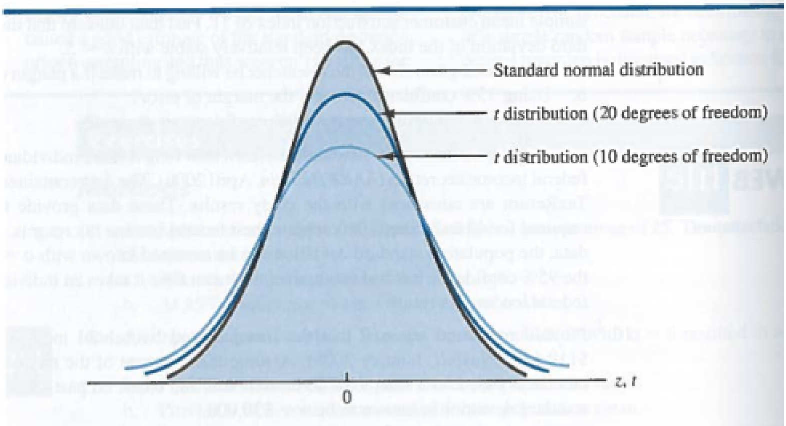
Interval Estimation: Unknown σ

- As we did before, we will work with the distribution of $\frac{\bar{X}-\mu}{\sigma/\sqrt{N}}$.
- Since we decided to replace σ with s, we will work with the distribution of $\frac{\bar{X}-\mu}{s/\sqrt{N}}$
- Since s is not exactly same as σ, the distribution of $\frac{\bar{X}-\mu}{s/\sqrt{N}}$ is not the standard normal any more.
- Now, $\frac{\bar{X}-\mu}{s/\sqrt{N}}$ **follows t-distribution.**

t-distribution

- t-distribution is similar to the standard normal distribution but not the same.
- t-distribution is centered around 0 as the standard normal distribution does.
- However, its shape will depend on a parameter referred to as **the degree of freedom.**
- The degree of freedom that is used for this interval estimation case here is $N - 1$ (Sample size -1)

t-distribution



t-distribution

- The t-distribution is always **centered around 0.**
- **Its central location does not change.**
- However, **its shape changes as the degree of freedom changes.**
- The other thing worth to notice is that t-distribution is getting closer to the standard normal distribution as the degree of freedom increases.

Interval Estimation: Unknown σ

- Recall that we have the following confidence interval when σ is known.

$$\bar{X} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{N}}$$

- When σ is unknown, we will replace σ with s and we will use t-distribution instead of the standard normal distribution.
- Thus, we also replace $z_{\alpha/2}$ with $t_{\alpha/2}$, where $t_{\alpha/2}$ is the value such that $P[T \geq t_{\alpha/2}] = \frac{\alpha}{2}$.

- Finally, we have a general form of interval estimate when σ is unknown as follows.

Interval estimate for μ: σ unknown

$$\bar{X} \pm t_{\alpha/2} \times \frac{s}{\sqrt{N}},$$

where s is the sample s.d., $(1-\alpha)$ is the confidence coefficient, and $t_{\alpha/2}$ is the value such that $P[T \geq t_{\alpha/2}] = \frac{\alpha}{2}$.

t-table

- Since we are going to use t-distribution, we need to use t-table to find the value $t_{\alpha/2}$.
- The way to read t-table is pretty much same as the way to read the standard normal table.
- There are two main difference.
  1. The probabilities given in t-table is $P[T \geq t]$ (Recall that probabilities in the standard normal table are $P[Z \leq z]$).
  2. The probabilities vary as the degree of freedom varies.

Construction of Interval Estimates: σ is unknown

- Suppose we have $N = 20$, $\bar{X} = 9312$ and $s = 4007$.
- If we want to construct 95% confidence interval, we need to find $t_{.025}$.
- In the table, we first need to find the appropriate degree of freedom in this example.
- Since we have $N = 20$, the degrees of freedom in this example will be $20 - 1 = 19$.
- Then, find the 19 in the first column in the table.

Construction of Interval Estimates: σ is unknown

- At the top row, we have area in upper tail (i.e. $P(T \geq t)$).
- Since we are interested in $0.025 (= \alpha/2)$, we need to read the cell where the column of $.025$ and the row of 19 intersect.
- Our $t_{.025}$ is 2.093.
- Finally, with $N = 20$, $\bar{X} = 9312$, $s = 4007$ and $t_{.025} = 2.093$, our 95% confidence interval is

$$9312 \pm 2.093 \times \frac{4007}{\sqrt{70}}$$

Interval Estimate: With Large N and unknown σ

- There are some evidence showing that, if $N \geq 50$, then the interval estimate we've learned can be used without assuming that the population distribution is a normal distribution.
- That is, if $N \geq 50$, we can safely use the interval estimate in this section without the normality assumption on the population distribution.
- Nowadays, it is getting easier to get more data. Hence, we usually have larger sample size than 50.
- While we start with assuming that the population distribution is a normal distribution, we don't need to make this assumption when we have larger sample size than 50.

Outline

Exercise 1

A simple random sample of 40 items resulted in a sample mean of 25. The population standard deviation is $\sigma = 5$.

a. What is the standard error of the sample mean?

b. Construct 95% confidence interval.

Exercise 2

The National Quality Research Center at the University of Michigan provides a quarterly measure of consumer opinions about products and services. A survey of 10 restaurants in the Fast Food group showed a sample mean customer satisfaction index of 71. The population standard deviation is known as $\sigma = 5$.

a. What assumption should the researcher make if he/she wants to construct a confidence interval?

b. Construct 95% confidence interval.

c. Construct 90% confidence interval.

Exercise 3

For a t-distribution with 16 degrees of freedom, find the area, or probability, in each region.

a. To the right of 2.120

b. To the left of 1.337

c. Between -2.120 and 2.120

d. Between -1.746 and 1.746

Exercise 4

The following sample data are from a normal population:10, 8, 12, 15, 13, 11, 6, 5.

- What is the point estimate for the population mean?
- What is the point estimate for the population standard deviation?
- Construct 95% confidence interval.

Exercise 5

The average cost per night of a hotel room in New York City is $273. Assume this estimate is based on sample of 45 hotels and the sample standard deviation is $64.

a. Construct 95% confidence interval.

b. Two years ago the average cost of a hotel room in New York City was $229. Discuss the change in cost over the two-year period.