# Lecture 10

## Inference for One Population Mean

Text: Chapter 5

*STAT 8010 Statistical Methods I*
February 18, 2020

Whitney Huang
Clemson University





Notes

## Agenda

1. **Statistical Inferences**
2. **Point/Interval Estimation**
3. **Confidence Intervals**
4. **Hypothesis Testing**





Notes

## Statistical Inference

For the rest of the semester, we will focus on conducting statistical inferences for the following tasks:

- Estimating one population mean
- Comparing two population means
- Comparing more than two population means
- Estimating population proportions
- Estimating relationship between two quantitative variables

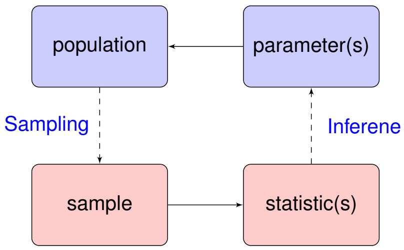




Notes

## Statistical Inference Cont'd

- We use parameters to describe the population
  **Example:** population mean $(\mu_X)$; population variance $(\sigma_X^2)$




- We use statistics of a sample to infer the population
  **Example:** sample mean $(\bar{X})$; sample variance $(s_X^2)$



Notes



## Estimating Population Mean $\mu$

**Goal:** To estimate the population mean using a (representative) sample:

- The sample mean, $\bar{X}_n = \frac{\sum_i^n X_i}{n}$, is a reasonable point estimate of the population mean $\mu_X$
- Need to quantify the level of uncertainty of the point estimate $\Rightarrow$ Interval estimation
- Need to figure out the sampling distribution of $\bar{X}_n$ in order to construct interval estimates $\Rightarrow$ Central Limit Theorem (CLT)



Notes



## Central Limit Theorem (CLT)

**CLT**

The **sampling distribution** of $\bar{X}_n$ will become approximately **normally distributed** as the **sample size** **($n$) becomes "large"**, **regardless of the shape of the population distribution!**

Let $X_1, X_2, \cdots, X_n$ be a random sample from a population $X$ with $\mu = \mathbb{E}[X]$ and $\sigma^2 = \text{Var}[X]$. Then $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} \xrightarrow{d} \text{N}(\mu, \frac{\sigma^2}{n})$ as $n \to \infty$.



Notes

## CLT In Action

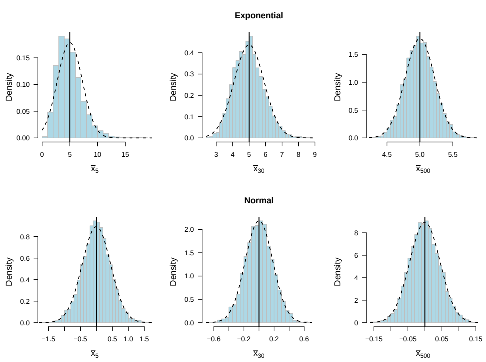
1. Generate 100 ($n$) random numbers from an Exponential distribution (population distribution)
2. Compute the sample mean of these 100 random numbers
3. Repeat this process 120 times





Notes

## CLT: Sample Size ($n$) and the Normal Approximation




Inference for One Population Mean
CLEMSON UNIVERSITY
Statistical Inferences
Point/Interval Estimation
Confidence Intervals
Hypothesis Testing



Notes

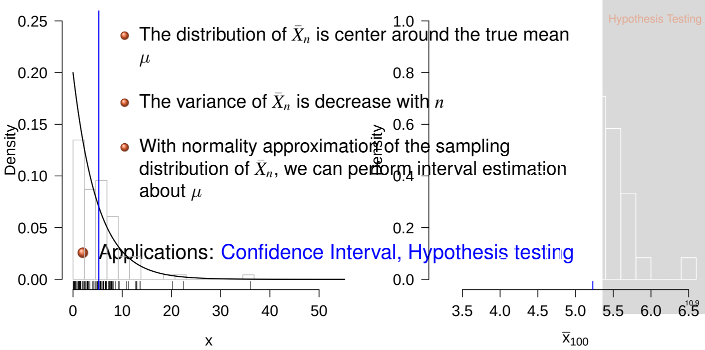
## Why CLT is important?

- CLT tells us the **distribution** of our estimator

$$\bar{X}_n \approx \mathrm{N}(\mu, \frac{\sigma^2}{n})$$

- The distribution of $\bar{X}_n$ is center around the true mean $\mu$
- The variance of $\bar{X}_n$ is decrease with $n$
- With normality approximation of the sampling distribution of $\bar{X}_n$, we can perform interval estimation about $\mu$

- Applications: Confidence Interval, Hypothesis testing




Inference for One Population Mean
CLEMSON UNIVERSITY
Statistical Inferences
Point/Interval Estimation
Confidence Intervals
Hypothesis Testing



Notes

## Confidence Intervals (CIs) for $\mu$

- Let's assume we know the population variance $\sigma^2$ (will relax this assumption later on)

- $(1-\alpha)\times 100\%$ CI for $\mu$:

$$\left[\bar{X}_n - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right],$$

  where $z_{(\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ percentile of $Z \sim \mathrm{N}(0,1)$

- $\frac{\sigma}{\sqrt{n}}$ is the standard error of $\bar{X}_n$, that is, the standard deviation of its sampling distribution




Notes

## Making Sense of Confidence Intervals

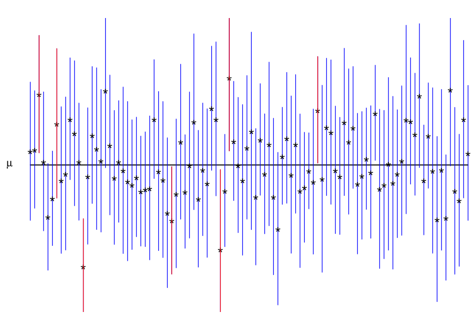
For any $\alpha \in (0,1)$:

$$\begin{aligned}
&\mathbb{P}\left(\bar{X}_n - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}_n + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)\\
=&\mathbb{P}\left(-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \bar{X}_n - \mu \le z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)\\
=&\mathbb{P}\left(-z_{\frac{\alpha}{2}} \le \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \le z_{\frac{\alpha}{2}}\right)\\
=&\mathbb{P}\left(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}\right)\\
=&\Phi(z_{\frac{\alpha}{2}}) - \Phi(-z_{\frac{\alpha}{2}})\\
=&1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha
\end{aligned}$$




Notes

## Making Sense of Confidence Intervals Cont'd







Notes

## Example: Average Height



We measure the heights of 40 randomly chosen men, and get a mean height of 5'9" ($\approx$175cm). Suppose we know the standard deviation of men's heights is 4" ($\approx$10cm). Find the 95% confidence interval of the true mean height of ALL men.




Notes

## Average Height Example Cont'd



1. Point estimate: $\bar{X}_{40} = \frac{\sum_{i=1}^{40} X_i}{40} = 69$ inches
2. Population standard deviation: $\sigma = 4$ inches
3. Standard error of $\bar{X}_{n=40} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{40}} = 0.63$ inches
4. 95%CI: Need to find $z_{0.05/2} = 1.96$ from the Z-table
5. 95% CI for $\mu_X$ is:

$$[69 - 1.96 \times 0.63, 69 + 1.96 \times 0.63]$$
$$= [67.77, 70.23]$$



Notes

## Properties of Confidence Intervals



- In contrast with the point estimate, $\bar{X}_n$, a $(1-\alpha)\%$ CI is an interval estimate, where the **length of CI** reflects our estimation uncertainty
- Typical $\alpha$ values: $0.01, 0.05, 0.1 \Rightarrow 99\%, 95\%, 90\%$ confidence intervals. **Interpretation**: If we were to take random samples over and over again, then $(1-\alpha)\%$ of these confidence intervals will contain the true $\mu$
- The length of a CI depends on
  - Population Standard Deviation: $\sigma$
  - Confidence Level: $1-\alpha$
  - Sample Size: $n$



Notes

## Sample Size Calculation



- We may want to estimate $\mu$ with a confidence interval with a predetermined margin of error (i.e. $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$)
- For example, in estimating the true mean height of All men we may want our CI to be just 0.5 inches in width
- The question is then, "**how many observations do we need to take** so that we have the desired margin of error?"



Notes

## Sample Size Calculation Cont'd



To compute the sample size needed to get a CI for $\mu$ with a specified margin of error, we use the formula below

$$n = \left(\frac{z_{\alpha/2} \times \sigma}{\text{Margin of error}}\right)^2$$

**Exercise**: Derive this formula using margin of error $= z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$



Notes

## Average Height Example Revisited



Compute the sample size needed in order to estimate the true mean height of All men such that the 95% CI to be 0.5 inches in width

1. Length of CI: $2 \times z_{0.025}\frac{\sigma}{\sqrt{n}} = 2\times$ margin of error
2. Want to find $n$ s.t. $z_{0.025}\frac{\sigma}{\sqrt{n}} = 0.25$
3. We have $n = \left(\frac{1.96\times 4}{0.25}\right)^2 = 983.4496$

Therefore, the required sample size is 984



Notes

## Confidence Intervals When $\sigma$ Unknown



- In practice, it is unlikely that $\sigma$ is available to us
- One reasonable option is to replace $\sigma$ with $s$, the sample standard deviation
- We need to account for this added uncertainty with a (slightly) different sampling distribution that has fatter tails

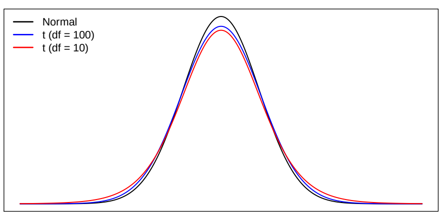
$\Rightarrow$ Student t Distribution (William Gosset, 1908)



Notes

## Student t Distribution

Inference for One Population Mean
Statistical Inferences
Point/Interval Estimation
Confidence Intervals
Hypothesis Testing




- Recall the standardize sampling distribution $\frac{\bar{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathrm{N}(0, 1)$
- Similarly , the studentized sampling distribution $\frac{\bar{X}_n-\mu}{\frac{s}{\sqrt{n}}} \sim \mathrm{t}_{df=n-1}$



Notes

## Confidence Intervals (CIs) for $\mu$ When $\sigma$ is Unknown

Inference for One Population Mean
Statistical Inferences
Point/Interval Estimation
Confidence Intervals
Hypothesis Testing

- $(1-\alpha) \times 100\%$ CI for $\mu$:

$$\left[\bar{X}_n - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}, \bar{X}_n + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}\right],$$

  where $t_{\frac{\alpha}{2},n-1}$ is the $1-\frac{\alpha}{2}$ percentile of a student t distribution with the degrees of freedom $= n-1$

- $\frac{s}{\sqrt{n}}$ is an estimate of the standard error of $\bar{X}_n$



Notes

## Average Height Example Revisited



We measure the heights of 40 randomly chosen men, and get a mean height of 5'9" (≈175cm), and a standard deviation of 4.5" (≈11.4cm). Find the 95% confidence interval of the true mean height of ALL men.



Notes

## Average Height Example Cont'd



1. Point estimate: $\bar{X}_{40} = \frac{\sum_{i=1}^{40} X_i}{40} = 69$ inches
2. Sample standard deviation: $s = 4.5$ inches
3. (Estimated) standard error of $\bar{X}_{n=40} = \frac{s}{\sqrt{n}} = \frac{4.5}{\sqrt{40}} = 0.71$ inches
4. 95%CI: Need to find $t_{0.05/2,39} = 2.02$ from a t-table (or using a statistical software)
5. 95% CI for $\mu_X$ is:

$$[69 - 2.02 \times 0.71, 69 + 2.02 \times 0.71]$$
$$= [67.57, 70.43]$$



Notes

## Hypothesis Testing



- **Hypothesis Testing**: A method for using sample data to decide between two competing claims (hypotheses) about a population characteristic (a parameter. e.g. $\mu$)
- **Examples**:
  - The true mean starting salary for graduates of four-year business schools is $4,500 per month $\Rightarrow \mu = 4,500$
  - The true mean monthly income for systems analysts is at least $6,000 $\Rightarrow \mu \geq 6,000$



Notes

## Hypotheses

- **Null Hypothesis**: A claim about a parameter that is initially assumed to be true. We use $H_0$ to denote a null hypothesis
- **Alternative Hypothesis**: The competing claim, denoted by $H_a$
- In carrying out a test of $H_0$ versus $H_a$ , the hypothesis $H_0$ will be rejected in favor of $H_a$ only if sample evidence strongly suggests that $H_0$ is false. If the sample data does not contain such evidence, $H_0$ will not be rejected
- Therefore, the two possible decisions in a hypothesis test are:
  - Reject $H_0$ (and go with $H_a$)
  - Fail to Reject $H_0$





Notes

## Hypotheses

Be careful setting up hypotheses:

- A statistical hypothesis test is only capable of demonstrating strong support for the alternative hypothesis $H_a$ (by rejecting the null hypothesis $H_0$)
- Failing to reject $H_0$ does not show strong support for the null hypothesis – only a lack of strong evidence against $H_0$, the null hypothesis





Notes

## The $2 \times 2$ Decision Paradigm for Hypothesis Testing

| True State | Decision | |
|---|---|---|
| | Reject $H_0$ | Fail to reject $H_0$ |
| $H_0$ is true | Type I error | Correct |
| $H_0$ is false | Correct | Type II error |

**Errors in Hypothesis Testing**

- The probability of a type I error is denoted by $\alpha$
- The probability of a type II error is denoted by $\beta$





Notes