GOV 2000/ E-2000 Section 5[1]

Sampling, Interval Estimation, and Hypothesis Testing

Solé Prillaman

Harvard University

October 3, 2013

[1]These notes and accompanying code draw on the notes from TF's from previous years.

Logistics - This Week

Problem Set 4- Due by 1pm on Tuesday to course dropbox.

Problem Set 3 Corrections- Due by 1pm on Tuesday to course dropbox.

No Reading Quiz

Practice Midterms- Not due, but recommended as midterm preparation.

Logistics - Next Week

Midterm- Posted October 8, due by 11:59pm on Sunday, October 13 with 5 hours to complete.

Office Hours- Cancelled Friday October 6. Rescheduled for Tuesday October 15 from 9:30am to 12pm.

Problem Set 5- Due by 1pm on Tuesday, October 15 to course dropbox.

Problem Set 4 Corrections- Due by 1pm on Tuesday, October 15 to course dropbox.

Reading Quiz (ALZ ch. 11 & 12)- Due by 9pm on Sunday, October 13 on course website.

Mid-semester Evaluation- Due by 9pm on Monday, October 14 on course website.

Problem Set Expectations

- Must be typeset using LaTeX or Word and submitted as one document containing graphics and explanation electronically in **pdf** form
- Must be accompanied by source-able, commented code
- Corrections do not need to include any of your original work, just 1-2 pages explaining where you made a mistake and how to correct it

MIDTERM

- Window of exam: Tuesday, October 8th (after class) - Sunday, October 13th at 11:59pm
- Needs to be completed in 5 hours
- Open note / book, but **no collaboration** allowed

ESTIMATION

ESTIMATION

- **Parameter-** characteristic of the population distribution (the distribution of X_i)

ESTIMATION

- **Parameter-** characteristic of the population distribution (the distribution of X_i)
  ex. μ or σ^2

ESTIMATION

- **Parameter-** characteristic of the population distribution (the distribution of X_i)

  ex. μ or σ^2
- **Estimand-** parameter being estimated

ESTIMATION

- **Parameter-** characteristic of the population distribution (the distribution of X_i)
  ex. μ or σ^2
- **Estimand-** parameter being estimated
  ex. μ

ESTIMATION

- **Parameter-** characteristic of the population distribution (the distribution of X_i)

  ex. μ or σ^2
- **Estimand-** parameter being estimated

  ex. μ
- **Statistic-** function of the sample used to estimate a parameter

ESTIMATION

- **Parameter-** characteristic of the population distribution (the distribution of X_i)
  ex. μ or σ^2
- **Estimand-** parameter being estimated
  ex. μ
- **Statistic-** function of the sample used to estimate a parameter
  ex. $\bar{X}_n$ or $\hat{\mu}$ or S^2

ESTIMATION

- **Parameter-** characteristic of the population distribution (the distribution of X_i)

  ex. μ or σ^2
- **Estimand-** parameter being estimated

  ex. μ
- **Statistic-** function of the sample used to estimate a parameter

  ex. $\bar{X}_n$ or $\hat{\mu}$ or S^2
- **Estimator-** a statistic (random variable) that describes the estimation procedure

Estimation

- **Parameter-** characteristic of the population distribution (the distribution of X_i)

  ex. μ or σ^2

- **Estimand-** parameter being estimated

  ex. μ

- **Statistic-** function of the sample used to estimate a parameter

  ex. $\bar{X}_n$ or $\hat{\mu}$ or S^2

- **Estimator-** a statistic (random variable) that describes the estimation procedure

  ex. $\bar{X}_n$

ESTIMATION

- **Parameter-** characteristic of the population distribution (the distribution of X_i)

  ex. μ or σ^2
- **Estimand-** parameter being estimated

  ex. μ
- **Statistic-** function of the sample used to estimate a parameter

  ex. $\bar{X}_n$ or $\hat{\mu}$ or S^2
- **Estimator-** a statistic (random variable) that describes the estimation procedure

  ex. $\bar{X}_n$
- **Estimate-** realized values of an estimator

Estimation

- **Parameter-** characteristic of the population distribution (the distribution of X_i)
  ex. μ or σ^2
- **Estimand-** parameter being estimated
  ex. μ
- **Statistic-** function of the sample used to estimate a parameter
  ex. $\bar{X}_n$ or $\hat{\mu}$ or S^2
- **Estimator-** a statistic (random variable) that describes the estimation procedure
  ex. $\bar{X}_n$
- **Estimate-** realized values of an estimator
  ex. $\bar{x}_n$

Estimation

Which are random variables?

ESTIMATION

Which are random variables?

Parameter?

ESTIMATION

Which are random variables?

Parameter?No

ESTIMATION

Which are random variables?

Parameter?No

Estimand?

ESTIMATION

Which are random variables?

Parameter?No

Estimand?No

ESTIMATION

Which are random variables?

Parameter?No

Estimand?No

Statistic?

Estimation

Which are random variables?

Parameter?No

Estimand?No

Statistic?Yes

ESTIMATION

Which are random variables?

Parameter?No

Estimand?No

Statistic?Yes

Estimator?

ESTIMATION

Which are random variables?

Parameter?No

Estimand?No

Statistic?Yes

Estimator?Yes

Estimation

Which are random variables?

Parameter?No

Estimand?No

Statistic?Yes

Estimator?Yes

Estimate?

ESTIMATION

Which are random variables?

Parameter?No

Estimand?No

Statistic?Yes

Estimator?Yes

Estimate?No

ESTIMATION

Which are random variables?

Parameter?No

Estimand?No

Statistic?Yes

Estimator?Yes

Estimate?No

Point Estimation

Point Estimation

Estimand μ

Point Estimation

Estimand μ

Estimator $\bar{X}_n$

Point Estimation

Estimand μ

Estimator $\bar{X}_n$

Estimate $\bar{x}_n$

Point Estimation

Estimand μ

Estimator $\bar{X}_n$

Estimate $\bar{x}_n$

Sampling Distribution of $\bar{X}_n$ with **Known** Population

1. Start with the population.

Sampling Distribution of $\bar{X}_n$ with **Known** Population

1. Start with the population.
2. Define the quantity of interest (the estimand). For us, it's μ.

Sampling Distribution of $\bar{X}_n$ with **Known** Population

1. Start with the population.
2. Define the quantity of interest (the estimand). For us, it's μ.
3. Choose a plausible estimator. For us, it's $\hat{\mu} = \bar{X}_n$.

Sampling Distribution of $\bar{X}_n$ with **Known** Population

1. Start with the population.
2. Define the quantity of interest (the estimand). For us, it's μ.
3. Choose a plausible estimator. For us, it's $\hat{\mu} = \bar{X}_n$.
4. Draw a sample of size n from the population and calculate the estimate, $\bar{x}_n$, using the estimator.

SAMPLING DISTRIBUTION OF $\bar{X}_n$ WITH **KNOWN** POPULATION

1. Start with the population.
2. Define the quantity of interest (the estimand). For us, it's μ.
3. Choose a plausible estimator. For us, it's $\hat{\mu} = \bar{X}_n$.
4. Draw a sample of size n from the population and calculate the estimate, $\bar{x}_n$, using the estimator.
5. Repeat step 4 many times (we will have $s = 1,000$).

Sampling Distribution of $\bar{X}_n$ with **Known** Population

1. Start with the population.
2. Define the quantity of interest (the estimand). For us, it's μ.
3. Choose a plausible estimator. For us, it's $\hat{\mu} = \bar{X}_n$.
4. Draw a sample of size n from the population and calculate the estimate, $\bar{x}_n$, using the estimator.
5. Repeat step 4 many times (we will have $s = 1,000$).
6. Create a density plot of your 1,000 estimates ($\bar{x}_n$) of $\bar{X}_n$.

SAMPLING DISTRIBUTION OF $\bar{X}_n$ WITH **KNOWN** POPULATION

1. Start with the population.
2. Define the quantity of interest (the estimand). For us, it's μ.
3. Choose a plausible estimator. For us, it's $\hat{\mu} = \bar{X}_n$.
4. Draw a sample of size n from the population and calculate the estimate, $\bar{x}_n$, using the estimator.
5. Repeat step 4 many times (we will have $s = 1,000$).
6. Create a density plot of your 1,000 estimates ($\bar{x}_n$) of $\bar{X}_n$.

Sampling Distribution of $\bar{X}_n$

The sampling distribution of $\bar{X}_n$ is the distribution of the column vector:

		Observations				$\hat{\mu} =$
		X_1	X_2	$\ldots$	X_n	$\bar{X}_n$
Samples (s)	1	$x_{1,1}$	$x_{2,1}$	$\ldots$	$x_{n,1}$	$\bar{x}_{n,1}$
	2	$x_{1,2}$	$x_{2,2}$	$\ldots$	$x_{n,2}$	$\bar{x}_{n,2}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	10,000	$x_{1,10000}$	$x_{2,10000}$	$\ldots$	$x_{n,10000}$	$\bar{x}_{n,10000}$

SAMPLING DISTRIBUTION OF $\bar{X}_n$ WITH **UNKOWN** POPULATION

In reality, we only get to observe one sample (not the entire population), so we cannot directly observe the sampling distribution of $\bar{X}_n$:

		Observations				$\hat{\mu} =$
		X_1	X_2	$\dots$	X_n	$\bar{X}_n$
Samples (s)	1	$x_{1,1}$	$x_{2,1}$	$\dots$	$x_{n,1}$	$\bar{x}_{n,1}$
	2	$x_{1,2}$	$x_{2,2}$	$\dots$	$x_{n,2}$	$\bar{x}_{n,2}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	10,000	$x_{1,10000}$	$x_{2,10000}$	$\dots$	$x_{n,10000}$	$\bar{x}_{n,10000}$

Central Limit Theorem

Remember that if our samples are sufficiently large ($n \geq 30$), the Central Limit Theorem states that the sample mean will be distributed as:

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

Estimated Sampling Distribution of $\bar{X}_n$ with **Unkown** Population

We could apply the CLT to get the sampling distribution of our estimator, except σ^2 is **unkown**.

ESTIMATED SAMPLING DISTRIBUTION OF $\bar{X}_n$ WITH **UNKOWN** POPULATION

We could apply the CLT to get the sampling distribution of our estimator, except σ^2 is **unkown**.

We can estimate σ^2 using the sample variance!

		Observations				$\hat{\mu} =$	$\hat{\sigma^2} =$	
		X_1	X_2	$\dots$	X_n	$\bar{X}_n$	S_n^2	$\hat{SE}[\bar{X}_n]$
	1	$x_{1,1}$	$x_{2,1}$	$\dots$	$x_{n,1}$	$\bar{x}_{n,1}$	$s_{n,1}^2$	$\frac{s_{n,1}}{\sqrt{n}}$
s	2	$x_{1,2}$	$x_{2,2}$	$\dots$	$x_{n,2}$	$\bar{x}_{n,2}$	$s_{n,2}^2$	$\frac{s_{n,2}}{\sqrt{n}}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	1000	$x_{1,1000}$	$x_{2,1000}$	$\dots$	$x_{n,1000}$	$\bar{x}_{n,1000}$	$s_{n,1000}^2$	$\frac{s_{n,1000}}{\sqrt{n}}$
	E[·]	μ	μ	$\dots$	μ	μ	σ^2	
	V[·]	σ^2	σ^2	$\dots$	σ^2	$\frac{\sigma^2}{n}$		

Estimated Sampling Distribution of $\bar{X}_n$ with **Unkown** Population

We could apply the CLT to get the sampling distribution of our estimator, except σ^2 is **unkown**.

We can estimate σ^2 using the sample variance!

		Observations				$\hat{\mu} =$	$\hat{\sigma}^2 =$	
		X_1	X_2	$\ldots$	X_n	$\bar{X}_n$	S_n^2	$\hat{SE}[\bar{X}_n]$
s	1	$x_{1,1}$	$x_{2,1}$	$\ldots$	$x_{n,1}$	$\bar{x}_{n,1}$	$s_{n,1}^2$	$\frac{s_{n,1}}{\sqrt{n}}$
	2	$x_{1,2}$	$x_{2,2}$	$\ldots$	$x_{n,2}$	$\bar{x}_{n,2}$	$s_{n,2}^2$	$\frac{s_{n,2}}{\sqrt{n}}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	1000	$x_{1,1000}$	$x_{2,1000}$	$\ldots$	$x_{n,1000}$	$\bar{x}_{n,1000}$	$s_{n,1000}^2$	$\frac{s_{n,1000}}{\sqrt{n}}$
	E[$\cdot$]	μ	μ	$\ldots$	μ	μ	σ^2	
	V[$\cdot$]	σ^2	σ^2	$\ldots$	σ^2	$\frac{\sigma^2}{n}$		

Central Limit Theorem

So, by the CLT again with n large we can estimate the sampling distribution of our estimator as :

$$\bar{X}_n \sim_{approx} N(\bar{x}_n, \frac{s_n^2}{n})$$

Generally, we report our point estimate $\bar{x}_n$ and the estimated standard error of our estimate $\frac{s_n}{\sqrt{n}}$.

BOOTSTRAPPED SAMPLING DISTRIBUTION OF $\bar{X}_n$ WITH **UNKNOWN** POPULATION

Or we can estimate the sampling distribution using boostrapping.

Bootstrapped Sampling Distribution of $\bar{X}_n$ with **Unknown** Population

Or we can estimate the sampling distribution using boostrapping.

1. Start with observed sample.

Bootstrapped Sampling Distribution of $\bar{X}_n$ with **Unknown** Population

Or we can estimate the sampling distribution using boostrapping.

1. Start with observed sample.
2. Define a quantity of interest (the estimand). For us, it's μ.

Bootstrapped Sampling Distribution of $\bar{X}_n$ with **Unknown** Population

Or we can estimate the sampling distribution using boostrapping.

1. Start with observed sample.
2. Define a quantity of interest (the estimand). For us, it's μ.
3. Choose a plausible estimator. For us, it's $\hat{\mu} = \bar{X}_n$.

Bootstrapped Sampling Distribution of $\bar{X}_n$ with **Unknown** Population

Or we can estimate the sampling distribution using boostrapping.

1. Start with observed sample.
2. Define a quantity of interest (the estimand). For us, it's μ.
3. Choose a plausible estimator. For us, it's $\hat{\mu} = \bar{X}_n$.
4. Take a resample of size n (with replacement) from the sample and calculate the estimate, $\bar{x}_n$, using the estimator.

Bootstrapped Sampling Distribution of $\bar{X}_n$ with **Unknown** Population

Or we can estimate the sampling distribution using boostrapping.

1. Start with observed sample.
2. Define a quantity of interest (the estimand). For us, it's μ.
3. Choose a plausible estimator. For us, it's $\hat{\mu} = \bar{X}_n$.
4. Take a resample of size n (with replacement) from the sample and calculate the estimate, $\bar{x}_n$, using the estimator.
5. Repeat step 4 many times (we will have $s = 1,000$).

BOOTSTRAPPED SAMPLING DISTRIBUTION OF $\bar{X}_n$ WITH **UNKNOWN** POPULATION

Or we can estimate the sampling distribution using boostrapping.

1. Start with observed sample.
2. Define a quantity of interest (the estimand). For us, it's μ.
3. Choose a plausible estimator. For us, it's $\hat{\mu} = \bar{X}_n$.
4. Take a resample of size n (with replacement) from the sample and calculate the estimate, $\bar{x}_n$, using the estimator.
5. Repeat step 4 many times (we will have $s = 1,000$).
6. Create a density plot of your 1,000 estimates ($\bar{x}_n$) of $\bar{X}_n$.

Bootstrapped Sampling Distribution of $\bar{X}_n$ with **Unknown** Population

Or we can estimate the sampling distribution using boostrapping.

1. Start with observed sample.
2. Define a quantity of interest (the estimand). For us, it's μ.
3. Choose a plausible estimator. For us, it's $\hat{\mu} = \bar{X}_n$.
4. Take a resample of size n (with replacement) from the sample and calculate the estimate, $\bar{x}_n$, using the estimator.
5. Repeat step 4 many times (we will have $s = 1,000$).
6. Create a density plot of your 1,000 estimates ($\bar{x}_n$) of $\bar{X}_n$.

Interval Estimation for **Large** Samples and σ **Known**

INTERVAL ESTIMATION FOR **LARGE** SAMPLES AND σ **KNOWN**

Estimand $[LB, UB]$ s.t. $P(LB \leq \mu \leq UB) = 100(1-\alpha)\%$

INTERVAL ESTIMATION FOR **LARGE** SAMPLES AND σ **KNOWN**

Estimand $[LB, UB]$ s.t. $P(LB \leq \mu \leq UB) = 100(1-\alpha)\%$

Estimator $\bar{X}_n \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

INTERVAL ESTIMATION FOR **LARGE** SAMPLES AND σ KNOWN

Estimand $[LB, UB]$ s.t. $P(LB \leq \mu \leq UB) = 100(1-\alpha)\%$

Estimator $\bar{X}_n \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

Estimate $\bar{x}_n \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

INTERVAL ESTIMATION FOR **LARGE** SAMPLES AND σ KNOWN

		Observations				
		X_1	X_2	$\dots$	X_n	$\bar{X}_n \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
	1	$x_{1,1}$	$x_{2,1}$	$\dots$	$x_{n,1}$	$\bar{x}_{n,1} \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
s	2	$x_{1,2}$	$x_{2,2}$	$\dots$	$x_{n,2}$	$\bar{x}_{n,2} \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	1000	$x_{1,1000}$	$x_{2,1000}$	$\dots$	$x_{n,1000}$	$\bar{x}_{n,1000} \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

Interval Estimation for **Large** Samples and σ **Known**

How did we choose our estimator?

Interval Estimation for **Large** Samples and σ **Known**

How did we choose our estimator?

By the CLT, if n is large, then $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$.

Interval Estimation for **Large** Samples and σ Known

How did we choose our estimator?

By the CLT, if n is large, then $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$.

If we standardize, we get that $Z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$

INTERVAL ESTIMATION FOR **LARGE** SAMPLES AND σ KNOWN

How did we choose our estimator?

By the CLT, if n is large, then $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$.

If we standardize, we get that $Z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$

Which implies that:

$$P(-z_{1-\alpha/2} \leq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\alpha/2}) = 1 - \alpha$$

Interval Estimation for **Large** Samples and σ **Known**

How did we choose our estimator?

By the CLT, if n is large, then $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$.

If we standardize, we get that $Z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

Which implies that:

$$
\begin{aligned}
& P(-z_{1-\alpha/2} \leq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\alpha/2}) = 1 - \alpha \\
\Rightarrow \quad & P(\bar{X}_n - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha
\end{aligned}
$$

Interval Estimation for **Large** Samples and σ Known

How did we choose our estimator?

By the CLT, if n is large, then $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$.

If we standardize, we get that $Z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$

Which implies that:

$$
\begin{aligned}
& P(-z_{1-\alpha/2} \leq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\alpha/2}) = 1 - \alpha \\
\Rightarrow \quad & P(\bar{X}_n - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha
\end{aligned}
$$

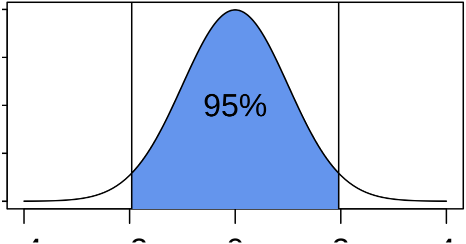
Interval Estimation for **Large** Samples and σ **Known**

How do we get $z_{1-\alpha/2}$?

Interval Estimation for **Large** Samples and σ **Known**

How do we get $z_{1-\alpha/2}$?

```
## alpha = .05
qnorm(1-.025, mean = 0, sd=1)
```



INTERVAL ESTIMATION FOR **LARGE** SAMPLES AND σ **KNOWN**

How do we interpret our estimated confidence interval?

INTERVAL ESTIMATION FOR **LARGE** SAMPLES AND σ **KNOWN**

How do we interpret our estimated confidence interval?
The true population mean μ will lie within the estimated intervals in 100(1-α)% of many repeated samples.

Quiz Break!

What is random in our estimator: $\bar{X}_n \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$?

Quiz Break!

What is random in our estimator: $\bar{X}_n \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$? $\bar{X}_n$

Quiz Break!

What is random in our estimator: $\bar{X}_n \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$? $\bar{X}_n$

What is random in our estimate: $\bar{x}_n \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$?

QUIZ BREAK!

What is random in our estimator: $\bar{X}_n \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$? $\bar{X}_n$

What is random in our estimate: $\bar{x}_n \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$? Nothing!

Interval Estimation for **Large** Samples and σ **Unknown**

INTERVAL ESTIMATION FOR **LARGE** SAMPLES AND σ **UNKNOWN**

Estimand $[LB, UB]$ s.t. $P(LB \leq \mu \leq UB) = 100(1-\alpha)\%$

Interval Estimation for **Large** Samples and σ **Unknown**

Estimand $[LB, UB]$ s.t. $P(LB \leq \mu \leq UB) = 100(1-\alpha)\%$

Estimator $\bar{X}_n \pm z_{1-\alpha/2} \cdot \frac{S_n}{\sqrt{n}}$

INTERVAL ESTIMATION FOR **LARGE** SAMPLES AND σ **UNKNOWN**

Estimand $[LB, UB]$ s.t. $P(LB \leq \mu \leq UB) = 100(1-\alpha)\%$

Estimator $\bar{X}_n \pm z_{1-\alpha/2} \cdot \frac{S_n}{\sqrt{n}}$

Estimate $\bar{x}_n \pm z_{1-\alpha/2} \cdot \frac{s_n}{\sqrt{n}}$

Interval Estimation for **Large** Samples and σ **Unknown**

		Observations				
		X_1	X_2	$\dots$	X_n	$\bar{X}_n \pm z_{1-\alpha/2} \cdot \frac{S_n}{\sqrt{n}}$
	1	$x_{1,1}$	$x_{2,1}$	$\dots$	$x_{n,1}$	$\bar{x}_{n,1} \pm z_{1-\alpha/2} \cdot \frac{s_{n,1}}{\sqrt{n}}$
s	2	$x_{1,2}$	$x_{2,2}$	$\dots$	$x_{n,2}$	$\bar{x}_{n,2} \pm z_{1-\alpha/2} \cdot \frac{s_{n,2}}{\sqrt{n}}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	1000	$x_{1,1000}$	$x_{2,1000}$	$\dots$	$x_{n,1000}$	$\bar{x}_{n,1000} \pm z_{1-\alpha/2} \cdot \frac{s_{n,1000}}{\sqrt{n}}$

Interval Estimation for **Large** Samples and σ **Unknown**

How did we choose our estimator?

INTERVAL ESTIMATION FOR **LARGE** SAMPLES AND σ **UNKNOWN**

How did we choose our estimator?

Same as before except we *estimated* $\sigma/\sqrt{n}$ with $\hat{SE}[\bar{X}_n] = \frac{S_n}{\sqrt{n}}$.

INTERVAL ESTIMATION FOR **LARGE** SAMPLES AND σ **UNKNOWN**

How did we choose our estimator?

Same as before except we *estimated* $\sigma/\sqrt{n}$ with $\hat{SE}[\bar{X}_n] = \frac{S_n}{\sqrt{n}}$.

Are we still getting 95% coverage over repeated samples?

Interval Estimation for **Large** Samples and σ **Unknown**

How did we choose our estimator?

Same as before except we *estimated* $\sigma/\sqrt{n}$ with $\hat{SE}[\bar{X}_n] = \frac{S_n}{\sqrt{n}}$.

Are we still getting 95% coverage over repeated samples?
No, but as n increases, this matters less.

Quiz Break!

What is random in our estimator: $\bar{X}_n \pm z_{1-\alpha/2} \cdot \frac{S_n}{\sqrt{n}}$?

Quiz Break!

What is random in our estimator: $\bar{X}_n \pm z_{1-\alpha/2} \cdot \frac{S_n}{\sqrt{n}}$? $\bar{X}_n$ **and** S_n

What is random in our estimate: $\bar{x}_n \pm z_{1-\alpha/2} \cdot \frac{s_n}{\sqrt{n}}$?

Quiz Break!

What is random in our estimator: $\bar{X}_n \pm z_{1-\alpha/2} \cdot \frac{S_n}{\sqrt{n}}$? $\bar{X}_n$ **and** S_n

What is random in our estimate: $\bar{x}_n \pm z_{1-\alpha/2} \cdot \frac{s_n}{\sqrt{n}}$? Nothing!

Interval Estimation for **Small** Samples and σ **Unknown**

INTERVAL ESTIMATION FOR **SMALL** SAMPLES AND σ **UNKNOWN**

Assuming $X_i \sim_{i.i.d} N(\mu, \sigma^2)$,

Estimand $[LB, UB]$ s.t. $P(LB \leq \mu \leq UB) = 100(1-\alpha)\%$

INTERVAL ESTIMATION FOR **SMALL** SAMPLES AND σ **UNKNOWN**

Assuming $X_i \sim_{i.i.d} N(\mu, \sigma^2)$,

Estimand $[LB, UB]$ s.t. $P(LB \leq \mu \leq UB) = 100(1-\alpha)\%$

Estimator $\bar{X}_n \pm t_{n-1,1-\alpha/2} \cdot \frac{S_n}{\sqrt{n}}$

Interval Estimation for **Small** Samples and σ **Unknown**

Assuming $X_i \sim_{i.i.d} N(\mu, \sigma^2)$,

Estimand $[LB, UB]$ s.t. $P(LB \leq \mu \leq UB) = 100(1-\alpha)\%$

Estimator $\bar{X}_n \pm t_{n-1,1-\alpha/2} \cdot \frac{S_n}{\sqrt{n}}$

Estimate $\bar{x}_n \pm t_{n-1,1-\alpha/2} \cdot \frac{s_n}{\sqrt{n}}$

INTERVAL ESTIMATION FOR **SMALL** SAMPLES AND σ **UNKNOWN**

		Observations				
		X_1	X_2	$\dots$	X_n	$\bar{X}_n \pm t_{n-1,1-\alpha/2} \cdot \frac{S_n}{\sqrt{n}}$
	1	$x_{1,1}$	$x_{2,1}$	$\dots$	$x_{n,1}$	$\bar{x}_{n,1} \pm t_{n-1,1-\alpha/2} \cdot \frac{s_{n,1}}{\sqrt{n}}$
s	2	$x_{1,2}$	$x_{2,2}$	$\dots$	$x_{n,2}$	$\bar{x}_{n,2} \pm t_{n-1,1-\alpha/2} \cdot \frac{s_{n,2}}{\sqrt{n}}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	1000	$x_{1,1000}$	$x_{2,1000}$	$\dots$	$x_{n,1000}$	$\bar{x}_{n,1000} \pm t_{n-1,1-\alpha/2} \cdot \frac{s_{n,1000}}{\sqrt{n}}$

Interval Estimation for **Small** Samples and σ **Unknown**

Why $t_{n-1,1-\alpha/2}$ instead of $z_{1-\alpha/2}$?

Interval Estimation for **Small** Samples and σ **Unknown**

Why $t_{n-1,1-\alpha/2}$ instead of $z_{1-\alpha/2}$?

When we standardize $\bar{X}_n$ and estimate σ with S_n,

$$\frac{\bar{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} \sim t_{n-1}$$

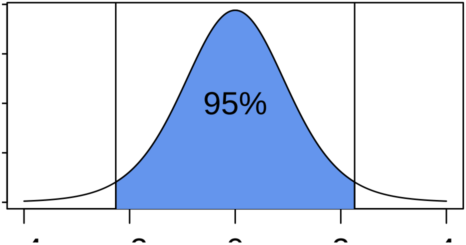
Interval Estimation for **Small** Samples and σ **Unknown**

How do we get $t_{n-1,1-\alpha/2}$?

Interval Estimation for **Small** Samples and σ **Unknown**

How do we get $t_{n-1,1-\alpha/2}$?

```
## alpha = .05
## 9 degrees of freedom
qt(1-.025, df=9)
```



Quiz Break

You are told that your estimand is $[LB, UB]$ s.t. $P(LB \leq \mu_1 - \mu_2 \leq UB) = 100(1-\alpha)\%$. Assuming n_1 and n_2 are both large and σ_1 and σ_2 are both unkown, what is your estimator and estimate?

Estimand $[LB, UB]$ s.t. $P(LB \leq \mu_1 - \mu_2 \leq UB) = 100(1-\alpha)\%$

Estimator

Quiz Break

You are told that your estimand is $[LB, UB]$ s.t. $P(LB \leq \mu_1 - \mu_2 \leq UB) = 100(1-\alpha)\%$. Assuming n_1 and n_2 are both large and σ_1 and σ_2 are both unkown, what is your estimator and estimate?

Estimand $[LB, UB]$ s.t. $P(LB \leq \mu_1 - \mu_2 \leq UB) = 100(1-\alpha)\%$

Estimator $\bar{X}_{n,1} - \bar{X}_{n,2} \pm z_{1-\alpha/2} \cdot \sqrt{\frac{S_{n,1}^2}{n_1} + \frac{S_{n,2}^2}{n_2}}$

Estimate

Quiz Break

You are told that your estimand is $[LB, UB]$ s.t. $P(LB \leq \mu_1 - \mu_2 \leq UB) = 100(1-\alpha)\%$. Assuming n_1 and n_2 are both large and σ_1 and σ_2 are both unkown, what is your estimator and estimate?

Estimand $[LB, UB]$ s.t. $P(LB \leq \mu_1 - \mu_2 \leq UB) = 100(1-\alpha)\%$

Estimator $\bar{X}_{n,1} - \bar{X}_{n,2} \pm z_{1-\alpha/2} \cdot \sqrt{\frac{S^2_{n,1}}{n_1} + \frac{S^2_{n,2}}{n_2}}$

Estimate $\bar{x}_{n,1} - \bar{x}_{n,2} \pm z_{1-\alpha/2} \cdot \sqrt{\frac{s^2_{n,1}}{n_1} + \frac{s^2_{n,2}}{n_2}}$

Hypothesis Testing

1. Set your **null hypothesis-** the assumed null state of the world.

Hypothesis Testing

1. Set your **null hypothesis-** the assumed null state of the world.

   ex. $H_0 : \mu = 0$

Hypothesis Testing

1. Set your **null hypothesis-** the assumed null state of the world.

   ex. $H_0 : \mu = 0$

2. Set your **alternative hypothesis-** the claim to be tested.

Hypothesis Testing

1. Set your **null hypothesis-** the assumed null state of the world.

   ex. $H_0 : \mu = 0$

2. Set your **alternative hypothesis-** the claim to be tested.

   ex. $H_A : \mu \neq 0$

Hypothesis Testing

1. Set your **null hypothesis-** the assumed null state of the world.

   ex. $H_0 : \mu = 0$
2. Set your **alternative hypothesis-** the claim to be tested.

   ex. $H_A : \mu \neq 0$
3. Choose your **test statistic (estimator)-** function of the sample and the null hypothesis used to test your hypothesis.

Hypothesis Testing

1. Set your **null hypothesis-** the assumed null state of the world.

   ex. $H_0 : \mu = 0$

2. Set your **alternative hypothesis-** the claim to be tested.

   ex. $H_A : \mu \neq 0$

3. Choose your **test statistic (estimator)-** function of the sample and the null hypothesis used to test your hypothesis.

   ex. $\frac{\bar{X}_n - \mu_0}{\frac{S_n}{\sqrt{n}}}$

Hypothesis Testing

1. Set your **null hypothesis-** the assumed null state of the world.

   ex. $H_0 : \mu = 0$

2. Set your **alternative hypothesis-** the claim to be tested.

   ex. $H_A : \mu \neq 0$

3. Choose your **test statistic (estimator)-** function of the sample and the null hypothesis used to test your hypothesis.

   ex. $\frac{\bar{X}_n - \mu_0}{\frac{S_n}{\sqrt{n}}}$

4. Determine the **null distribution-** the sampling distribution of the test statistic assuming that the null hypothesis is true.

HYPOTHESIS TESTING

1. Set your **null hypothesis-** the assumed null state of the world.

   ex. $H_0 : \mu = 0$

2. Set your **alternative hypothesis-** the claim to be tested.

   ex. $H_A : \mu \neq 0$

3. Choose your **test statistic (estimator)-** function of the sample <u>and</u> the null hypothesis used to test your hypothesis.

   ex. $\frac{\bar{X}_n - \mu_0}{\frac{S_n}{\sqrt{n}}}$

4. Determine the **null distribution-** the sampling distribution of the test statistic assuming that the null hypothesis is true.

   ex. $\frac{\bar{X}_n - 0}{\frac{S_n}{\sqrt{n}}} \sim N(0, 1)$

Hypothesis Testing

5. Set your **decision rule-** the function that specifies where we accept or reject the null hypothesis for a given value of the test statistic; a function of α.

Hypothesis Testing

5. Set your **decision rule-** the function that specifies where we accept or reject the null hypothesis for a given value of the test statistic; a function of α.

   ex. reject null if $\left|\frac{\bar{X}_n - 0}{\frac{S_n}{\sqrt{n}}}\right| \geq z_{1-\alpha/2}$ or ≥ 1.96

Hypothesis Testing

5. Set your **decision rule-** the function that specifies where we accept or reject the null hypothesis for a given value of the test statistic; a function of α.

   ex. reject null if $\left|\frac{\bar{X}_n - 0}{\frac{S_n}{\sqrt{n}}}\right| \geq z_{1-\alpha/2}$ or ≥ 1.96
6. Calculate the **rejection region-** the set of values for which we will reject the null hypothesis.

HYPOTHESIS TESTING

5. Set your **decision rule-** the function that specifies where we accept or reject the null hypothesis for a given value of the test statistic; a function of α.

   ex. reject null if $\left|\frac{\bar{X}_n - 0}{\frac{S_n}{\sqrt{n}}}\right| \geq z_{1-\alpha/2}$ or ≥ 1.96

6. Calculate the **rejection region-** the set of values for which we will reject the null hypothesis.

   ex. $[-\infty, -z_{1-\alpha/2}]$ and $[z_{1-\alpha/2}, \infty]$
   or $[-\infty, -1.96]$ and $[1.96, \infty]$

Hypothesis Testing

5. Set your **decision rule-** the function that specifies where we accept or reject the null hypothesis for a given value of the test statistic; a function of α.

   ex. reject null if $\left|\frac{\bar{X}_n - 0}{\frac{S_n}{\sqrt{n}}}\right| \geq z_{1-\alpha/2}$ or ≥ 1.96

6. Calculate the **rejection region-** the set of values for which we will reject the null hypothesis.

   ex. $[-\infty, -z_{1-\alpha/2}]$ and $[z_{1-\alpha/2}, \infty]$
   or $[-\infty, -1.96]$ and $[1.96, \infty]$

7. Using your sample, calculate your observed estimate.

Hypothesis Testing

5. Set your **decision rule-** the function that specifies where we accept or reject the null hypothesis for a given value of the test statistic; a function of α.

   ex. reject null if $\left|\frac{\bar{X}_n - 0}{\frac{S_n}{\sqrt{n}}}\right| \geq z_{1-\alpha/2}$ or ≥ 1.96

6. Calculate the **rejection region-** the set of values for which we will reject the null hypothesis.

   ex. $[-\infty, -z_{1-\alpha/2}]$ and $[z_{1-\alpha/2}, \infty]$
   or $[-\infty, -1.96]$ and $[1.96, \infty]$

7. Using your sample, calculate your observed estimate.

   ex. $\frac{\bar{x}_n - \mu_0}{\frac{s_n}{\sqrt{n}}}$

Hypothesis Testing

5. Set your **decision rule-** the function that specifies where we accept or reject the null hypothesis for a given value of the test statistic; a function of α.

   ex. reject null if $\left|\frac{\bar{X}_n - 0}{\frac{S_n}{\sqrt{n}}}\right| \geq z_{1-\alpha/2}$ or ≥ 1.96

6. Calculate the **rejection region-** the set of values for which we will reject the null hypothesis.

   ex. $[-\infty, -z_{1-\alpha/2}]$ and $[z_{1-\alpha/2}, \infty]$
   or $[-\infty, -1.96]$ and $[1.96, \infty]$

7. Using your sample, calculate your observed estimate.

   ex. $\frac{\bar{x}_n - \mu_0}{\frac{s_n}{\sqrt{n}}}$

8. Determine if your estimate is within the rejection region and draw conclusion about hypothesis.

Hypothesis Testing

5. Set your **decision rule-** the function that specifies where we accept or reject the null hypothesis for a given value of the test statistic; a function of α.

   ex. reject null if $\left|\frac{\bar{X}_n - 0}{\frac{S_n}{\sqrt{n}}}\right| \geq z_{1-\alpha/2}$ or ≥ 1.96

6. Calculate the **rejection region-** the set of values for which we will reject the null hypothesis.

   ex. $[-\infty, -z_{1-\alpha/2}]$ and $[z_{1-\alpha/2}, \infty]$
   or $[-\infty, -1.96]$ and $[1.96, \infty]$

7. Using your sample, calculate your observed estimate.

   ex. $\frac{\bar{x}_n - \mu_0}{\frac{s_n}{\sqrt{n}}}$

8. Determine if your estimate is within the rejection region and draw conclusion about hypothesis.

   ex.

$$\begin{array}{lll} \text{Reject null} & \text{if} & \frac{\bar{x}_n - \mu_0}{\frac{s_n}{\sqrt{n}}} \in [-\infty, -1.96] \cup [1.96, \infty] \\ \text{Fail to reject null} & \text{if} & \frac{\bar{x}_n - \mu_0}{\frac{s_n}{\sqrt{n}}} \notin [-\infty, -1.96] \cup [1.96, \infty] \end{array}$$

P-Values

- **p-value-** Assuming that the null hypothesis is true, the probability of getting something at least as extreme as our observed test statistic, where extreme is defined in terms of the alternative hypothesis.

P-Values

- **p-value-** Assuming that the null hypothesis is true, the probability of getting something at least as extreme as our observed test statistic, where extreme is defined in terms of the alternative hypothesis.

How do we find p-values?

P-Values

- **p-value-** Assuming that the null hypothesis is true, the probability of getting something at least as extreme as our observed test statistic, where extreme is defined in terms of the alternative hypothesis.

How do we find p-values?

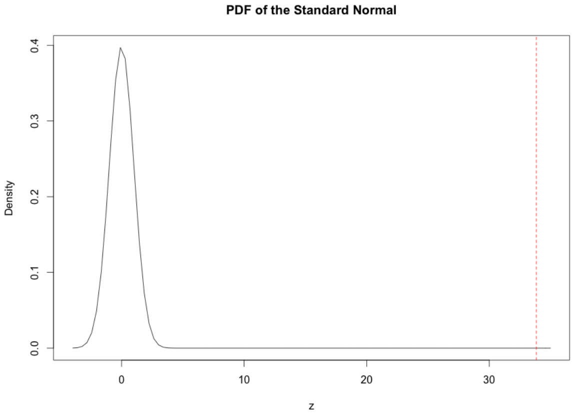
```
# create a population with mean 12, sd 4
pop <- rnorm(n=1e6, mean=12, sd=4)

# draw a single sample of size 100 from population
my.sample <- sample(pop, size=100, replace=T)

# calculate our test statistic
test.statistic <- mean(my.sample) /(sd(my.sample)/10)

# find the p-value
p.value <- 2*(pnorm(test.statistic))
```

P-Values



Questions

Questions?