ECON 340
Economic Research Methods

Div Bhagia

Lecture 1: Introduction

So many questions...

Always have questions that need answers

- Do electric vehicle subsidies increase sales?
- Does the use of phones inhibit classroom learning?
- Is there racial discrimination in the labor market?
- Does raising interest rates lead to inflation?
- Who will win the next US election?

Quantitive Empirical Research

- A *research question* is any question you plan to answer by conducting research
- *Empirical research* is based on real-world observations
- *Quantitative empirical research*: empirical research that uses quantitative measurements
- In this class, we will learn to answer a research question using quantitative empirical research

Quantitive Empirical Research

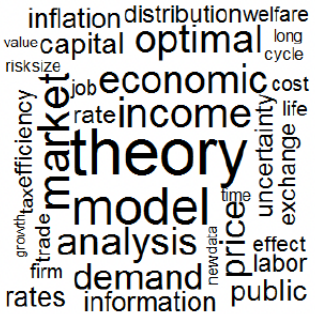
Everyone is using it (for good reason)

- Economists, other social scientists
- Think tanks, governments, policymakers
- Businesses

Our world is becoming more and more data-oriented.

Most used words in economic papers

1970s



1990s

2010-2015

This Course

Introduce you to tools used in quantitative research

Main goals:

- Understand statistical and econometric methods
- Be able to implement these methods in R
- Carry out a research project

Course Components

- Active Engagement (10%)
- Problem Sets (20%)
- Research Paper: Interim Submission (10%)
- Research Paper: Final Submissions (20%)
- Midterm (20%)
- Final Exam (20%)

Research Project

- As a part of this class, you will write an empirical research paper *using R*
- You will pick a question and a dataset and use the tools from this class to answer your question
- You can pick a dataset from the list of datasets provided on Canvas or use an external dataset
- If you pick an external dataset, please run it by me well in advance of your submissions so I can make sure it works

Research Project: Dates

- March 28: Interim Submission worth 10% (pick dataset and question and perform preliminary data analysis)
- April 9: Feedback on your research question
- May 7: Final paper due worth 20%

Introductions

- preferred name and pronouns
- major and year at CSUF
- what is your comfort food?
- what do you want to get out of this class?

Who likes greek letters?

Summation Notation

$$\sum_{i=1}^{N} X_i = X_1 + X_2 + ... + X_N$$

Example:

$X = \{2, 9, 6, 8, 11, 14\}$

$\sum_{i=1}^{4} X_i = X_1 + X_2 + X_3 + X_4 = 2 + 9 + 6 + 8 = 25$

Summation Notation

Another way of using a summation sign is to write

$$\sum_{x \in A} x$$

which refers to summing up all elements in A.

To sum up x for all possible values x, we can simply write

$$\sum_{x} x$$

Things you CAN do

1. Pull constants out of or into the summation sign.

$$\sum_{i=1}^{N} bX_i = b \sum_{i=1}^{N} X_i$$

Things you CAN do

2. Split apart (or combine) sums (addition) or differences (subtraction)

$$\sum_{i=1}^{N}(bX_i + cY_i) = b\sum_{i=1}^{N} X_i + c\sum_{i=1}^{N} Y_i$$

Things you CAN do

3. Multiply through constants by the number of terms in the summation

$$\sum_{i=1}^{N}(a + bX_i) = aN + b\sum_{i=1}^{N} X_i$$

Things you CANNOT do

1. Split apart (or combine) products (multiplication) or quotients (division).

$$\sum_{i=1}^{N} X_i Y_i \neq \sum_{i=1}^{N} X_i \times \sum_{i=1}^{N} Y_i$$

Things you CANNOT do

2. Move the exponent out of or into the summation.

$$\sum_{i=1}^{N} X_i^a \neq \left(\sum_{i=1}^{N} X_i \right)^a$$

Things To Do Until Next Class

1. Review the syllabus carefully
2. Make sure you can access the course content on Canvas
3. Install R and R Studio on your computer (how to handout on Canvas)
4. Work on Class Handout 1

ECON 340
Economic Research Methods

Div Bhagia

Lecture 2
Empirical Distribution & Measures of Central Tendency

Describing Data

A dataset is a collection of variables. Each variable contains multiple observations of the same measurement.

Types of variables:

- *Categorical*: gender, race, education (*binary*: two categories)
- *Continuous*: income, age, GPA

How do we summarize the information contained in a variable?

The Empirical Distribution

How often do different values occur?

For categorical variables:

$$f_k = \frac{n_k}{n} = \frac{\text{observations in category } k}{\text{total observations}}$$

f_k captures the relative frequency of outcome k.

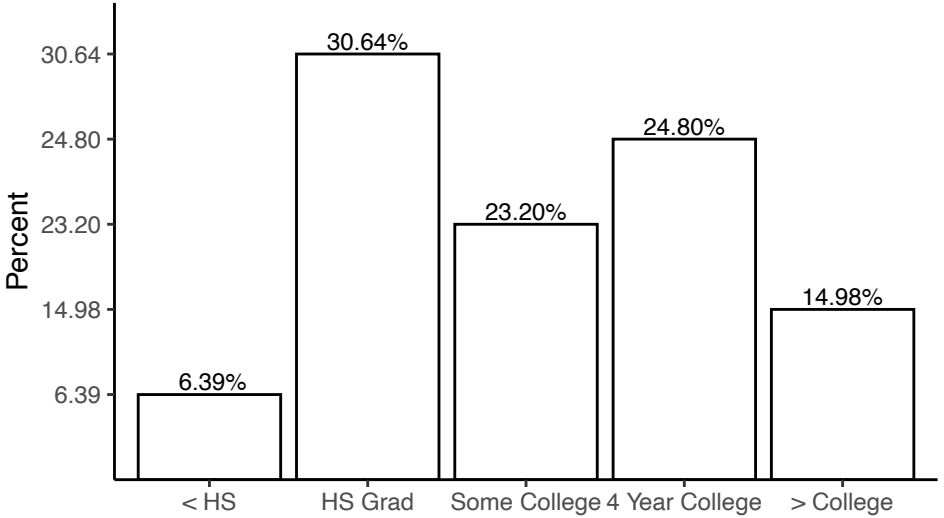
Frequency Distribution Table

Education	Count	Percent
$<$ HS	1540	6.39
HS Grad	7388	30.64
Some College	5595	23.20
4 Year College	5979	24.80
$>$ College	3611	14.98
Total	24113	100

Frequency Distribution Table

Education	Count	Percent	Cumulative
$<$ HS	1540	6.39	6.39
HS Grad	7388	30.64	37.03
Some College	5595	23.20	60.23
4 Year College	5979	24.80	85.02
$>$ College	3611	14.98	100.00
Total	24113	100	

Histogram: Education



The Empirical Distribution

What about continuous variables?

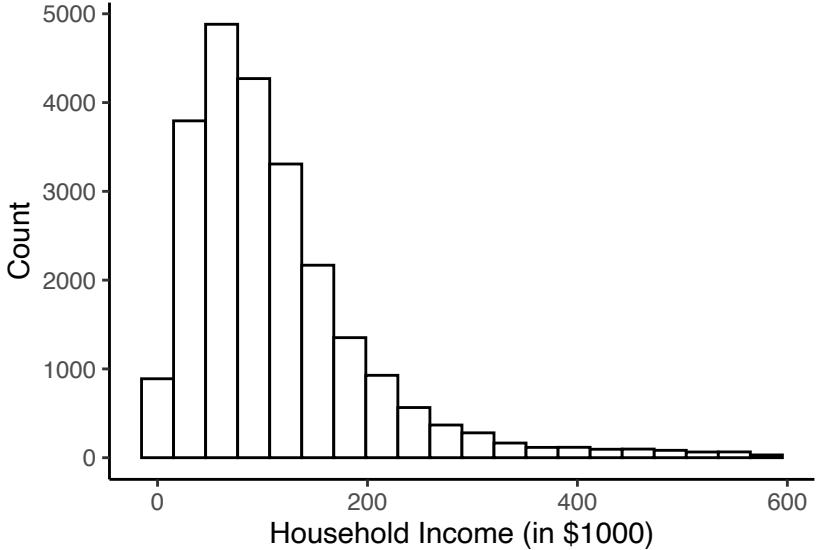
The Empirical Distribution

What about continuous variables?

How often do different values occur in a particular interval?

$$f_k = \frac{\text{observations in } \textit{interval } k}{\text{total observations}}$$

Histogram: Household Income



Source: American Community Survey (ACS) 2019

Measures of Central Tendency

Mean: is the average value

Median: is the middle value

Mode: is the number that is repeated more often than any other

Example: 5, 5, 10, 10, 10, 10, 20

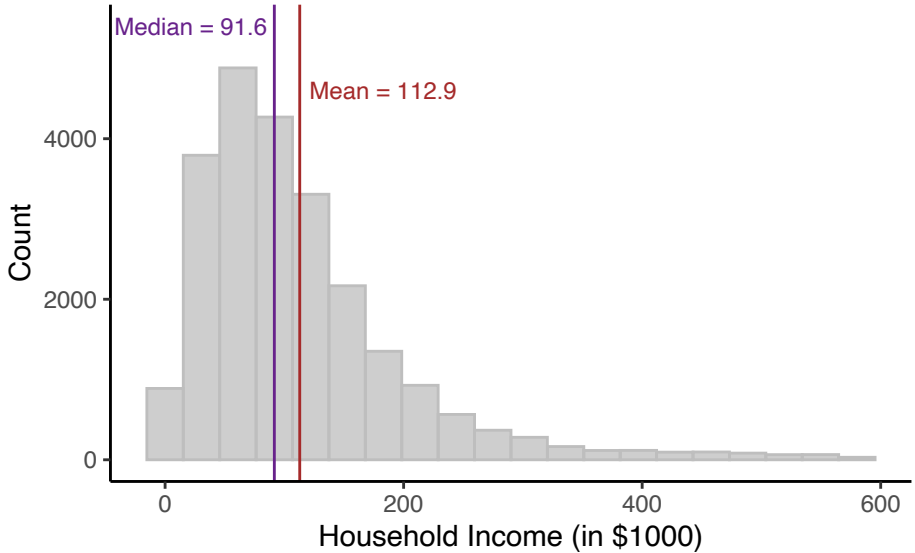
Mean

To calculate the mean:

$$\bar{X} = \frac{\text{sum of all observations}}{\text{number of observations}} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Use $\bar{X}$ to denote the sample mean and μ to denote the population mean.

Mean vs Median



Source: American Community Survey (ACS) 2019

Mean vs Median

- Mean household income: $112,900
- Median household income: $91,600

Why are mean earnings higher than the median?

Percentiles

The P^{th} **percentile** is a value such that P% of observations are at or below that number.

25th percentile a.k.a 1st quartile
75th percentile a.k.a 3rd quartile

What is the 50th percentile called?

More about Mean

- $\sum_{i=1}^{n} X_i = n\bar{X}$

More about Mean

- $\sum_{i=1}^{n} X_i = n\bar{X}$
- Deviations from the mean are always zero

$$\sum_{i=1}^{n}(X_i - \bar{X}) = \sum_{i=1}^{n} X_i - n\bar{X} = n\bar{X} - n\bar{X} = 0$$

More about Mean

- $\sum_{i=1}^{n} X_i = n\bar{X}$
- Deviations from the mean are always zero

$$\sum_{i=1}^{n}(X_i - \bar{X}) = \sum_{i=1}^{n} X_i - n\bar{X} = n\bar{X} - n\bar{X} = 0$$

- We can always write

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \frac{X_i}{n}$$

An easier way to calculate mean

- If data is grouped, we can use the frequency distribution table to calculate the mean:

$$\bar{X} = \frac{\sum_{k=1}^{K} n_k X_k}{n} = \sum_{k=1}^{K} f_k X_k$$

- Previous example: 5, 5, 10, 10, 10, 10, 20

X_k	n_k	f_k	$X_k f_k$
5	2		
10	4		
20	1		
Total	7		

Weighted Mean

The weighted mean of a set of data is

$$\bar{X} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

where w_i is the weight of the i^{th} observation.

Why might we want to use a weighted mean?

2016 Election Predictions

The New York Times

TheUpshot

POLITICAL CALCULUS

A 2016 Review: Why Key State Polls Were Wrong About Trump

By Nate Cohn

May 31, 2017

193

Education weighting seems to explain a lot

Education was a huge driver of presidential vote preference in the 2016 election, but many pollsters did not adjust their samples — a process known as weighting — to make sure they had the right number of well-educated or less educated respondents.

It's no small matter, since well-educated voters are much likelier to take surveys than less educated ones. About 45 percent of respondents in a typical national poll of adults will have a bachelor's degree or higher, even though the census says that only 28 percent of adults (those 18 and over) have a degree. Similarly, a bit more than 50 percent of respondents who say they're likely to vote have a degree, compared with 40 percent of voters in newly released 2016 census voting data.

Things to do next

- Review this week's material; handouts, notes, and reading (NYT article) on Canvas
- You may be asked to summarize what you got out of the reading in the next class

ECON 340
Economic Research Methods

Div Bhagia

Lecture 3
Variance, Standard Deviation, Z-Score

NYT Article: 2016 Election Predictions

- Summarize the main issue being discussed in the article.
- What were the three types of errors identified in the article? What is the common thread across these errors?
- One of the fixes suggested in the article was “education weighting”. Which of the three errors would this fix and how?

NYT Article: 2016 Election Predictions

- Summarize the main issue being discussed in the article.
- What were the three types of errors identified in the article? What is the common thread across these errors?
- One of the fixes suggested in the article was “education weighting”. Which of the three errors would this fix and how?
- In general, how can we pick a sample that is representative of the population to avoid having to reweight?

Another Example

- We want to estimate the average starting salary of students at a university that has only two majors
- Half of the students are *Business* majors, while the other half are *Engineering* majors
- Randomly select 100 Business students and 100 Engineering for a survey
- Response rate among Business students is 100%, while it 50% for engineering students

How can we use weighting to adjust for this?

Last Class

How to describe variables?

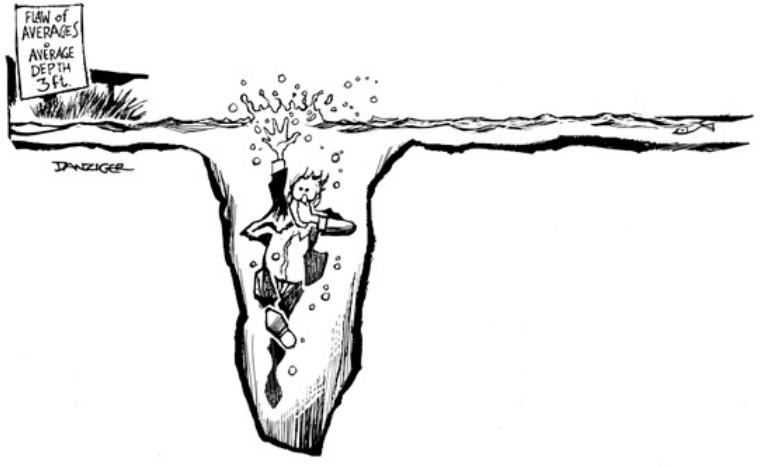
- Empirical Distribution
- Measures of central tendency: mean and median

μ : population mean, $\bar{X}$: sample mean

Two equivalent formulas:

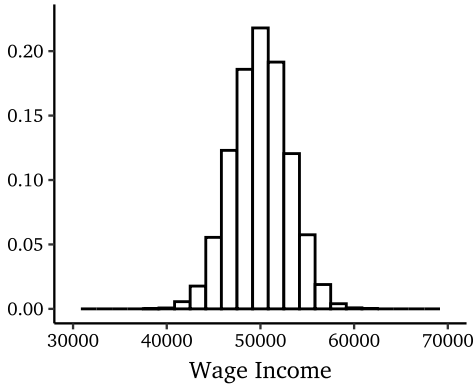
$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \bar{X} = \sum_{k=1}^{K} f_k X_k$$

Measures of central tendency are not enough!



Where would you want to live?

Mushroom Kingdom



Mean = Median= $50,000

Bowser's Kingdom

Mean = Median= $50,000

Deviations from the Mean

- Even with identical mean and median, the two countries are not identical.
- There is certainly more *dispersion* or *variability* in income in Bowser's Kingdom.
- More observations are *further from the mean* in Bowser's Kingdom.
- *What could be a potential statistic that could capture this?*

Deviations from the Mean

One option: average deviations from the mean. *Will this work?*

X_i	$X_i - \mu$
5	
5	
10	
10	
20	

Deviations from the Mean

Why does this not work? Remember from the last class:

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X}) &= \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \bar{X} \qquad (Why?)\\ &= \sum_{i=1}^{n} X_i - n\bar{X}\\ &= n\bar{X} - n\bar{X} = 0 \qquad (Why?)\end{aligned}$$

Deviations from the Mean

Why does this not work? Remember from the last class:

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X}) &= \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \bar{X} \qquad (Why?) \\ &= \sum_{i=1}^{n} X_i - n\bar{X} \\ &= n\bar{X} - n\bar{X} = 0 \qquad (Why?)\end{aligned}$$

Can you think of a way to construct a statistic that would capture variation around the mean?

Variance and Standard Deviation

Population Variance

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_x)^2$$

Sample Variance

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

Standard Deviation

$$\sigma_X = \sqrt{\sigma_X^2} \qquad S_X = \sqrt{S_X^2}$$

Variance and Standard Deviation

Back to our example.

X_i	$(X_i - \mu)$	$(X_i - \mu)^2$
5	-5	
5	-5	
10	0	
10	0	
20	10	
50	**0**	

Variance with Grouped Data

Population Variance

$$\sigma_X^2 = \sum_{k=1}^{K} f_k (X_k - \mu_X)^2$$

Sample Variance

$$S_X^2 = \frac{n}{n-1} \sum_{k=1}^{K} f_k (X_k - \bar{X})^2$$

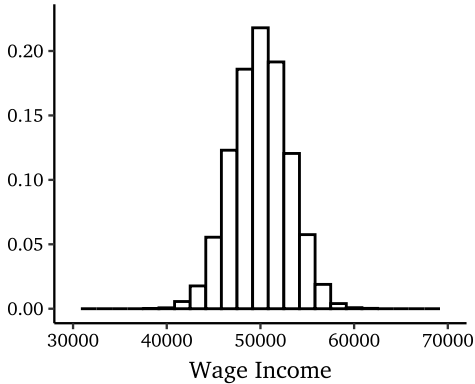
Variance with Grouped Data

In our example: 5, 5, 10, 10, 20. Present this as:

X_k	f_k	$f_k X_k$	$(X_k - \mu)^2$	$f_k(X_k - \mu)^2$
5	2/5			
10	2/5			
20	1/5			
Total				

Where would you want to live?

Mushroom Kingdom



Mean = Median= $50,000
SD= $3,000

Bowser's Kingdom

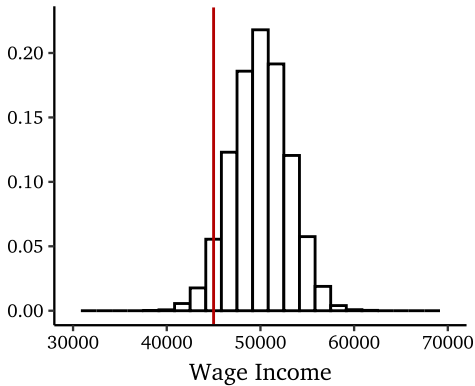
Mean = Median= $50,000
SD= $5,000

Where would you want to live?

- If we don't know where we will end up in the income distribution, some of us might prefer the Mushroom Kingdom since it is unlikely we would earn very little.
- For the same reason, some of us might like Bowser, as it is more likely that one could make a lot.
- But what if Luigi has a job for you as a plumber in both locations, and you will earn $45,000 regardless of where you end up? Are you now indifferent between the two?

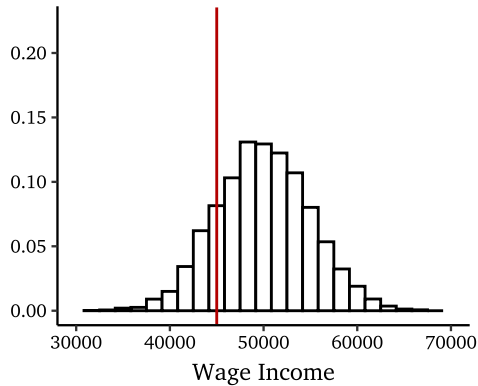
Where would you want to live?

Mushroom Kingdom



Mean = Median= $50,000
SD= $3,000

Bowser's Kingdom



Mean = Median= $50,000
SD= $5,000

Z-Score

We can calculate the Z-Score to capture how many standard deviations (σ) away from the mean (μ) a specific observation is.

$$Z = \frac{X - \mu}{\sigma} \quad \rightarrow \quad X = \mu + Z.\sigma$$

Example: $\sigma_{MK} = 3000$, $\sigma_{BK} = 5000$

$$Z_{MK} = \frac{45000 - 50000}{3000} = -1.66 \qquad Z_{BK} = \frac{45000 - 50000}{5000} = -1$$

Z-Score

- Someone who earns $45,000 in the Mushroom Kingdom is 1.66 *standard deviations* below the mean.
- While someone who earns $45,000 in the Bowser's Kingdom is 1 *standard deviation* below the mean.
- Here, Z-score is informative about how many people are there between someone who earns $45,000 and the average person
- More generally, Z-score tells us the relative position of an observation in the distribution

Things to do next

- Make sure you are staying up to date with the class; notes complement the slides
- Please utilize my office hours
- Coming up: Problem Set 1 (Due next week on Tues, 02/06)

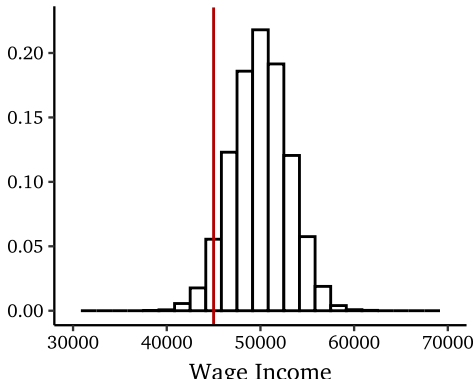
ECON 340
Economic Research Methods

Div Bhagia

Lecture 4
Covariance and Correlation

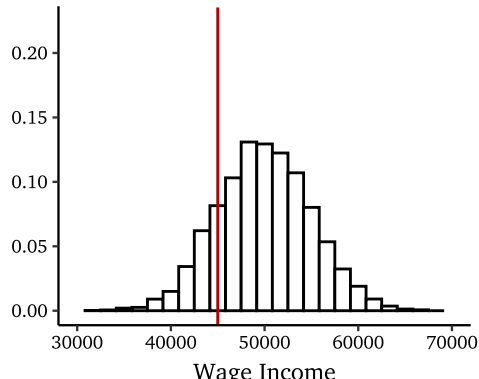
Where would you want to live?

Mushroom Kingdom



Mean = Median= $50,000
SD= $3,000

Bowser's Kingdom



Mean = Median= $50,000
SD= $5,000

Z-Score

We can calculate the Z-Score to capture how many standard deviations (σ) away from the mean (μ) a specific observation is.

$$Z = \frac{X - \mu}{\sigma} \quad \rightarrow \quad X = \mu + Z.\sigma$$

Example: $\sigma_{MK} = 3000$, $\sigma_{BK} = 5000$

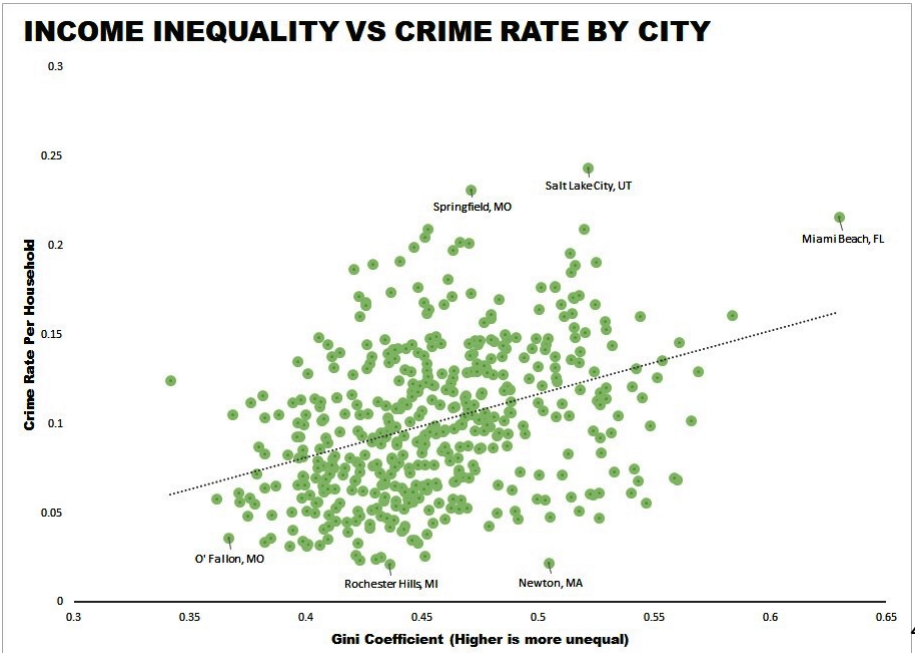
$$Z_{MK} = \frac{45000 - 50000}{3000} = -1.66 \qquad Z_{BK} = \frac{45000 - 50000}{5000} = -1$$

Describing Data

How do we summarize the information contained in a variable?

- Empirical distribution, histogram, percentiles
- Measures of central tendency: mean, median, mode
- Measures of variance: range, variance, standard deviation

How do we summarize the relationship between two variables?


INCOME INEQUALITY VS CRIME RATE BY CITY
Crime Rate Per Household
0.3
0.25
0.2
0.15
0.1
0.05
0
Salt Lake City, UT
Springfield, MO
Miami Beach, FL
O' Fallon, MO
Rochester Hills, MI
Newton, MA
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
Gini Coefficient (Higher is more unequal)

Describing Relationships

- Scatterplot: a graph where each point represents an observation of two variables
- Can see the relationship between two variables
- Positive relationship if when X is high Y is high (and when X is low Y is low)
- Negative relationship if when X is high Y is low (and when X is low Y is high)
- *How to construct a statistic to capture this?*

Covariance

Covariance indicates whether there is a positive or negative relationship between two variables.

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y) \quad (Population)$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \quad (Sample)$$

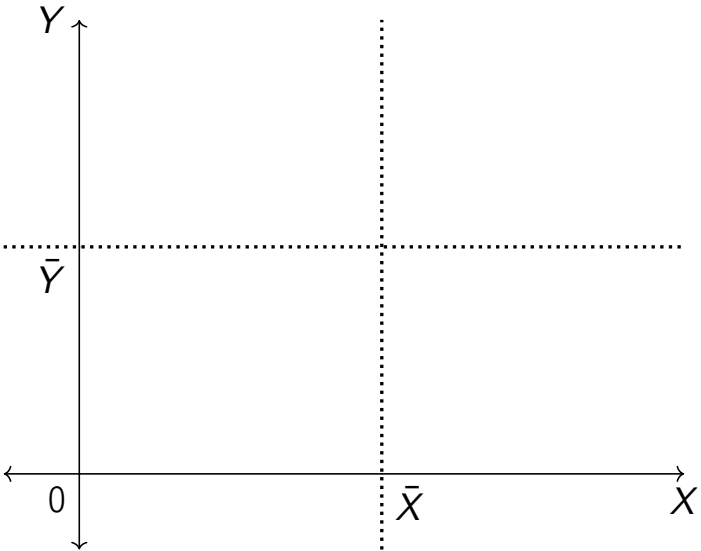
Example

X_i: sleep in hours, Y_i: exercise in hours

Week	X_i	Y_i	$(X_i - \mu_X)(Y_i - \mu_Y)$
1	6	0.5	
2	9	0.3	
3	9	1	
Total			

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y) =$$

Why does the formula work?



Correlation

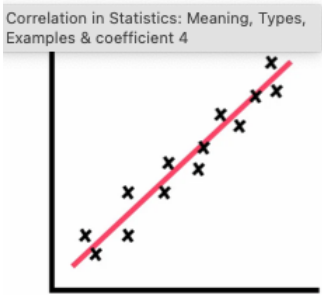
Correlation also indicates the *strength* of the relationship in addition to the *direction*.

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (Population) \qquad r_{XY} = \frac{S_{XY}}{S_X S_Y} \quad (Sample)$$

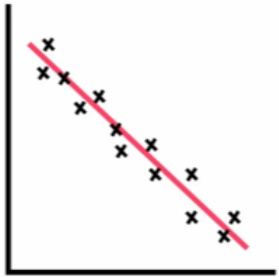
Bounded between -1 and 1.

- $\rho = 0$, no linear relationship
- $\rho = 1$, perfect positive linear relationship
- $\rho = -1$, perfect negative linear relationship

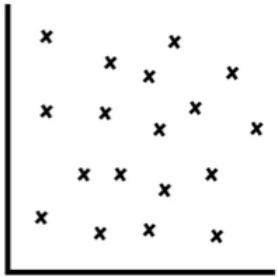
Correlation



Positive Correlation



Negative Correlation



No Correlation

Example

X_i: sleep in hours, Z_i: exercise in minutes

Week	X_i	Z_i	$(X_i - \mu_X)(Z_i - \mu_Z)$
1	6	30	
2	9	18	
3	9	60	
Total			

$$\sigma_{XZ} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Z_i - \mu_Z) =$$

Finally... Correlation is not causation

A positive correlation between inequality and crime doesn't suggest that inequality → crime. This is for two reasons:

- *Reverse causality*: crime → inequality (unlikely here but a concern in many situations)
- *Other confounding factors*: larger, more congested cities tend to be more unequal and also have higher crime rates

Things to do next

- Install R and R Studio on your computers if you haven't already (How to handout on Canvas)
- Please read Chapter 1 from Stock & Watson (uploaded on Canvas). We will discuss it on Tuesday.
- Coming up: Problem Set 1 (Due on Tues, 02/06)

ECON 340
Economic Research Methods

Div Bhagia

Lecture 5
Research Questions and Data

Discussion: Chapter 1, Stock and Watson

1. Does reducing class size improve elementary school education?

- Using data on 420 California school districts in 1999, we find that *students in districts with small class sizes tend to perform better on standardized tests.*
- Can we conclude that small class sizes → better test scores?
- What could be some *confounding* factors?

Discussion: Chapter 1, Stock and Watson

2. Is there racial discrimination in the market for home loans?

- Researchers at the Federal Reserve Bank found that 28% of black applicants are denied mortgages, while only 9% of white applicants are denied.
- Does this indicate there is racial bias in mortgage lending?
- What could be some *confounding* factors?

Discussion: Chapter 1, Stock and Watson

3. Does healthcare spending improve health outcomes?

   - What are the two challenges identified in the reading when trying to answer this question using cross-country data on healthcare expenditures and mortality rates?

Discussion: Chapter 1, Stock and Watson

4. By how much will US GDP grow next year?

   - How is this question different from the first three?

Causal Questions

- Causality: specific action leads to specific, measurable consequences
- The first three questions we discussed are causal
- More examples:
  - Does sleep affect productivity?
  - Will the Fed's interest rate hike lower inflation?
  - Do higher minimum wages decrease employment?

Prediction

- Prediction: using the information on some variables to predict the value of another variable
- You do not need to know a causal relationship to make a good prediction
- A good way to “predict” whether it is raining is to observe whether pedestrians are using umbrellas, but using an umbrella does not cause it to rain.
- Forecasting: predictions about the future

Where are we headed?

- Conceptual framework we will build up to in this class—multiple regression model
- Multiple regression model can be used to answer both types of questions
- The multiple regression model is very useful because it gives us a mathematical way to *quantify how a change in one variable affects another variable, holding other things constant*.
- You will be utilizing this model for your research project

Multiple Regression Model

For example, using the multiple regression model, we can answer questions such as:

> What effect does a change in class size have on test scores, holding constant or *controlling* for student characteristics (such as family income)?
>
> What effect does your race have on your chances of having a mortgage application granted, holding constant other factors such as your ability to repay the loan?

Dependent vs Independent Variable

- The outcome variable is often called the *dependent variable*
- The variable(s) that affect the dependent variable are called *independent variable*(s)
- Other variables that might confound the effect of an independent variable on the dependent variable are called *control variables*

Your Research Project

- Pick a question that can be answered using one of the datasets compiled for this class or an external dataset
- The question should be *well-defined* and *feasible* using the data you picked

Not well-defined: *Are smaller classes better?*

Well-defined: *Does a smaller class size lead to better scores on standardized tests?*

Your Research Project

Say, you picked this question: *Does a smaller class size lead to better scores on standardized tests?*

- Identify your dependent variable
- Identify your independent variable
- Identify some other variables in the dataset that are potential confounders, which will be your control variables.

If all the needed variables are available in your dataset, your question is feasible.

Types of Data

Experimental versus observational (mostly what we will use)

- **Cross-Sectional:** many entities, single time period
- **Time Series:** single entity, multiple time periods
- **Panel/Longitudinal Data:** multiple entities, multiple time periods

Examples?

Establishing Causality

- As we have said, establishing causality is hard.
- Think about the following question:

  Does the use of electronic devices inhibit classroom learning?
- Say, I give you data from all classes held at CSUF in the last five years. The data contains average grades for each class and whether the instructor allowed electronic devices in the classroom.
- Could you answer the above question? If yes, how?

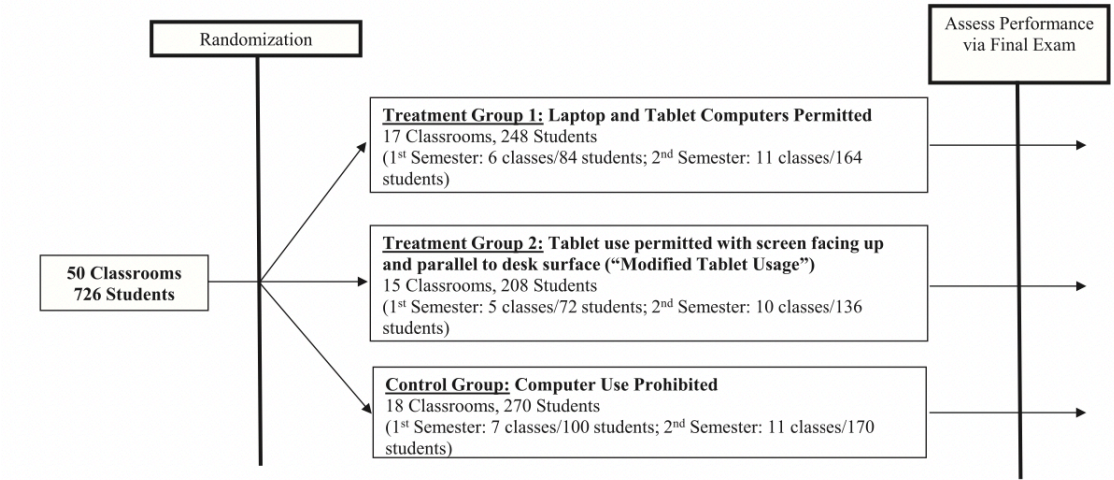
Establishing Causality

- We could compare grades for classes in which devices are permitted vs. others
- But what if better instructors don't care about students using devices?
- This would lead us to find little impact of device use, as classes where devices are allowed also happen to be taught by better instructors
- Here, the instructor quality is an example of a *confounding* factor

Establishing Causality

- How to establish causality?
  - Gold standard: experiments/randomized controlled trials (RCTs)
  - Assign randomly to *treatment* or *control* group
- Carter, Greenberg, and Walker (2017): randomly allowed students to access their laptop and tablet computers during an introductory economics course at the United States Military Academy at West Point

Carter, Greenberg, and Walker (2017)



Why does randomization help?

- Which classes allow electronic devices or not is random
- There should be no differences in instructor effectiveness because the classes were chosen randomly
- RCTs are widely employed in clinical research, e.g., experimental drug trials, other treatments
- Increasingly used in economics and other social sciences, e.g., PROGRESSA, free deworming program in Kenya (Miguel and Kremer, 2004), microcredits (Banerjee et al., 2013)

Aside: Experiments in Development Economics

- Development economists were early adopters of experiments in economics
- The 2019 Nobel prize in economics was awarded to Abhijit Banerjee, Esther Duflo, and Michael Kremer “for their experimental approach to alleviating global poverty.”
- (Duflo is the youngest person and the second woman to win the award.)
- If you are intrigued, check out Poor Economics by Banerjee and Duflo

But, we can't all run experiments...

- Hard to conduct experiments for a lot of Economics questions
- Economists have developed a whole toolkit of *quasi-experimental* methods. This is the heart of Econometrics.
- Starting point: multiple regression model that allows us to *control* for variables
- Focus on being able to come as close as possible to an idealized experiment

Another aside

- The 2021 Nobel prize in economics was awarded to David Card, Joshua D. Angrist, and Guido W. Imbens for their contributions to answering causal questions using *natural experiments*
- In the last two decades, *quasi-experimental* methods have been used to quantify the labor market effects of minimum wages, immigration, and education, amongst other questions.
- We will talk more about this towards the end of the semester.

Things To Do Next

- For the rest of this week and next week, we will learn how to prepare and analyze data in R
- Install R and R Studio on your computer if you haven't (how to handout on Canvas)
- Download the dataset "caschool.csv" from our Dropbox data folder and save it in a new folder called "Econ340_R" on your laptop (don't forget the location of the folder).
- Bring your laptop to the next class. Alternatively, you can use the computers in the lab. In the latter case, try to be here 5 minutes before class to set up.

ECON 340
Economic Research Methods

Div Bhagia

Lecture 9: Distribution, Expectation, and Variance

Taking stock

We have learned how to describe variables in the data using:

- Empirical distribution
- Mean and median
- Variance and standard deviation
- Correlation and covariance

Random Sampling

- Often, data is available only for a sample of the population
- Ideally, we want a sample *representative* of the population we are interested in and not a *biased* sample
- We can achieve this by taking a *random sample* from the population
- Random sample: each unit from the population has an equal probability of being chosen

Simple Example

Say, we want to take a random sample of 2 from a population of 5.

Population:
$X_1 = \$60,000$
$X_2 = \$40,000$
$X_3 = \$40,000$
$X_4 = \$50,000$
$X_5 = \$60,000$

Population mean:
$\mu = \$50,000$

Pick a sample randomly: Spin a Wheel

Attempt 1
$\bar{X}_1 =$

Attempt 2
$\bar{X}_2 =$

Attempt 3
$\bar{X}_3 =$

Random Sampling

- We can get different values of the sample mean depending on the sample we pick
- Sample mean is a *random variable*!
- Then what can we infer about the population mean from the sample mean?
- But before that, what is a random variable?

Random Variables

- *Random variable* is a numerical summary of a random outcome.
- Examples: outcome from a coin toss or a die roll, or number of times your wireless network fails before a deadline, etc.
- Random variables can be *discrete* or *continuous*
- Discrete random variables take a discrete set of values, like $0, 1, 2, \ldots$
- Continuous random variables take on a continuum of possible values

Discrete Random Variables

- *Probability distribution* of a discrete random variable: all possible values of the variable and their probabilities.

$$f(x) = Pr(X = x)$$

where $0 \leqslant f(x) \leqslant 1$ for all x and $\sum_x f(x) = 1$.

- *Cumulative probability distribution* gives the probability that the random variable is less than or equal to a particular value.

$$F(x) = Pr(X \leqslant x) = \sum_{x' \leqslant x} f(x')$$

Discrete Random Variable: Example

X: outcome from rolling a die

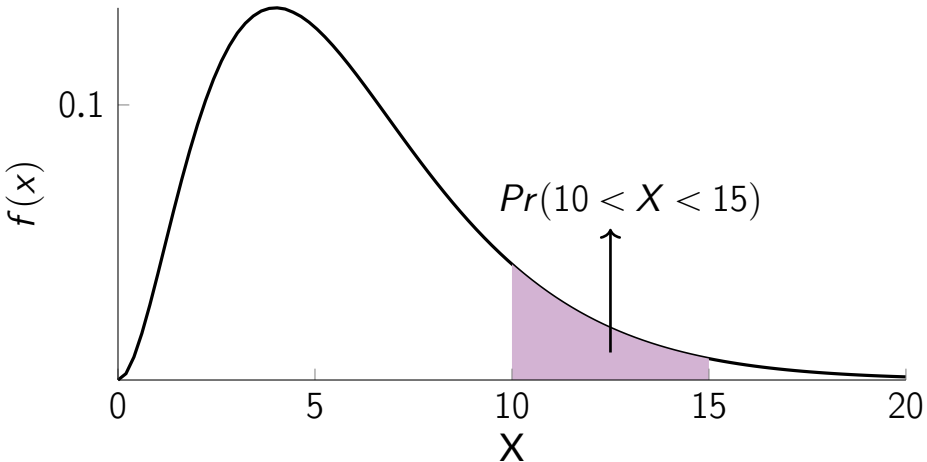
X	$f(X)$	$F(X)$
1	$1/6$	$1/6$
2	$1/6$	$2/6$
3	$1/6$	$3/6$
4	$1/6$	$4/6$
5	$1/6$	$5/6$
6	$1/6$	1

Also referred to as probability distribution function (PDF) and cumulative distribution function (CDF).

Continuous Random Variable

- For continuous random variables, due to a continuum of possible values, it is not feasible to list the probability of each possible value.
- So instead, the area under the *probability density function* $f(x)$ between any two points gives the probability that the random variable falls between those two points.
- *Cumulative probability distribution* for continuous RVs is defined as before $F(x) = Pr(X \leqslant x)$.

Probability Density Function



How to calculate the area under the curve?

- Area under the curve is calculated using an *integral* (just like a sum but for continuous variables)
- However, don't sweat, in most cases a statistical program or old-school tables (in the back of textbooks) can help us find these areas for commonly used distributions
- We will now define expectation and variance for the discrete case, but it is generalizable to continuous RVs

Expectation

- *Expectation*: average value of the random variable over many repeated trials or occurrences
- Computed as a weighted average of the possible outcomes, where the weights are the probabilities
- The expectation of X is also called the *expected value* or the *mean* and is denoted by μ_X or $E(X)$

$$\mu_X = E(X) = \sum_x f(x)x$$

Example

An outdoor market vendor sells handmade crafts.

Weather	**Probability**	**Sales ($)**
No rain	0.6	300
Light rain	0.3	150
Heavy rain	0.1	50

$$E(Sales) = (0.6 \times 300) + (0.3 \times 150) + (0.1 \times 50) = \$230$$

Example

An outdoor market vendor sells handmade crafts.

Weather	Probability	Sales ($)
No rain	0.6	300
Light rain	0.3	150
Heavy rain	0.1	50

$$E(Sales) = (0.6 \times 300) + (0.3 \times 150) + (0.1 \times 50) = \$230$$

Daily sales will fluctuate due to randomness of the weather. However, if this day were to repeat itself multiple times, the vendor would, on average, expect to achieve daily sales of $230.

Variance and Standard Deviation

The variance and standard deviation measure the dispersion or the “spread”.

$$\sigma_X^2 = Var(X) = E[(X - \mu_X)^2] = \sum_x (x - \mu_X)^2 f(x)$$

Variance is the expected value of the squared deviation of X from its mean.

In our example, the variance will capture the potential variability or fluctuation in sales on any given day compared to the average (expected value).

Example

An outdoor market vendor sells handmade crafts.

Weather	Probability	Sales ($)
No rain	0.6	300
Light rain	0.3	150
Heavy rain	0.1	50

$$E(Sales) = 0.6 \cdot (300) + 0.3 \cdot (150) + 0.1 \cdot (50) = \$230$$

$$\begin{aligned} Var(Sales) &= 0.6 \cdot (300 - 230)^2 + 0.3 \cdot (150 - 230)^2 + 0.1 \cdot (50 - 230)^2 \\ &= 8100 \end{aligned}$$

Variance and Standard Deviation

Alternative formula for the variance:

$$Var(X) = E[(X - \mu_X)^2] = E(X^2) - \mu_X^2$$

Since variance is in units of the square of X, therefore we use *standard deviation* which is the square-root of variance

$$\sigma_X = \sqrt{\sigma_X^2}$$

For our example, standard deviation of sales is $\sqrt{8100} = \$90$.

Transformations of Random Variables

If you shift every outcome by some constant a,

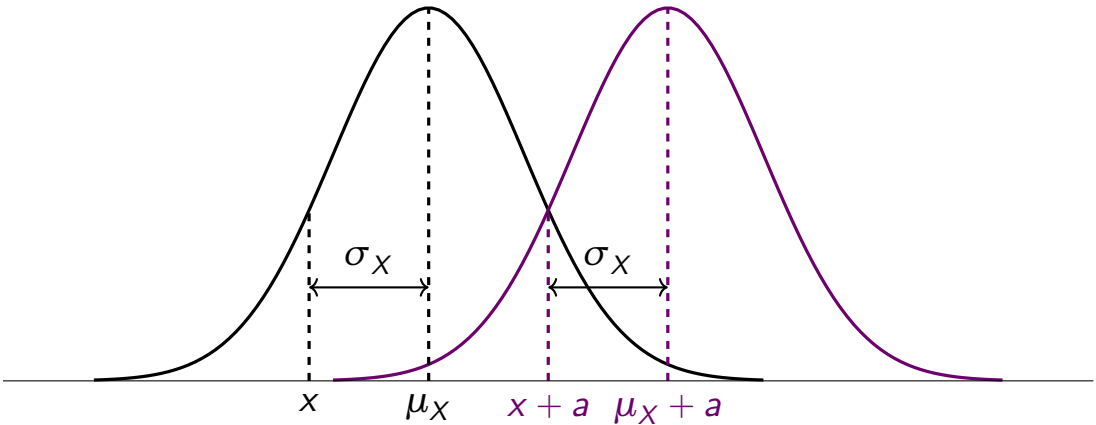
- Mean also shifts by a:

$$E(X + a) = E(X) + a$$

- Variance is unchanged:

$$Var(X + a) = Var(X)$$

Shifting the Distribution



Transformations of Random Variables

If you scale every outcome by some constant b,

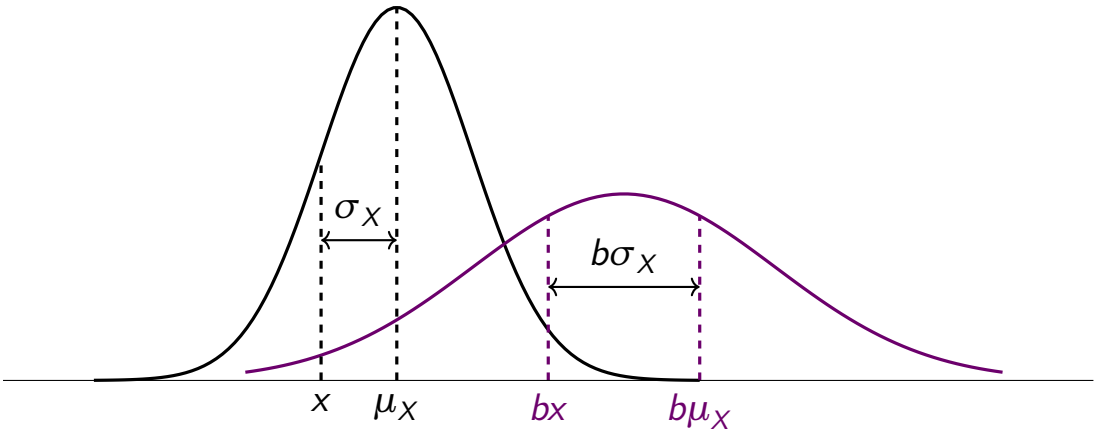
- Mean is also scaled by b:

$$E(bX) = bE(X)$$

- Variance is scaled by b^2:

$$Var(bX) = b^2 Var(X)$$

Scaling the Distribution



Transformations of Random Variables

More generally, if X is a random variable and

$$Y = a + bX$$

Then Y is also a random variable with

$$E(Y) = a + bE(X) \qquad Var(Y) = b^2 Var(X)$$

In addition, a linear transformation of a random variable does not change the shape of the distribution.

Example

The market vendor offered you a job, you get \$20 and 10% commission (Y) on sales (X) per day.

$$Y = 20 + 0.1X$$

Weather	Probability	X	Y
No rain	0.6	300	20 + 0.1(300) = 50
Light rain	0.3	150	20 + 0.1(150) = 35
Heavy rain	0.1	50	20 + 0.1(50) = 25

$$E(Y) = 20 + 0.1E(X) = 20 + 0.1(230) = 43$$
$$Var(Y) = 0.1^2 \cdot Var(X) = 0.01(8100) = 81$$
$$SD(Y) = \sqrt{81} = 9 \ (= 0.1SD(X))$$

Standardized Random Variables

A random variable can be transformed into a random variable with mean 0 and variance 1 by subtracting its mean and then dividing by its standard deviation, a process called standardization.

$$Z = \frac{X - \mu_X}{\sigma_X}$$

Here, $E(Z) = 0$ and $\sigma_Z = 1$.

ECON 340
Economic Research Methods

Div Bhagia

Lecture 10: Normal Distribution and Z-Score

Random Variables

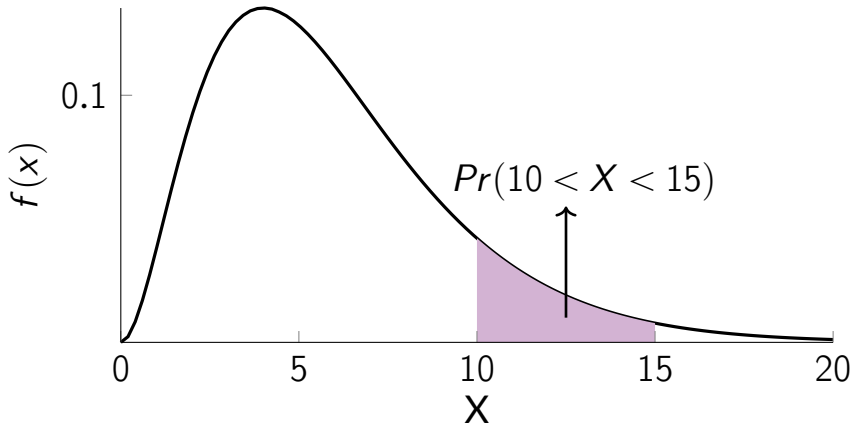
- *Random variables* take different values under different scenarios.
- Examples: outcome from a coin toss or a die roll, or number of times your wireless network fails before a deadline, etc.
- The likelihood of these scenarios is summarized by the probability distribution.
- Random variables can be *discrete* or *continuous*

Distribution of a Random Variable

- For a discrete random variable, probability distribution given by the probability of each outcome.
- Continuous random variables summarized by the *probability density function*, where area under the curve gives us the probability of an outcome being in an interval.

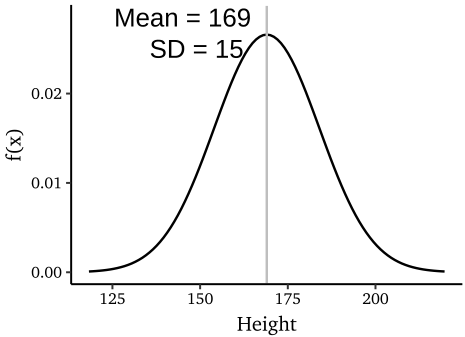
Probability Density Function

The area under the curve tells us the probability of an outcome being in a particular interval.



Normal Distribution

One distribution appears more than others — Normal Distribution



$height \sim N(169, 225)$

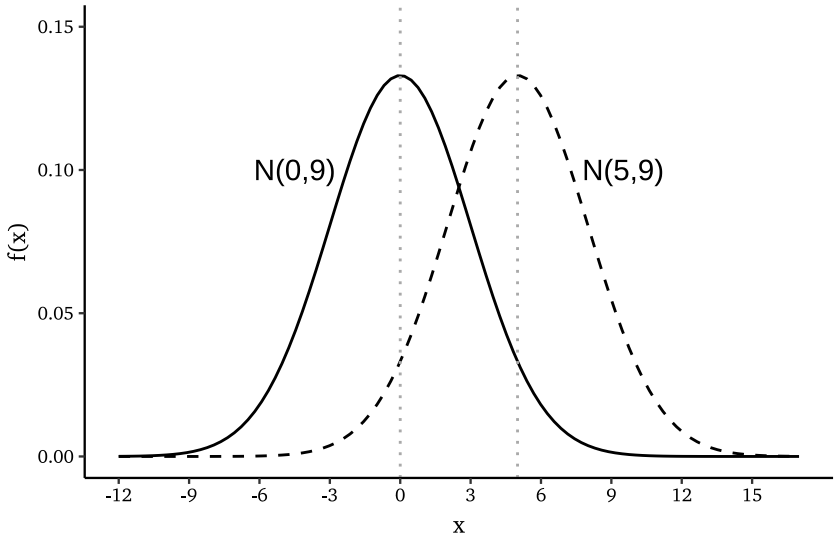
What's special about it?

- Symmetric (no skew, mean=median, bell-shaped)
- Height, birthweight, SAT scores, etc., normally distributed
- Sampling distribution approximately normal

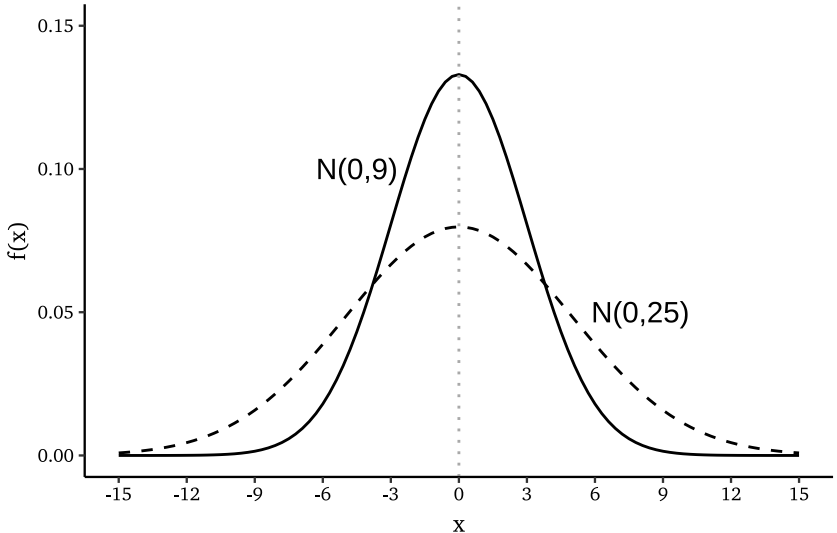
Normal Distribution

- Normal distribution with mean μ and variance σ^2 is expressed as $N(\mu, \sigma^2)$
- So if I write $X \sim N(12, 4)$, it means X is normally distributed with mean 12 and variance 4
- The standard normal distribution is the normal distribution with mean 0 and variance 1, denoted by $N(0, 1)$
- Random variables that have a $N(0, 1)$ distribution are often denoted by Z

Normal Distribution



Normal Distribution



How to find the area under the curve?

- Often interested in finding the probability that a random variable lies in a particular interval
- Cumbersome to take the integral each time
- Since the normal distribution is so commonly used, one can find these probabilities easily for the *standard normal variable*:

$$Z \sim N(0, 1)$$

- We can use the standard normal probabilities to get the probabilities for any normally distributed variable

Standardized Random Variables

A random variable can be transformed into a random variable with mean 0 and variance 1 by subtracting its mean and then dividing by its standard deviation, a process called standardization.

$$Z = \frac{X - \mu_X}{\sigma_X}$$

Transformations of Random Variables

- Say, X is a random variable with mean μ_X, variance σ_X^2, and standard deviation σ_X
- If we transform X to create a new random variable

$$Y = a + bX$$

- Then Y is a random variable that has a distribution with the same shape as X and with
  - Mean: $\mu_Y = a + b \cdot \mu_X$
  - Variance: $\sigma_Y^2 = b^2 \cdot \sigma_X^2$
  - SD: $\sigma_Y = b \cdot \sigma_X$

Z-Score

- We can rearrange the terms in the Z-score

$$Z = \frac{X - \mu_X}{\sigma_X} \quad \rightarrow \quad Z = \frac{-\mu_X}{\sigma_X} + \frac{1}{\sigma_X} \cdot X$$

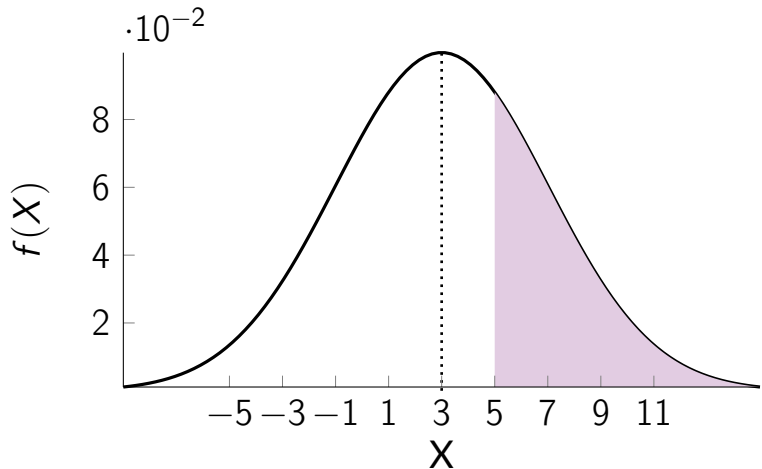
- So here $Z = a + b \cdot X$ with $a = \frac{-\mu_X}{\sigma_X}$ and $b = \frac{1}{\sigma_X}$
- Since Z is a transformed random variable:

$$\mu_Z = a + b \cdot \mu_X = \frac{-\mu_X}{\sigma_X} + \frac{1}{\sigma_X} \cdot \mu_X = 0$$

$$\sigma_Z^2 = b^2 \cdot \sigma_X^2 = \left(\frac{1}{\sigma_X}\right)^2 \cdot \sigma_X^2 = 1$$

How to find the area under the curve?

For example, say $X \sim N(3, 16)$. We want to calculate $Pr(X \geqslant 5)$.



How to find the area under the curve?

Given $X \sim N(3, 16)$,

$$Z = \frac{X - \mu_X}{\sigma_X} = \frac{X - 3}{4} \sim N(0, 1)$$

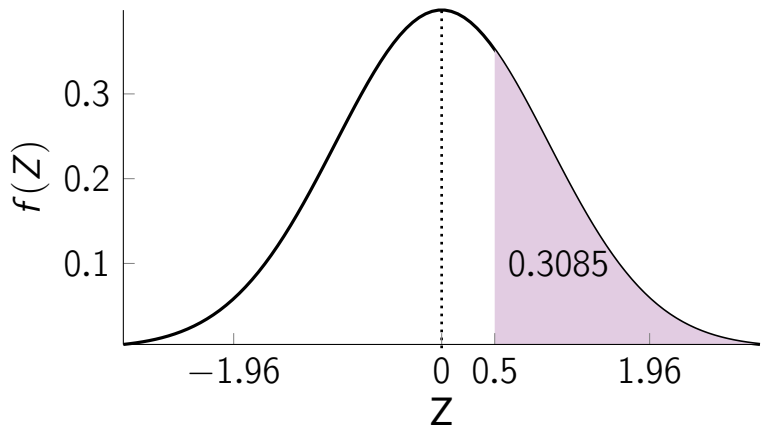
Note that,

$$Pr(X \geqslant 5) = Pr\left(\frac{X - 3}{4} \geqslant \frac{5 - 3}{4}\right) = Pr(Z \geqslant 0.5)$$

We can now refer to the standard normal table and find that $Pr(Z \geqslant 0.5)$

How to find the area under the curve?

Find $Pr(Z \geqslant 0.5)$ from the standard normal table.



Recipe

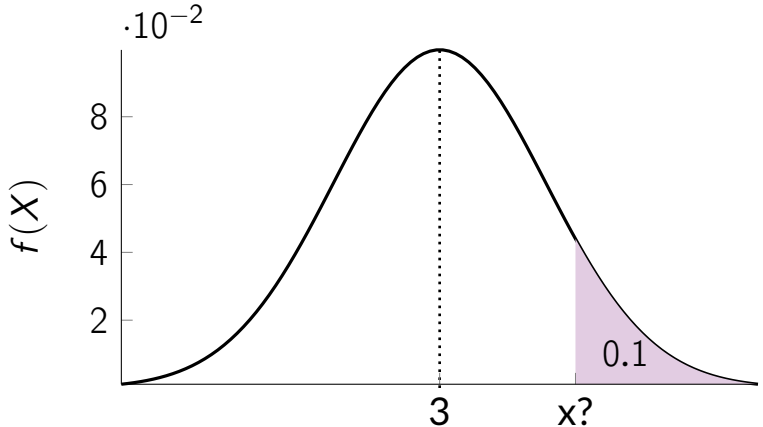
Given $X \sim N(\mu, \sigma^2)$, general recipe to find $Pr(x_0 < X < x_1)$:

- Find $z_0 = (x_0 - \mu)/\sigma$ and $z_1 = (x_1 - \mu)/\sigma$
- Use standard normal table to find $Pr(z_0 < Z < z_1)$

Example. Given $X \sim N(3, 16)$, find $Pr(2 < X < 5)$.

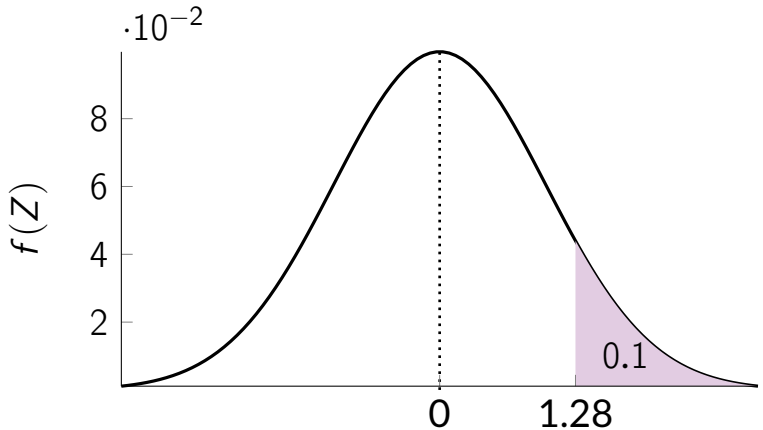
Finding points from area under the curve

For example, say $X \sim N(3, 16)$ and we are given $Pr(X > x) = 0.10$. How to find x?



Finding points from area under the curve

Start by finding z, such that $Pr(Z > z) = 0.1$.



Finding points from area under the curve

Now note that,

$$Z = \frac{X - \mu}{\sigma} \rightarrow X = \mu + Z.\sigma$$

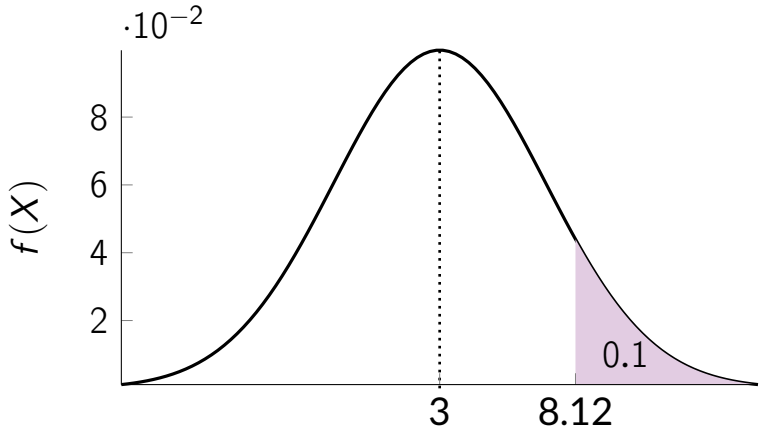
We found $Pr(Z > 1.28) = 0.1$, we can find the corresponding x for 1.28 as follows:

$$3 + 1.28 \times 4 = 8.12$$

So we have that $Pr(X > 8.12) = 0.1$.

Finding points from area under the curve

Transforming z back to $x = 3 + 1.28 \times 4 = 8.12$.



Finding points from area under the curve

- Sometimes we are given $Pr(X < x)$ or $Pr(X > x)$ and we need to find x.
- *Example*: Given $X \sim N(3, 16)$ and $Pr(X > x) = 0.10$, find x.
- Start by finding z, such that $Pr(Z > z) = 0.1$. From the standard normal table, $z = 1.28$.
- Now we just need to convert z to x.
- Since $Z = \frac{X-\mu}{\sigma} \rightarrow X = \mu + Z.\sigma$, so $x = 3 + 1.28 \times 4 = 8.12$

ECON 340
Economic Research Methods

Div Bhagia

Lecture 11: Independence & Correlation

Random Variables

- *Random variables* take different values under different scenarios.
- Examples: outcome from a coin toss or a die roll, or number of times your wireless network fails before a deadline, etc.
- The likelihood of these scenarios is summarized by the probability distribution.
- Random variables can be *discrete* or *continuous*

Two Random Variables

The *joint probability distribution* of two discrete random variables is the probability that the random variables simultaneously take on certain values.

$$f(x, y) = Pr(X = x, Y = y)$$

	Rain $(X = 1)$	No Rain $(X = 0)$	Total
60-min commute $(Y = 60)$	0.3	0.2	
30-min commute $(Y = 30)$	0.1	0.4	
Total			

Marginal Distribution

The *marginal probability distribution* of a random variable Y is just another name for its probability distribution.

$$f(y) = Pr(Y = y) = \sum_{x} Pr(X = x, Y = y)$$

Conditional Distribution

The distribution of a random variable Y conditional on another random variable X taking on a specific value is called the conditional distribution of Y given X.

$$f(y|x) = Pr(Y = y|X = x) = \frac{Pr(X = x, Y = y)}{Pr(X = x)} = \frac{f(x, y)}{f(x)}$$

Commute Times

	Rain $(X = 1)$	No Rain $(X = 0)$	Total
60-min commute $(Y = 60)$	0.3	0.2	
30-min commute $(Y = 30)$	0.1	0.4	
Total			

Conditional Expectation

The *conditional expectation* of Y given X is the mean of the conditional distribution of Y given X.

$$E(Y|X=x) = \sum_y yPr(Y=y|X=x) = \sum_y y \cdot f(y|x)$$

Calculate $E(Y|X=1)$ and $E(Y|X=0)$ in the last example. Comparing these tells us how X affects Y.

Can define conditional variance similarly.

Independence

Two random variables X and Y are independently distributed, or independent, if knowing the value of one of the variables provides no information about the other.

$$Pr(Y = y|X = x) = Pr(Y = y)$$

Example: Two consecutive coin tosses.

Note: We can equivalently say that X and Y are independent if $E(Y|X) = E(Y)$.

Covariance and Correlation

Covariance is a measure of the extent to which two random variables move together.

Let X and Y be a pair of random variables, then the *covariance* of X and Y is given by:

$$\sigma_{XY} = Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y$$

The *correlation* between X and Y is given by:

$$\rho_{XY} = corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \quad \text{where } -1 \leqslant \rho \leqslant 1$$

Uncorrelated vs Independence

If X and Y are independent, then they are also uncorrelated.

$$E(Y|X) = E(Y) \rightarrow \rho_{XY} = 0$$

However, it is not necessarily true that if X and Y are uncorrelated, then they are also independent.

Sums of Random Variables

X and Y is a pair of random variables, then

$$E(aX + bY) = aE(X) + bE(Y)$$

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$$

If X and Y are independent:

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$$

Portfolio Diversification

You are now contemplating between two stocks with the same average return and spread.

$$\mu_X = \mu_Y \qquad \sigma_X^2 = \sigma_Y^2$$

Should you pick any one stock at random or invest equally in both?

What's next?

- Problem Set 3 is now posted on Canvas (due next week). You can attempt Questions 1 and 2.
- Thursday: Start with Sampling and Estimation.

ECON 340
Economic Research Methods

Div Bhagia

Lecture 12: Good Estimators, Sample Mean Distribution, Confidence Intervals

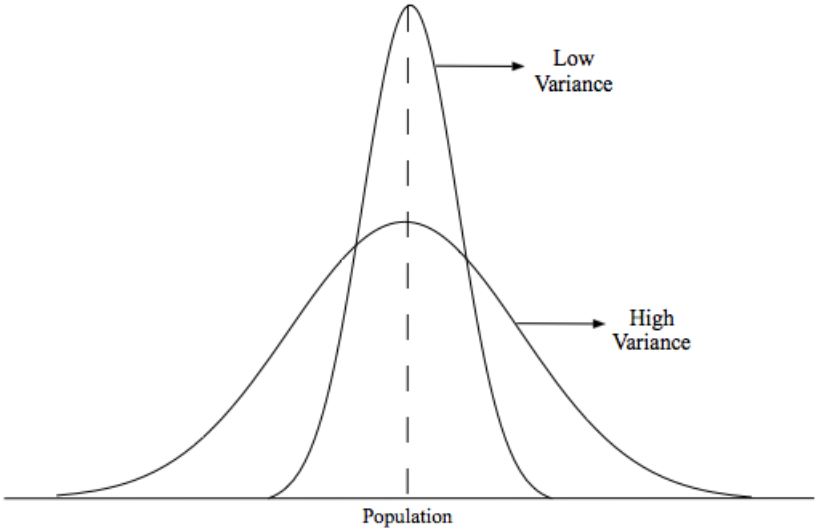
Sampling and Estimation

- We want to learn something about the population
- But often, we can collect data only for a sample of the population
- Good news: if the sample is drawn *randomly* we can use statistical methods to reach *tentative* answers
- Use sample quantities to *estimate* population parameters
- Sample *estimators* are random variables

Estimators

- Denote the population parameter of interest by θ
- And let's denote its sample estimator by $\hat{\theta}$
- Three desirable properties for an estimator:
  - *Unbiasedness*: $E(\hat{\theta}) = \theta$
  - *Efficiency*: lower variance is better
  - *Consistency*: as the sample size becomes infinitely large, $\hat{\theta} \rightarrow \theta$

What is a good estimator?



Expectation and Variance of $\bar{X}$

Let $X_1, X_2, \ldots, X_n$ denote independent random draws (random sample) from a population with mean μ and variance σ^2.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Then $\bar{X}$ is also a random variable with:

$$E(\bar{X}) = \mu \qquad Var(\bar{X}) = \frac{\sigma^2}{n}$$

So $\bar{X}$ is an unbiased and consistent estimator for μ.

Sample Mean Distribution

The distribution of the sample mean is normal if *either* of the following is true:

- The underlying population is normal
- The sample size is large, say $n \geqslant 100$

The first one follows from the sample mean being a linear combination of normally distributed variables.

The latter is implied by the *Central Limit Theorem*.

Central Limit Theorem

If $X_1, X_2, .., X_n$ are drawn randomly from a population with mean μ and variance σ^2, sample mean $\bar{X}$ is normally distributed with mean μ and variance σ^2/n as long as n is large.

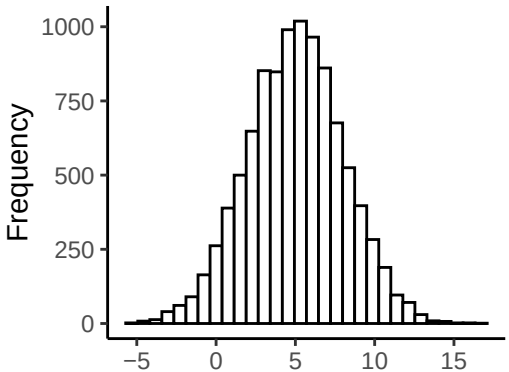
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Simulation

Normal Population

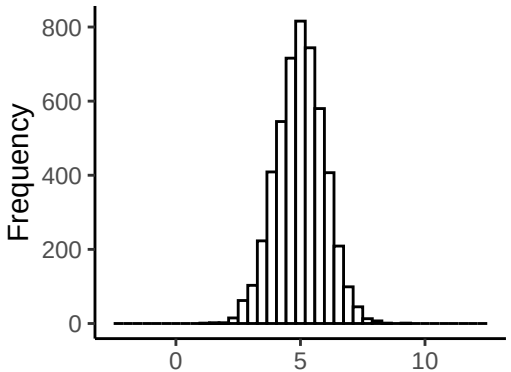
Normal Population

$\mu = 5, \sigma^2 = 9$



Sample Mean Distribution

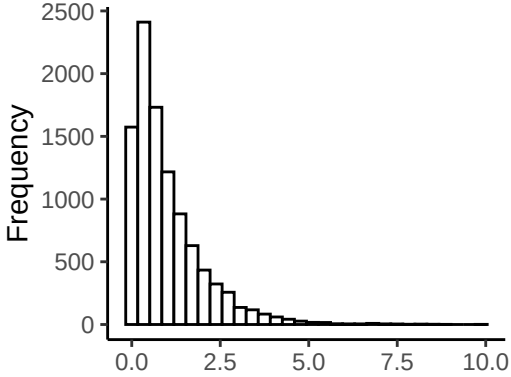
$n = 10, E(\bar{X}) = 5, Var(\bar{X}) = 0.9$



Non-Normal Population

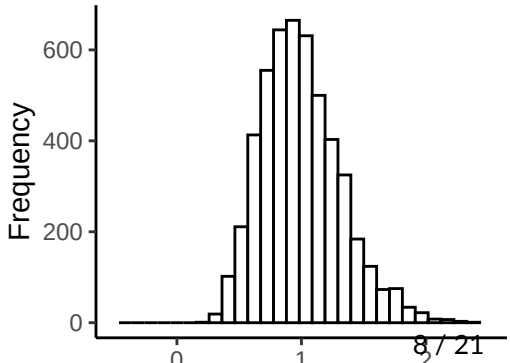
Non-Normal Population

$\mu = 1, \sigma^2 = 1$



Sample Mean Distribution

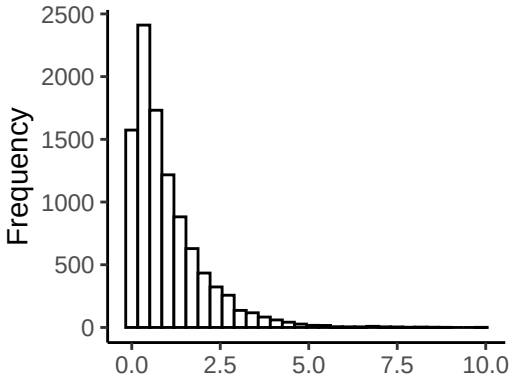
$n = 10, E(\bar{X}) = 0.92, Var(\bar{X}) = 0.1$



Central Limit Theorem

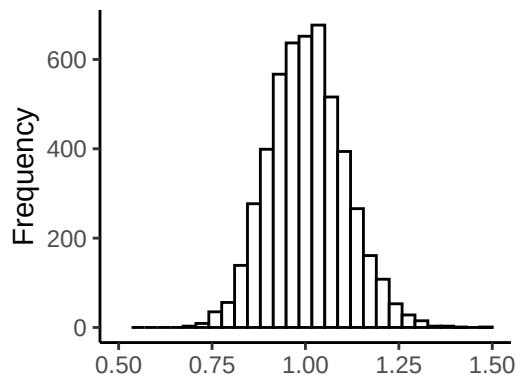
Non-Normal Population

$\mu = 1, \sigma^2 = 1$

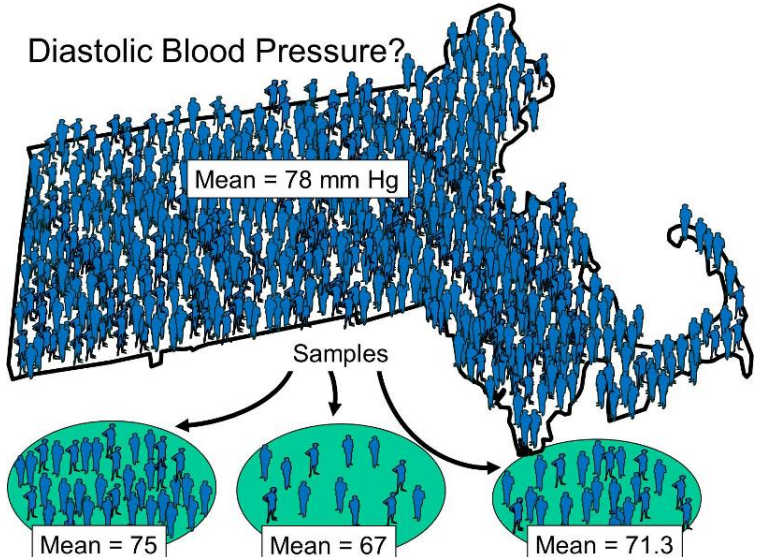


Sample Mean Distribution

$n = 100, \bar{X} = 1, Var(\bar{X}) = 0.01$



Example: Blood Pressure in Massachusetts



Confidence Intervals

- Let's say we picked a random sample of 100 people from Massachusetts and took their blood pressure and found $\bar{x} = 75$.
- Given this estimate of 75, what can we say about the true mean?
- Here $n = 100$ so by CLT, $\bar{X} \sim N(\mu, \sigma^2/n)$. For now assume we know $\sigma^2 = 552.25$.
- Then we should be able to say with some certainty that the true mean lies *somewhere* around 75.

Confidence Intervals

- Create an interval around the sample mean that gives us a range of plausible values for the population mean.
- We can have confidence intervals of varying levels of confidence, most common are 90%, 95%, or 99%.
- The level of confidence is the probability that a calculated confidence interval contains the true population parameter.

How to construct a confidence interval?

Say we want to construct a 90% confidence interval for the true mean.

So far we have established that $\bar{X} \sim N(\mu, \sigma^2/n)$.

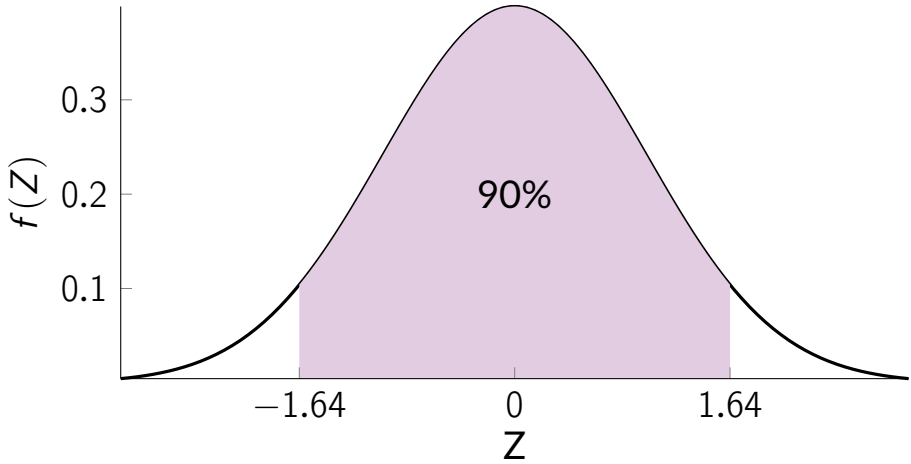
Note that then,

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

From the Standard Normal table, we can find that

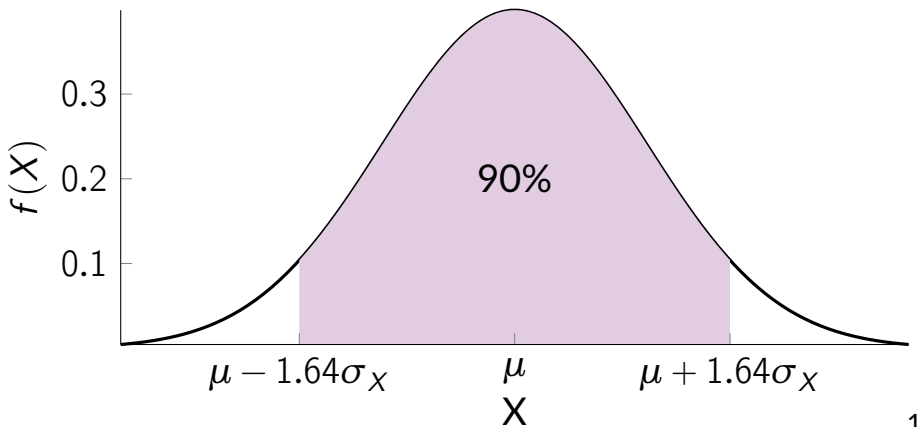
$$P(-1.64 < Z < 1.64) = 0.90$$

Standard Normal Distribution



Normal Distribution

90% of the area under the curve lies within 1.64 standard deviations of the mean.



90% Confidence Intervals

$$Pr(-1.64 < Z < 1.64) = 0.90$$

$$Pr\left(-1.64 < \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} < 1.64\right) = 0.90$$

$$Pr\left(\mu - 1.64\sigma_{\bar{X}} < \bar{X} < \mu + 1.64\sigma_{\bar{X}}\right) = 0.90$$

$$Pr\left(\bar{X} - 1.64\sigma_{\bar{X}} < \mu < \bar{X} + 1.64\sigma_{\bar{X}}\right) = 0.90$$

90% Confidence Intervals

$$Pr\left(\bar{X} - 1.64\sigma_{\bar{X}} < \mu < \bar{X} + 1.64\sigma_{\bar{X}}\right) = 0.90$$

Note that $\sigma_{\bar{X}} = \sigma/\sqrt{n}$, so the 90% confidence interval here is given by:

$$\bar{x} \pm 1.64 \cdot \frac{\sigma}{\sqrt{n}}$$

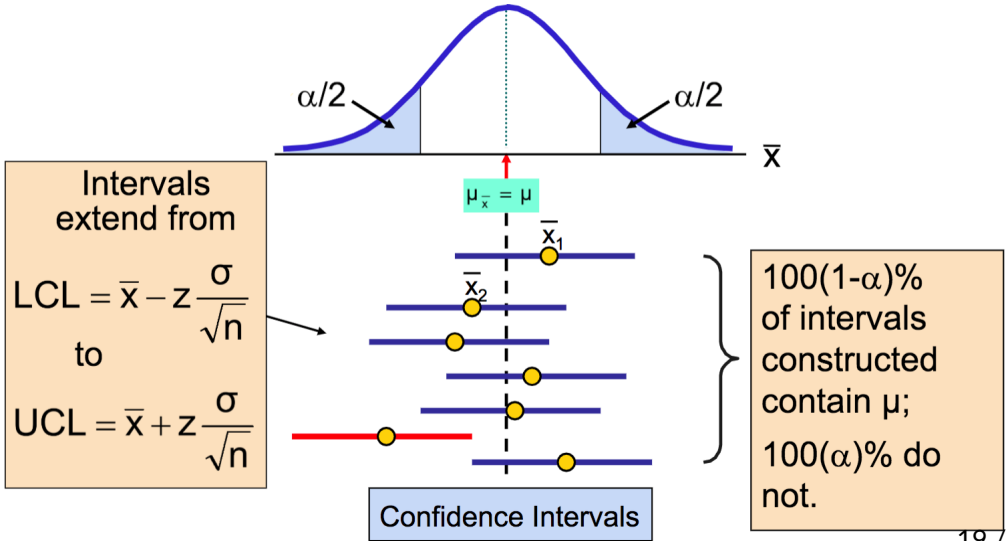
Plugging in $\sigma = \sqrt{552.25}$ and $n = 100$. We get $[71.15, 78.85]$.

Confidence Intervals: Interpretation

There is a 90% chance that the true population average for blood pressure lies in this interval.

What this really means is that if we took 100 random samples from the population and calculated 90% confidence intervals for each sample, we would expect 90 out of 100 intervals to contain the true population mean.

Confidence Intervals: Interpretation



Confidence Intervals: Recipe

Let $z_{\alpha/2}$ be the z-value that leaves area $\alpha/2$ in the upper tail of the normal distribution.

Then $1-\alpha$ confidence interval is given by

$$\bar{x} \pm \underbrace{z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{\text{Margin of Error}}$$

Next up

- Problem Set 3 is due next Tuesday
- Next week: Continue with sampling and estimation
- Week after: Review class and midterm

ECON 340
Economic Research Methods

Div Bhagia

Lecture 13: Confidence Intervals

Expectation and Variance of $\bar{X}$

Let $X_1, X_2, \ldots, X_n$ denote independent random draws (random sample) from a population with mean μ and variance σ^2.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Then $\bar{X}$ is also a random variable with:

$$E(\bar{X}) = \mu \qquad \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

So $\bar{X}$ is an unbiased and consistent estimator for μ.

Sample Mean Distribution

The distribution of the sample mean is normal if *either* of the following is true:

- The underlying population is normal
- The sample size is large, say $n \geqslant 100$

The first one follows from the sample mean being a linear combination of normally distributed variables.

The latter is implied by the *Central Limit Theorem*.

Central Limit Theorem

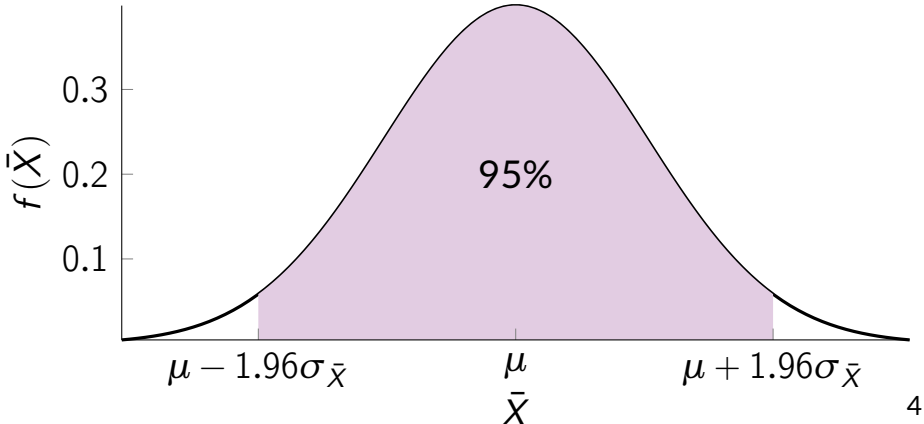
If $X_1, X_2, .., X_n$ are drawn randomly from a population with mean μ and variance σ^2, sample mean $\bar{X}$ is normally distributed with mean μ and variance σ^2/n as long as n is large.

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

Simulation

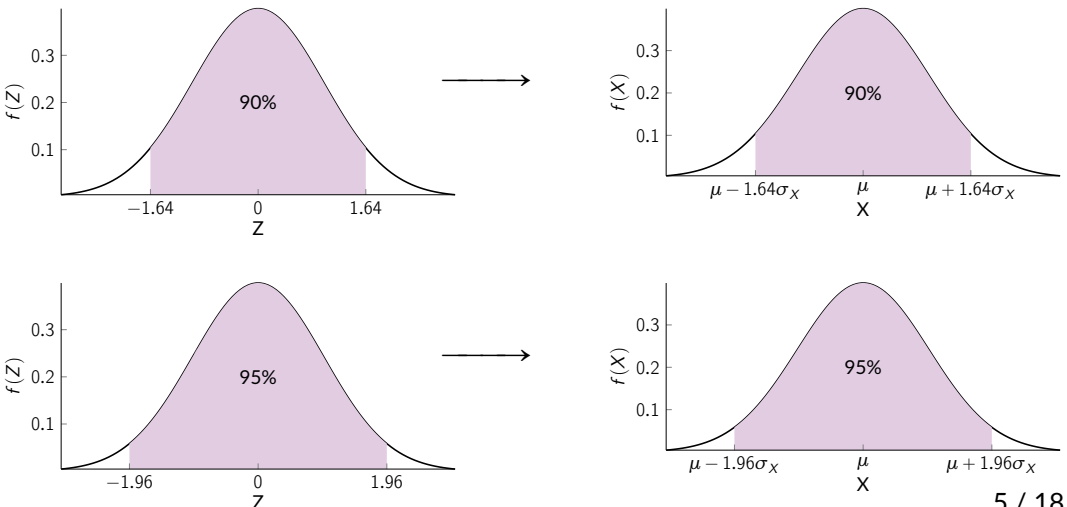
Normal Distribution

95% of the area under the curve lies within 1.96 standard deviations of the true mean.



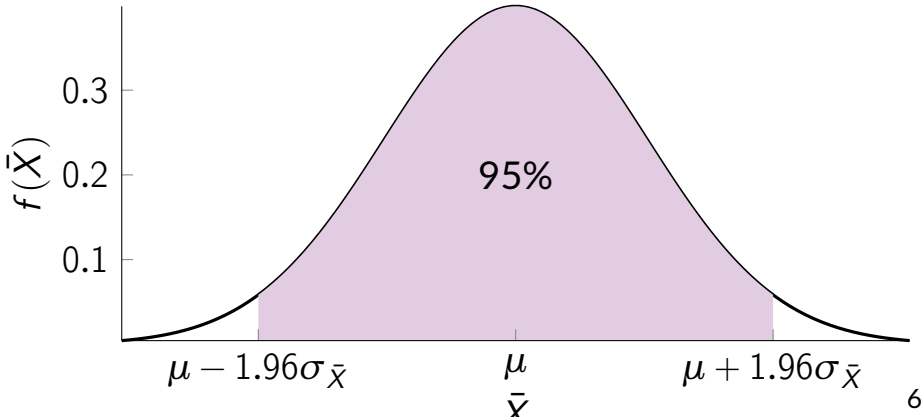
Aside

Standard Normal → Normal Distribution



Normal Distribution

95% of the time that we take a sample and calculate the sample mean, it will be within 1.96 standard deviations of μ.



Confidence Intervals

- If 95% of the time, the sample mean $\bar{X}$ will be within 1.96 standard deviations of the true mean μ.
- Then 95% of the time, the true mean μ will be within 1.96 standard deviations of the sample mean $\bar{X}$.
- Use this logic to create a 95% confidence interval for the true population mean μ.
- Say in your sample you found sample mean $\bar{x}$, then 95% confidence interval for μ:

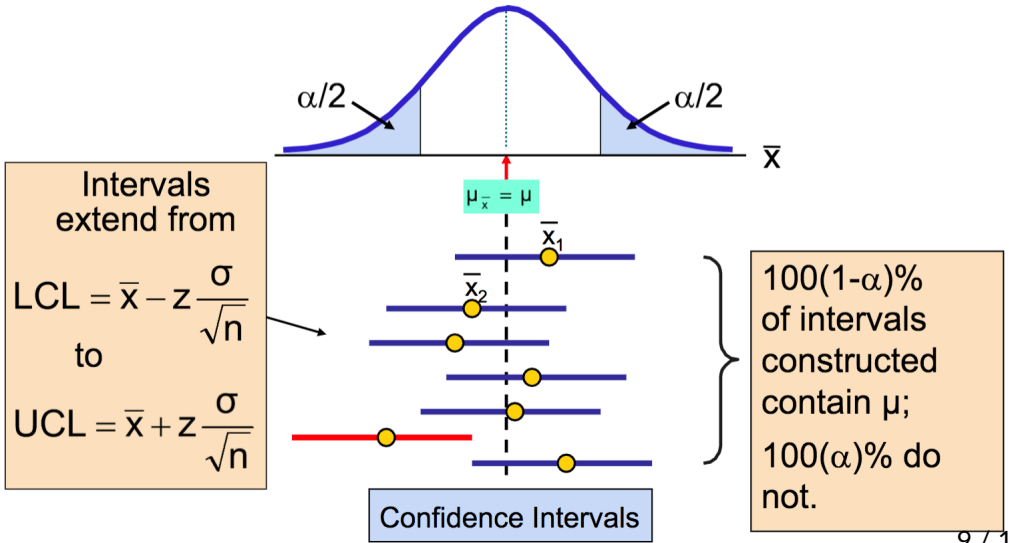
$$\bar{x} \pm 1.96\sigma_{\bar{X}}$$

Confidence Intervals: Interpretation

There is a 95% chance that the true population average lies in the interval $\bar{x} \pm 1.96\sigma_{\bar{X}}$.

What this really means is that if we took 100 random samples from the population and calculated 95% confidence intervals for each sample, we would expect 95 out of 100 intervals to contain the true population mean.

Confidence Intervals: Interpretation



Recipe: Confidence Intervals

Let $z_{\alpha/2}$ be the z-value that leaves area $\alpha/2$ in the upper tail of the normal distribution.

Then $1-\alpha$ confidence interval is given by

$$\bar{x} \pm \underbrace{z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{\text{Margin of Error}}$$

Margin of Error

The margin of error will be reduced if

- Population standard deviation is reduced $(\downarrow \sigma)$
- The sample size is increased $(\uparrow n)$
- The confidence level is decreased $(\downarrow (1-\alpha))$

Population variance is not known!

- So far, we have assumed that we know the true population variance σ^2
- This is obviously not realistic!
- Most times we have to use the sample variance S^2 instead of σ^2.
- How do we create a confidence interval in this case?

Population variance is not known

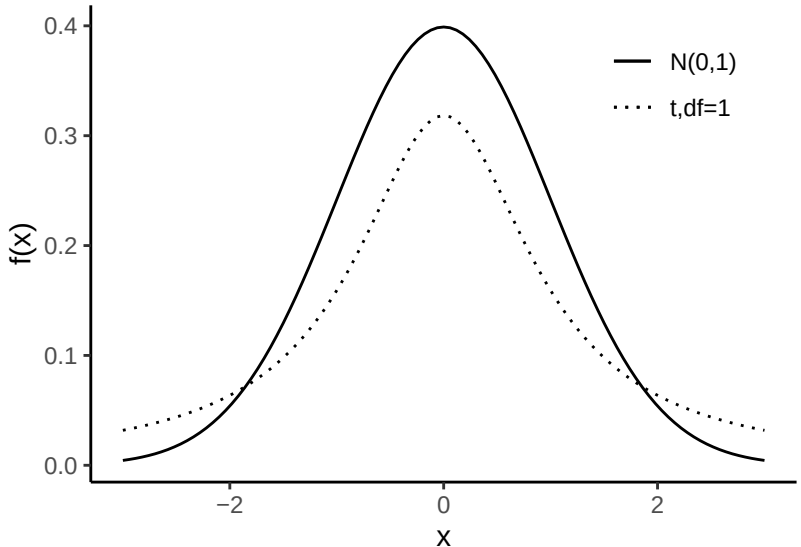
Instead of the Z statistic, we can use the T statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

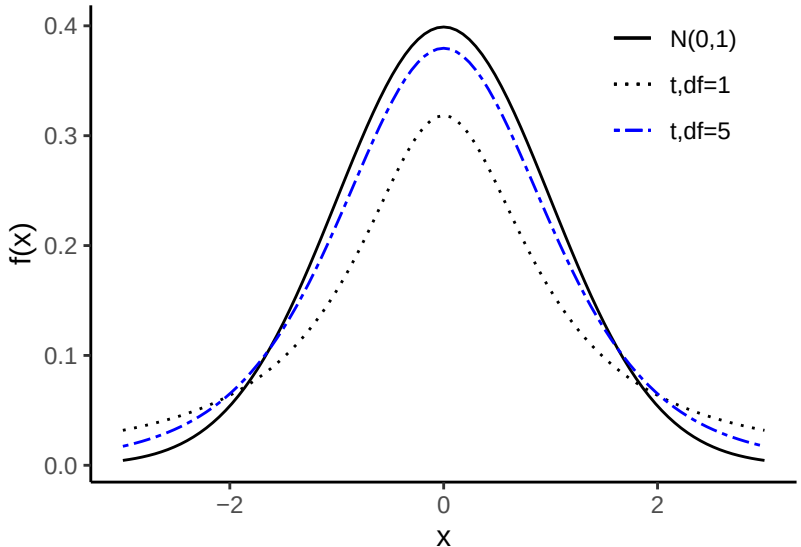
It can be shown that this statistic follows a t distribution with $n - 1$ degrees of freedom.

The t-distribution is similar to the normal distribution but has thicker tails to account for the greater uncertainty in smaller samples. However, in large samples t-distribution can be approximated by the standard-normal.

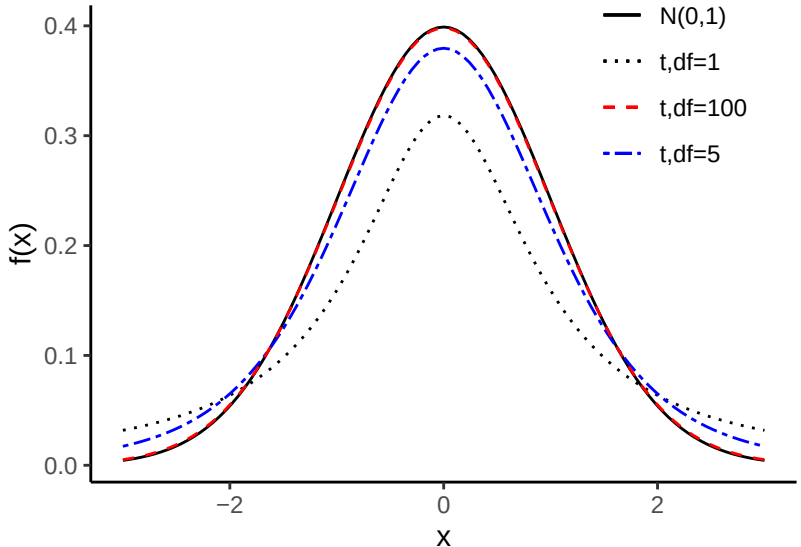
Student's T Distribution



Student's T Distribution



Student's T Distribution



Confidence Intervals: Unknown Variance

- Construct the confidence interval as before but now use T statistic instead of Z
- So need to use critical value for t

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}$$

- But since we said that in large samples, t is approximated by z, we can continue using the standard normal table for the critical values if $n \geqslant 100$.

Next up

- Problem Set 3 is due by the end of the day today
- Next class: Hypothesis testing and p-values
- Next week:
  - Review class on Tuesday
  - Midterm exam on Thursday