STATISTICS

FOR DATA ANALYSIS

SESSION OUTLINE

1	**Why Statistics?**	*Discuss the role of statistics in the context of business intelligence and decision-making, and introduce the statistics workflow*
2	**Descriptive Statistics**	*Understand data using descriptive statistics, including frequency distributions and measures of central tendency & variability*
3	**Probability Distributions**	*Model data with probability distributions, and use the normal distribution to calculate probabilities and make value estimates*
4	**Central Limit Theorem**	*Introduce the Central Limit Theorem, which leverages the normal distribution to make inferences on populations with any distribution*
5	**Confidence Intervals**	*Make estimates with confidence intervals, which use sample statistics to define a range where an unknown population parameter likely lies*
6	**Hypothesis Tests**	*Draw conclusions with hypothesis tests, which let you evaluate assumptions about population parameters using sample statistics*
7	**Regression Analysis**	*Make predictions with regression analysis, and estimate the values of a dependent variable via its relationship with independent variables*

SETTING EXPECTATIONS

This course is about **introducing** & **demystifying** essential statistics concepts

- *Our goal is to break down seemingly complex techniques using simple and intuitive explanations that will help you develop an intuition into when, why, and how to use them in the real world*

It's also about **applying** those concepts to real-world use cases

- *As we introduce each topic, we'll use Microsoft Excel as a tool to apply them through hands-on demos & assignments, and include additional projects to test your knowledge in different scenarios*

We'll be using **Excel** for **Office 365** on a **PC** for the course demos

- *What you see on your screen may not always match what you see on mine, especially if you are running a different operating system or following along with an older version of Excel*

You do **NOT** need a math or stats background to take this course

- *Although we will cover many statistical equations (and their equivalent Excel functions), the focus will be placed on the meaning behind them and not in the technical details or proof*

THE PROJECT

THE SITUATION

You've just been hired as a Recruitment Analyst by **Maven Business School**, an online startup that's looking to disrupt the postgraduate programs offered by traditional universities

THE BRIEF

You have data from the first graduating class of their MBA program, including details & scores from their application, the program itself, and their employment status 2 months later

Your goal is to **leverage statistics** to evaluate the results of this class, predict the performance of future classes, and propose changes in recruitment to improve graduate outcomes

THE OBJECTIVES

- Understand the data with descriptive statistics
- Model the data with probability distributions
- Make estimates with confidence intervals
- Draw conclusions with hypothesis tests
- Make predictions with regression analysis

MAVEN
BUSINESS SCHOOL

THE MAVEN BUSINESS SCHOOL DATASET

Student ID	Undergrad Degree	Undergrad Grade	MBA Grade	Work Experience	Employability (Before)	Employability (After)	Status	Annual Salary
1	Business	68.4	90.2	No	252	276	Placed	$111,000
2	Business	62.1	92.8	No	423	410	Not Placed	
3	Computer Science	70.2	68.7	Yes	101	119	Placed	$107,000
4	Engineering	75.1	80.7	No	288	334	Not Placed	
5	Finance	60.9	74.9	No	248	252	Not Placed	
6	Computer Science	74.5	80.7	No	145	209	Not Placed	
7	Finance	76.4	83.3	No	401	462	Placed	$109,000
8	Business	82.6	88.7	No	287	342	Placed	$148,000
9	Finance	76.9	75.4	No	275	347	Placed	$255,500
10	Computer Science	83.3	82.1	No	254			
11	Business	75.8	87.5	No	182			
12	Engineering	76	66.9	No	117			
13	Business	62.8	71.3	No	130			
14	Engineering	82.8	76.8	No	219			
15	Business	76	72.3	No	152			
16	Finance	76.9	72.4	No	228			
17	Computer Science	75.8	72	Yes	62			
18	Art	78	81	No	393			
19	Business	82.4	96.1	No	277			
20	Computer Science	76.2	76.7	No	206			
21	Business	62.5	80.3	No	229			
22	Art	78	77.8	No	182			
23	Engineering	66.5	62.6	No	98			
24	Computer Science	63.5	80.2	No	125			
25	Business	82.6	79.1	No	164			
26	Computer Science	79.2	77.8	No	186			
27	Computer Science	75	75.1	No	235			
28	Art	74.4	82.2	No	184			
29	Finance	67.9	70.5	No	76			
30	Art	76.8	70.8	No	126			
31	Business	83	87.5	No	183			
32	Computer Science	88.9	79.5	No	242			
33	Business	76.5	80.8	No	207			
34	Finance	79.9	79.6	Yes	181			
35	Business	70.4	88.9	No	239			

n=95

Field	Description
Student ID	*A unique identifier for each Maven Business School student*
Undergrad Degree	*The student's undergraduate degree*
Undergrad Grade	*The student's final grade average from their undergraduate degree (0-100)*
MBA Grade	*The student's final grade average from our master's degree program (0-100)*
Work Experience	*Indicator of the student's work experience prior to the program (Yes/No)*
Employability (Before)	*The student's score from a third-party test that measures their appeal to employers in selected industries, taken during their admissions process (0-500)*
Employability (After)	*The student's score from the same test, taken after obtaining their Master's*
Status	*Indicator of the student's employment status (Placed/Not Placed)*
Annual Salary	*The student's annual salary (USD)*

HELPFUL RESOURCES

Learn

Books

- *Naked Statistics – Charles Wheelan*
- *The Art of Statistics – David Spiegelhalter*

Websites

- *scribbr.com/category/statistics/*

YouTube

- *youtube.com/c/Kozyrkov – Statistical Thinking*
- *youtube.com/user/ExcelIsFun – Statistical Analysis*

Practice

Data Playground

- *mavenanalytics.io/data-playground*

Online Datasets

- *kaggle.com/datasets*
- *data.world/datasets/open-data*
- *vincentarelbundock.github.io/Rdatasets/articles/data*

WHY STATISTICS?

WHY STATISTICS FOR BUSINESS INTELLIGENCE?

In this section we'll discuss the **role of statistics** in the context of business intelligence and the decision-making process, review key terms, and introduce the statistics workflow

TOPICS WE'LL COVER:

- Why Statistics?
- Populations
- Statistics Workflow

GOALS FOR THIS SECTION:

- *Identify scenarios when statistics helps use data to make smart decisions, and when it's not needed*
- *Understand the concepts of populations & samples*
- *Review the statistics workflow and the concepts that will be covered throughout the course*

WHY STATISTICS?

Business intelligence is about using data to make smart decisions

Statistics is about *evaluating* those decisions under *uncertain* circumstances

When do you need statistics?

1) You don't have all the data you're interested in
   - You can only analyze some of the data you need to make your decision
   - There's **uncertainty** involved

2) The decision you're making is important
   - You don't want to make the wrong one based on your limited data
   - There's something specific to **evaluate**

POPULATION & SAMPLES

Why Statistics?

Populations

Statistics Workflow

A **population** contains all the data you're interested in to make your decision

- It's the data you wish you had, but are unlikely to get
- Any figure that summarizes a population is called a **parameter**

A **sample** contains some of the data from the population

- It's the data you have (which should ideally represent the population)
- Any figure that summarizes a sample is called a **statistic**

Statistics lets you make reasonable estimates about **parameters** using **statistics**

HEY THIS IS IMPORTANT!

Statistics can't create certainty out of uncertainty, it just helps you make controlled decisions under it!

THE STATISTICS WORKFLOW

Descriptive Statistics → **Probability Distributions** → **Confidence Intervals** → **Hypothesis Tests** → **Regression Analysis**

Descriptive Statistics: **Understand** what your sample data looks like

Probability Distributions: If the sample data fits a probability distribution, use it as a **model** for the entire population

Confidence Intervals: If the sample doesn't fit a distribution, use the central limit theorem to make **estimates** about population parameters

Hypothesis Tests: Continue to leverage the central limit theorem to draw **conclusions** about what a population looks like based on a sample

Regression Analysis: Use additional variables to increase the accuracy of your estimates and make **predictions** based on their relationships

HEY THIS IS IMPORTANT!

If you have all the population data, or simply need a bit of inspiration to make an "unimportant" decision, then descriptive statistics may be all you need!

DESCRIPTIVE STATISTICS

DESCRIPTIVE STATISTICS

In this section we'll cover understanding data with **descriptive statistics**, including frequency distributions, measures of central tendency, and measures of variability

TOPICS WE'LL COVER:

- Statistics Basics
- Distributions
- Central Tendency
- Variability

GOALS FOR THIS SECTION:

- *Identify the different types of variables in a dataset, along with their use cases*
- *Create frequency tables and plot the distributions of numerical variables using histograms*
- *Calculate the mean, median, mode, and standard deviation of a numerical variable*
- *Visualize the key descriptive statistics of a numerical variable using a box plot*

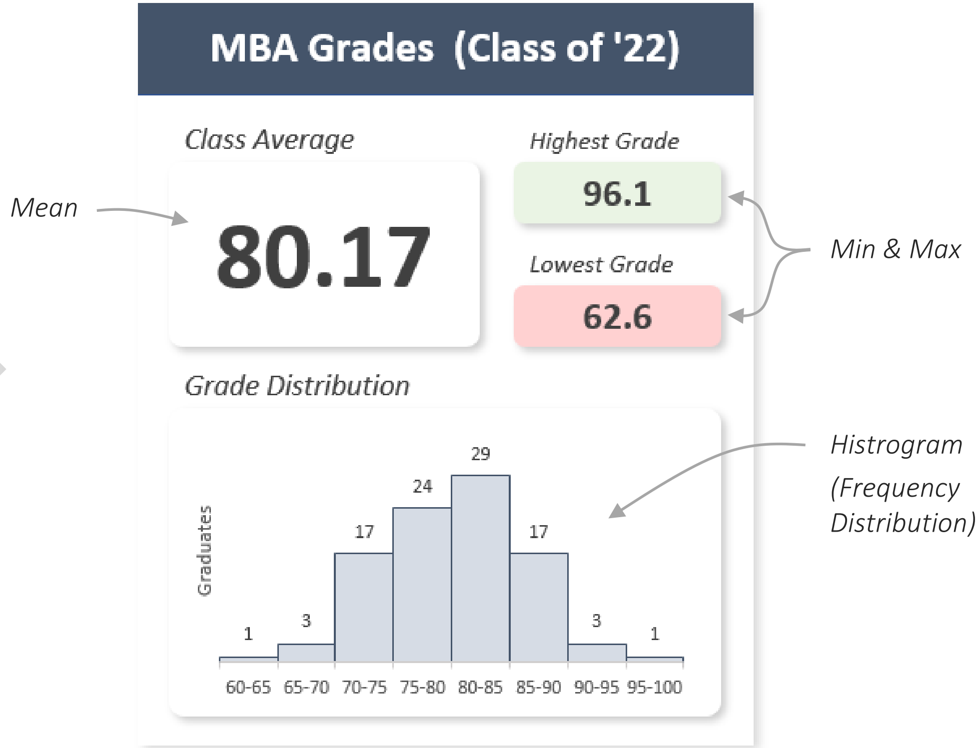
DESCRIPTIVE STATISTICS

The purpose of **descriptive statistics** is to summarize the characteristics of a variable

- They reduce a large array of numbers into a handful of figures that describe it accurately

Student ID	MBA Grade
1	90.2
2	92.8
3	68.7
4	80.7
5	74.9
6	80.7
7	83.3
8	88.7
9	75.4
10	82.1
11	87.5
12	66.9
13	71.3
14	76.8
15	72.3
16	72.4
17	72
18	81
19	96.1
20	76.[illegible]

n=95



TYPES OF VARIABLES

Statistics Basics

Distributions

Central Tendency

Variability

There are two main **types of variables** in a dataset: Numerical & Categorical

- **Numerical** variables represent numbers that are meant to be *aggregated*
- **Categorical** variables represent groups that can be used to *filter* numerical values

NUMERICAL:

Undergrad Grade	MBA Grade	Employability (Before)	Employability (After)	Annual Salary
68.4	90.2	252	276	$111,000
62.1	92.8	423	410	
70.2	68.7	101	119	$107,000
75.1	80.7	288	334	
60.9	74.9	248	252	
74.5	80.7	145	209	
76.4	83.3	401	462	$109,000
82.6	88.7	287	342	$148,000
76.9	75.4	275	347	$255,500
83.3	82.1	254	313	$103,500

CATEGORICAL:

Student ID	Undergrad Degree	Work Experience	Status
1	Business	No	Placed
2	Business	No	Not Placed
3	Computer Science	Yes	Placed
4	Engineering	No	Not Placed
5	Finance	No	Not Placed
6	Computer Science	No	Not Placed
7	Finance	No	Placed
8	Business	No	Placed
9	Finance	No	Placed
10	Computer Science	No	Placed

Even though these are numbers, this is a categorical variable (they won't be aggregated)

Possible question:

*"What's the mean **annual salary** by **work experience**?"*

aggregation *filter*

TYPES OF DESCRIPTIVE STATISTICS

There are 3 main **types of descriptive statistics** that can be applied to a variable:

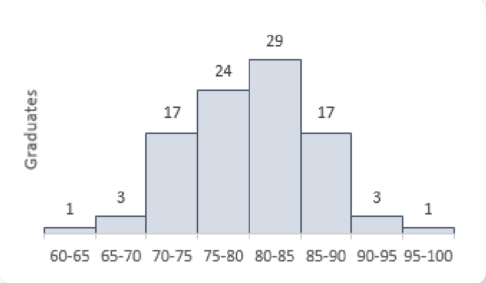
Distribution

*Represents the **frequency** of each value*

Examples:

- Frequency Tables
- Histograms

Grade Distribution



Central Tendency

*Represents the **middle** of the values*

Examples:

- Mean, Median, and Mode
- Skew

Class Average

80.17

Variability

*Represents the **dispersion** of the values*

Examples:

- Min, Max, and Range
- Quartiles & Interquartile Range
- Box & Whisker Plots
- Variance & Standard Deviation

Highest Grade	*Lowest Grade*
96.1	**62.6**

HEY THIS IS IMPORTANT!

Most measures of central tendency and variability can only be applied to numerical variables

FREQUENCY DISTRIBUTIONS

A **frequency distribution** counts the observations of each possible value in a variable

- They are commonly depicted using frequency tables

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	93.6
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

n=17

FREQUENCY TABLE:

Undergrad Degree	Frequency	Relative Frequency
Art	1	6%
Business	8	47%
Computer Science	2	12%
Engineering	4	24%
Finance	2	12%

The relative frequency shows the count of each value as a % of the total

PRO TIP: Use a PivotTable or the COUNTIFS() function to calculate frequencies for categorical variables in Excel

FREQUENCY DISTRIBUTIONS

Statistics Basics

Distributions

Central Tendency

Variability

For numerical variables, a frequency distribution typically counts the number of observations that fall into defined ranges or "bins" *(1-5, 6-10, etc.)*

- They are commonly depicted using grouped frequency tables or histograms

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	93.6
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

FREQUENCY TABLE:

Undergrad Grade	Frequency
63.3	1
66	1
67.5	1
67.7	1
68.1	1
68.7	1
70.1	1
74	1
74.6	1
75.3	1
75.6	1
76	1
78.9	1
79.3	1
82.9	1
88.8	1
93.6	1

This isn't a meaningful representation of the distribution of the data

FREQUENCY DISTRIBUTIONS

For numerical variables, a frequency distribution typically counts the number of observations that fall into defined ranges or "bins" *(1-5, 6-10, etc.)*

- They are commonly depicted using grouped frequency tables or histograms

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	93.6
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

GROUPED FREQUENCY TABLE:

Undergrad Grade	Frequency	Cumulative Relative Frequency
60-65	1	6%
65-70	5	35%
70-75	3	53%
75-80	5	82%
80-85	1	88%
85-90	1	94%
90-95	1	100%
Grand Total	**17**	

The cumulative relative frequency shows the running total of the relative frequencies

PRO TIP: Group the numerical values in a PivotTable or use the FREQUENCY() function with the upper limits to calculate frequencies for each bin in Excel

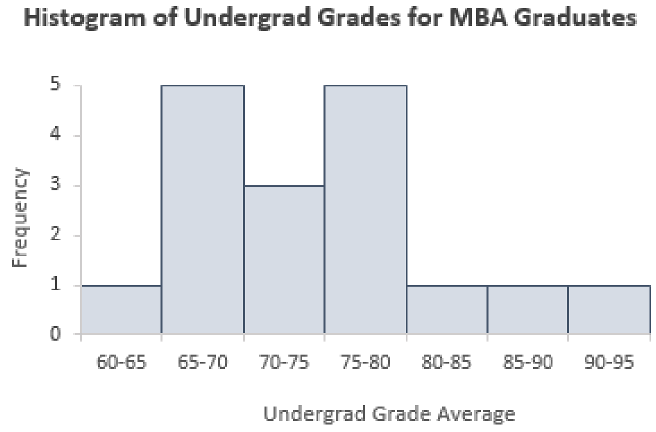
HISTOGRAMS

Histograms are used to visualize the distribution of a numerical variable

- They also provide a glimpse of the variable's central tendency and variability

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	93.6
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

n=17



PRO TIP: Create a histogram by using a column chart to plot the variable's frequency table, instead of using Excel's native histogram chart type (not as customizable)

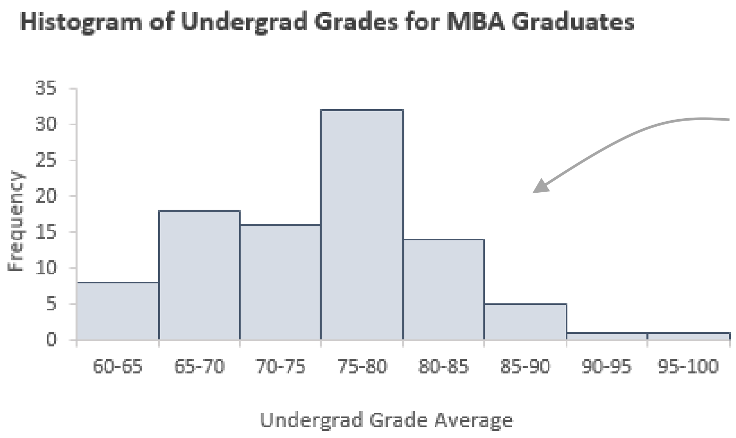
HISTOGRAMS

Histograms are used to visualize the distribution of a numerical variable

- They also provide a glimpse of the variable's central tendency and variability

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	93.6
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

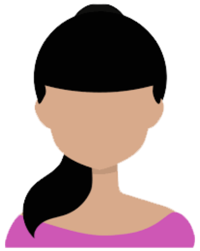
n=95



Histograms are best suited for variables with many observations, to reflect the true population distribution

PRO TIP: Bin size can significantly change the shape and "smoothness" of a histogram, so select a bin width that accurately shows the data distribution

ASSIGNMENT: FREQUENCY DISTRIBUTIONS



NEW MESSAGE
October 1, 2022

From: **Molly Mean** *(Director of Education)*

Subject: **First Graduate Class Results**

Welcome to Maven Business School!

As you know, we just had our first ever batch of MBA graduates.

I've looked at their grades individually already, but I'm not getting much insight from them – too many numbers!

I really need to get a clear picture of their grade averages to see if I need to make any tweaks to the program's curriculum.

Do you think you could give me a hand?

Thanks!

Reply   Forward

Key Objectives

1. Create a **frequency table** for the "MBA Grade" variable
2. Visualize the results using a **histogram**

MEAN

The **mean** is the calculated "average" value in a set on numbers

- It is calculated by dividing the sum of all values by the count of all observations
- It can only be applied to numerical variables *(not categorical)*

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	93.6
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

$$mean = \frac{sum\ of\ all\ values}{count\ of\ observations}$$

$$= \frac{1{,}270.4}{17} = \mathbf{74.73}$$

PRO TIP: Use the AVERAGEIFS() function if you want to calculate the mean for values that meet a specified criteria *(i.e., Mean by Undergrad Degree)*

LIMITATIONS OF THE MEAN

The main **limitation of the mean** is that it is sensitive to outliers *(extreme values)*

- *"The average income in America is not the income of the average American"*

$35,000 *$50,000* *$65,000* *$100,000,000*

mean = ***$50,000***

mean = ***$25,037,500***

HEY THIS IS IMPORTANT!

While the mean is typically great for making a "best-guess" estimate of a value, it's important to complement this value with other descriptive statistics like the distribution, median, and mode to see if the mean value is being distorted by outliers

MEDIAN

Statistics Basics

Distributions

Central Tendency

Variability

The **median** is the “middle value” in a sorted set of numbers

- Unlike the mean, the median is NOT sensitive to outliers
- When there are two middle-ranked values, the median is the average of the two

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	93.6
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

Undergrad Grade
63.3
66
67.5
67.7
68.1
68.7
70.1
74
74.6
75.3
75.6
76
78.9
79.3
82.9
88.8
93.6

n=17

Median = **74.6**

MEDIAN

Statistics Basics

Distributions

Central Tendency

Variability

The **median** is the "middle value" in a sorted set of numbers

- Unlike the mean, the median is NOT sensitive to outliers
- When there are two middle-ranked values, the median is the average of the two

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

Undergrad Grade
63.3
66
67.5
67.7
68.1
68.7
70.1
74
74.6
75.3
75.6
76
78.9
79.3
82.9
88.8

n=16

Median = **74.3**

*(average of **74** and **74.6**)*

MODE

Statistics Basics

Distributions

Central Tendency

Variability

The **mode** is the "most frequent" value in a variable

- It can be applied to both numerical and categorical variables

Mode = "**Business**"

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	93.6
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

Mode = **N/A**

MODE

Statistics Basics

Distributions

Central Tendency

Variability

The **modal class** is the group with the highest frequency

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	93.6
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

GROUPED FREQUENCY TABLE:

Undergrad Grade	Frequency
60-65	1
65-70	5
70-75	3
75-80	5
80-85	1
85-90	1
90-95	1
Grand Total	**17**

Mode = **65-70**, **75-80**

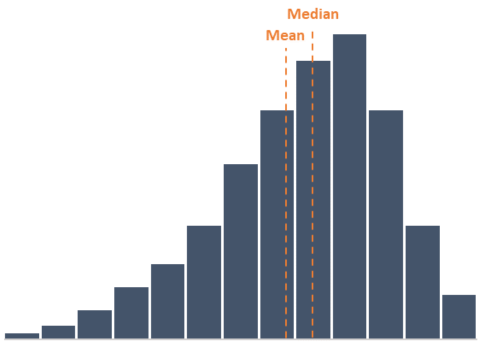
*This is a **multi-modal** distribution, which indicates that there may be another variable impacting the undergrad grades*

SKEW

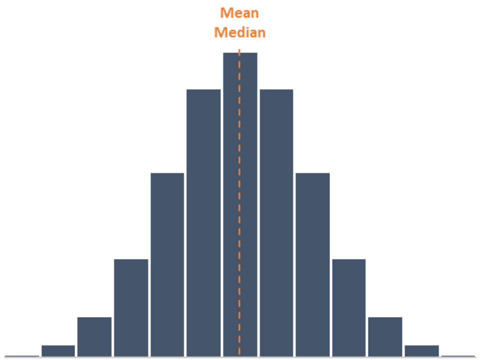
The **skew** represents the asymmetry of a distribution around its mean

- In a **zero-skewed** distribution, the mean and median are equal
- In a **right-skewed** *(or positive)* distribution, the mean is typically greater than the median
- In a **left-skewed** *(or negative)* distribution, the mean is typically smaller than the median

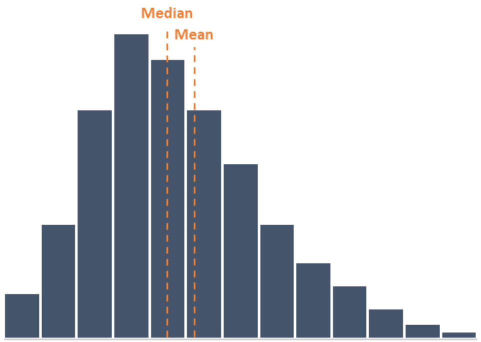
Left skew



Zero skew

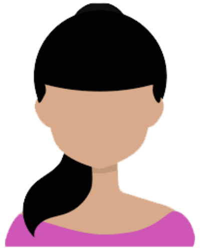


Right skew



*This is one of the properties of a **normal distribution***
(more on that later!)

ASSIGNMENT: MEASURES OF CENTRAL TENDENCY



NEW MESSAGE

October 1, 2022

From: **Molly Mean** *(Director of Education)*

Subject: **RE: First Graduate Class Results**

Thanks for visualizing those grades for me!

It's interesting to see that not many students scored above 90.

I wonder if, since this is a Masters in *Business* Administration, Business Undergrads tend to do better than others.

Could you give me a quick summary that shows the average MBA grades for Business Undergrads vs Other Undergrads?

I'd appreciate if you could interpret the results for me as well.

Thanks!

Reply   Forward

Key Objectives

1. Calculate the **mean**, **median**, **mode**, and **skew** for the "MBA Grade" variable by "Undergrad Degree"

RANGE

Statistics Basics

Distributions

Central Tendency

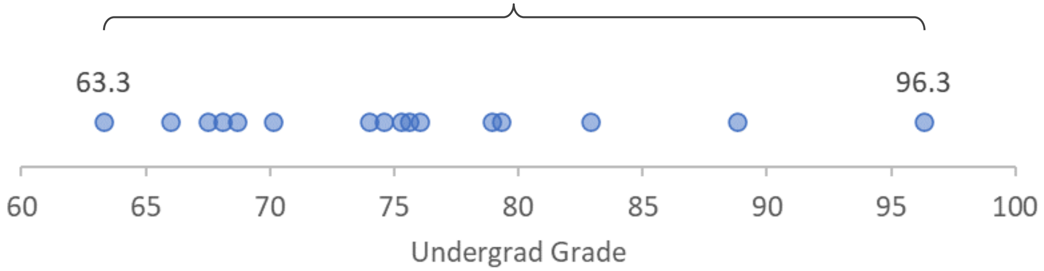
Variability

The **range** is the spread from the lowest *(min)* to the highest *(max)* value in a variable

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	96.3
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Engineering	75.3
Engineering	68.1
Finance	63.3

n=16

$$range = 96.3 - 63.3 = \mathbf{33}$$



HEY THIS IS IMPORTANT!

While the range is generally a good indicator of the variability in a numerical variable, a single outlier can cause it to change significantly

INTERQUARTILE RANGE

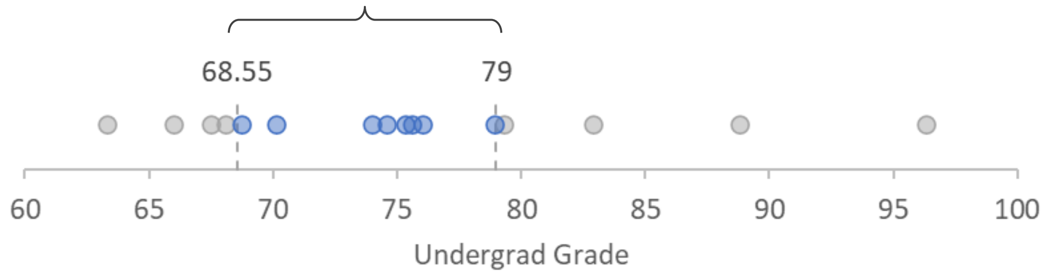
The **interquartile range** is the spread of the *middle half* of the values in a variable

- In other words, it's the spread from the **first quartile** to the **third quartile**

	Undergrad Grade	
25%	63.3	
	66	
	67.5	
	68.1	
		Q1 = 68.55
25%	68.7	
	70.1	
	74	
	74.6	
		Median
25%	75.3	
	75.6	
	76	
	78.9	
		Q3 = 79
25%	79.3	
	82.9	
	88.8	
	96.3	

n=16

$$IQR = 79 - 68.55 = \mathbf{10.45}$$



HEY THIS IS IMPORTANT!

The quartiles in this example are calculated by *including* the median, but Excel also lets you use the *exclusive* method

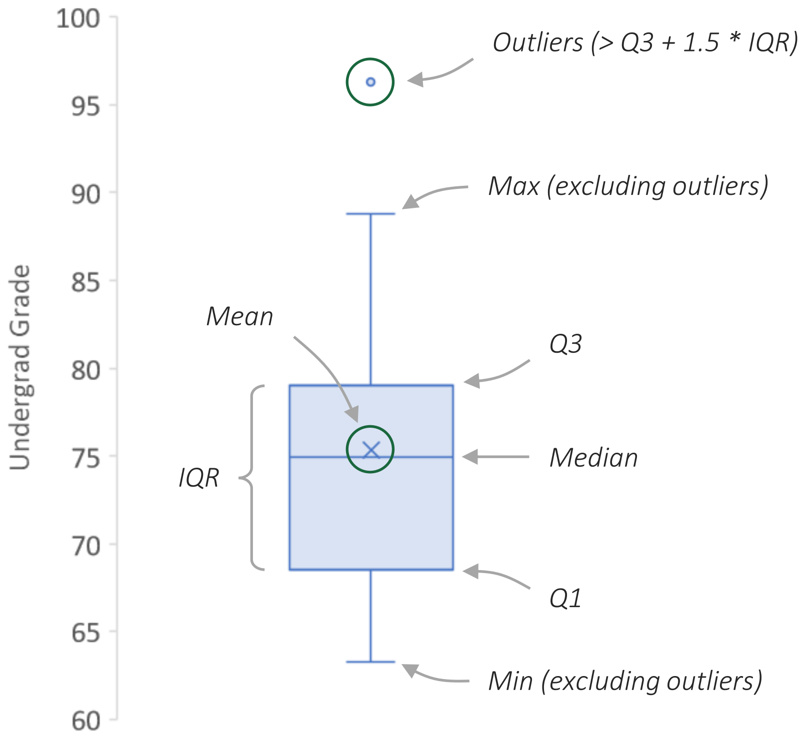
There is no right or wrong method to use, but many prefer the inclusive method, as it leads to a narrower range

BOX & WHISKER PLOTS

Box & whisker plots are used to visualize key descriptive statistics

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	96.3
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Engineering	75.3
Engineering	68.1
Finance	63.3

n=16



BOX & WHISKER PLOTS

Statistics Basics

Distributions

Central Tendency

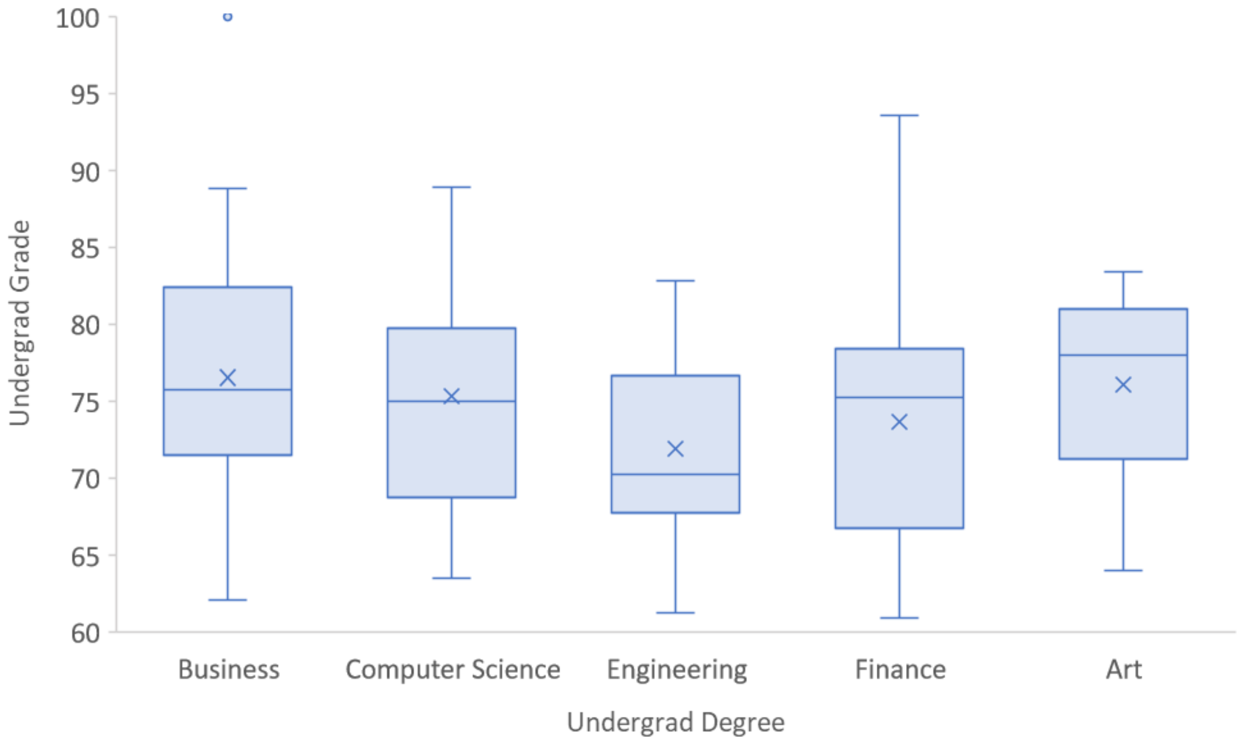
Variability

Box & whisker plots are used to visualize key descriptive statistics

- They can be used to quickly compare statistical characteristics between categories

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	96.3
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Engineering	75.3
Engineering	68.1
Finance	63.3

n=95



STANDARD DEVIATION

Statistics Basics

Distributions

Central Tendency

Variability

The **standard deviation** measures, on average, how far each value lies from the mean

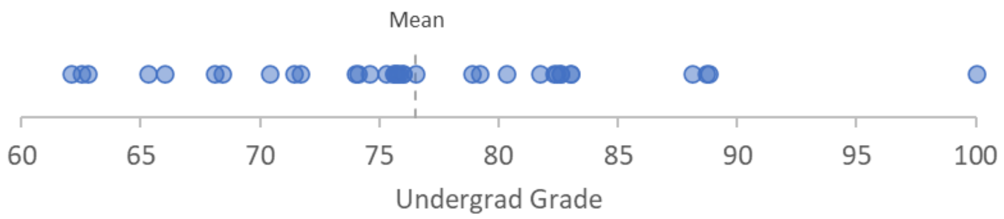
- The *higher* the standard deviation, the *wider* a distribution is *(and vice versa)*

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	96.3
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Engineering	75.3
Engineering	68.1
Finance	63.3

n=95

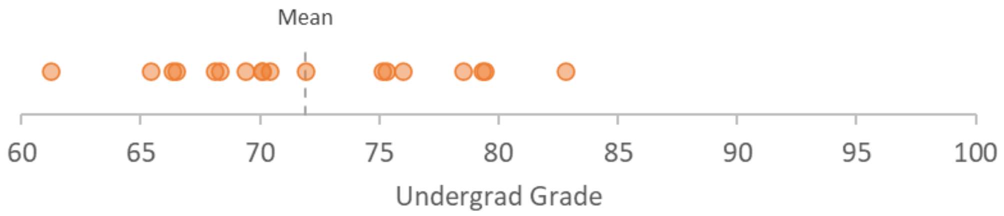
Std Dev

Business Undergrads



= 8.17

Engineering Undergrads



= 5.79

VARIANCE

Statistics Basics

Distributions

Central Tendency

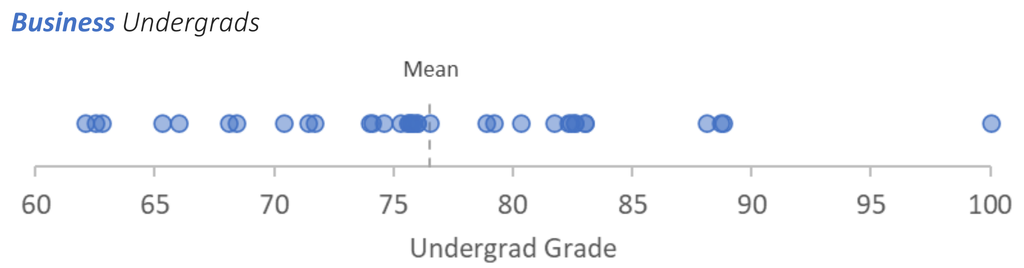
Variability

The **variance** is the square of the standard deviation

- Since its units are on a larger scale than the variable it's based on, it's not intuitive to interpret

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	96.3
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Engineering	75.3
Engineering	68.1
Finance	63.3

n=95



Variance

= **66.82**

HEY THIS IS IMPORTANT!

The variance does have its place in some statistical tests, so it shouldn't be discarded, but as a single numerical measure of a variable's dispersion the standard deviation is more effective

PRO TIP: COEFFICIENT OF VARIATION

The **coefficient of variation** measures the standard deviation relative to the mean

- It is used to compare the standard deviations of variables with significantly different means

Undergrad Grade	Employability (Before)
74	133
74.6	122
79.3	236
70.1	143
88.8	354
66	214
82.9	225
96.3	261
75.6	277
67.5	282
68.7	322
76	326
67.7	421
75.3	368
68.1	279
63.3	268

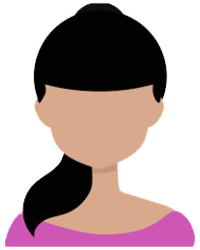
n=95

$$CV = \frac{standard\ deviation}{mean}$$

	Undergrad Grade	Employability (Before)
Standard Deviation:	7.42	85.94
Mean:	74.9	239.9
Coefficient of Variation:	**0.099**	**0.358**

*On average, undergrad grades differ from the mean by **~10%** of its value, while employability scores differ by **~ 35%***

ASSIGNMENT: MEASURES OF VARIABILITY



NEW MESSAGE
October 1, 2022

From: **Molly Mean** *(Director of Education)*

Subject: **RE: First Graduate Class Results**

Interesting observation on the “skew” there – I hadn’t even heard that word before!

So… if it looks like it’s the Business Undergrads in our program that are getting the uncommonly high scores, could it be that their grades are just more dispersed altogether?

I would hate to make a wrong assumption here.

It would help if you could provide some sort of visual as well, especially if I’m going to end up taking this to the board.

Thanks again!

Reply Forward

Key Objectives

1. Calculate the **range**, **interquartile range**, and **standard deviation** for the “MBA Grade” variable by “Undergrad Degree”
2. Compare the “MBA Grade” by “Undergrad Degree” using a **box plot**

KEY TAKEAWAYS: DESCRIPTIVE STATISTICS

There are two main types of variables: **numerical** & **categorical**

- *Numerical variables are meant to be aggregated, and categorical variables are used to create groups*

The **distribution** represents the "shape" of a variable

- *Histograms are a great way to visualize this "shape" by plotting the frequency of each value (or class)*

The **mean** & **median** locate the "center" of a distribution

- *Don't focus on using one instead of the other, rather on using both to complement each other*

The **standard deviation** measures the dispersion around the mean

- *Use a box plot alongside the standard deviation to provide additional context on the variability and center*

MAVEN PIZZA PARLOR | PROJECT BRIEF

You are a BI Consultant that has just been approached by **Maven Pizza Parlor**, a new pizza place in New Jersey that needs help with their demand planning

From: **Mary Margherita** *(Owner)*
Subject: **Daily Pizza Sales**

Hi!

We we're extract our daily pizza sales from our POS system, and we want to use this for planning, but none in the team is data savvy.

Is that something you could help us with?

We want to know how many pizza sales to expect every day, how much they typically vary, and if they fluctuate by day of the week.

Thank you!

Pizza_Sales.xlsx

Reply Forward

Key Objectives

1. Summarize the daily pizza sales by using descriptive statistics

PROBABILITY DISTRIBUTIONS

PROBABILITY DISTRIBUTIONS

In this section we'll cover modeling data with **probability distributions**, and use the normal distribution to calculate probabilities and make estimates about normal populations

TOPICS WE'LL COVER:

- Distribution Basics
- Distribution Types
- Normal Distribution
- Z-Scores
- Probabilities
- Value Estimates

GOALS FOR THIS SECTION:

- *Understand the concept of a probability distribution, and its relationship with frequency distributions*
- *Learn about the different types of probability distributions, and their main differences*
- *Identify the properties of the normal distribution*
- *Calculate probabilities, values, and z-scores from normal distributions using Excel functions*

PROBABILITY DISTRIBUTIONS

Distribution Basics

Distribution Types

Normal Distribution

Z-Scores

Probabilities

Value Estimates

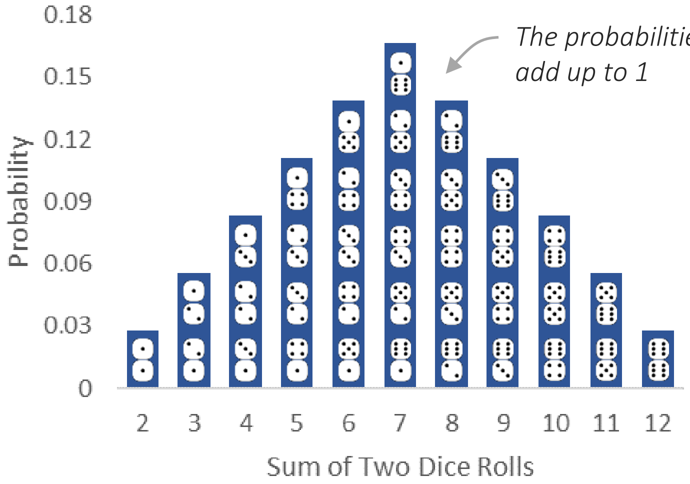
A **probability distribution** represents a variable's idealized frequency distribution

It shows all the possible values a variable can take, and their chances of occurring

- Frequencies in a sample are based on the underlying probabilities of those values occurring

EXAMPLE | *Results of rolling two dice*

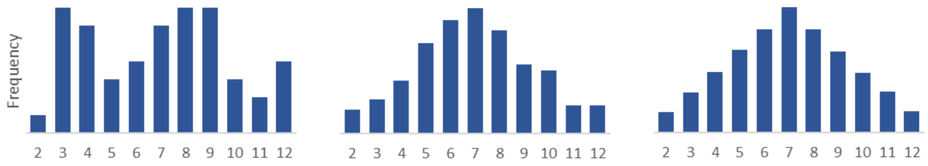
***PROBABILITY DISTRIBUTION** (Population):*



The probabilities add up to 1

***FREQUENCY DISTRIBUTION** (Sample):*

*Sample size = **50*** *n = **500*** *n = **50,000***



In an infinite sample, a variable's relative frequency distribution is equal to its probability distribution!

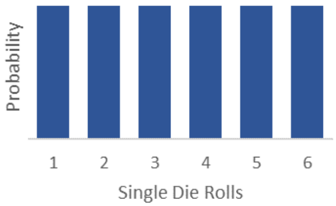
*This is known as a **binomial distribution**, and it can be used to calculate probabilities on the outcome of rolling two dice (without rolling them fifty thousand times!)*

TYPES OF PROBABILITY DISTRIBUTIONS

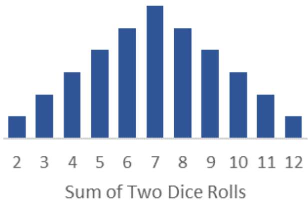
There are two **types of probability distributions**: Discrete & Continuous

1) **Discrete** probability distributions

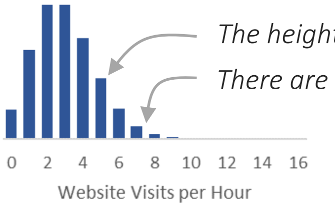
Uniform



Binomial



Poisson

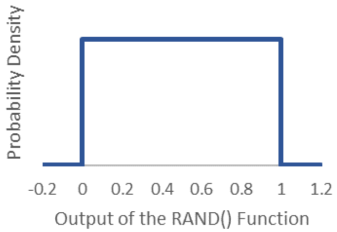


The height of each bar is its probability

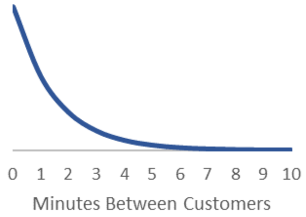
There are "gaps" between the numbers

2) **Continuous** probability distributions

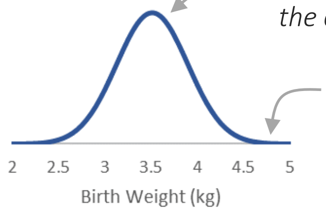
Uniform



Exponential



Normal



The height of the curve is NOT its probability, the ***area under the curve*** *is (more on this later!)*

The numbers can take any value

THE NORMAL DISTRIBUTION

Distribution Basics

Distribution Types

Normal Distribution

Z-Scores

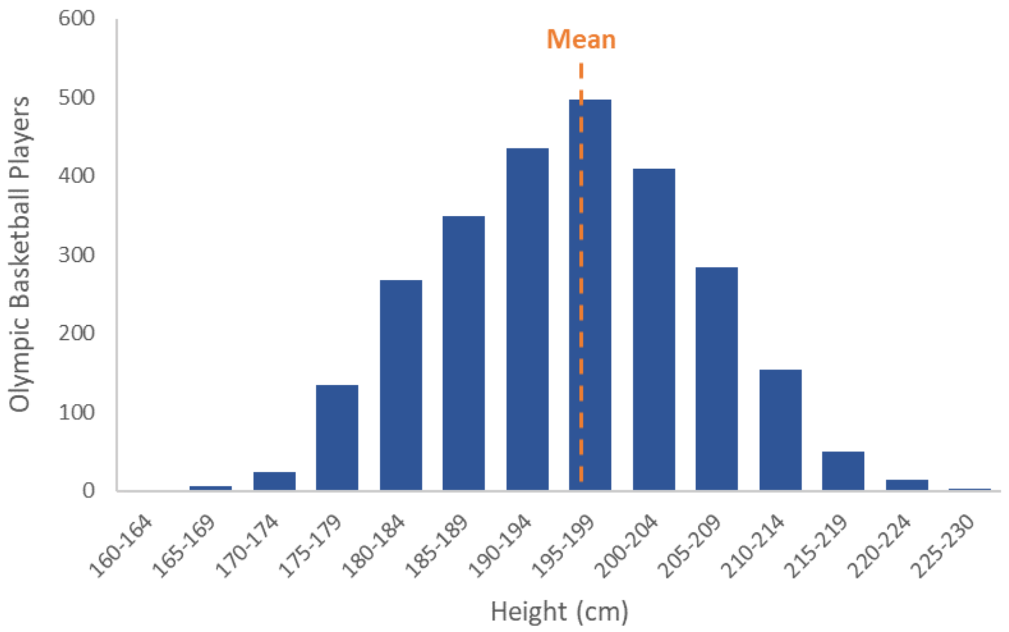
Probabilities

Value Estimates

Many numerical variables naturally follow a **normal distribution**, or "bell curve"

- Normal distributions are symmetrical around the mean and have no skew *(mean = median)*, with most data concentrated around its center and flaring out in "tails" on both ends

EXAMPLE | *Olympic Basketball Player Heights*



HEY THIS IS IMPORTANT!

Since they are so common, many statistical tests are designed for normally distributed populations, which is why we'll mostly focus on the normal distribution in the course

THE NORMAL DISTRIBUTION

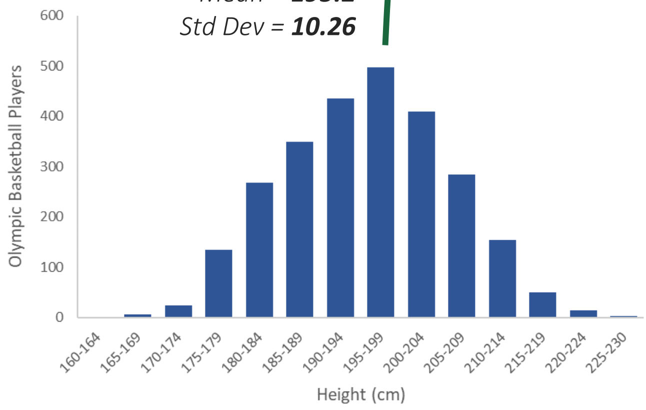
The normal distribution is described by two values: the **mean** & **standard deviation**

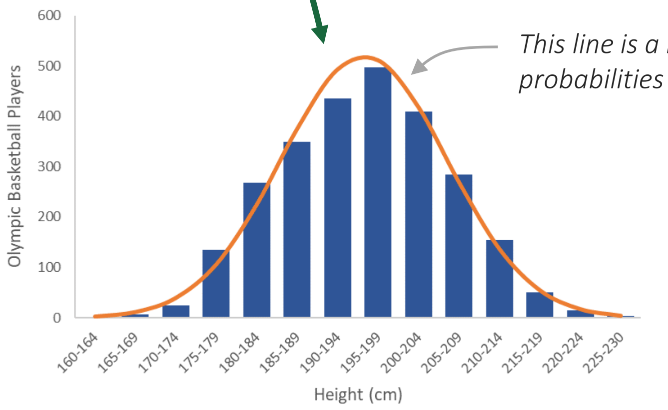
$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

*This is the **probability density function** for the normal distribution, which determines the height of the normal curve at any value (x) given a mean (μ) and a standard deviation (σ)*

(don't worry, there's an Excel function for it!)

Mean = ***195.2***
Std Dev = ***10.26***





This line is a model that we can use to calculate probabilities of Olympic Basketball player heights!

THE NORMAL DISTRIBUTION

Distribution Basics

Distribution Types

Normal Distribution

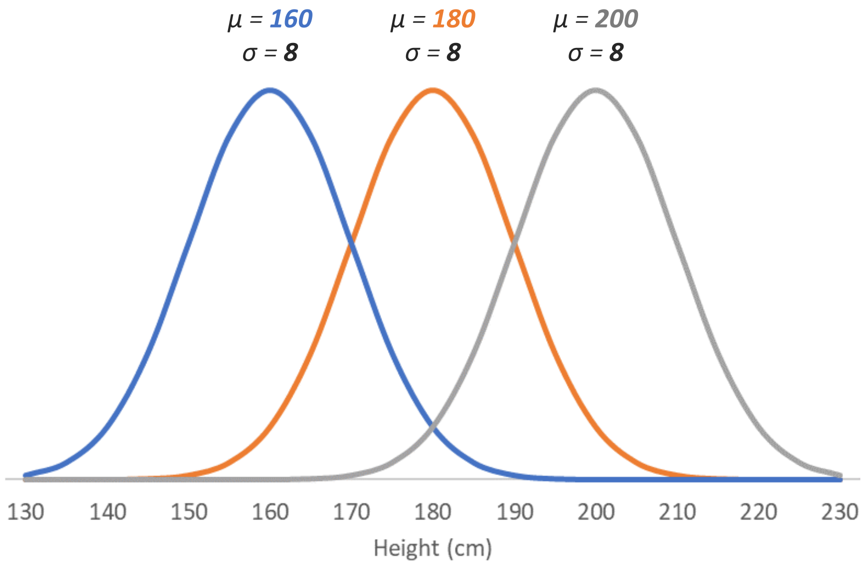
Z-Scores

Probabilities

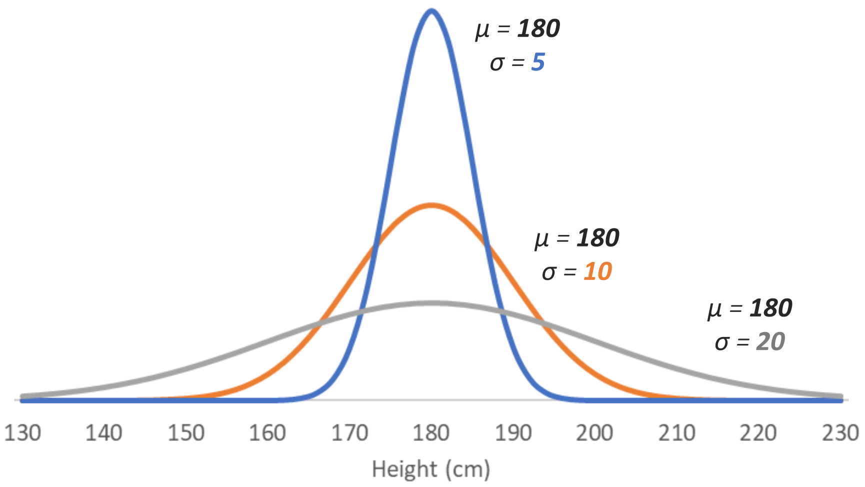
Value Estimates

The normal distribution is described by two values: the **mean** & **standard deviation**

- The mean determines the *center* of the distribution, and the standard deviation its *width*



*Changing the mean **shifts** the curve along the x axis*



*Changing the standard deviation **squeezes** or **stretches** the curve*

Z-SCORES

A **z-score** indicates how many standard deviations away from the mean a value lies

$$z = \frac{x - \mu}{\sigma}$$

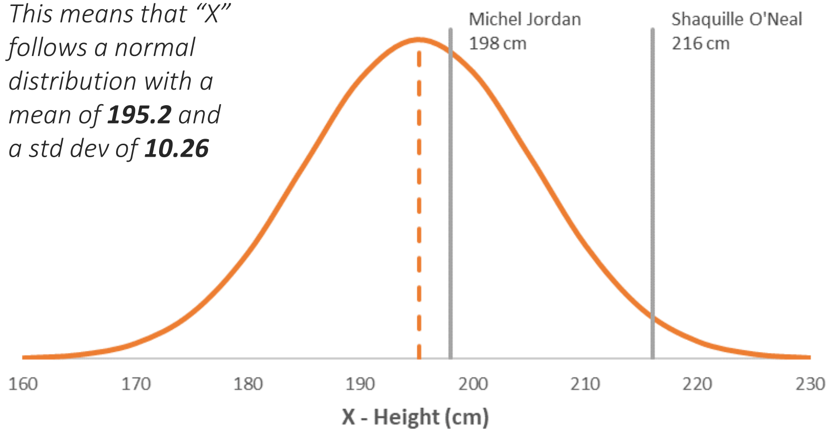
To calculate a z-score for a value, simply subtract the mean and divide by the standard deviation

(or use the STANDARDIZE function)

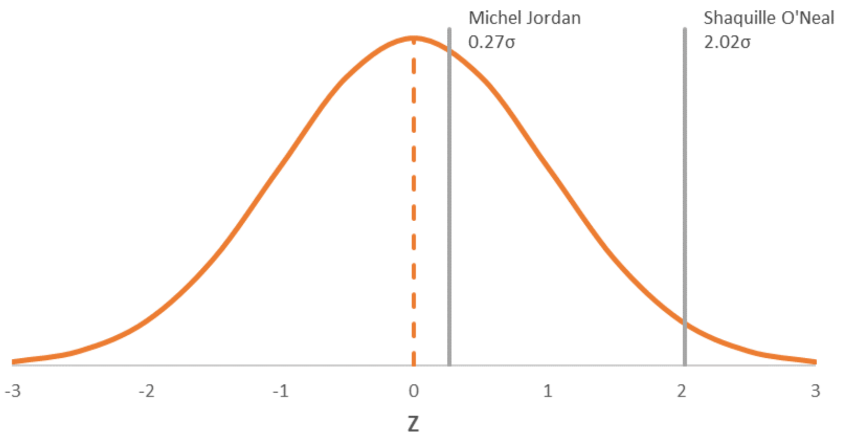
$$z = \frac{198 - 195.2}{10.26} = 0.27$$

X ~ N (195.2, 10.26)

This means that "X" follows a normal distribution with a mean of ***195.2*** *and a std dev of* ***10.26***



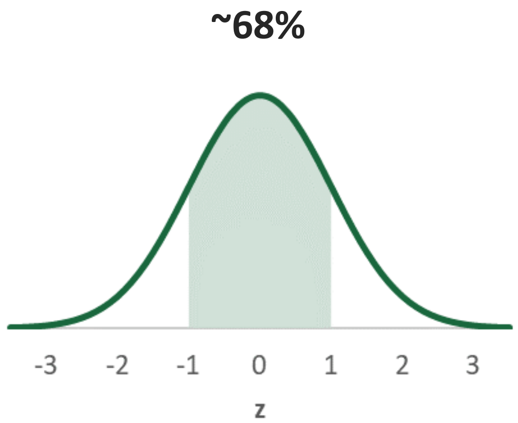
Z ~ N (0, 1)



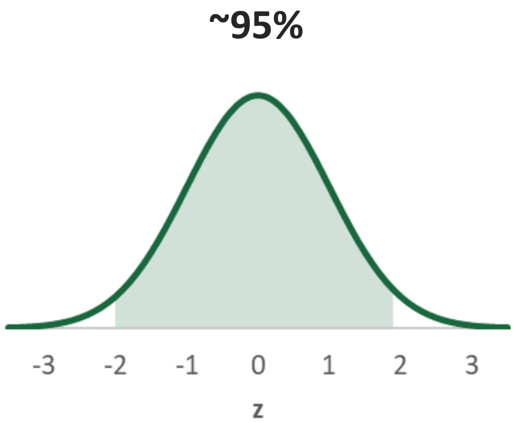
This is known as the ***standard normal distribution****, or z-distribution, and has a mean of 0 and a standard deviation of 1*

THE EMPIRICAL RULE

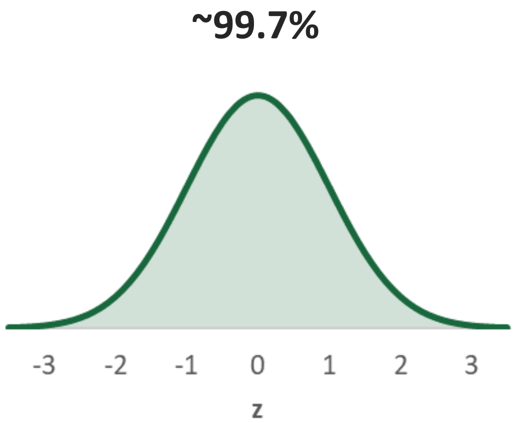
The **empirical rule** outlines where most values fall in a normal distribution



1σ *away from the mean*



2σ *away from the mean*



3σ *away from the mean*

PRO TIP: Beyond using a histogram to determine whether your data is distributed normally, check if it follows the empirical rule

ASSIGNMENT: NORMAL DISTRIBUTIONS

NEW MESSAGE
October 7, 2022

From: **Nick Normal** *(Head of Student Placement)*

Subject: **Student Salaries**

Hey, nice to meet you!

I just spoke to Molly, and she mentioned that you were going to be able to make some predictions on student grades since they were “normal”, or something like that.

Could you do the same for graduate salaries?

It sounds like something that could be really beneficial for me.

Looking forward to hearing back from you,

Thanks!

Reply Forward

Key Objectives

1. Plot the distribution of “Annual Salary” to see if it **resembles a bell curve**
2. Check if the **mean & median are equal**
3. Calculate the percentage of salaries that lie 1, 2, and 3 standard deviations from the mean to see if the variable **follows the empirical rule**

EXCEL NORMAL DISTRIBUTION FUNCTIONS

These **Excel functions** help make calculations related to the normal distribution:

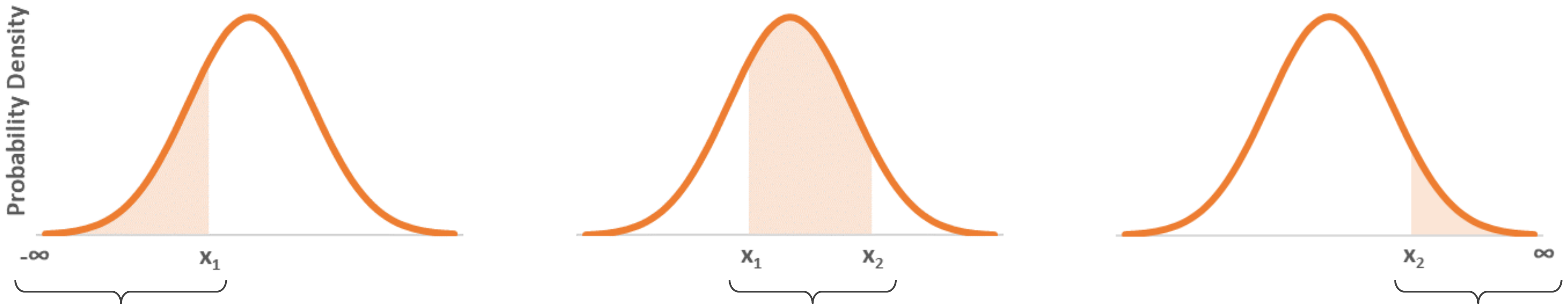
NORM.DIST()	*Returns the cumulative probability or the probability density at an x value from a given normal distribution*	=**NORM.DIST**(x, μ, σ, cumulative)
NORM.INV()	*Returns the x value in a given normal distribution at a specified cumulative probability*	=**NORM.INV**(probability, μ, σ)
STANDARDIZE()	*Returns the z-score for a specified x value in a given normal distribution*	=**STANDARDIZE**(x, μ, σ)
NORM.S.DIST()	*Returns the cumulative probability or the probability density at a z-score from the standard normal distribution*	=**NORM.S.DIST**(z, cumulative)
NORM.S.INV()	*Returns the z-score in the standard normal distribution at a specified cumulative probability*	=**NORM.S.INV**(probability)

CALCULATING PROBABILITIES

If a variable follows a normal distribution, you can **calculate the probability** of randomly obtaining a value within a specified range

- This is determined by the area under the curve in that range

Less than or equal to "x_1" | **Between "x_1" and "x_2"** | **Greater than or equal to "x_2"**



*The area from negative infinity to "x_1" is the **cumulative probability***

This is the cumulative probability of "x_2" minus the cumulative probability of "x_1"

This is 1 (the entire area under the curve) minus the cumulative probability of "x_2"

HEY THIS IS IMPORTANT!

You CANNOT calculate the probability of obtaining an x value *exactly* – there's no area under a single point!

THE NORM.DIST FUNCTION

NORM.DIST()

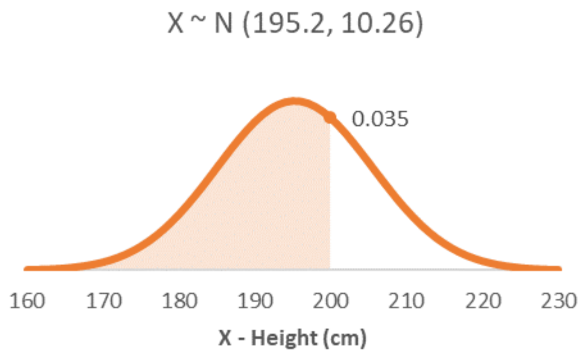
Returns the cumulative probability or the probability density at "x" from a normal distribution

=**NORM.DIST**(x, mean, standard_dev, cumulative)

- *The **value** to calculate the probability for* (x)
- *The **mean** & **standard deviation** for the normal distribution of the population* (mean, standard_dev)
- ***TRUE**: The area under the curve*
***FALSE:** The height of the curve* (cumulative)

Possible question:

*"What's the probability of an Olympic Basketball Player being **2 meters tall or shorter**?"*



This is the probability!

=**NORM.DIST**(200, 195.2, 10.26, TRUE) **= 0.68**

=**NORM.DIST**(200, 195.2, 10.26, FALSE) **= 0.035**

This is just the height of the curve

THE NORM.DIST FUNCTION

NORM.DIST()

Returns the cumulative probability or the probability density at "x" from a normal distribution

=**NORM.DIST**(x, mean, standard_dev, cumulative)

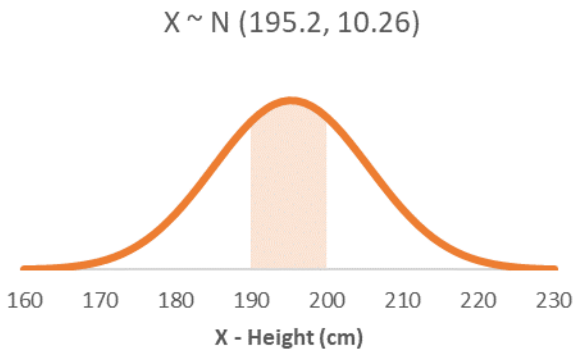
*The **value** to calculate the probability for*

*The **mean** & **standard deviation** for the normal distribution of the population*

***TRUE**: The area under the curve*
***FALSE:** The height of the curve*

Possible question:

*"What's the probability of an Olympic Basketball Player being **between 1.9 and 2 meters tall**?"*



=**NORM.DIST**(200, 195.2, 10.26, TRUE) = **0.68**

=**NORM.DIST**(190, 195.2, 10.26, TRUE) = **0.3061**

=0.68-0.306 = **0.3739**

This is the probability!

THE NORM.DIST FUNCTION

NORM.DIST()

Returns the cumulative probability or the probability density at "x" from a normal distribution

=**NORM.DIST**(x, mean, standard_dev, cumulative)

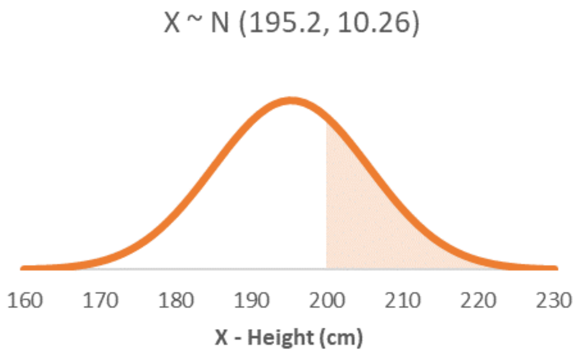
*The **value** to calculate the probability for*

*The **mean** & **standard deviation** for the normal distribution of the population*

***TRUE**: The area under the curve*
***FALSE:** The height of the curve*

Possible question:

*"What's the probability of an Olympic Basketball Player being **at least 2 meters tall**?"*



=**NORM.DIST**(200, 195.2, 10.26, TRUE) = **0.68**

=1-**NORM.DIST**(190, 195.2, 10.26, TRUE) = **0.32**

The cumulative probability under the entire curve is equal to 1

(it's every value possible!)

This is the probability!

THE NORM.S.DIST FUNCTION

Distribution Basics

Distribution Types

Normal Distribution

Z-Scores

Probabilities

Value Estimates

NORM.S.DIST()

Returns the cumulative probability or the probability density at “z” from the z-distribution

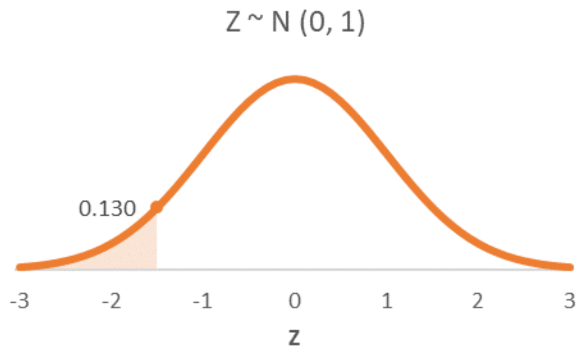
=**NORM.S.DIST**(z, cumulative)

*The **z-score** to calculate the probability for*

***TRUE**: The area under the curve*
***FALSE:** The height of the curve*

Possible question:

*“What’s the probability of an Olympic Basketball Player being **at least 1.5 standard deviations shorter than the mean**?”*



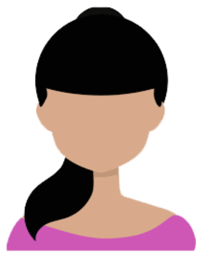
=**NORM.S.DIST**(-1.5, TRUE) **= 0.066**

This is the probability!

=**NORM.S.DIST**(-1.5, FALSE) **= 0.130**

This is just the height of the curve

ASSIGNMENT: CALCULATING PROBABILITIES



NEW MESSAGE
October 8, 2022

From: **Molly Mean** *(Director of Education)*

Subject: **Honor Students**

Hi again!

I keep thinking about the possibilities now that we know the grade averages for our graduates follow a normal distribution.

For example, I'd love to consider anyone that graduates with an average of 90 or higher an "honor student".

What would be the probability of someone getting that grade?

I'd hate for it to be more than 10% of students.

Thanks!

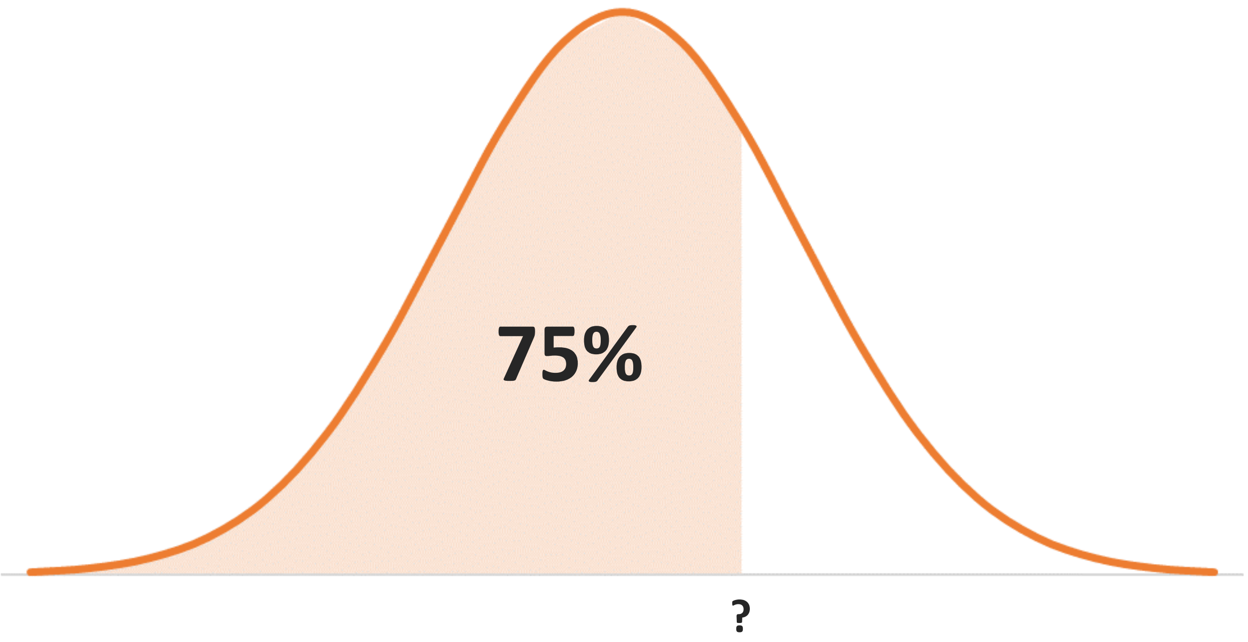
Reply Forward

Key Objectives

1. Use the **NORM.DIST** function to calculate the probability of getting an "MBA Grade" greater than or equal to 90

ESTIMATING VALUES

If a variable follows a normal distribution, you can **estimate the value of "x" or "z"** at a specified cumulative probability



THE NORM.INV FUNCTION

- Distribution Basics
- Distribution Types
- Normal Distribution
- Z-Scores
- Probabilities
- **Value Estimates**

NORM.INV()

Returns the x value in a normal distribution at a specified cumulative probability

=**NORM.INV**(probability, mean, standard_dev)

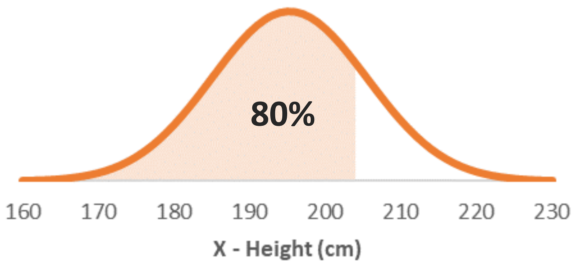
*The **cumulative probability** for the value you want*

*The **mean** & **standard deviation** for the normal distribution of the population*

Possible question:

*"How tall do you need to be to be **taller than 80%** of Olympic Basketball Players?"*

X ~ N (195.2, 10.26)



=**NORM.INV**(0.8, 195.2, 10.26) = **203.8** cm

THE NORM.S.INV FUNCTION

NORM.S.INV()

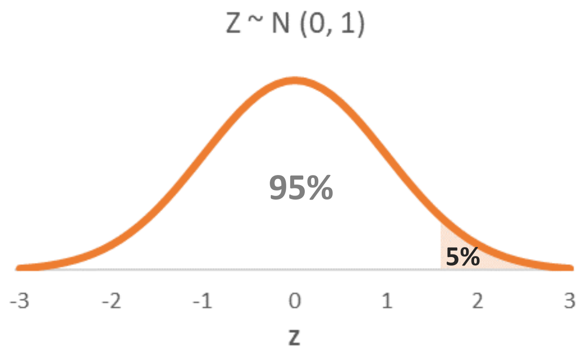
Returns the z-score in the standard normal distribution at a specified cumulative probability

=**NORM.S.INV**(probability)

*The **cumulative probability** for the z-score you want*

Possible question:

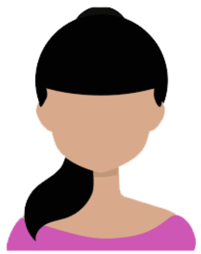
*"The **top 5%** of Olympic Basketball Players are how many standard deviations taller than the mean?"*



=**NORM.S.INV**(1-0.05) **= 1.64** σ

Remember that the cumulative probability starts from negative infinity, so for the "top 5%" the probability is 95% (1-5%)

ASSIGNMENT: ESTIMATING VALUES



NEW MESSAGE
October 8, 2022

From: **Molly Mean** *(Director of Education)*

Subject: **RE: Honor Students**

Hi there, thanks again!

I'll stick with 90 as the threshold for honor students.

Just out of curiosity though… what grade would put students in the top 10% of the class?

And how many standard deviations away from the average student would that be?

Looking forward to hearing back from you.

P.S. You're crushing it!

Reply Forward

Key Objectives

1. Use the **NORM.INV** function to calculate the "MBA Grade" for the top 10%
2. Use the **NORM.S.INV** function to calculate the z-score for the top 10%

KEY TAKEAWAYS: PROBABILITY DISTRIBUTIONS

A probability distribution is an **idealized frequency distribution**

- *It shows all the possible values the variable can take, and the probability of each value occurring*

Many variables naturally follow a **normal distribution**

- *The data is symmetrical around its mean, and flares out in "tails" (the width depends on the standard deviation)*

The probability in a normal distribution is the **area under its curve**

- *It can only be calculated in intervals, not for exact values!*

There are **Excel functions** to solve normal probability problems

- *NORM.DIST and NORM.S.DIST let you calculate the probability of randomly obtaining values in specified ranges*
- *NORM.INV and NORM.S.INV let you estimate values or z-scores based on their cumulative probabilities*

MAVEN MEDICAL CENTER | PROJECT BRIEF

You are a Data Analyst at the **Maven Medical Center** in Springfield, MA and just received a project request from the chief gynecologist

From: **Betty Birth** *(Chief Gynecologist)*

Subject: **Need some probability figures**

Good morning!

We've had over 30% of the babies born this year weigh under 2.5kg, which is considered low. The percentage itself seems a little high to me though. Is there any way you could check what the probability of a baby weighing under 2.5kg is with the data we have?

I could also use the number of births we've had so far in the top & bottom 1% if possible.

Thank you!

Birth_Weights.xlsx

Reply   Forward

Key Objectives

1. Check if the weights can be assumed to follow a normal distribution
2. If so, calculate the probability of a baby weighing 2.5kg or less
3. Estimate the values at the 1% and 99% cumulative probabilities
4. Count the number of births under and over those thresholds

THE CENTRAL LIMIT THEOREM

THE CENTRAL LIMIT THEOREM

In this section we'll cover **the central limit theorem** (CLT), which will allow us to apply the concepts we learned on the normal distribution to populations that follow any distribution

TOPICS WE'LL COVER:

- CLT Basics
- Standard Error
- Implications
- Applications

GOALS FOR THIS SECTION:

- *Understand the concept of a sampling distribution, and its relationship with the central limit theorem*
- *Identify the impact of the sample size on the normality & variability of the sampling distribution*
- *Calculate the standard error of a sampling distribution*
- *Review the implications & applications of the CLT*

SAMPLING DISTRIBUTION OF THE MEAN

The **sampling distribution of the mean** is obtained by taking many samples from a population, calculating the mean for each, and plotting their frequencies

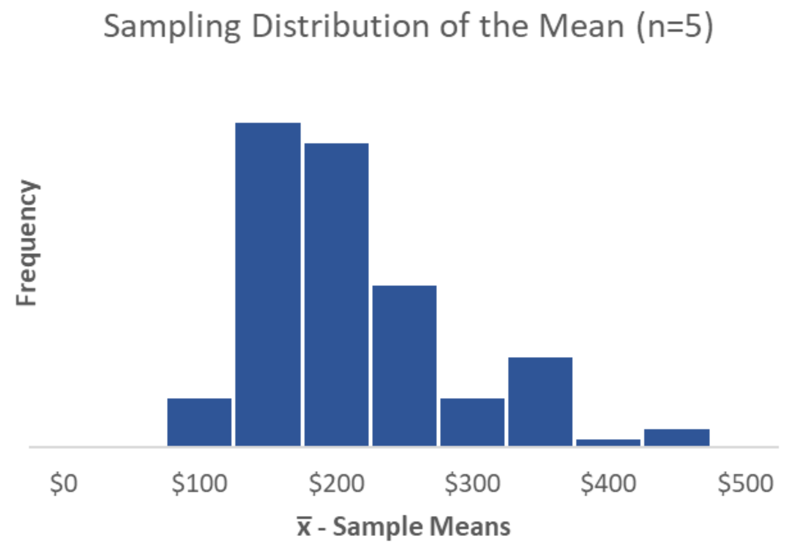
EXAMPLE *Daily Airbnb Rates in New York City*

Population

Price
$150
$95
$156
$259
$250
$59
$330
$85
$125
$285
$95
$171
$60

n>39,000

	Sample (n=5)					*Mean*
1	$60	$305	$45	$272	$201	**$177**
2	$400	$322	$160	$233	$120	**$247**
3	$200	$174	$159	$356	$70	**$192**
4	$297	$50	$60	$55	$91	**$111**
5	$71	$112	$43	$41	$80	**$69**
6	$229	$345	$195	$774	$225	**$354**
7	$150	$90	$314	$80	$190	**$165**
8	$97	$170	$128	$225	$71	**$138**
...						
100	$312	$50	$99	$165	$213	**$168**



THE CENTRAL LIMIT THEOREM

Central Limit Theorem Basics

Standard Error

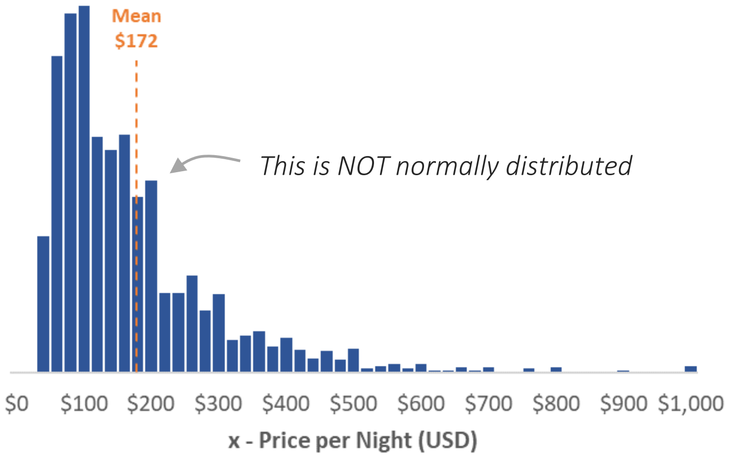
Implications

Applications

The **central limit theorem** states that the means of large enough samples of *any* population will be normally distributed around the population mean

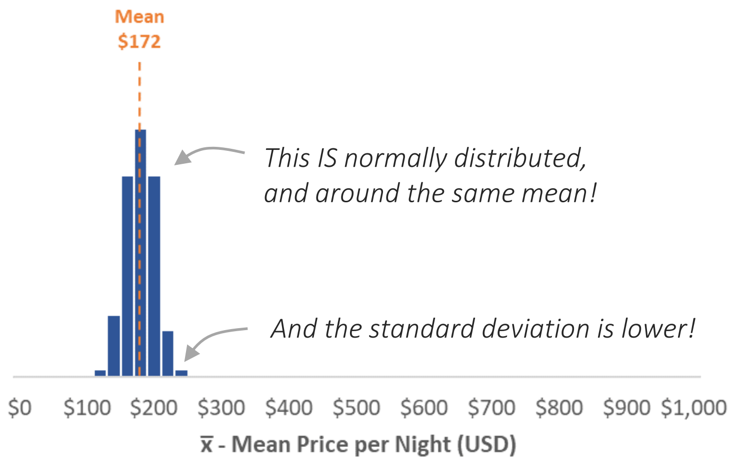
EXAMPLE | *Daily Airbnb Rates in New York City*

POPULATION DISTRIBUTION:



*These are **individual values** (x)*

SAMPLING DISTRIBUTION** (100 samples, n=50)**:



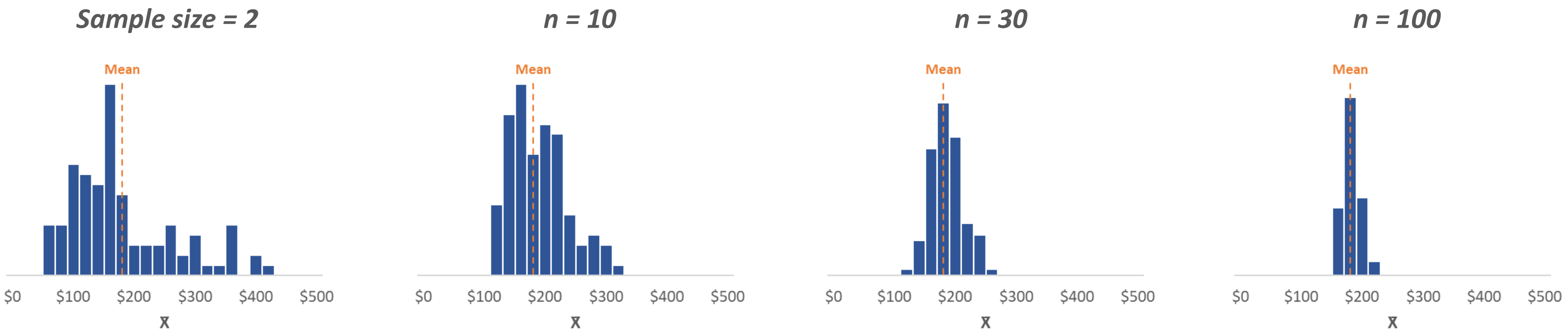
*These are **sample means** ($\bar{x}$)*

THE CENTRAL LIMIT THEOREM

The **central limit theorem** states that the means of large enough samples of *any* population will be normally distributed around the population mean

- A sample size of **30** or more is typically required *(n > 30)*

SAMPLING DISTRIBUTION *(100 samples)***:**



HEY THIS IS IMPORTANT!

As sample size increases, the sampling distribution approximates a normal distribution

STANDARD ERROR

Central Limit Theorem Basics
Standard Error
Implications
Applications

As you know, normal distributions are described by their mean & standard deviation

For the normal distribution of the sample means, the mean is the same as its population's mean, but the standard deviation is known as the **standard error**

- The standard error is the standard deviation of the sample means around the population mean

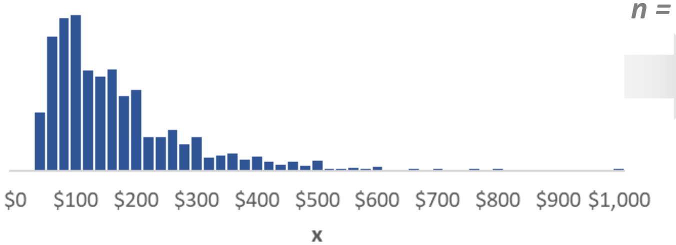
$$SE = \frac{\sigma}{\sqrt{n}}$$

To calculate the standard error, simply divide the standard deviation of the population by the square root of the sample size

HEY THIS IS IMPORTANT!

As sample size increases, the standard error decreases

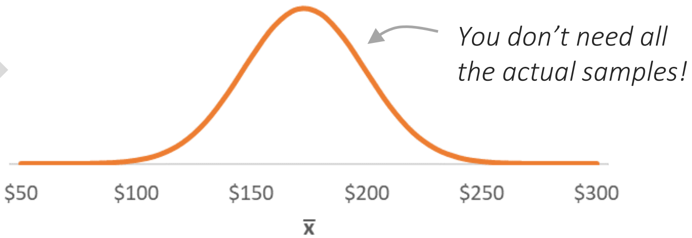
Population of Airbnb Prices in NYC
μ=172, σ=144



n = 30

$$SE = \frac{144}{\sqrt{30}} = \mathbf{26.3}$$

Sampling Distribution of the Mean (n=30)
$\bar{x} \sim N(172, 26.3)$



IMPLICATIONS OF THE CENTRAL LIMIT THEOREM

The central limit theorem has these important **implications**:

1) If you have data on a population (mean & standard deviation), you can make inferences about any sufficiently large sample from that population

2) If you have data on a population and a sufficiently large sample, you can infer whether the sample belongs to that population

3) If you have data on a sufficiently large sample, you can make inferences about the population from which the sample was drawn

4) If you have data on two sufficiently large samples, you can infer whether they belong to the same population

HEY THIS IS IMPORTANT!

This is the basis for **inferential statistics**, which let you come to conclusions about a population from a sample!

APPLICATIONS OF THE CENTRAL LIMIT THEOREM

Central Limit Theorem Basics

Standard Error

Implications

Applications

The central limit theorem has two key **applications** we'll cover:

1. **Making estimates** with confidence intervals
   - For example, you can use the mean & standard deviation from a sample to estimate a range where the population mean likely lies

2. **Drawing conclusions** with hypothesis tests
   - For example, you can use the mean & standard deviation from a sample to conclude whether it was likely drawn from a population with a certain mean

HEY THIS IS IMPORTANT!

This can all be done using the same theory we've learned on the normal distribution!

KEY TAKEAWAYS: THE CENTRAL LIMIT THEOREM

Sample means **are normally distributed** around their population mean, no matter the distribution of the population

- *As the sample size increases, the normality increases (a sample size of at least 30 is required)*

The **standard error** is the standard deviation of the sample means

- *As the sample size increases, the standard error decreases*

The Central Limit Theorem enables **inferential statistics**

- *You can make inferences about unknown populations based on large enough samples!*

CONFIDENCE INTERVALS

MAKING ESTIMATES WITH CONFIDENCE INTERVALS

←μ→

In this section we'll cover making estimates with **confidence intervals**, which use sample statistics to define a range where an unknown population parameter likely lies

TOPICS WE'LL COVER:

- Estimation Basics
- Types of Intervals
- T Distribution
- Proportions
- Two Populations

GOALS FOR THIS SECTION:

- *Understand the main components of a confidence interval, the point estimate & margin of error*
- *Identify the impact of the setting the confidence level on the margin of error*
- *Use the t distribution for confidence intervals when the population standard deviation is unknown*
- *Calculate confidence intervals for the difference in mean and proportions between two populations*

CONFIDENCE INTERVALS

A **confidence interval** is an estimate of an unknown population value using a sample

- It is a range defined by a point estimate, like the sample mean, plus/minus a margin of error
- It includes a confidence level, or probability of including the population value *(can't be certain!)*

Employability (Before)
252
423
101
288
248
145
401
287
275
254
182
117
130
219
152
[illegible]

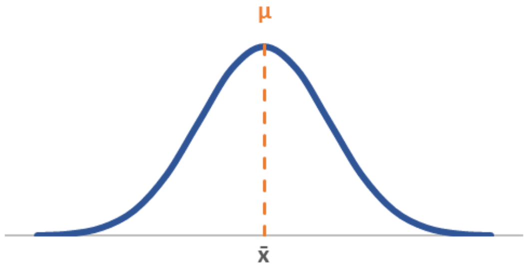
n=95

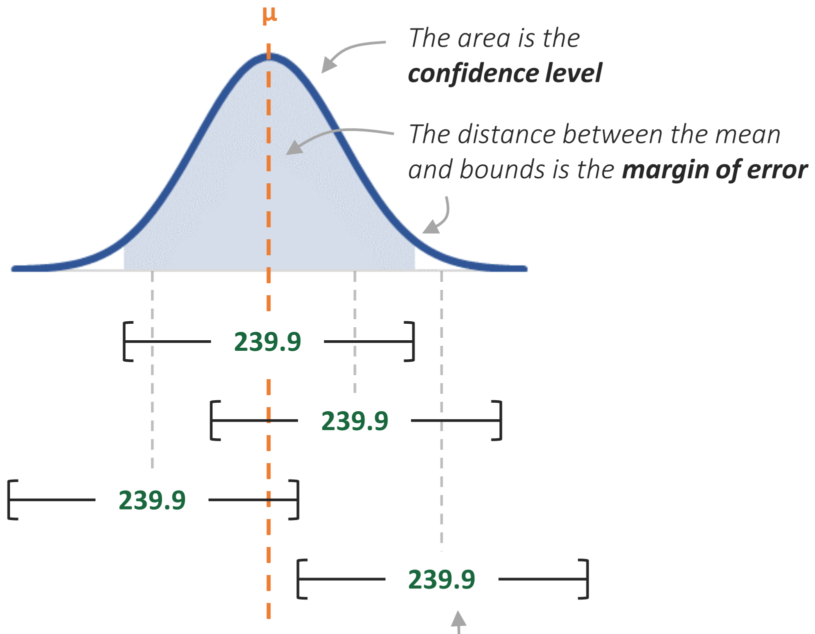
Estimating the population mean:

$$\bar{x} = 239.9$$

$$\mu = ?$$

Remember, the sample means are normally distributed around the population mean



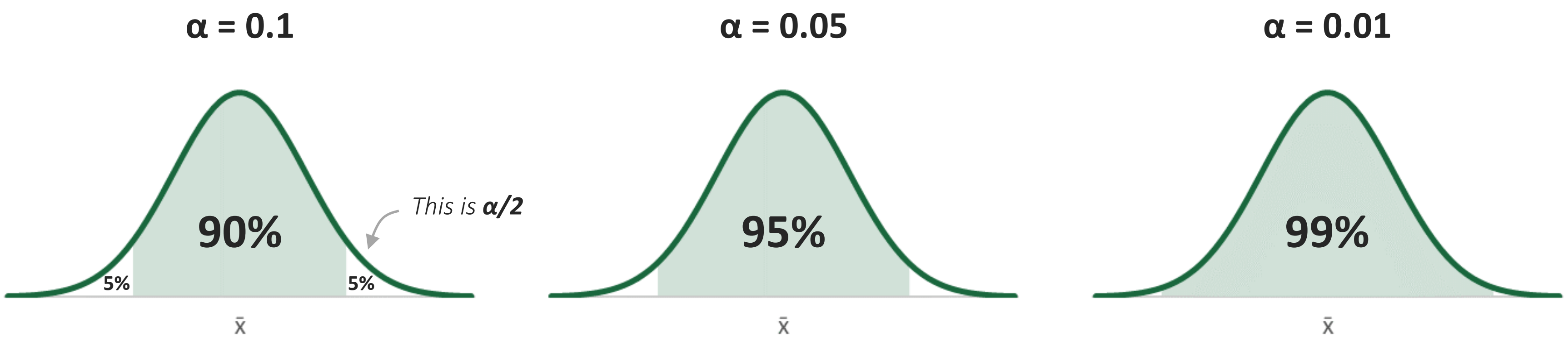


It's possible, but not probable, that the interval won't include the mean!

CONFIDENCE LEVEL

The **confidence level** represents the probability that your confidence interval includes the population parameter

- This is established by **alpha (α)**, which is 1 minus the confidence level
- Typical alpha values are **0.1**, **0.05**, and **0.01**, but you can use any value you'd like!



HEY THIS IS IMPORTANT!

As you increase the confidence level, the confidence interval also increases, so take time to establish an accepted probability of error (α) in favor of a narrower interval

MARGIN OF ERROR

- **Estimation Basics**
- Types of Intervals
- T Distribution
- Proportions
- Two Populations

The **margin of error** represents the value to add to each side of your sample statistic, or point estimate, to generate the confidence interval

- This is determined by the confidence level and the standard error

Employability (Before)
252
423
101
288
248
145
401
287
275
254
182
117
130
219
152
[illegible]

n=95

$$\bar{x} = \mathbf{239.9}$$
$$n = \mathbf{95}$$
$$\sigma = \mathbf{90}$$
$$\alpha = \mathbf{0.05}$$

We want to calculate the margin of error from this sample with a 95% confidence level

NOTE: *We're assuming that we know the standard deviation of the population, which won't always be the case; more on that later*

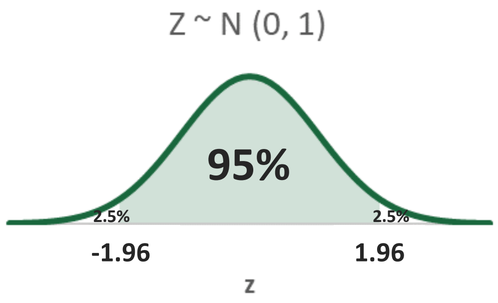
$$CI = 239.9 \pm \boxed{Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}}$$

This is the margin of error!

This is the z-score that the confidence level represents, known as ***critical value***

This is the cumulative probability up to 1-α/2

Z = **NORM.S.INV**(1-0.025) = **1.96** σ



How many standard deviations away from the population mean does my confidence level allow my sample mean to be?

MARGIN OF ERROR

Estimation Basics

Types of Intervals

T Distribution

Proportions

Two Populations

The **margin of error** represents the value to add to each side of your sample statistic, or point estimate, to generate the confidence interval

- This is determined by the confidence level and the standard error

Employability (Before)
252
423
101
288
248
145
401
287
275
254
182
117
130
219
152
228

n=95

$$\underbrace{\bar{x} = \mathbf{239.9} \quad n = \mathbf{95} \quad \sigma = \mathbf{90} \quad \alpha = \mathbf{0.05}}$$

We want to calculate the margin of error from this sample with a 95% confidence level

NOTE: *We're assuming that we know the standard deviation of the population, which won't always be the case; more on that later*

$$CI = 239.9 \pm 1.96 * \boxed{\frac{\sigma}{\sqrt{n}}}$$

This is the ***standard error****, or standard deviation of the sample means*

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{90}{\sqrt{95}} = \mathbf{9.23}$$

$$CI = 239.9 \pm 1.96 * 9.23$$

$$CI = 239.9 \pm \mathbf{18.09}$$

This is the margin of error!

THE CONFIDENCE.NORM FUNCTION

CONFIDENCE.NORM()

Returns the margin of error for a specified confidence level and standard error

=**CONFIDENCE.NORM**(alpha, standard_dev, size)

*The **alpha** (α) for the confidence level*

*The **standard deviation** of the population (σ) and **sample size** (n) for the standard error*

Employability (Before)
252
423
101
288
248
145
401
287
275
254

n=95

$\bar{x} = \mathbf{239.9}$

$n = \mathbf{95}$

$\sigma = \mathbf{90}$

$\alpha = \mathbf{0.05}$

=**CONFIDENCE.NORM**(0.05, 90, 95) = **18.09**

$CI = 239.9 \pm \mathbf{18.09}$

ASSIGNMENT: CONFIDENCE INTERVALS

NEW MESSAGE
October 13, 2022

From: **Nick Normal** *(Head of Student Placement)*

Subject: **RE: Student Salaries**

Hi again,

I know you said our salary data doesn’t follow a normal distribution, but just wanted to try to see if you can produce some sort of expected annual salary from our graduates.

I just read a survey online where they found that the average salary for recent MBA graduates in the US is $101,000.

The standard deviation is $76k, if that means anything to you.

Hope you can come up with something,

Thanks!

Reply Forward

Key Objectives

1. Calculate the **mean** and **sample size** from the sample of graduates
2. Set a **confidence level**
3. Calculate the **margin of error**
4. Set the limits for the **confidence interval**

TYPES OF CONFIDENCE INTERVALS

Estimation Basics

Types of Intervals

T Distribution

Proportions

Two Populations

There are two **types of confidence intervals** for estimating the mean:

1. **Z-intervals**: the population standard deviation (σ) is KNOWN
   - These are uncommon in real life – if you don't know μ, why would you know σ?
   - The population standard deviation is used to calculate the standard error
   - The standard normal distribution *(z distribution)* is used to calculate the critical value

2. **T-intervals** : the population standard deviation (σ) is UNKNOWN
   - These are more realistic – you only have data from the sample
   - The sample standard deviation is used to calculate the standard error
   - The student's t distribution is used to calculate the critical value

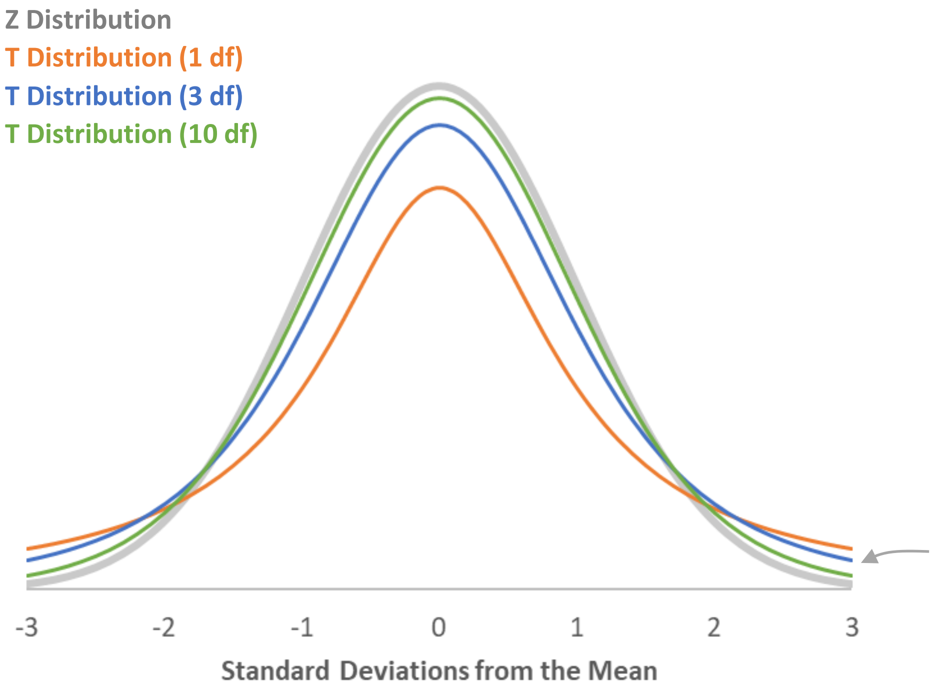
HEY THIS IS IMPORTANT!

Both require the original populations to be assumed normal, or the sample size to be greater than or equal to 30 (so that the central limit theorem applies)

STUDENT'S T DISTRIBUTION

T distributions are like the standard normal distribution, but with "fatter tails"

- They are described by their **degrees of freedom** (sample size – 1)



More data at the tails!
In the Z Distribution, 99.7% of the data lies within 3σ of the mean
In the T Distribution with 1 degree of freedom, only 79.5% does

HEY THIS IS IMPORTANT!

As the degrees of freedom increase, the t distribution approximates the z distribution

They are practically identical for samples of 100 or more observations

EXCEL T DISTRIBUTION FUNCTIONS

These **Excel functions** help make calculations related to the t distribution:

T.DIST()	*Returns the cumulative probability or the probability density at a t-score from a given t distribution*	=**T.DIST**(t, df, cumulative)
T.INV()	*Returns the t-score from a given t distribution at a specified cumulative probability*	=**T.INV**(probability, df)
CONFIDENCE.T()	*Returns the margin of error for a specified confidence level and standard error*	=**CONFIDENCE.T**(α, std_dev, n)

CONFIDENCE INTERVALS WITH THE T DISTRIBUTION

Estimating a **confidence interval with the t distribution** is like the z distribution

- You use the sample standard deviation (s) instead of the population standard deviation (σ)
- You use the t distribution to calculate the critical value instead of the z distribution

$$CI = \bar{x} \pm Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} \quad \rightarrow \quad CI = \bar{x} \pm t_{\alpha/2} * \frac{s}{\sqrt{n}}$$

Employability (Before)
252
423
101
288
248
145
401
287
275
254

n=95

$\bar{x} = \mathbf{239.9}$

$s = \mathbf{85.9}$

$n = \mathbf{95}$

$df = \mathbf{94}$

$\alpha = \mathbf{0.05}$

$$CI = 239.9 \pm \underbrace{t_{\alpha/2}}_{} * 8.81$$

T.INV(1-0.025) **= 1.98** σ

$$CI = 239.9 \pm 1.98 * 8.81$$

$$CI = 239.9 \pm \mathbf{17.5}$$

CONFIDENCE INTERVALS WITH THE T DISTRIBUTION

Estimating a **confidence interval with the t distribution** is like the z distribution

- You use the sample standard deviation (s) instead of the population standard deviation (σ)
- You use the t distribution to calculate the critical value instead of the z distribution

$$CI = \bar{x} \pm Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} \quad \rightarrow \quad CI = \bar{x} \pm t_{\alpha/2} * \frac{s}{\sqrt{n}}$$

Employability (Before)
252
423
101
288
248
145
401
287
275
254

n=95

$\bar{x} = \mathbf{239.9}$

$s = \mathbf{85.9}$

$n = \mathbf{95}$

$df = \mathbf{94}$

$\alpha = \mathbf{0.05}$

$$CI = 239.9 \pm \underbrace{t_{\alpha/2} * \frac{s}{\sqrt{n}}}$$

=**CONFIDENCE.T**(0.05, 85.9, 95) **= 17.5**

$$CI = 239.9 \pm \mathbf{17.5}$$

ASSIGNMENT: CONFIDENCE INTERVALS (T DISTRIBUTION)

NEW MESSAGE
October 14, 2022

From: **Nick Normal** *(Head of Student Placement)*

Subject: **RE: RE: Student Salaries**

Hey,

Thanks for getting me that estimate!

I'm curious though… do we need the data from the study?

I would think that with the amount of our graduates that have been placed already we could an estimate ourselves.

Think you're up for it?

Thanks again!

Reply Forward

Key Objectives

1. Calculate the **standard deviation** from the sample of graduates
2. Set a **confidence level**
3. Calculate the **margin of error**
4. Set the limits for the **confidence interval**

CONFIDENCE INTERVALS FOR PROPORTIONS

A **proportion** is a percentage of a population that meets a certain criteria

You can use the z distribution to calculate **confidence intervals for proportions**

- This is for categorical variables with two possible values *(it meets the criteria, or it doesn't)*

*This is the **sample proportion***

*This is the number of **successful outcomes***

$$\hat{p} = \frac{x}{n}$$

*This is the **sample size***

*This is the **margin of error***

$$CI = \hat{p} \pm \underbrace{Z_{\alpha/2} * \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}}}_{\text{margin of error}}$$

*This is the **critical value** for the confidence level*

*This is the **standard error***

HEY THIS IS IMPORTANT!

Both **p̂*n** and **(1-p̂)*n** must be greater than 5 for the central limit theorem to apply

CONFIDENCE INTERVALS FOR PROPORTIONS

A **proportion** is a percentage of a population that has a certain property

You can use the z distribution to calculate **confidence intervals for proportions**

- This is for categorical variables with two possible values *(it has the property, or it doesn't)*

EXAMPLE *Percentage of Graduates with Previous Work Experience*

Work Experience
No
No
Yes
No
No
No
Yes
Yes
No
No
No

n=95 (23 Yes, 72 No)

$$\hat{p} = \frac{23}{95} = \mathbf{0.242}$$

$$1 - \hat{p} = \mathbf{0.758}$$

$$\alpha = \mathbf{0.1}$$

*We want to calculate the confidence interval from this sample with a **90%** confidence level*

$$CI = 0.242 \pm \underbrace{Z_{\alpha/2}} * \sqrt{\frac{0.242 * (0.758)}{95}}$$

NORM.S.INV(1-0.05) = **1.64** σ

$$CI = 0.242 \pm 1.64 * 0.04$$

$$CI = 0.242 \pm 0.07 = (17\%, 31\%)$$

ASSIGNMENT: CONFIDENCE INTERVALS FOR PROPORTIONS

NEW MESSAGE
October 14, 2022

From: **Nick Normal** *(Head of Student Placement)*

Subject: **Graduate Placement**

Hey again,

Loved the work on the salary data, thank you!

The problem now is getting these students to land jobs.

We've had 53 placed so far, which is 55%. That's not terrible, but I'd hate to be getting numbers under 50% in future classes.

Are you able to get me an estimate with the data we have?

I'd like to be 95% certain this time around.

Thanks

Reply Forward

Key Objectives

1. Calculate the **sample proportion**
2. Check if the **central limit theorem** applies
3. Calculate the **margin of error**
4. Set the limits for the **confidence interval**

CONFIDENCE INTERVALS FOR TWO POPULATIONS

You can create confidence intervals for the **difference between two population means**

There are two possible scenarios for this:

Dependent Samples

*The sample subjects are **directly related** to each other*

Examples:

- People's weight before & after a diet
- Patients' blood pressure before & after taking a certain pill
- Same operators' performance with process A vs. B

$\underbrace{\hspace{6cm}}$

*The measurements come in **pairs***
(both samples are the same size)

Independent Samples

*The sample subjects have **no relationship** to each other*

Examples:

- Students' test scores for two different schools
- Employee satisfaction for in-person vs. remote companies
- Salaries for men and women at the same company

$\underbrace{\hspace{6cm}}$

*The measurements come from **separate groups***
(the samples can be different sizes)

HEY THIS IS IMPORTANT!

You'll see these split up further into population standard deviation (σ) known & unknown, but we'll focus on σ unknown, since it's the most common in real life

DEPENDENT SAMPLES

Estimation Basics

Types of Intervals

T Distribution

Proportions

Two Populations

You can create confidence intervals for **dependent samples** by calculating the difference from each pair in the samples, and then treating the difference as one population

Student ID	Undergrad Grade	MBA Grade	Difference
1	68.4	90.2	21.8
2	62.1	92.8	30.7
3	70.2	68.7	-1.5
4	75.1	80.7	5.6
5	60.9	74.9	14
6	74.5	80.7	6.2
7	76.4	83.3	6.9
8	82.6	88.7	6.1
9	76.9	75.4	-1.5
10	83.3	82.1	-1.2
11	75.8	87.5	11.7
12	76	66.9	-9.1
13	62.8	71.3	8.5
14	82.8	76.8	-6
15	76	72.3	-3.7
16	76.9	72.4	-4.5

n=95

This is the difference in MBA Grade and Undergrad Grade ***for the same student****!*

Confidence interval (σ unknown):

$$CI = \bar{d} \pm t_{\alpha/2} * \frac{s}{\sqrt{n}}$$

95% confidence level

$$CI = 5.19 \pm 1.98 * \frac{9.22}{\sqrt{95}}$$

$$CI = 5.19 \pm 1.87$$

$$CI = (\mathbf{3.3}, \mathbf{7.1})$$

This means that MBA grades are higher on average!

ASSIGNMENT: DEPENDENT SAMPLES

NEW MESSAGE
October 15, 2022

From: **Nick Normal** *(Head of Student Placement)*

Subject: **Employability Scores**

Hi,

It's a shame that we can't be sure that at least 50% of students will be placed 2 months from graduation, so I'm trying to see what factors to dig into deeper.

Looking at the employability scores, it looks like on average our graduates are improving by 50 points on their results.

Can you get me a confidence interval with 90% confidence?

Thanks again!

Reply Forward

Key Objectives

1. Calculate the **difference** between the dependent samples
2. Calculate the sample **mean** and **standard deviation** from the difference
3. Calculate the **margin of error**
4. Set the limits for the **confidence interval**

INDEPENDENT SAMPLES

Estimation Basics

Types of Intervals

T Distribution

Proportions

Two Populations

Confidence intervals for **independent samples** have two key calculation differences:

1. The **standard error** uses the variance (and sample size) from *both* samples
2. The **degrees of freedom** for the t-score (critical value) also consider the sample variances

*This is the **margin of error***

*This is the **point estimate**, or difference in sample means*

$$CI = \boxed{(\bar{x}_1 - \bar{x}_2)} \pm \overbrace{\boxed{t_{\alpha/2}} * \boxed{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}}$$

*This is the **critical value** for the confidence level*

*This is the **standard error***

INDEPENDENT SAMPLES

Confidence intervals for **independent samples** have two key calculation differences:

1. The **standard error** uses the variance (and sample size) from *both* samples
2. The **degrees of freedom** for the t-score (critical value) also consider the sample variances

$t_{\alpha/2}$ =**T.INV**(α/2, degrees of freedom)

For one population, or dependent samples, this is ***n-1***

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

This looks scary, but it's using values we know how to calculate!

- *The sample variances*
- *The sample sizes*

HEY THIS IS IMPORTANT!

You might see independent samples divided into separate calculations if the population standard deviations can or can't be assumed to be equal, but this formula works for both!

ASSIGNMENT: INDEPENDENT SAMPLES

NEW MESSAGE
October 16, 2022

From: **Nick Normal** *(Head of Student Placement)*

Subject: **RE: Employability Scores**

Hey!

Thanks for the data on the employability improvement, I'm going to send that over to Tommy in Admissions so he can factor that into his process.

Final though though… can we find a positive difference in the employability scores for graduates that have been placed so far vs. those that haven't? That could be a good indicator for me.

I think sticking with the same 90% confidence should work.

Thanks in advance!

Reply Forward

Key Objectives

1. Calculate the **mean** and **variance** from both samples
2. Calculate the **point estimate**, or difference in sample means
3. Calculate the **margin of error**
4. Set the limits for the **confidence interval**

PRO TIP: DIFFERENCE BETWEEN PROPORTIONS

Estimation Basics

Types of Intervals

T Distribution

Proportions

Two Populations

You can also calculate confidence intervals for **difference in population proportions**

*This is the **margin of error***

*This is the **point estimate**, or difference in sample proportions*

$$CI = (\hat{p}_1 - \hat{p}_2) \pm \underbrace{Z_{\alpha/2} * \sqrt{\frac{\hat{p}_1 * (1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2 * (1-\hat{p}_2)}{n_2}}}_{\text{margin of error}}$$

*This is the **critical value** for the confidence level*

*This is the **standard error***

KEY TAKEAWAYS: CONFIDENCE INTERVALS

Confidence intervals use samples to **estimate population values**

- *The estimate is tied to a confidence level, which is the probability that the interval includes the population value*

The interval size is based on a **critical value** and a **standard error**

- *The critical value defines the number of standard deviations the sample mean can be from the population mean*
- *The standard error is the standard deviation of the sample means*

Use the t distribution when **σ is unknown**

- *In most real-life scenarios, you won't know the standard deviation of the population*

The same concepts apply when comparing **two populations**

- *Dependent samples can be converted into a single population*
- *Independent samples simply have different calculations for the critical value and standard error*

MAVEN PHARMA | PROJECT BRIEF

You are the Lead Statistician at the **Maven Pharma**, a pharmaceutical company that is in the final testing stage for a new drug to treat arthritis

From: **Patty Pill** *(Head of R&D)*

Subject: **Treatment results**

Hello!

We have the data from the trial on the new arthritis treatment we're developing. We had 84 subjects for the trial, 41 which did take the medication and 43 "placebos" which didn't.

Can you use a 99% confidence interval to see if the percentage of patients with "Marked" improvements is significantly higher for those that took the treatment (vs. the placebos)?

Thank you!

Treatment_Results.xlsx

Reply Forward

Key Objectives

1. Check if the central limit theorem applies
2. Estimate the difference in population proportions with a 99% confidence level
3. Reach a conclusion from the results

HYPOTHESIS TESTS

DRAWING CONCLUSIONS WITH HYPOTHESIS TESTS

In this section we'll cover drawing conclusions with **hypothesis tests**, which let you evaluate assumptions about population parameters based on sample statistics

TOPICS WE'LL COVER:

- Hypothesis Testing
- Types of Errors
- Types of Tests
- Proportions
- Two Populations

GOALS FOR THIS SECTION:

- *Understand the concepts of a null and alternative hypothesis, and how to frame them correctly*
- *Perform hypothesis tests for the mean & proportions for one and two populations*
- *Review the two types of errors in a hypothesis test, and how you can influence them in their design*
- *Draw the correct conclusions from hypothesis tests*

HYPOTHESIS TESTS

A **hypothesis test** lets you evaluate how well a sample supports an assumption

- More specifically, it is a process of evaluating whether a sample provides clear enough evidence that an initial assumption about a population was wrong

Steps for a hypothesis test:

1) State your assumption
2) Define an accepted probability of error
3) Check how well the data supports your assumption
4) Translate that into a probability that it supports it
5) Is it worse than your accepted probability of error?
   a) Yes – your assumption was wrong!
   b) No – your assumption was right!*

HEY THIS IS IMPORTANT!

Remember when we said statistics let you evaluate decisions under uncertain circumstances?

The hypothesis test is the tool that accomplishes that!

**This isn't technically accurate*

HYPOTHESIS TESTS

A **hypothesis test** lets you evaluate how well a sample supports an assumption

- More specifically, it is a process of evaluating whether a sample provides clear enough evidence that an initial assumption about a population was wrong

Steps for a hypothesis test:

1) State the **null** and **alternative hypotheses**
2) Set a **significance level**
3) Calculate the **test statistic** for the sample
4) Calculate the **p-value**
5) Draw a **conclusion** from the test
   a) Reject the null hypothesis
   b) Fail to reject the null hypothesis

HEY THIS IS IMPORTANT!

Remember when we said statistics let you evaluate decisions under uncertain circumstances?

The hypothesis test is the tool that accomplishes that!

NULL & ALTERNATIVE HYPOTHESIS

The **null hypothesis (H_o)** is the assumption about a population you'd like to evaluate

The **alternative hypothesis (H_a)** is any scenario in which that assumption is wrong

- The null hypothesis should be tied to a decision you'd be most comfortable making *(the "status quo")*
- That way, if the test "proves" the null hypothesis wrong, you'll be more comfortable NOT making it

EXAMPLE *Evaluating the need for a new soda filling machine*

H_o	$\mu = 355$	*(our machine fills each can with 355ml on average – we don't need a new one)*
H_a	$\mu \neq 355$	*(our machine doesn't fill each can with 355ml on average – we need a new one)*

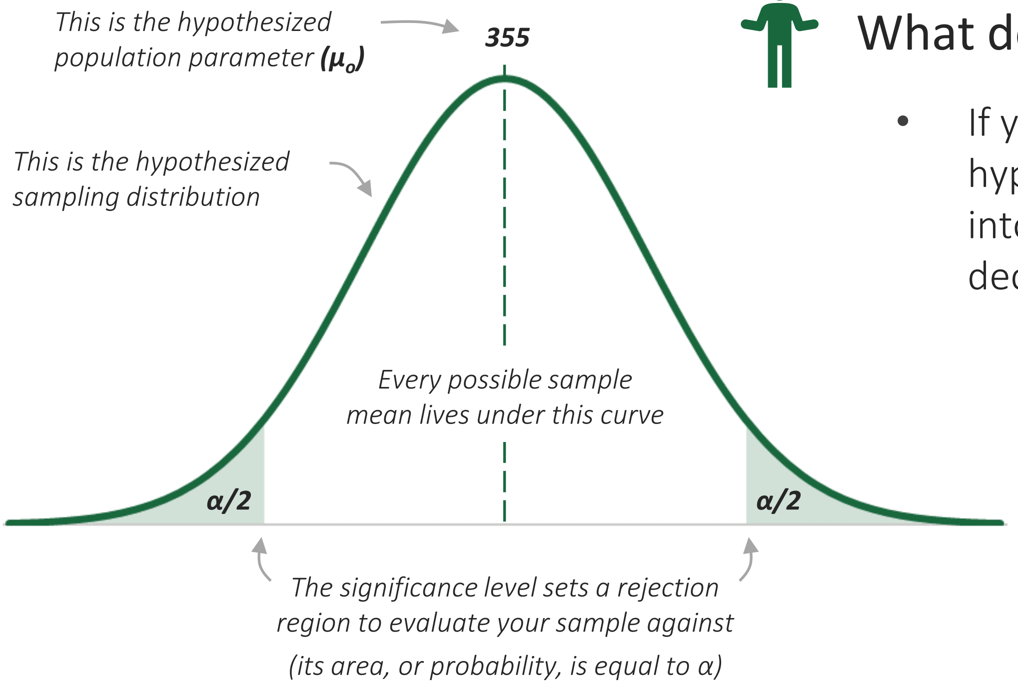
HEY THIS IS IMPORTANT!

You're not proving either of the hypotheses right, you're only testing to see if the sample data makes the null hypothesis look wrong enough to make you feel comfortable taking the alternative action

SIGNIFICANCE LEVEL

The **significance level** is the threshold you set to determine when the evidence against your null hypothesis is considered “strong enough” to prove it wrong

- This is set by **alpha (α)**, which is the accepted probability of error



What does this mean?

- If your sample statistic is so far from your hypothesized population parameter that it falls into the rejection region, then you’re comfortable declaring the null hypothesis as wrong

HEY THIS IS IMPORTANT!

Set the significance level before knowing where your sample lies in the distribution (otherwise what’s the point?)

TEST STATISTIC

Hypothesis Testing

Types of Errors

Types of Tests

Proportions

Two Populations

The **test statistic** is your sample's t-score from the hypothesized sampling distribution

- It's the standard deviations between what your sample is and what you're saying it should be

*This is the **test statistic** for your sample*

$$t = \frac{\bar{x} - \mu_o}{\frac{s}{\sqrt{n}}}$$

This is the difference between your sample mean and your hypothesized population mean

*This is the **standard error**, or standard deviation of the sample means*

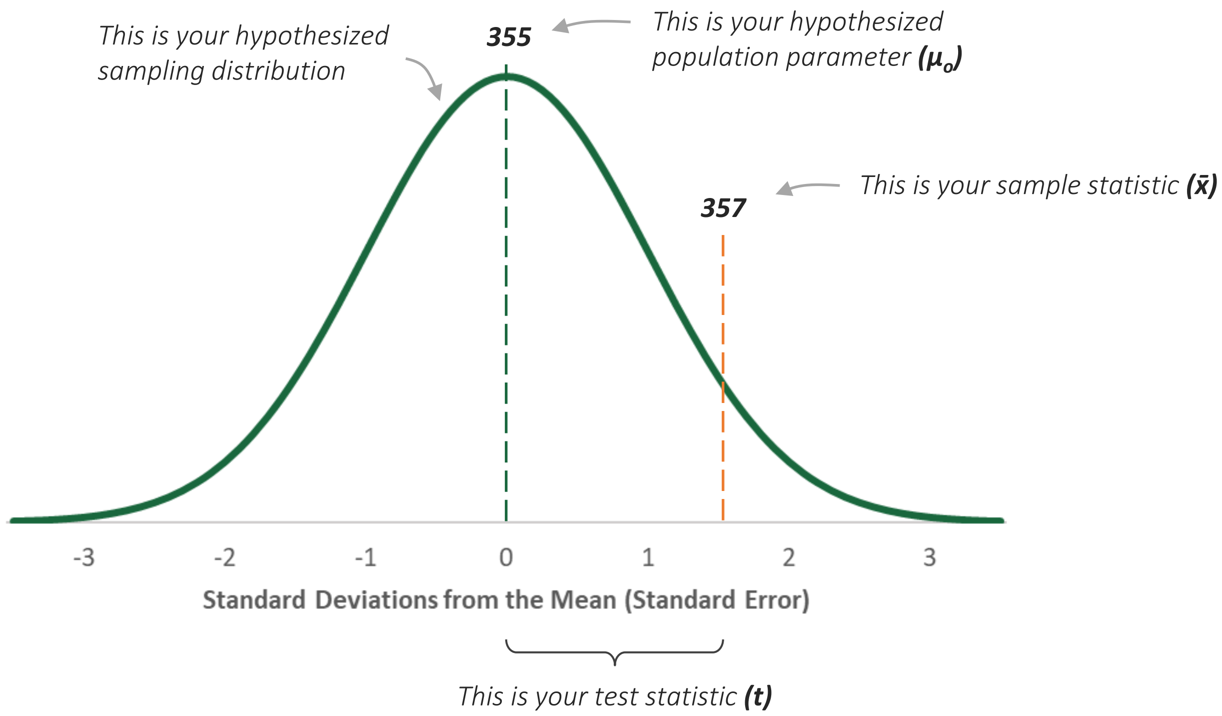
HEY THIS IS IMPORTANT!

This is assuming that the population standard deviation (σ) is unknown, since it's more common, but you can use the z-score if it is known and swap out "s" for "σ"

TEST STATISTIC

The **test statistic** is your sample's t-score from the hypothesized sampling distribution

- It's the standard deviations between what your sample is and what you're saying it should be



P-VALUE

Hypothesis Testing

Types of Errors

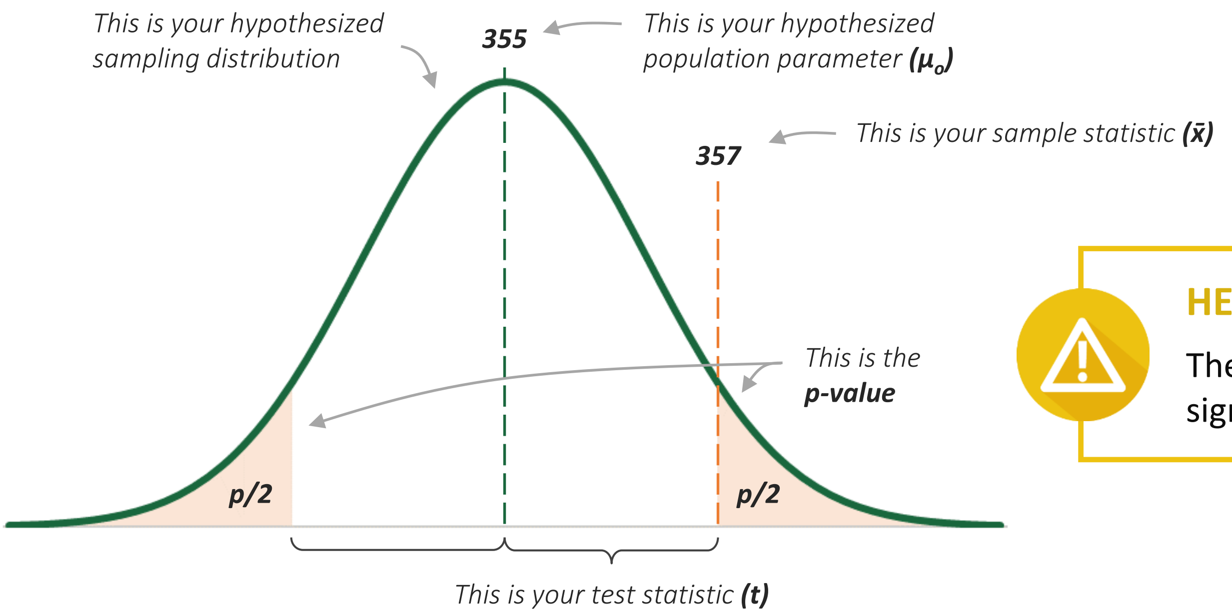
Types of Tests

Proportions

Two Populations

The **p-value** is the probability that your sample supports the null hypothesis

More specifically, it's the probability of obtaining a test statistic at least as large as the one from your sample *(negative or positive)* if your null hypothesis is true



HEY THIS IS IMPORTANT!

The p-value in itself is meaningless, it's the significance level that puts it into context!

CONCLUSIONS

Hypothesis Testing

Types of Errors

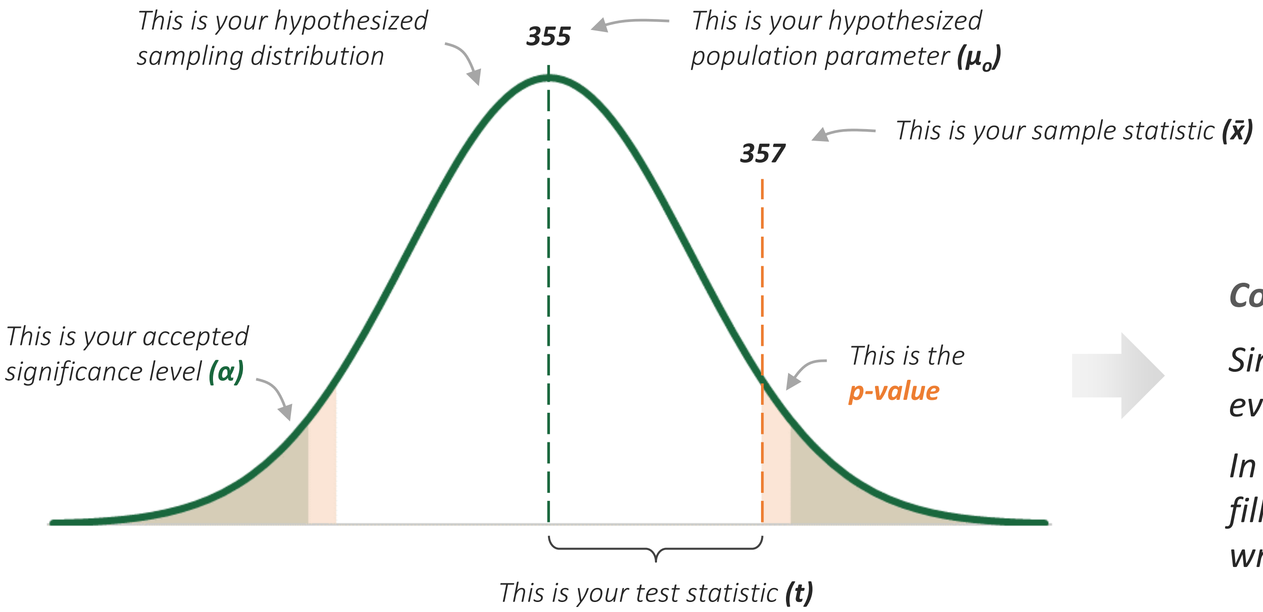
Types of Tests

Proportions

Two Populations

There are two possible **conclusions** in a hypothesis test:

- If **p>α**, then you **fail to reject** the null hypothesis *(not enough evidence!)*
- If **p≤α**, then you **reject** the null hypothesis *(strong enough evidence!)*



Conclusion:

Since p>α, we don't have sufficient evidence to reject our null hypothesis

In other words, assuming the machine fills 355ml on average doesn't seem wrong – let's keep using it!

CONCLUSIONS

Hypothesis Testing

Types of Errors

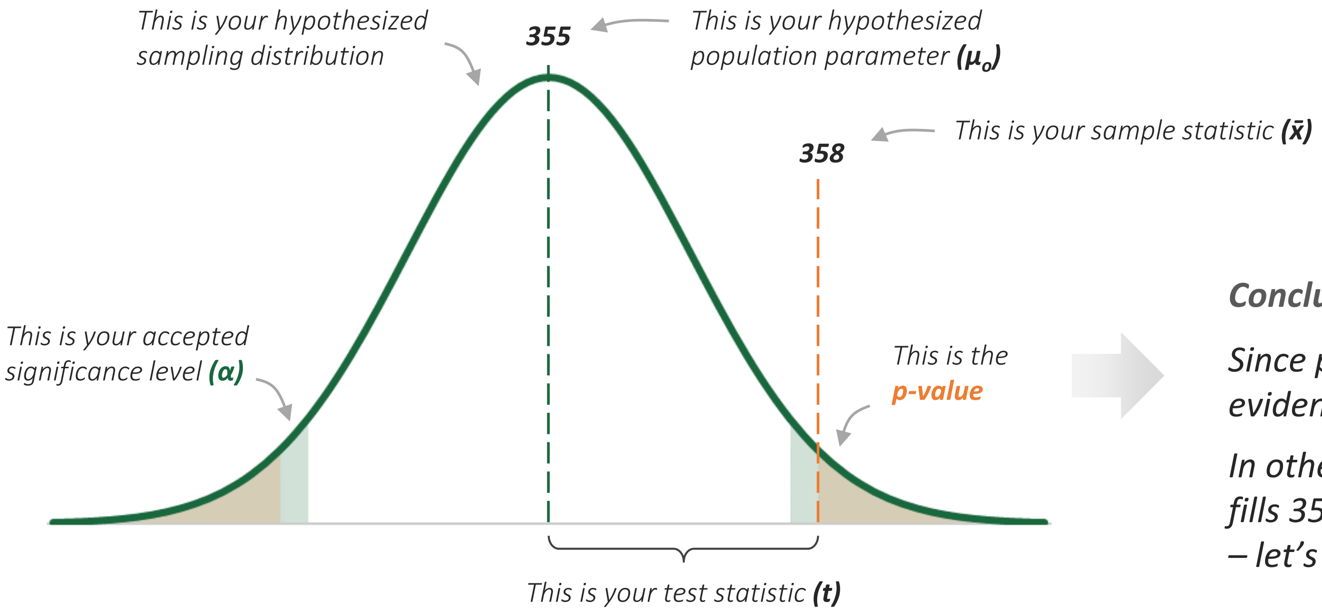
Types of Tests

Proportions

Two Populations

There are two possible **conclusions** in a hypothesis test:

- If **p>α**, then you **fail to reject** the null hypothesis *(not enough evidence!)*
- If **p≤α**, then you **reject** the null hypothesis *(strong enough evidence!)*

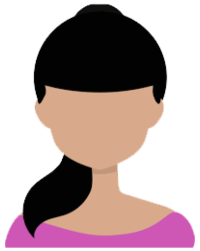


Conclusion:

Since p≤α, we do have sufficient evidence to reject our null hypothesis

In other words, assuming the machine fills 355ml on average seems wrong – let's buy a new one!

ASSIGNMENT: HYPOTHESIS TESTS



NEW MESSAGE
October 18, 2022

From: **Molly Mean** *(Director of Education)*

Subject: **Curriculum Planning**

Hi again!

We planned the “difficulty” of our curriculum so that our students would graduate with an average grade of 80.

It looks like that was the case this time around, we had an average of 80.2, but I don’t want to leave it to random chance.

I’d say that if there’s less than a 20% chance of 80 being the real average with the current curriculum, we need to make some modifications to it.

Thank you!

Reply   Forward

Key Objectives

1. State the **null** & **alternative hypotheses**
2. Set a **significance level**
3. Calculate the **test statistic** for the sample
4. Calculate the **p-value**
5. Draw a **conclusion** from the test

RELATIONSHIP WITH CONFIDENCE INTERVALS

Hypothesis tests have a **direct relationship with confidence intervals**

- If you use the same alpha (α), a confidence interval WILL include the hypothesized population mean when failing to reject the null hypothesis, and WON'T include it when rejecting it

Employability (Before)
252
423
101
288
248
145
401
287
275
254
182
117
130
219
152
[illegible]

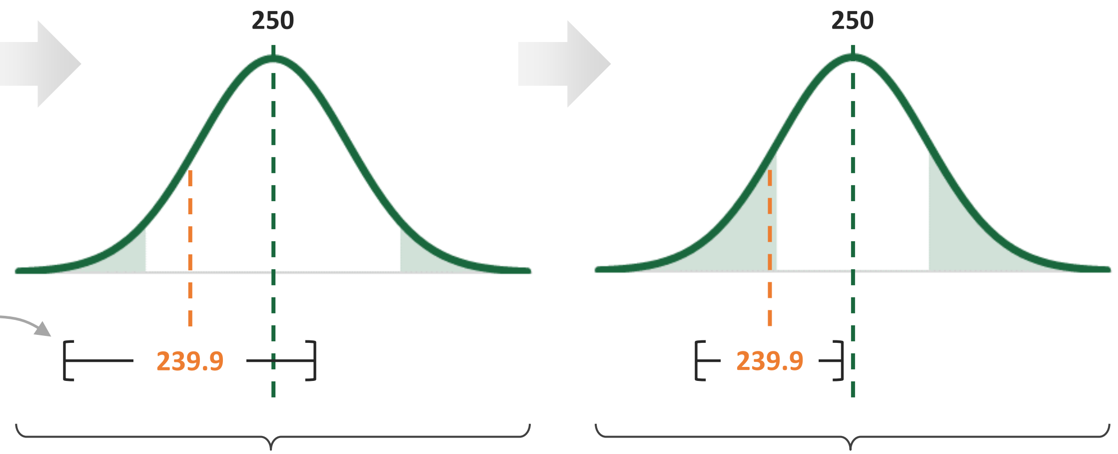
n=95

H_o: $\mu = 250$

H_a: $\mu \neq 250$

*Remember that the size of the interval is **1-α***

$\alpha = 0.1$ $\qquad$ $\alpha = 0.3$



*The sample does not cross into the rejection region, so we **fail to reject** the null hypothesis*

*The confidence interval **includes** the hypothesized population mean*

*The sample does cross into the rejection region, so we **reject** the null hypothesis*

*The confidence interval **doesn't include** the hypothesized population mean*

TYPE I & TYPE II ERRORS

There are two errors you can make in hypothesis tests: **Type I** & **Type II** errors

1) **Type I**: rejecting a true null hypothesis
2) **Type II**: failing to reject a false null hypothesis

Null hypothesis is…	**True**	**False**
Rejected	**Type I Error**	Correct Conclusion
Not Rejected	Correct Conclusion	**Type II Error**

HEY THIS IS IMPORTANT!

The significance level (α) is the probability of making a **type I error**, so the lower it is the less likely you are to make it – but the more likely you are to make a type II error!

TYPE I & TYPE II ERRORS

Hypothesis Testing

Types of Errors

Types of Tests

Proportions

Two Populations

There are two errors you can make in hypothesis tests: **Type I** & **Type II** errors

1) **Type I**: rejecting a true null hypothesis
2) **Type II**: failing to reject a false null hypothesis

EXAMPLE *Evaluating the need for a new soda filling machine*

H_o	*The machine works as expected*
H_a	*The machine doesn't work as expected*

What type of error is worse?

I. Buying a new machine when you didn't need one
II. Not buying a new machine when you needed it

TYPE I & TYPE II ERRORS

Hypothesis Testing

Types of Errors

Types of Tests

Proportions

Two Populations

There are two errors you can make in hypothesis tests: **Type I** & **Type II** errors

1) **Type I**: rejecting a true null hypothesis
2) **Type II**: failing to reject a false null hypothesis

EXAMPLE *Evaluating if an email you received is spam*

H_o	*The email isn't spam*
H_a	*The email is spam*

What type of error is worse?

I. Screening an email that isn't spam
II. Getting a spam email in your inbox

HEY THIS IS IMPORTANT!

You'll never know what the right choice is with 100% certainty *(that's statistics!)*, so it's critical to think about how comfortable you are in making a type I or II error when setting the significance level

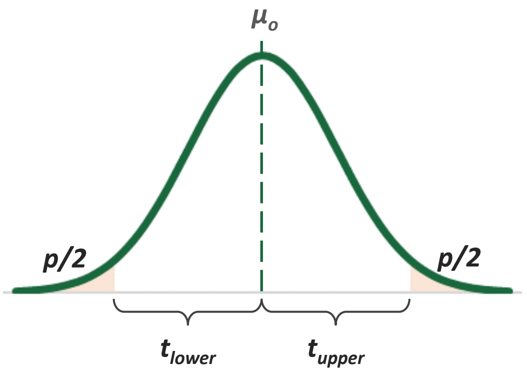
TYPES OF HYPOTHESIS TESTS

- Hypothesis Testing
- Types of Errors
- **Types of Tests**
- Proportions
- Two Populations

There are 3 **types of hypothesis tests** that you can make:

Two Tail

$H_o: \mu = \mu_o$
$H_a: \mu \neq \mu_o$



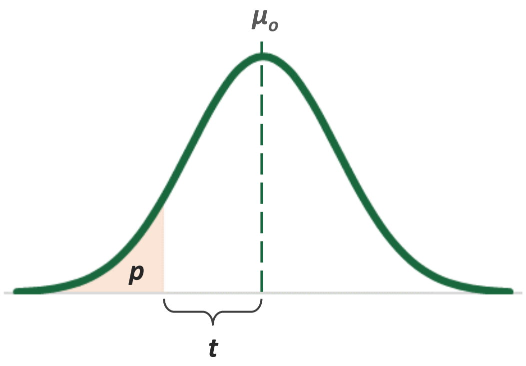
Excel p-value formulas:

=**T.DIST**(t_{lower}, df, TRUE)*2

=**T.DIST.2T**(t_{upper}, df)

One Tail to the Left

$H_o: \mu \geq \mu_o$
$H_a: \mu < \mu_o$

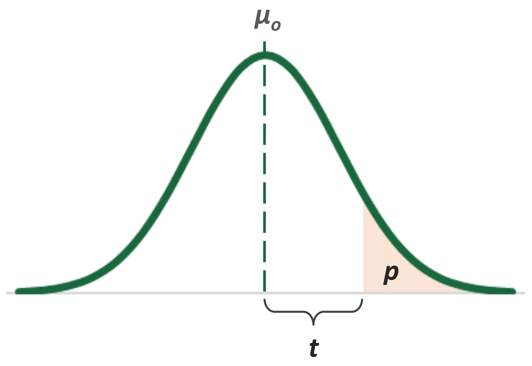


Excel p-value formulas:

=**T.DIST**(t, df, TRUE)

One Tail to the Right

$H_o: \mu \leq \mu_o$
$H_a: \mu > \mu_o$



Excel p-value formulas:

=1-**T.DIST**(t, df, TRUE)

=**T.DIST.RT**(t, df)

HYPOTHESIS TESTS FOR PROPORTIONS

- Hypothesis Testing
- Types of Errors
- Types of Tests
- **Proportions**
- Two Populations

You can use the z distribution to calculate **hypothesis tests for proportions**

- The only thing that changes is the standard error calculation in the test statistic

*This is the **test statistic** for your sample*

$$Z = \frac{\hat{p} - p_o}{\sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}}}$$

This is the difference between the sample proportion and the hypothesized population proportion

*This is the **standard error**, or standard deviation of the sample proportions*

HEY THIS IS IMPORTANT!

Both **p̂*n** and **(1-p̂)*n** must be greater than 5 for the central limit theorem to apply

ASSIGNMENT: PROPORTIONS

NEW MESSAGE
October 25, 2022

From: **Nick Normal** *(Head of Student Placement)*

Subject: **RE: RE: Student Salaries**

Hi again,

I keep thinking about the confidence interval for our graduate salaries you sent me estimating the mean to be between $111,000 and $127,000.

Are you able to check if more than half of our placed graduates earn at least $100,000?

That could be a huge promotional piece to publish!

I think a 5% risk of publishing if it turns out to be false is fine.

Looking forward to hearing from you!

Reply   Forward

Key Objectives

1. Calculate the **proportion** of placed graduates that earn at least $100,000
2. Check if the **central limit theorem** applies
3. Select the right **type of hypothesis test**
4. State the **null** & **alternative hypotheses**
5. Set the **significance level**
6. Calculate the **test statistic** for the sample
7. Calculate the **p-value**
8. Draw a **conclusion** from the test

DEPENDENT SAMPLES

You can make hypothesis tests for **dependent samples** by calculating the difference from each pair in the samples, and then treating the difference as one population

Student ID	Undergrad Grade	MBA Grade	Difference
1	68.4	90.2	21.8
2	62.1	92.8	30.7
3	70.2	68.7	-1.5
4	75.1	80.7	5.6
5	60.9	74.9	14
6	74.5	80.7	6.2
7	76.4	83.3	6.9
8	82.6	88.7	6.1
9	76.9	75.4	-1.5
10	83.3	82.1	-1.2
11	75.8	87.5	11.7
12	76	66.9	-9.1
13	62.8	71.3	8.5
14	82.8	76.8	-6
15	76	72.3	-3.7
16	76.9	72.4	-4.5

n=95

This is the difference in MBA Grade and Undergrad Grade ***for the same student****!*

Do a normal hypothesis test for the mean!

PRO TIP: Do a two tail test with **Ho: μ=0** if you want to check if there is any significant difference in the means of the samples, or a one tail test if you want to check if one is greater than the other

ASSIGNMENT: DEPENDENT SAMPLES

NEW MESSAGE
October 26, 2022

From: **Tommy Test** *(Head of Admissions)*

Subject: **Employability Improvements**

Hi,

I spoke with Nick, and he mentioned that he's confident we can build in a "50-point improvement" on employability scores into our recruitment process, and I just want to double check that it's not an incorrect assumption to make.

Do you think you could run a quick test?

Honestly, it's not a HUGE deal so unless you're really confident that's not the case, I'll just stick to his number.

Thanks – and nice to finally speak to you!

Reply   Forward

Key Objectives

1. Calculate the **difference** between the dependent samples
2. Calculate the sample **mean** and **standard deviation** from the difference
3. State the **null** & **alternative hypotheses**
4. Set the **significance level**
5. Calculate the **test statistic** for the sample
6. Calculate the **p-value**
7. Draw a **conclusion** from the test

INDEPENDENT SAMPLES

Hypothesis tests for **independent samples** have two key calculation differences:

1. The **standard error** in the test statistic uses the variance (and sample size) from *both* samples
2. The **degrees of freedom** for the p-value also consider the sample variances

This is the difference between the sample means

*This is the **test statistic** for your sample*

This is the hypothesized difference in population means

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

*This is the **standard error**, or standard deviation of the sample means*

INDEPENDENT SAMPLES

Hypothesis tests for **independent samples** have two key calculation differences:

1. The **standard error** in the test statistic uses the variance (and sample size) from *both* samples
2. The **degrees of freedom** for the p-value also consider the sample variances

p-value =**T.DIST**(t, degrees of freedom, 1)

*For one population, or dependent samples, the degrees of freedom are **n-1***

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Remember that we already used this for the confidence intervals of independent samples!

ASSIGNMENT: INDEPENDENT SAMPLES

NEW MESSAGE
October 27, 2022

From: **Tommy Test** *(Head of Admissions)*

Subject: **Prior Work Experience**

Hi again,

I took a quick look at the salary data for our first batch of graduates, and it looks like those with previous work experience are earning a bit more on average.

Can we assume that this will always be the case? If so, we may have to start screening some applicants based on this.

I don't want to potentially impact our student numbers on a hunch though, so lets only take a 1% risk.

Thanks!

Reply Forward

Key Objectives

1. Calculate the **mean** and **variance** from both samples
2. Calculate the **difference in sample means**
3. State the **null** & **alternative hypotheses**
4. Set the **significance level**
5. Calculate the **test statistic** for the sample
6. Calculate the **degrees of freedom**
7. Calculate the **p-value**
8. Draw a **conclusion** from the test

KEY TAKEAWAYS: HYPOTHESIS TESTS

Hypothesis tests let you **evaluate assumptions** about a population

- *The null hypothesis is the assumption to evaluate, and the alternative hypothesis is any other possibility*
- *You're not looking to confirm this assumption (it's already the status quo), just testing to see if it's wrong*

The **significance level** sets the threshold for "sufficient" evidence

- *It draws a "probability line" that says, if a sample is at least this improbable, then the assumption is wrong*
- *The lower the significance level, the lower the chance of a Type I error, but the higher the chance of a Type II*

The **p-value** is the probability of the sample fitting the assumption

- *If it's greater than the significance level, the you don't have sufficient evidence to reject the assumption*
- *If it's less than the significance level, then you reject the assumption (null hypothesis)*

MAVEN SAFETY COUNCIL | PROJECT BRIEF

You are a freelance Data Scientist working on a project for the **Maven Safety Council**, an initiative looking to educate the public on safe driving practices

From: **Stuart Stop** *(Council President)*

Subject: **Warning Sign Results**

Hi!

We put up a sign asking drivers to slow down and warning of the dangers of speeding and took three sets of measurements. We recorded the speed of 100 cars before putting up the sign, 100 more shortly after putting up the sign, and a final 100 after the sign had been in place for a longer period.

Could you check if the sign significantly reduced the average speed?

Thank you!

Car_Speeds.xlsx

Reply Forward

Key Objectives

1. Identify the type of test needed
2. Perform the hypothesis test
3. Draw a conclusion from the results

REGRESSION ANALYSIS

MAKING PREDICTIONS WITH REGRESSION ANALYSIS

In this section we'll cover making predictions with **regression analysis**, which helps estimate the values of a dependent variable by leveraging its relationship with independent variables

TOPICS WE'LL COVER:

- Linear Relationships
- Regression Basics
- Model Evaluation
- Multiple Regression

GOALS FOR THIS SECTION:

- *Identify linear relationships between variables*
- *Understand the difference between correlation and causation, and its implications on regression analysis*
- *Create linear regression models in Excel and use them to make predictions for dependent variables*
- *Evaluate the accuracy of linear regression models*

LINEAR RELATIONSHIPS

Linear Relationships

Regression Basics

Model Evaluation

Multiple Linear Regression

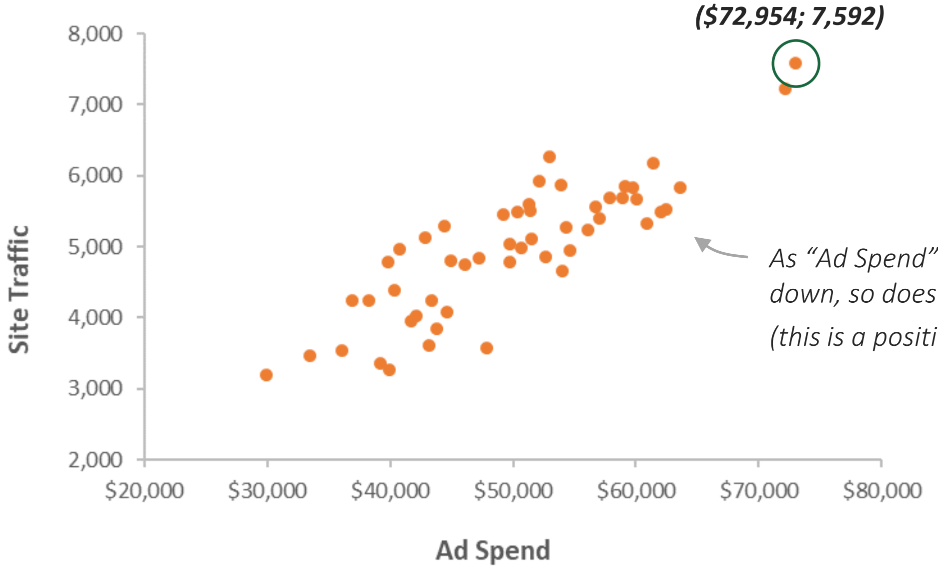
It's common for numerical variables to have **linear relationships** between them

- When one variable changes, so does the other *(they co-variate!)*
- This relationship is commonly visualized with a scatterplot

Week	Ad Spend	Site Traffic
1	$46,125	4,751
2	$60,007	5,661
3	$50,314	5,491
4	$44,432	5,293
5	$72,954	7,592
6	$47,288	4,835
7	$40,830	4,962
8	$43,760	3,850
9	$62,487	5,517
10	$33,480	3,456
11	$59,110	5,851
12	$72,150	7,225
13	$56,740	5,565
14	$42,106	4,033
15	$42,857	5,129

n=52

SCATTERPLOT:



As "Ad Spend" moves up or down, so does "Site Traffic"

(this is a positive relationship)

LINEAR RELATIONSHIPS

There are two possible linear relationships: **positive** & **negative**

- Variables can also have **no relationship**

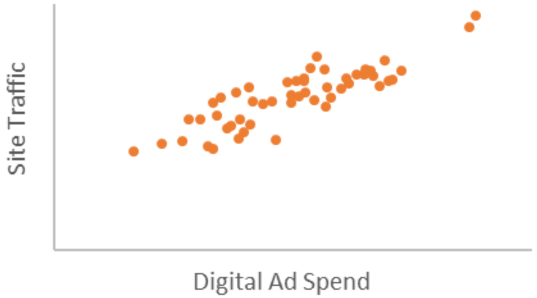
Linear Relationships

Regression Basics

Model Evaluation

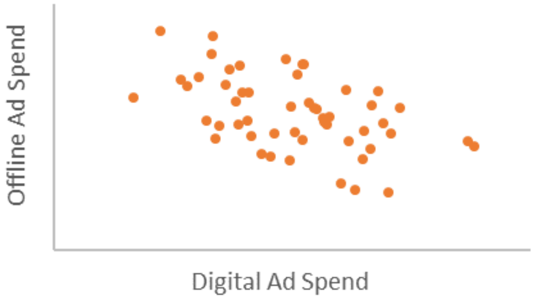
Multiple Linear Regression

Positive Relationship



As one changes, the other changes in the ***same direction***

Negative Relationship



As one changes, the other changes in the ***opposite direction***

No Relationship

No association can be found between the changes in one variable and the other

CORRELATION

Linear Relationships
Regression Basics
Model Evaluation
Multiple Linear Regression

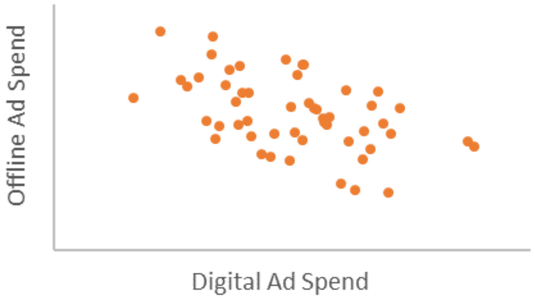
The **correlation (r)** measures the strength & direction of a linear relationship *(-1 to 1)*

- **-1** is a perfect negative correlation, **0** is no correlation, and **1** is a perfect positive correlation

r = 0.858

Strong positive correlation

r = -0.499



Moderate negative correlation

r = 0.008

No correlation

PRO TIP: Use **CORREL()** or **PEARSON()** to calculate the correlation between variables in Excel

CORRELATION & CAUSATION

Linear Relationships

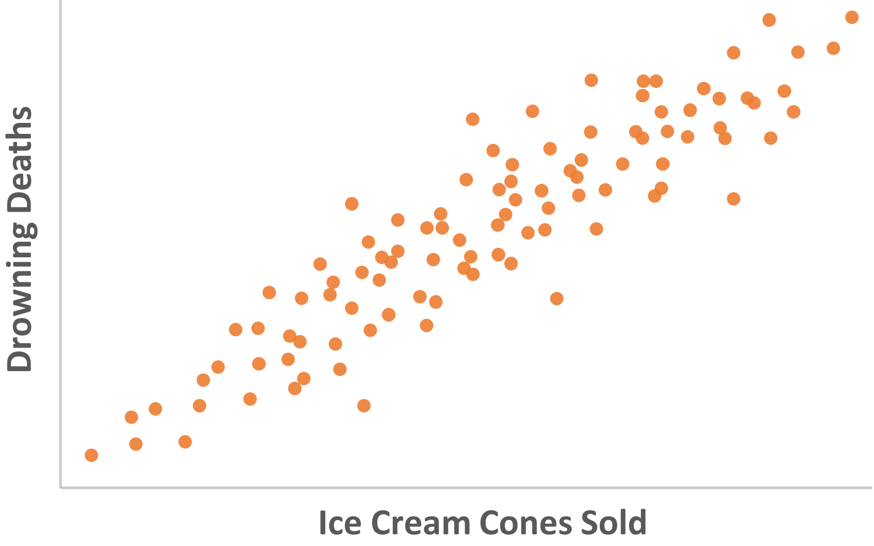
Regression Basics

Model Evaluation

Multiple Linear Regression

Correlation between variables means that there is a relationship between them

Causation means that changes in one variable *cause* the other one to change



These two variables are clearly correlated, but…

Do ice cream cones CAUSE people to drown?
Do drowning deaths CAUSE a surge in ice cream sales?

Of course not!

HEY THIS IS IMPORTANT!

Correlation does NOT imply causation!

Keep in mind that variables can be related without causing a change in one another

ASSIGNMENT: LINEAR RELATIONSHIPS

NEW MESSAGE
October 31, 2022

From: **Nick Normal** *(Head of Student Placement)*

Subject: **RE: RE: Student Salaries**

Hey!

The article on student salaries is crushing, thanks again!

I was wondering if there's something we can do to estimate the potential salaries for the students that haven't gotten jobs yet.

Could you check if the annual salaries for our placed graduates have any relationship with the other data we have on them?

If so, any way you could visualize it for me?

Thanks

Reply Forward

Key Objectives

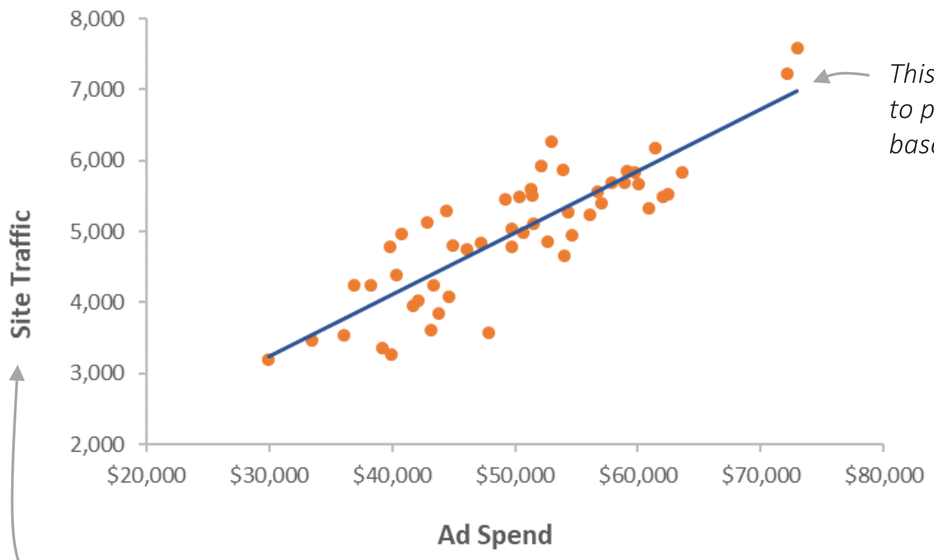
1. Calculate the **correlation** between "Annual Salary" and each of the other numerical variables
2. Create a **scatterplot** to visualize the relationship for the variables with the highest correlation

The goal of **regression** is to predict a dependent variable using independent variables

- This is achieved by fitting a line through the sample data points that models the population

Linear Relationships
Regression Basics
Model Evaluation
Multiple Linear Regression

EXAMPLE *Predicting Site Traffic based on Advertising Spend*



*This line is a **model** that can be used to predict site traffic in a given month based on the advertising budget!*

*This is the **dependent variable (y)**, which is what you're trying to predict*

*This is the **independent variable (x)**, which helps you predict the dependent variable*

HEY THIS IS IMPORTANT!

It is not recommended to predict values outside the range of the sample, as the relationship is not guaranteed to continue

LINEAR REGRESSION MODEL

The **linear regression model** is an equation that best describes a linear relationship

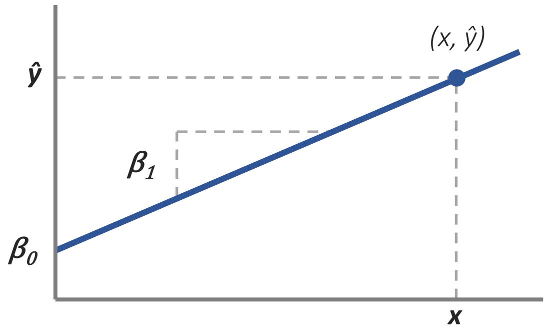
*This is the **predicted** value for the dependent variable*

This is the value for the independent variable

$$\hat{y} = \beta_0 + \beta_1 x$$

*This is the **y-intercept***

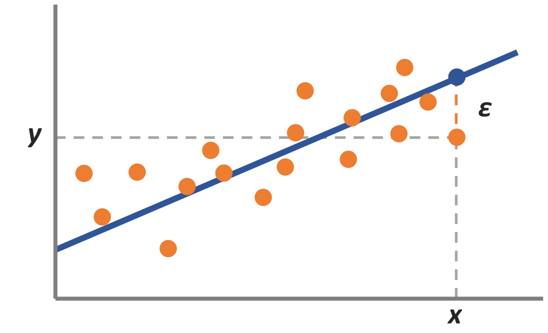
*This is the **slope**, or size of the relationship*



*This is the **real** value for the dependent variable*

$$y = \beta_0 + \beta_1 x + \epsilon$$

*This is an **error**, or residual, caused by the difference between the real & predicted values*



LEAST SQUARED ERROR

Linear Relationships

Regression Basics

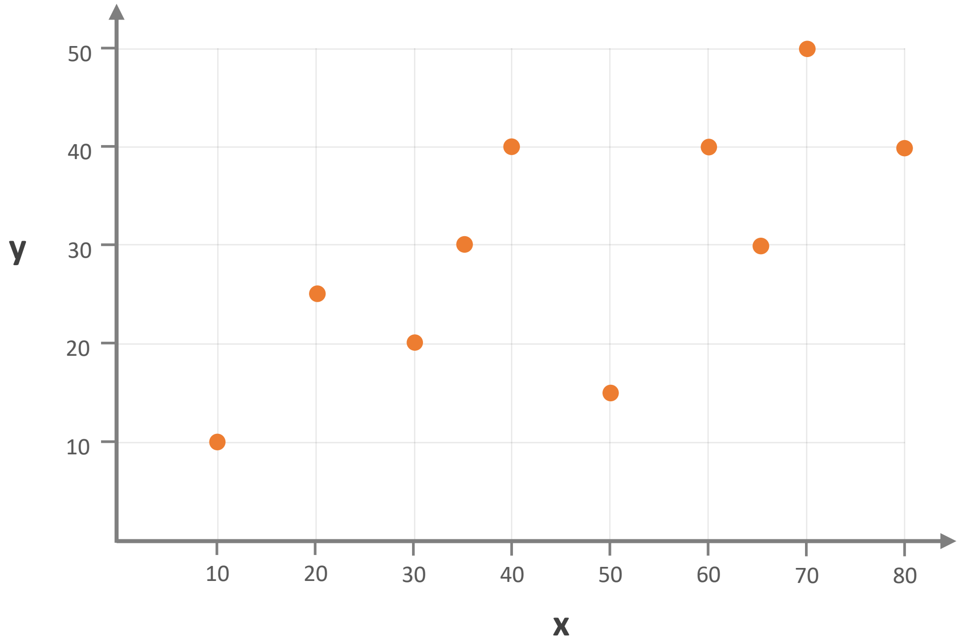
Model Evaluation

Multiple Linear Regression

The **least squared error** method finds the line that best fits through the sample points

It works by squaring the residuals, adding them up, and minimizing that sum

- **NOTE:** Squaring the residuals removes the negatives, but also gives more weight to outliers!



x	y
10	10
20	25
30	20
35	30
40	40
50	15
60	40
65	30
70	50
80	40

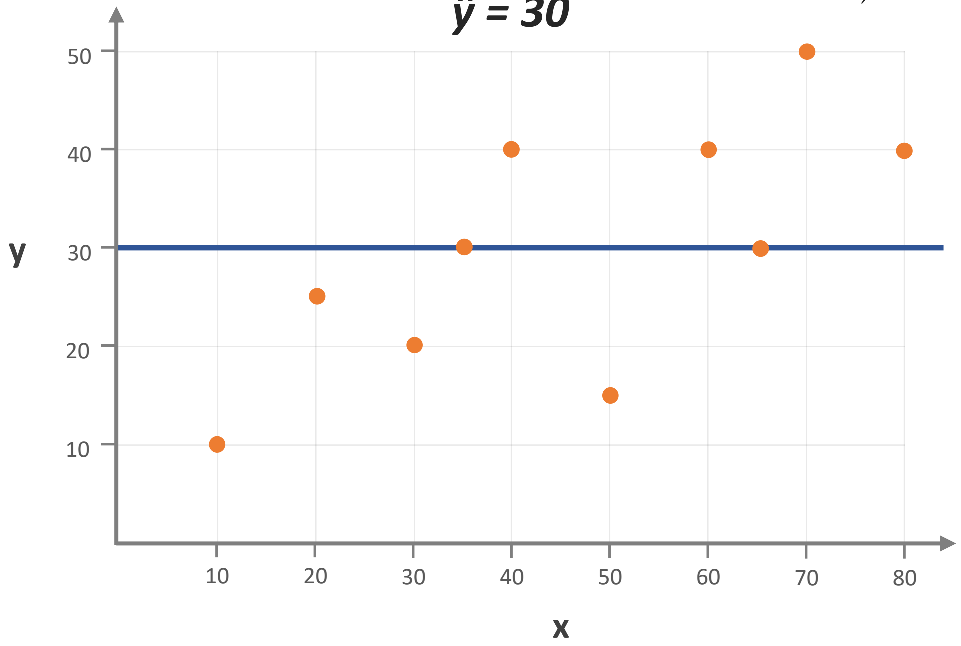
LEAST SQUARED ERROR

The **least squared error** method finds the line that best fits through the sample points

It works by squaring the residuals, adding them up, and minimizing that sum

- **NOTE:** Squaring the residuals removes the negatives, but also gives more weight to outliers!

ŷ = 30 *This is the **mean**, which used to be our best guess!*



x	y	ŷ
10	10	30
20	25	30
30	20	30
35	30	30
40	40	30
50	15	30
60	40	30
65	30	30
70	50	30
80	40	30

LEAST SQUARED ERROR

Linear Relationships

Regression Basics

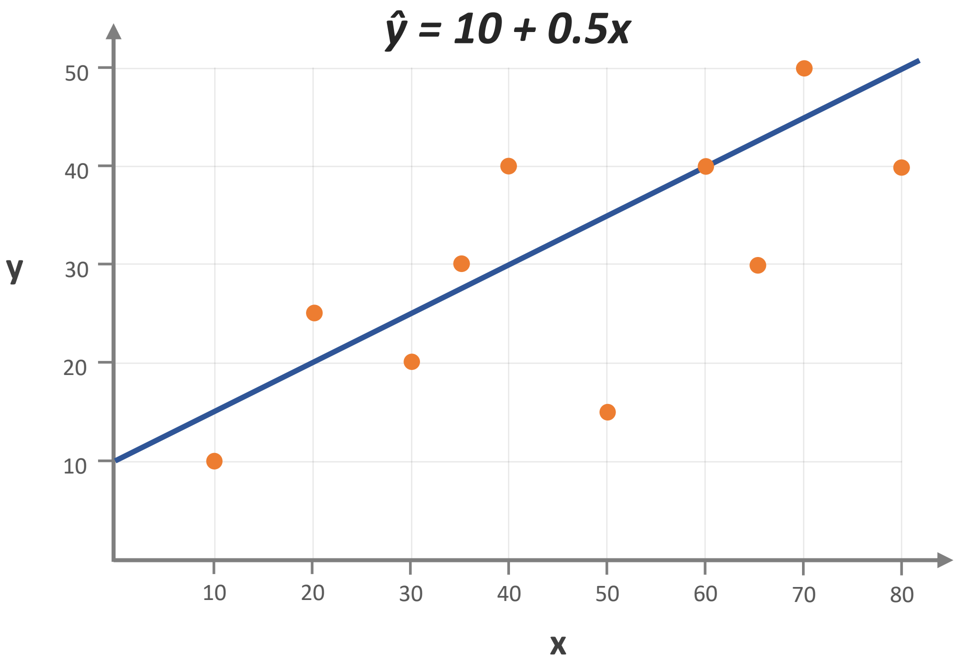
Model Evaluation

Multiple Linear Regression

The **least squared error** method finds the line that best fits through the sample points

It works by squaring the residuals, adding them up, and minimizing that sum

- **NOTE:** Squaring the residuals removes the negatives, but also gives more weight to outliers!



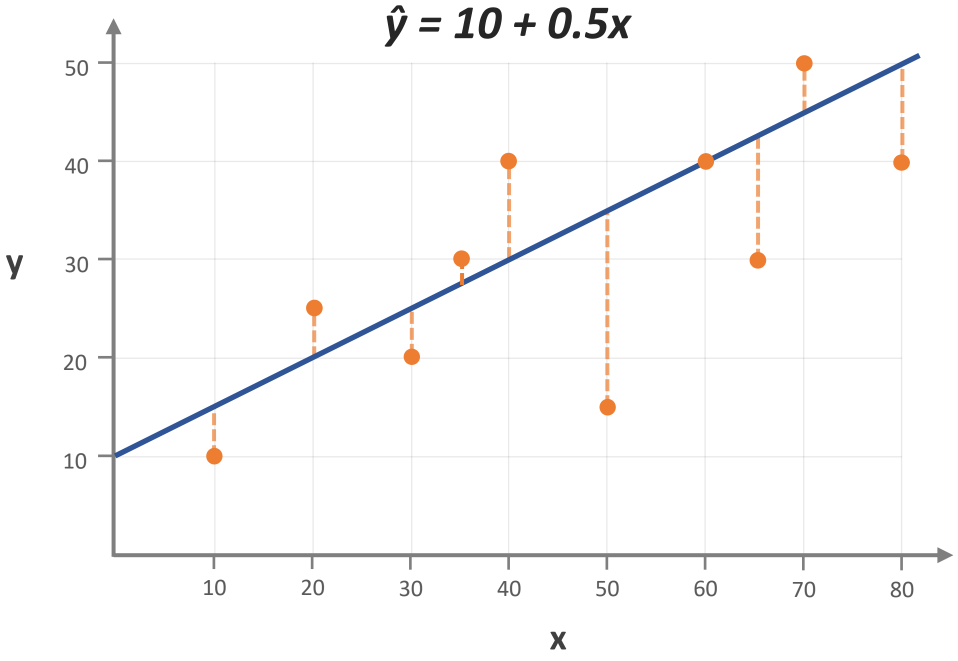
x	y	ŷ
10	10	15
20	25	20
30	20	25
35	30	27.5
40	40	30
50	15	35
60	40	40
65	30	42.5
70	50	45
80	40	50

LEAST SQUARED ERROR

The **least squared error** method finds the line that best fits through the sample points

It works by squaring the residuals, adding them up, and minimizing that sum

- **NOTE:** Squaring the residuals removes the negatives, but also gives more weight to outliers!



x	y	ŷ	ε	ε²
10	10	15	5	**25**
20	25	20	-5	**25**
30	20	25	5	**25**
35	30	27.5	-2.5	**6.25**
40	40	30	-10	**100**
50	15	35	20	**400**
60	40	40	0	**0**
65	30	42.5	12.5	**156.25**
70	50	45	-5	**25**
80	40	50	10	**100**

=

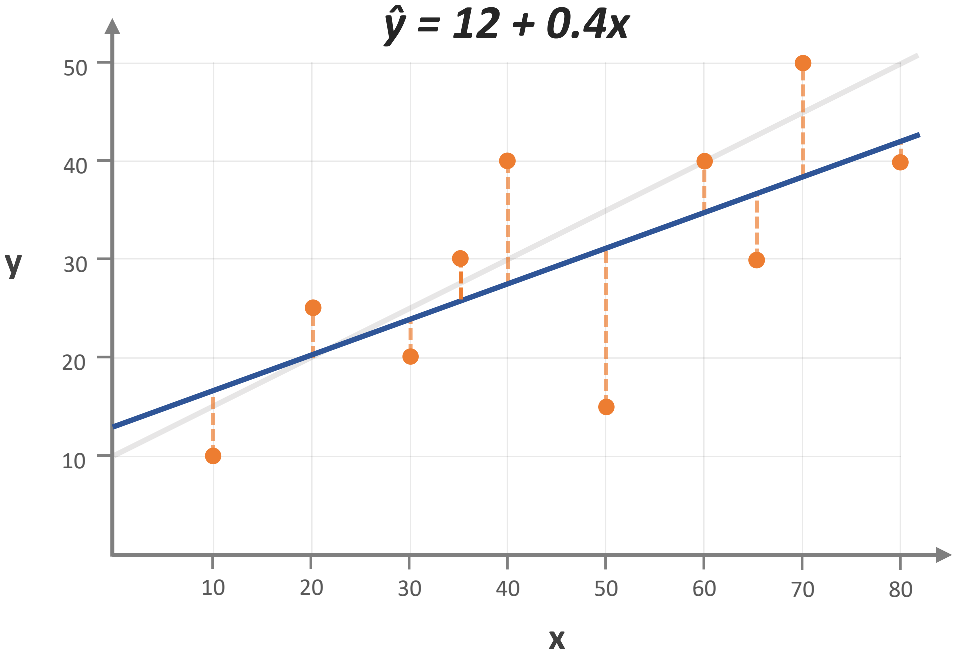
SUM OF SQUARED ERROR: **862.5**

LEAST SQUARED ERROR

The **least squared error** method finds the line that best fits through the sample points

It works by squaring the residuals, adding them up, and minimizing that sum

- **NOTE:** Squaring the residuals removes the negatives, but also gives more weight to outliers!



x	y	ŷ	ε	ε²
10	10	16	6	**36**
20	25	20	-5	**25**
30	20	24	4	**16**
35	30	26	-4	**16**
40	40	28	-12	**144**
50	15	32	17	**289**
60	40	36	-4	**16**
65	30	38	8	**64**
70	50	40	-10	**100**
80	40	44	4	**16**

SUM OF SQUARED ERROR: = **722**

EXCEL'S LINEAR REGRESSION FUNCTIONS

These **Excel functions** help make calculations related to linear regression:

CORREL()	*Returns the coefficient of correlation (r) between two numeric variables*	=**CORREL**(array1, array2)
INTERCEPT()	*Returns the y-intercept (β_0) from a linear regression given a dependent & independent variable*	=**INTERCEPT**(known_ys, known_xs)
SLOPE()	*Returns the slope (β_1) from a linear regression given a dependent & independent variable*	=**SLOPE**(known_ys, known_xs)
FORECAST()	*Returns the predicted value ($\hat{y}$) at "x" from a linear regression given a dependent & independent variable*	=**FORECAST**(x, known_ys, known_xs)
RSQ()	*Returns the coefficient of determination (r^2) between a dependent & independent variable*	=**RSQ**(known_ys, known_xs)
STEYX()	*Returns the standard error of the linear regression model given a dependent & independent variable*	=**STEYX**(known_ys, known_xs)

ASSIGNMENT: SIMPLE LINEAR REGRESSION

NEW MESSAGE
November 5, 2022

From: **Nick Normal** *(Head of Student Placement)*

Subject: **Employability Improvement**

Hi,

By now we know that our program improves student's employability scores by 50 on average.

But could there be another variable that explains by how much we can expect each individual student to improve by?

That would be huge!

Looking forward to hearing back about this,

Thanks

Reply Forward

Key Objectives

1. Calculate the **correlation** between "Employability Improvement" and any relevant numerical variables
2. Create a **scatterplot** to visualize the relationship for the variables with the highest correlation
3. If applicable, build a **regression model** to predict "Employability Improvement"

R-SQUARED

Linear Relationships

Regression Basics

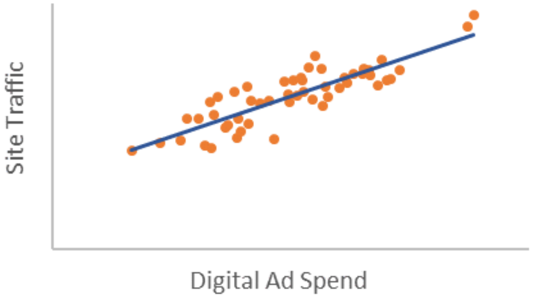
Model Evaluation

Multiple Linear Regression

R-Squared, or coefficient of determination, measures how much better the regression model is at estimating “y” values than the previous best estimate (the mean)

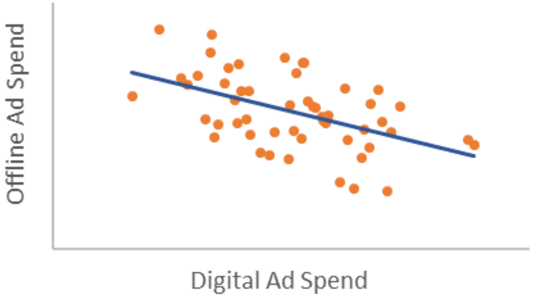
- The higher R-Squared is (0-1), the more confident you can be in the accuracy of your predictions

R^2 = 0.736



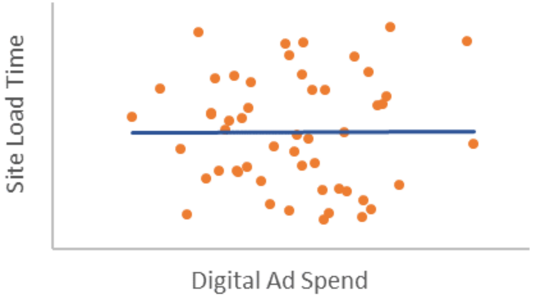
*Digital Ad Spend explains **73%** of the variation in Site Traffic, so it can be used to predict it*

R^2 = 0.249



*Digital Ad Spend only explains **25%** of the variation in Offline Ad Spend, so it likely won’t predict it very well*

R^2 = 0.000



Digital Ad Spend doesn’t explain any of the variation in Site Load Time, and shouldn’t be used to predict it

R-SQUARED

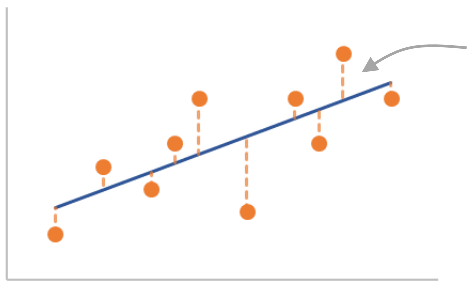
R-Squared, or coefficient of determination, measures how much better the regression model is at estimating "y" values than the previous best estimate (the mean)

- The higher R-Squared is (0-1), the more confident you can be in the accuracy of your predictions

*This is the **sum of squared error** (distance between values & line)*

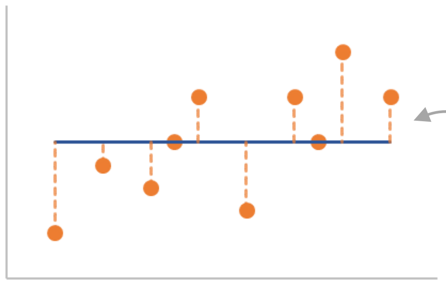
$$R^2 = 1 - \frac{SSE}{SST}$$

*This is the **sum of squared total** (distance between values & mean)*



This is the variation of "y" not explained by "x"

As this decreases relative to the regular measure, then the value of R^2 increases



This is the variation of "y" around its mean

STANDARD ERROR

In regression, the **standard error** is the average distance between the line and the data

- It is the standard deviation of the sample values around the regression line
- This is a good, intuitive measure of how well your model predicts

*This is the **standard error***

$$S = \sqrt{MSE}$$

*This is the **mean square error***
(the square root is the mean error!)

$$MSE = \frac{SSE}{\boxed{n - q - 1}}$$

*This is the **sum of squared error***
(squared distance between values & line)

*These are the **degrees of freedom***

*This is the **sample size***

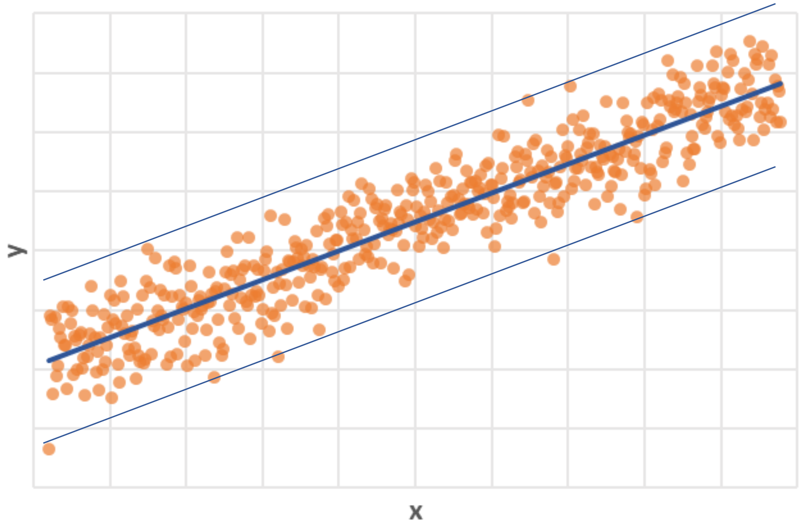
*This is the **# of independent variables***

HOMOSKEDASTICITY

Homoskedasticity simply means that the "scatter" is the same along the entire line

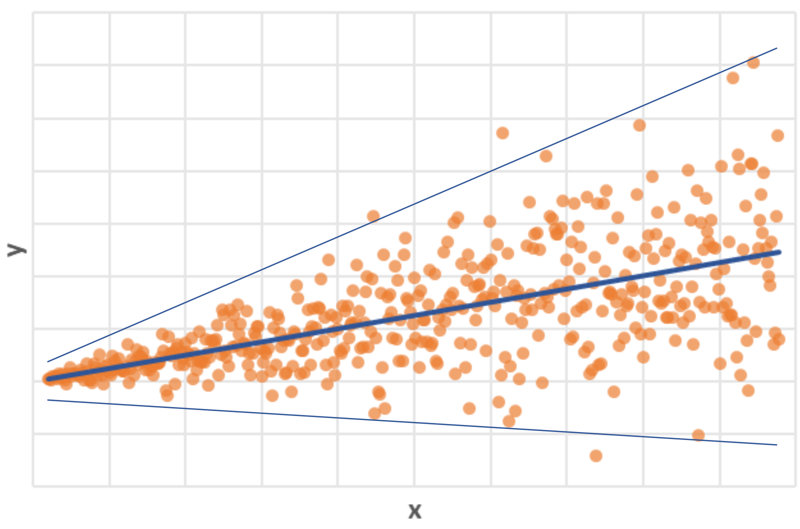
- This is necessary in order to make accurate predictions across the full range of "x" values

Homoskedasticity



The "scatter" is consistent over the entire "x" range

Heteroskedasticity



The "scatter" spreads out as "x" increases

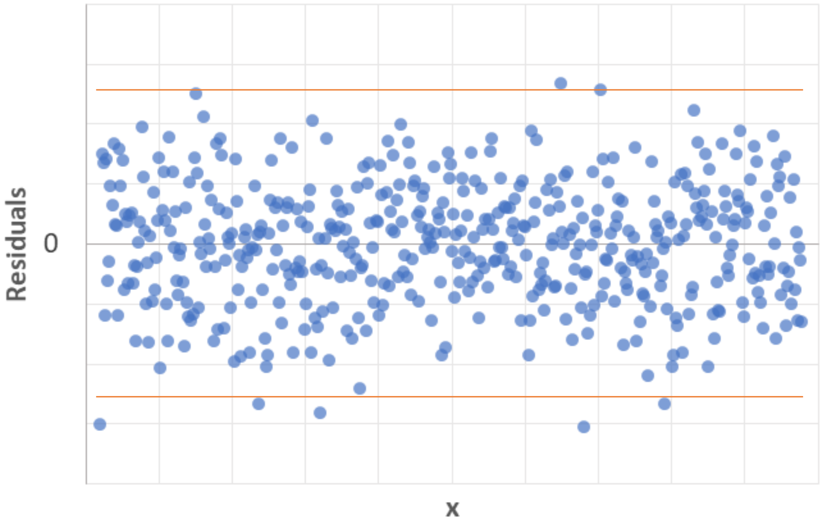
HOMOSKEDASTICITY

Homoskedasticity simply means that the "scatter" is the same along the entire line

- This is necessary in order to make accurate predictions across the full range of "x" values

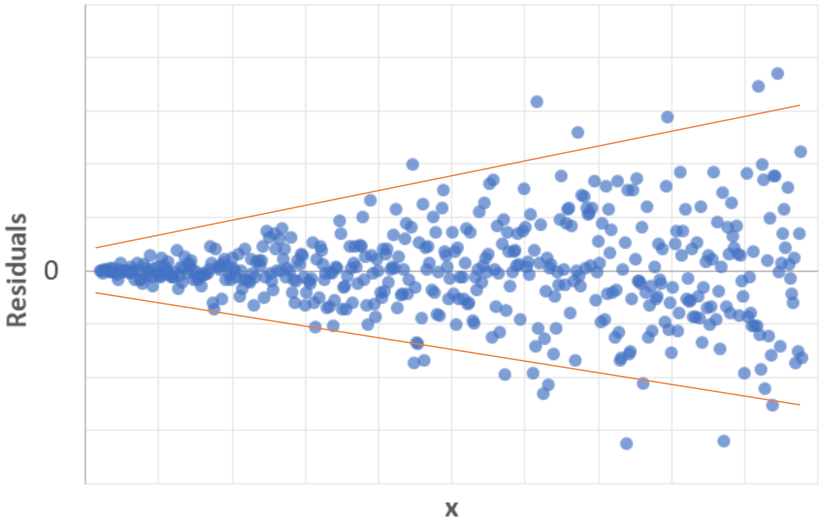
*These are known as **residual plots***

Homoskedasticity



The residuals have consistent variance

Heteroskedasticity



The residuals have inconsistent variance

HYPOTHESIS TEST

Regressions include an implied **hypothesis test** in which the null hypothesis states that no relationship exists between the dependent and independent variables

- In other words, you are trying to find significant evidence that your regression model isn't useless

Steps for a hypothesis test:

☑ 1) State the **null** and **alternative hypotheses**

☑ 2) Set a **significance level**

3) Calculate the **test statistic** for the sample

4) Calculate the **p-value**

☑ 5) Draw a **conclusion** from the test

a) If $p \leq \alpha$, reject the null hypothesis *(you're confident the model isn't useless – use it!)*

b) If $p > \alpha$, don't reject it *(you can't confirm the model isn't useless – don't use it)*

TEST STATISTIC

Linear Relationships

Regression Basics

Model Evaluation

Multiple Linear Regression

The **test statistic** is the f-score for your sample in your regression model

- It's the standard deviations between your model and the hypothesized "useless" model
- More specifically, it's the ratio of the variability that your model explains vs the variability it doesn't

*This is the **test statistic** for your sample*

$$F = \frac{MSR}{MSE}$$

*This is the **mean squared regression** (how much variability the model explains)*

*This is the **mean squared error** (how much variability the model doesn't explain)*

*This is the **sum of squared total** (distance between values & mean)*

*This is the **sum of squared error** (distance between values & line)*

$$MSR = \frac{SST - SSE}{q}$$

*This is the **# of independent variables***

P-VALUE

In regression, the **p-value** is the probability that your model is useless

- In other words, the likelihood that you got a relationship this "strong" when no relationship exists
- The lower the p-value, the stronger the evidence that there is a relationship between the variables

p-value = **F.DIST.RT**(x, deg_freedom1, deg_freedom2)

*The **test statistic** for your regression (F)*

*The **# of independent variables** (q)*

*The **degrees of freedom** (n-q-1)*

HEY THIS IS IMPORTANT!

Remember to set the significance level (α) before calculating the p-value!

ASSIGNMENT: MODEL EVALUATION

NEW MESSAGE
November 11, 2023

From: **Nick Normal** *(Head of Student Placement)*

Subject: **RE: Employability Improvement**

Hi!

I can't believe what I'm seeing here – Tommy will be thrilled!

Can we trust really trust this?

We did all sorts of confidence intervals and hypothesis tests on the mean, and I just want to be confident here as well.

Is there any way you can check that with 95% confidence?

Great work once again,

Thanks!

Reply Forward

Key Objectives

1. Calculate the **r-squared** value
2. Calculate the **standard error**
3. Confirm the **homoskedasticity**
4. Run a **hypothesis test**
5. Draw a **conclusion** on the accuracy of the regression model's predictions

PRO TIP: MULTIPLE LINEAR REGRESSION

Multiple linear regression is used for predicting a single dependent variable based on *multiple* independent variables

- In other words, it's the same linear regression model, but with additional "x" variables

SIMPLE LINEAR REGRESSION MODEL:

$$y = \beta_0 + \beta_1 x + \epsilon$$

MULTIPLE LINEAR REGRESSION MODEL:

$$y = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n}_{} + \epsilon$$

Instead of just one "x", we have a ***whole set of independent variables***
(and associated coefficients) to help predict the dependent variable (y)

PRO TIP: MULTIPLE LINEAR REGRESSION

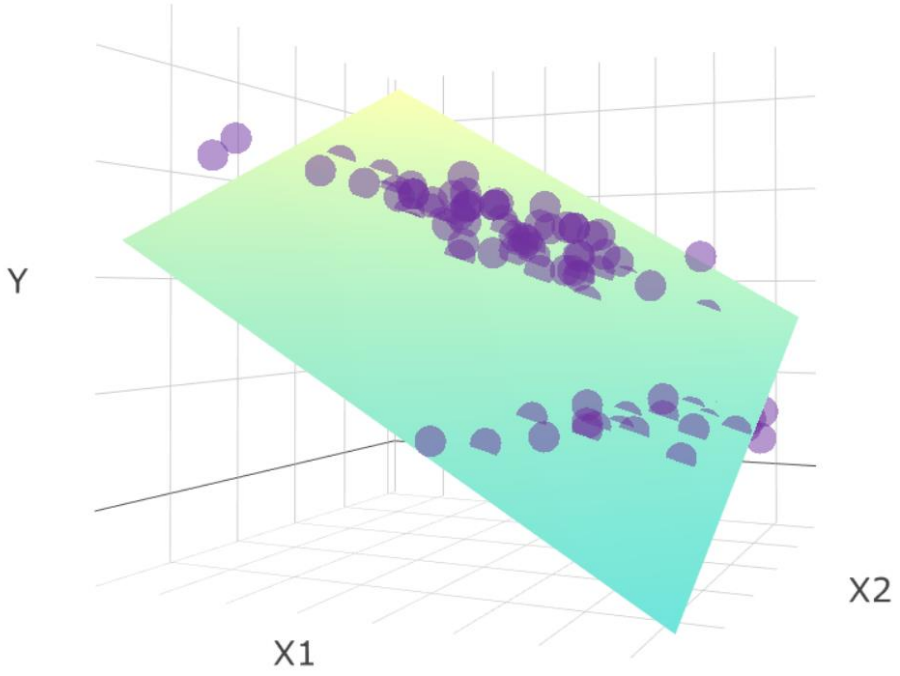
Linear Relationships

Regression Basics

Model Evaluation

Multiple Linear Regression

To visualize how multiple linear regression works with 2 independent variables, imagine fitting a plane *(instead of a line)* through a 3D scatterplot:



HEY THIS IS IMPORTANT!

Multiple linear regression can scale well beyond two variables, so the visual analysis is no longer feasible

That's why it's so important to understand the different model evaluation metrics!

PRO TIP: MULTIPLE LINEAR REGRESSION

The multiple linear regression model has two additional metrics to evaluate:

- The **Adjusted R-Squared** "penalizes" the R-Squared value based on the number of variables
- The **Coefficient P-Values** show the probability that each independent variable is meaningless

EXAMPLE *Predicting Employability (After) based on Undergrad Grade & Employability (Before)*

Output from Excel's Analysis ToolPak

Regression Statistics	
Multiple R	0.9927911
R Square	0.985634168
Adjusted R Square	0.985321868
Standard Error	11.33038767
Observations	95

*This is the **"goodness of fit"** and **"standard error"** for the model*

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	810330.7898	405165.3949	3156.042231	1.7267E-85
Residual	92	11810.74701	128.3776848		
Total	94	822141.5368			

*This is the **p-value** for the whole model*

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-222.4261978	11.8174638	-18.8218218	2.63607E-33	-245.8967009	-198.9556946	-245.8967009	-198.9556946
Undergrad Grade	3.650040497	0.156180213	23.37069732	1.81322E-40	3.339853112	3.960227882	3.339853112	3.960227882
Employability (Before)	0.992544221	0.013705637	72.41868594	6.35519E-83	0.965323643	1.0197648	0.965323643	1.0197648

*This is the **regression model***

*These are the **p-values** for each coefficient (this helps remove useless coefficients)*

KEY TAKEAWAYS: REGRESSION ANALYSIS

★ Numerical variables commonly have **linear relationships**

- *The correlation (r) measures the strength and direction of the relationship, but does NOT imply causation!*

★ Regression lets you **predict "y" values** for any given "x" values

- *Correlation should exist between the variables, and causality should be logically possible*

★ The regression model is the **line that best fits** through the data

- *It's described by an equation with a y-intercept, slope coefficients for each "x" value, and a residual (error)*

★ You can **evaluate the accuracy** of the model with several metrics

- *The R^2 value measures how well the line fits the data, the standard error measures the average distance between the line and the data, and the p-value helps you confirm that the model can be used for prediction*

MAVEN AIRLINES | PROJECT BRIEF

You are a BI Analyst at **Maven Airlines** and were just put in charge of a major project that could potentially get you promoted to Senior BI Analyst

From: **Peter Plane** *(Senior BI Analyst)*

Subject: **Cost Formula**

Hey there!

We managed to get our hands on some data that could help us forecast our annual costs. It includes the fuel price, load factor (what percentage of seats are filled on average), and an index for the output (revenue per passenger mile), which we have forecasts for.

Could you use that to produce a reliable formula we could use?

Thank you!

Airline_Costs.xlsx

Reply Forward

Key Objectives

1. Check if a linear relationship exists between the variables
2. Build a simple linear regression model for the variable with the best correlation
3. Build a multiple linear regression model using all the variables
4. Compare the models' performance
5. Select the best model for the forecast