MAVEN
BUSINESS SCHOOL

Essential Statistics for Data Analysis

Analysis of Master's Business School Graduates with Statistics

Alfikri Ramadhan

DATA ANALYTICS

Hello!

I am Alfikri Ramadhan

- Github: github.com/fikrionii
- LinkedIn: linkedin.com/in/alfikri-ramadhan
- Email: alfikri12@gmail.com

Table of Contents

Setting Expectation for This Project

⚠ **This project is performed with Microsoft Excel**

⚠ **I will list the functions used in this course in the Appendix**

⚠ **This project WILL NOT going too deep into the math behind it.**

If you want to learn the concept behind those excel formulas or the math concept behind it, you may check these sources:

- Statquest with Josh Stammer (**YouTube**)
- Essential Statistics for Data Analysis by Maven Analytics (**Udemy**)
- Statistics for Data Science and Business Analysis by 365Careers (**Udemy**)

DATA ANALYTICS

01

WHY STATISTICS?

Data Analytics is about using data to make smart decision

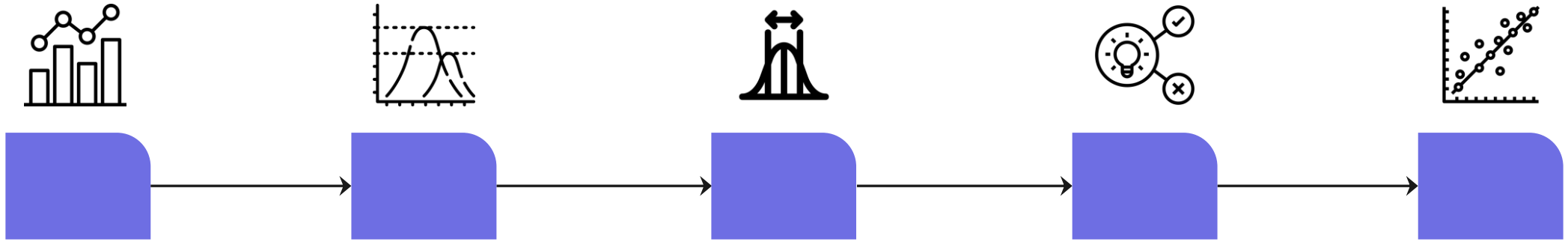
What is Statistics?

The study of how to collect, analyse, summarize, and present data

Why Learn Statistics?

Statistics helps us make estimates and predictions of population using a sample

The Statistics Workflow



Descriptive Statistics

Understand our sample data using mean, median, histogram and other descriptive statistics

Probability Distributions

Use the sample as a model for the entire population **if it fits a probability distribution**

Confidence Intervals

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

Hypothesis Testing

Draw conclusions about what a population looks like based on a sample

Regression Analysis

Use relationships based on available variable to **make prediction**

In this project, we will focus on hypothesis testing and regression analysis

DATA ANALYTICS

02

PROJECT INTRODUCTION

The Maven Business School

Maven Business School is an online startup that's looking to disrupt the postgraduate programs offered by traditional universities.

As a **Recruitment Analyst** in the startup, we have data from the first graduating class of the MBA program, including details & scores from their application, the program itself, and their employment status 2 months later.

Our Goal

Leverage statistics to evaluate the result of this class, predict the performance of future classes, and propose changes in the recruitment to improve graduate outcomes

Our Objective

- Understand our data with descriptive statistics
- Draw conclusions with hypothesis tests
- Make predictions with regression analysis

The Project Dataset

	A	B	C	D	E	F	G	H	I
1	Student ID	Undergrad Degree	Undergrad Grade	MBA Grade	Work Experience	Employability (Before)	Employability (After)	Status	Annual Salary
2	1	Business	68.4	90.2	No	252	276	Placed	$111,000
3	2	Business	62.1	92.8	No	423	410	Not Placed	
4	3	Computer Science	70.2	68.7	Yes	101	119	Placed	$107,000
5	4	Engineering	75.1	80.7	No	288	334	Not Placed	
6	5	Finance	60.9	74.9	No	248	252	Not Placed	
7	6	Computer Science	74.5	80.7	No				
8	7	Finance	76.4	83.3	No				
9	8	Business	82.6	88.7	No				
10	9	Finance	76.9	75.4	No				
11	10	Computer Science	83.3	82.1	No				
12	11	Business	75.8	87.5	No				
13	12	Engineering	76	66.9	No				
14	13	Business	62.8	71.3	No				
15	14	Engineering	82.8	76.8	No				
16	15	Business	76	72.3	No				
17	16	Finance	76.9	72.4	No				
18	17	Computer Science	75.8	72	Yes				
19	18	Art	78	81	No				
20	19	Business	82.4	96.1	No				
21	20	Computer Science	76.2	76.7	No				
22	21	Business	62.5	80.3	No				
23	22	Art	78	77.8	No				
24	23	Engineering	66.5	62.6	No				
25	24	Computer Science	63.5	80.2	No				
26	25	Business	82.6	79.1	No				
27	26	Computer Science	79.2	77.8	No				
28	27	Computer Science	75	75.1	No				
29	28	Art	74.4	82.2	No				
30	29	Finance	67.9	70.5	No				

The dataset consist of 95 rows and 9 features

The Dataset is provided by Maven Analytics

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)

DATA ANALYTICS

03

DESCRIPTIVE STATISTICS

Statistical Summary

	Undergrad Grade	MBA Grade	Employability (Before)	Employability (After)	Annual Salary
Count	95	95	95	95	53
Unique	75	81	83	81	34
Mean	75.0	80.2	240	289	$119,387
Std	7.5	6.2	86	94	$45,547
Min	60.9	62.6	62	102	$75,500
25%	68.7	76.1	182	228	$99,000
50%	75.6	80.2	236	286	$104,500
75%	79.4	84.7	287	348	$124,000
Max	100.0	96.1	423	481	$340,000

What can we learn from these numbers?

- We can see the increase in the mean of graduate's employability score before and after obtaining their master's degree.
- From all our graduates, not all of them has annual salary or has been placed in a job (95 vs. 53)
- By seeing the mean and median, we can expect all grades and employability scores to have normal distribution, while Annual Salary distribution is skewed to the right. We will draw histograms to confirm this.

What's our graduates looks like?

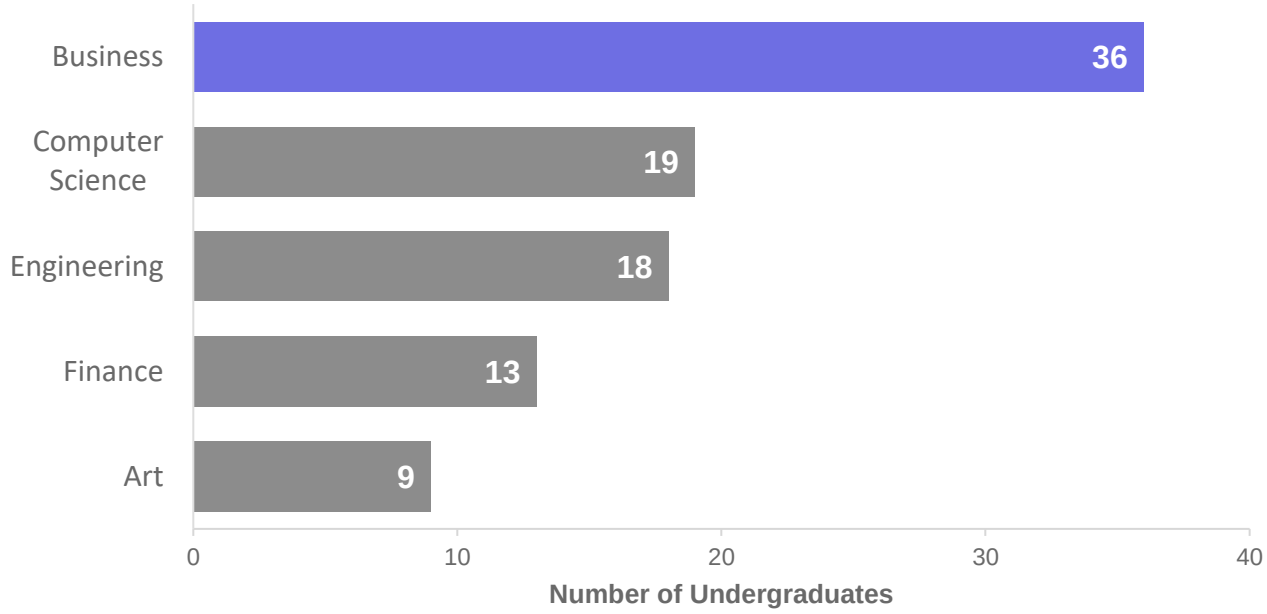
There are

95

Master's Graduates

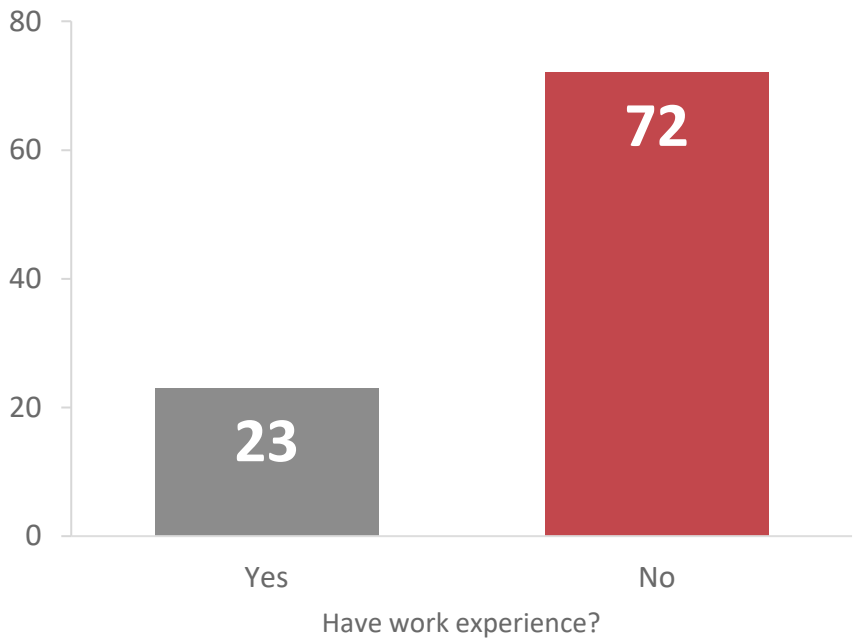
from Maven Business School (MBS)…

…where most of them are **Business Undergraduates**

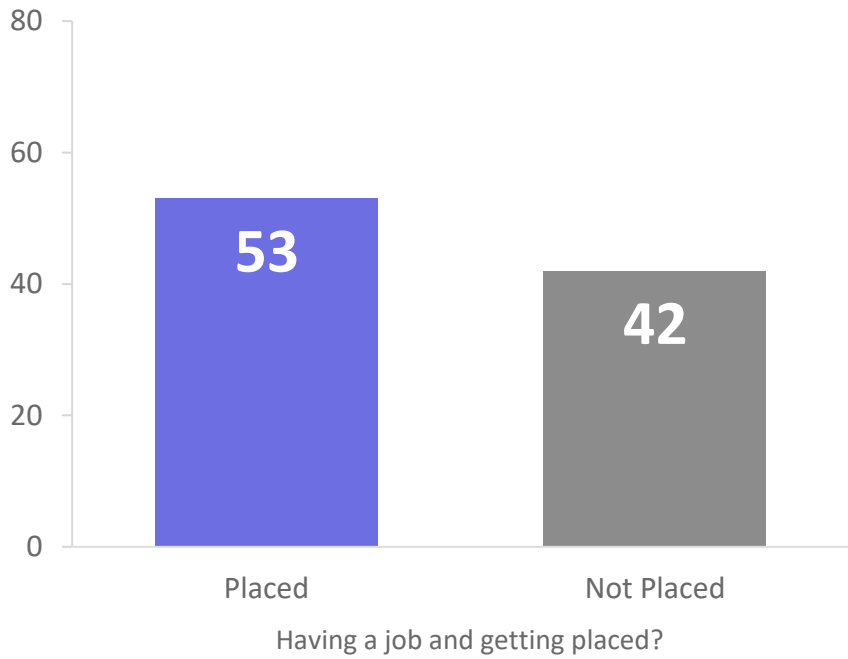


Did they have a job before and after graduating?

While most of our students **don't have work experience** prior to entering MBS...



... more than **half of the graduates are getting a job** after graduating

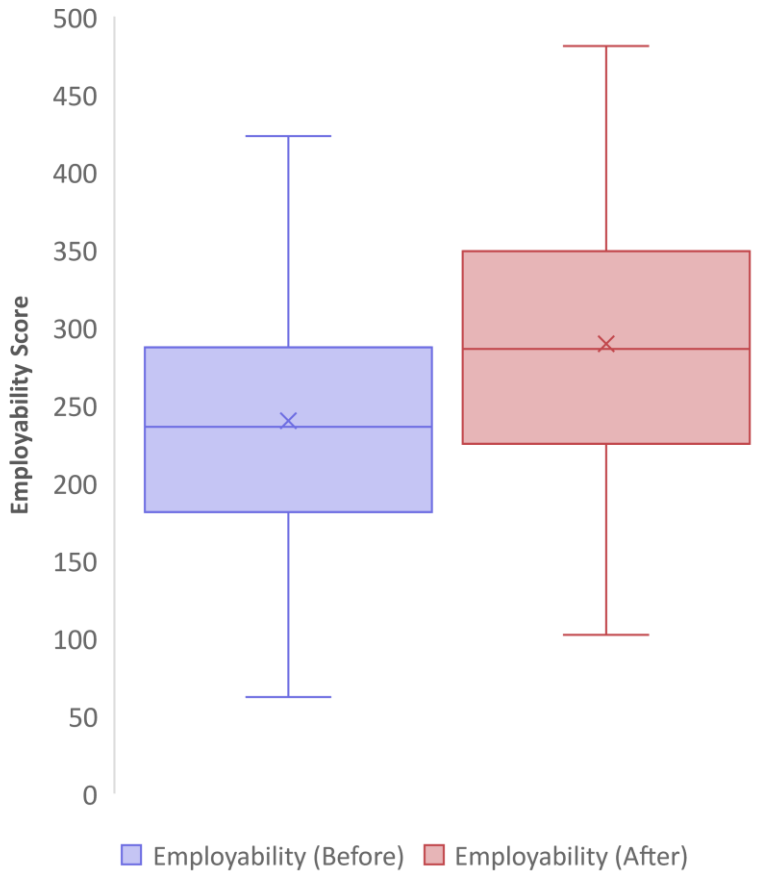


Grade and Employability Score Histogram

Feature	Skewness
Undergrade Grade	0.417
MBA Grade	-0.120
Employability (Before)	0.265
Employability (After)	0.075

Since the skewness is between -0.5 and 0.5, **the distribution is approximately symmetric**

Can we see improvement on graduate's Employability Score?



In the boxplot, we can see the graduate's **employability score are improving after they obtained their MBA.**

Can we assume the student's employability score are improving by an average of 50 points?

Question like this is one of the example of assumption that could arise from our sample. By defining a significance level and calculating p-value**, we can answer this question with hypothesis test!**

DATA ANALYTICS

04

HYPOTHESIS TESTING

What is Hypothesis Test?

👩‍🔬 A **hypothesis test** is a method of statistical inference used to decide whether the data sufficiently support a particular hypothesis. Hypothesis testing allows us to make probabilistic statements about population parameters.

💡 Steps for a hypothesis test:

1. State the null and alternative hypotheses
2. Set a significance level and α (accepted probability of error)
3. Calculate the test statistics for the sample
4. Calculate the p-value
5. Draw a conclusion from the test

How to Evaluate Hypothesis Test?

Null and Alternative Hypothesis

- The null hypothesis (H_o) is the assumption about a population we'd like to evaluate
- The alternative hypothesis (H_a) is any scenario in which that assumption is wrong

α (accepted probability of error)

The significance level or alpha level is the probability of making the wrong decision when the null hypothesis is true. Usually, these tests are run with an alpha level of .05 (5%), but other levels commonly used are .01 and .10.

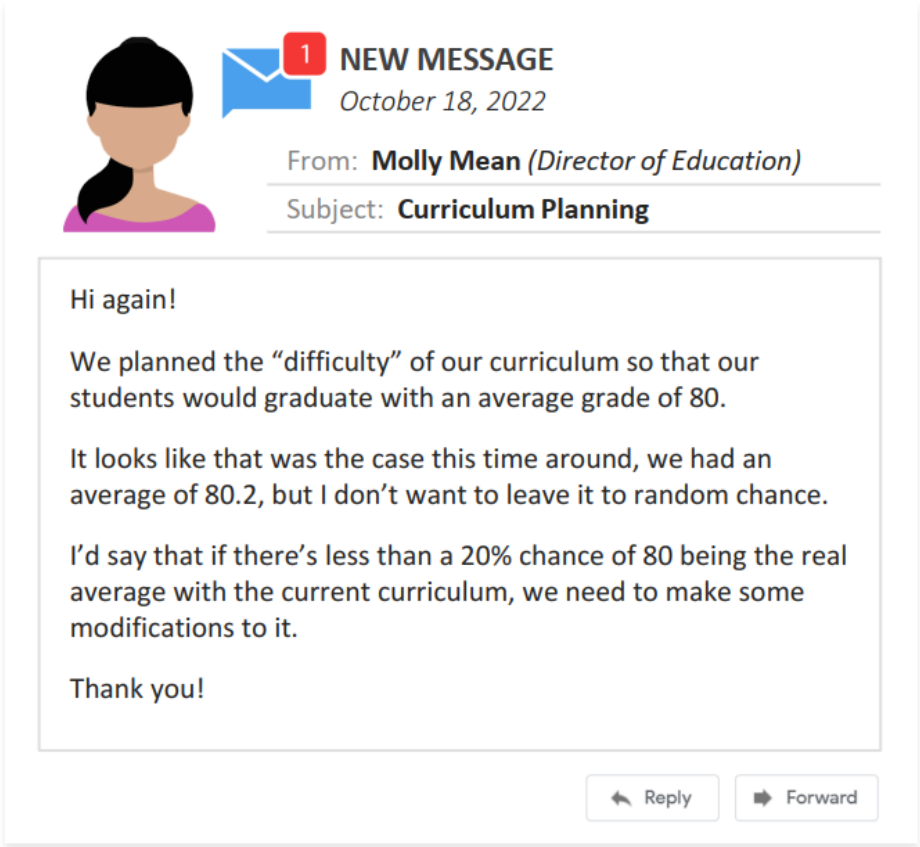
P-Value (p)

The P-value is the probability that our sample supports the null hypothesis. Without going too much into the math behind it, the p-value is calculated with test statistics, standard error, and probability distributions.

Conclusion of a hypothesis test ⇨

- If $p>\alpha$, then we **fail to reject** the null hypothesis (*not enough evidence!)*
- If $p<\alpha$, then we **reject** the null hypothesis (*strong enough evidence!)*

Hypothesis Test #1:
Getting 80 as Average MBA Grade



Steps for Hypothesis Test

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

Is there less than an 20% chance of 80 being the real MBA grade average with the current curriculum?

Student ID	Undergrad Degree	Undergrad Grade	MBA Grade
1	Business	68.4	90.2
2	Business	62.1	92.8
3	Computer Science	70.2	68.7
4	Engineering	75.1	80.7
5	Finance	60.9	74.9
6	Computer Science	74.5	80.7
7	Finance	76.4	83.3
8	Business	82.6	88.7
9	Finance	76.9	75.4
10	Computer Science	83.3	82.1
11	Business	75.8	87.5
12	Engineering	76	66.9
13	Business	62.8	71.3
14	Engineering	82.8	76.8
15	Business	76	72.3

n = 95

HYPOTHESES			
Ho:	μ	=	80
Ha:	μ	≠	80

SIGNIFICANCE LEVEL	
Alpha (α):	**0.2**

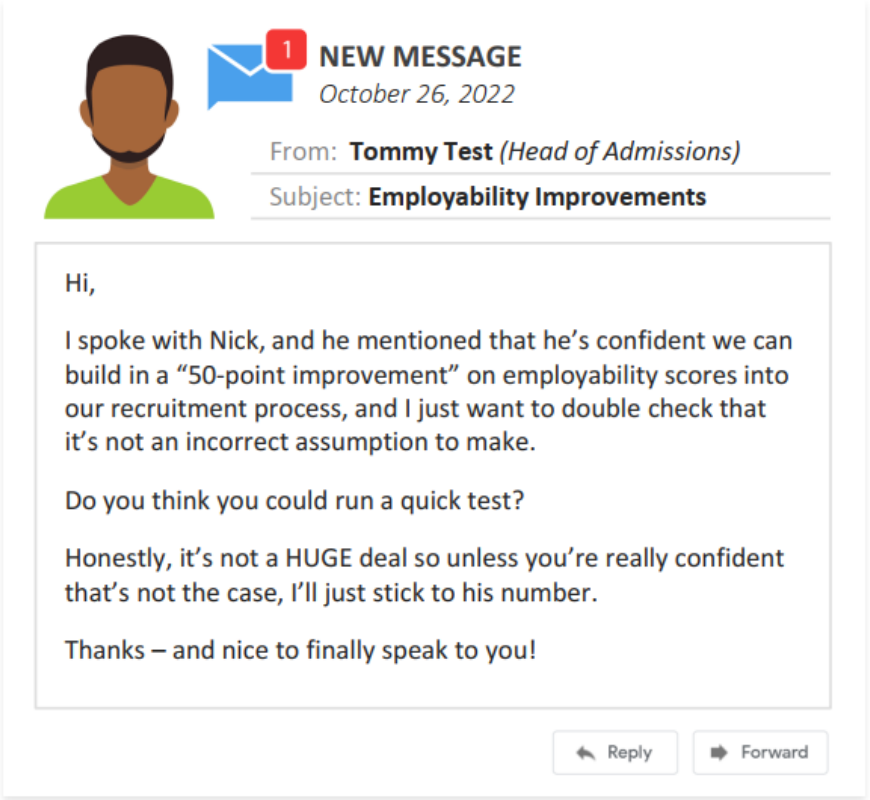
SAMPLE DATA	
Sample Size:	95
Mean:	80.2
Std Dev:	6.17

HYPOTHESIS TEST	
Standard Error:	0.63
Test Statistic (t):	0.27
P-Value:	**0.790**

Conclusion

Since p>α, we fail to reject the null hypothesis. We don't have sufficient evidence to prove that the average MBA grades are any different than 80. Therefore, there is no need to change our current curriculum.

Hypothesis Test #2:
50-point Improvement on Employability Scores



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

Steps for Hypothesis Test

1. Calculate the **difference** between employability score before and after graduation
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 statistics** for the sample
6. Calculate the **p-value**
7. Draw a **conclusion** from the test

Can we assume that student's employability scores are improving by an average of 50 points after the MBA program?

Student ID	Employability (Before)	Employability (After)	Difference
1	252	276	24
2	423	410	-13
3	101	119	18
4	288	334	46
5	248	252	4
6	145	209	64
7	401	462	61
8	287	342	55
9	275	347	72
10	254	313	59
11	182	232	50
12	117	163	46
13	130	119	-11
14	219	304	85

n = 95

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

HYPOTHESES			
Ho:	$\mu_1-\mu_2$	$\geq$	50
Ha:	$\mu_1-\mu_2$	$<$	50

SIGNIFICANCE LEVEL	
Alpha (α):	**0.02**

SAMPLE DATA	
Sample Size:	95
Mean:	49.4
Std Dev:	29.66

HYPOTHESIS TEST	
Standard Error:	3.04
Test Statistic (t):	-0.18
P-Value:	**0.427**

Conclusion

Since p>α, we fail to reject the null hypothesis. We don't have sufficient evidence to prove the average improvement in employability score is lower than 50. Let's use the "50-score improvement" in our recruitment process!

DATA ANALYTICS

05

REGRESSION ANALYSIS

What is Regression Analysis?

Regression analysis is **a statistical method that shows the relationship between two or more variables**. Usually expressed in a graph, the method tests the relationship between a dependent variable against independent variables.

Linear Regression Model

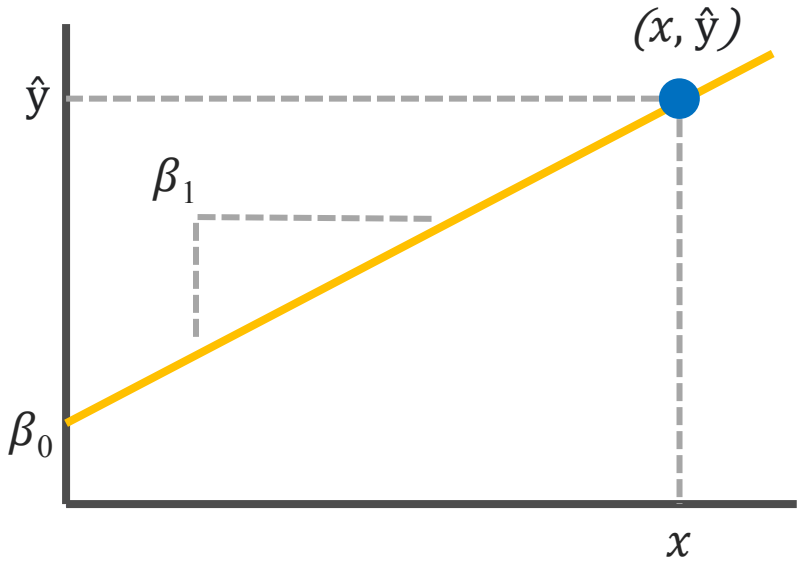
*This is the **predicted** value for the dependent variable*

This is the value for the independent variable

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

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

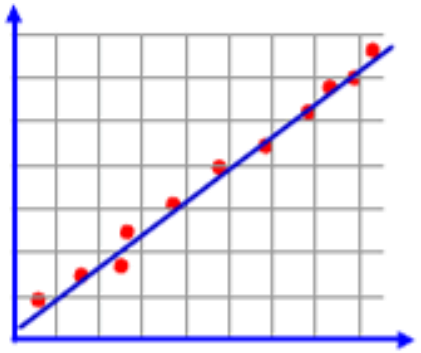
*This is the **slope,** or size of the relationship*



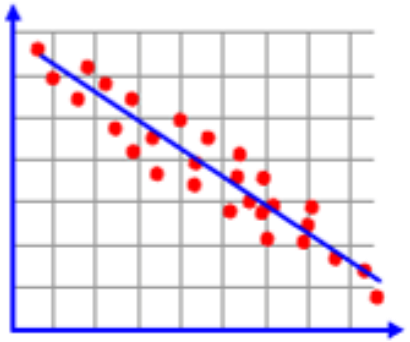
Correlation

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

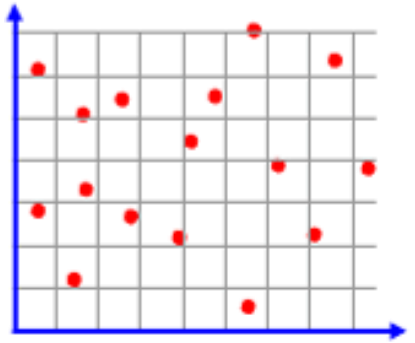
-1 is a perfect negative correlation, 0 is no correlation, and 1 is a perfect positive correlation



Strong positive correlation



Moderate negative correlation



No correlation

!!

Correlation does NOT imply causation!

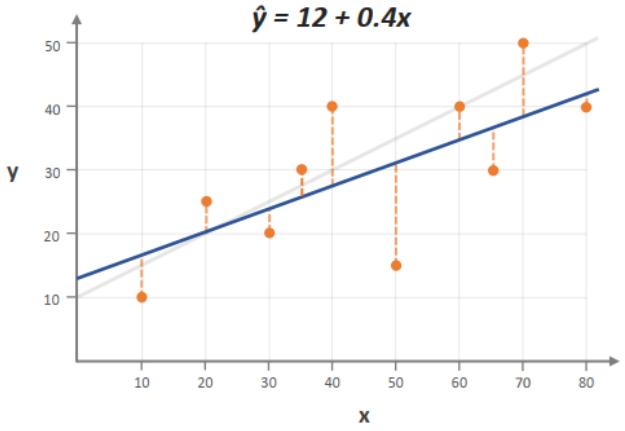
Just because two things correlate does not necessarily mean that one causes the other.

Least Squared Error & R-Squared

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



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

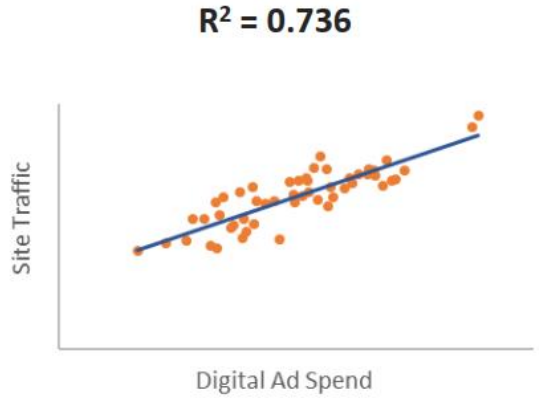
||

SUM OF SQUARED ERROR: **722**

R-Squared

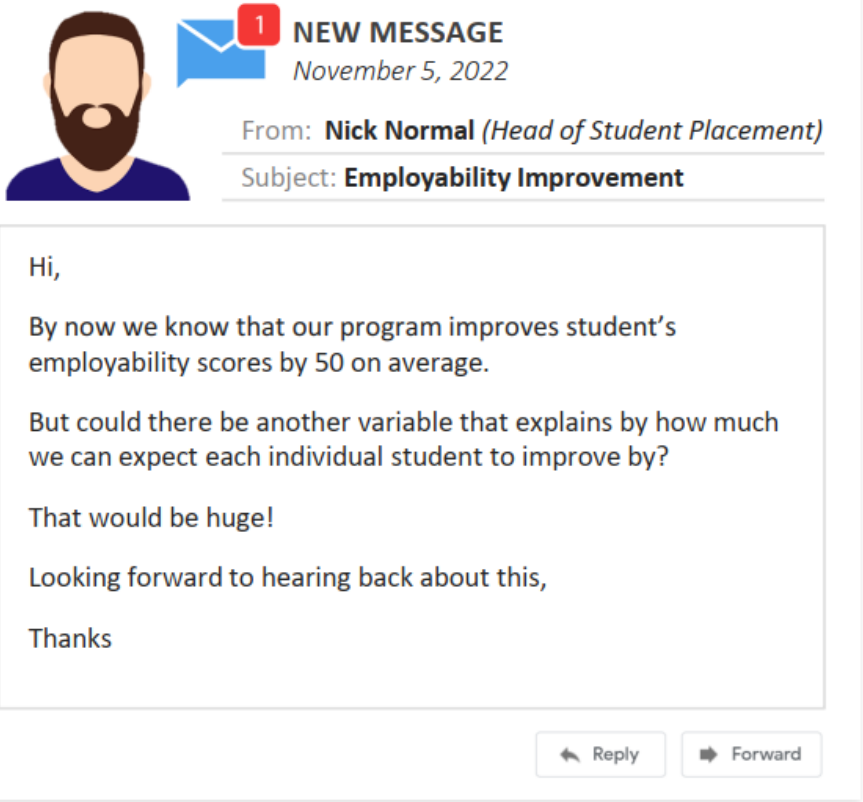
R-squared, or coefficient of determination, measures how much better the regression model is at estimating “y” values.

The higher R-Squared is (0-1), the more confident we can be in our predictions.



Digital Ad spend explains **73%** of the variation in Site Traffic. Based on the R^2 value, **we can use Digital Ad Spend to predict the Site Traffic.**

Regression Analysis:
Predicting Employability Score Improvement



Steps for Regression Analysis

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. Build a regression model to predict "Employability Improvement"
4. Check the **R-Squared** value to measure how well the model fits the data

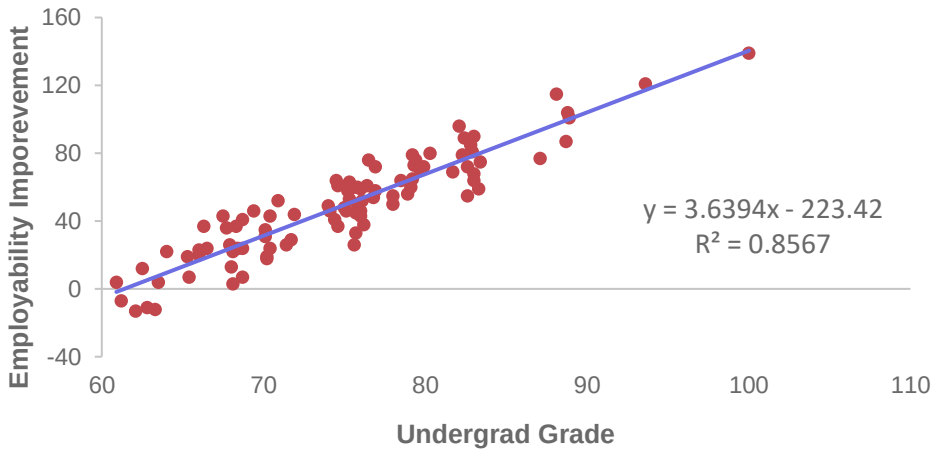
Regression Analysis

Employability (Before)	Employability (After)	Employability Improvement	Undergrad Grade	MBA Grade
252	276	24	68.4	90.2
423	410	-13	62.1	92.8
101	119	18	70.2	68.7
288	334	46	75.1	80.7
248	252	4	60.9	74.9
145	209	64	74.5	80.7
401	462	61	76.4	83.3
287	342	55	82.6	88.7
275	347	72	76.9	75.4
254	313	59	83.3	82.1
182	232	50	75.8	87.5
117	163	46	76	66.9

First, let's check the correlation between these variables to Employability Improvement:

Correlation	
Undergrad Grade:	0.9256
MBA Grade:	0.1149

Since Undergrad Grade has higher correlation, we will use it in our Linear Regression Model



What does this mean?

- $y = 3.6394x - 223.42$ is our Linear Regression Model. We can input **x value** (Undergrad Grade) to predict the **y value** (Employability improvement).
- R^2 value of 0.8567 means Undergrade grade **explains 85%** of the variation in Employability Improvement, which is quite high and gives us more confident in our prediction.

Analysis Toolpak

Excel has a built-in **Analysis Toolpak** that allows us to perform regression analysis quickly and efficiently.

In this example, we predict the Employability Score (After) using Undergrad Grade and Employability Score (Before)

SUMMARY OUTPUT

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"** 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			

	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**

Prediction using the Regression Model

Undergrade Grade:	70
Employability (Before):	285
Employability (After):	315.95

The prediction result

Thank You!

Feel free to contact me and let's discuss!

github.com/fikrionii
linkedin.com/in/alfikri-ramadhan
alfikri12@gmail.com



EXCEL FUNCTIONS

APPENDIX

Excel Functions for Statistics

Statistics Descriptive #1

COUNT()	*Returns the number of cells that contain numbers and count them within the list of range of array.*
UNIQUE()	*Returns a list of unique values in a list or range. Combine it with COUNT() to count unique values within a list or range.*
AVERAGE()	*Returns the arithmetic mean, calculated by adding all numbers of given data set and then dividing the sum by the total number of values (or count) of given items.*
MEDIAN()	*Returns the median of the given numbers. The median is the number in the middle of a set of numbers.*
MODE()	*Returns he most frequently occurring number in a group of numbers*

Statistics Descriptive #2

MIN()	*Returns the smallest number in a set of values.*
MAX()	*Returns the biggest number in a set of values.*
QUARTILE()	*Returns the quartile for a given set of data. This function divides the data set into four equal groups: first quartile, second quartile, third quartile, and maximum value.*
SKEW()	*Returns the skewness of a distribution. Skewness characterizes the degree of asymmetry of a distribution around its mean.*
SQRT()	*Returns a positive square root.*

Hypothesis Testing

T.DIST()

Returns the cumulative probability or the probability density at a t-score from the given t distribution.

depends on the type of hypothesis test (Two Tail, One Tail to the left, or One Tail to the Right) the **T.DIST() function will differ slightly*

Linear Regression

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