
Statistics Made Simple
Data Science Bootcamp

We brought curriculum from Harvard Business School Online

- Interpret data to inform business decisions
- Recognize trends, detect outliers, and summarize data sets
- Analyze relationships between variables
- Develop and test hypotheses
- Craft sound survey questions and draw conclusions from population samples
- Implement regression analysis and other analytical techniques in Excel

Go to Course Syllabus

About the Professor

Janice Hammond is the Jesse Philips Professor of Manufacturing and Senior Associate Dean for Culture and Community at Harvard Business School. She currently teaches Supply Chain Management in the MBA program and is program chair for the HBS Executive Education International Women's Foundation and Women's Leadership Programs. Her research focuses on speed and flexibility in manufacturing and logistics systems to respond to changing customer demand.

Online Business Analytics Course | HBS Online

We'll learn five weeks in 3 hours :D

Business Analytics

Syllabus

Business Analytics introduces quantitative methods used to analyze data and make better management decisions. This course is not based on rote memorization of equations or facts, but focuses on honing your understanding of key concepts, your managerial judgment, and your ability to apply course concepts to real business problems.

Modules		Lessons	Learning Objectives	
Module 1	Describing and Summarizing Data	• Visualizing Data • Descriptive Statistics • Relationships Between Two Variables	• Create visual representations of data that allow you to recognize trends and detect outliers • Define and calculate descriptive statistics to summarize data sets concisely • Analyze relationships between two variables by creating scatter plots and calculating the correlation coefficient	Quiz
Module 2	Sampling and Estimation	• Creating Representative and Unbiased Samples • The Normal Distribution • Confidence Intervals • Amazon's Inventory Sampling	• Determine an adequate sample size, explain the importance of random sampling, and craft sound survey questions to create representative samples • Draw conclusions about the larger population by calculating sample statistics and applying the properties of the normal distribution • Estimate the accuracy of statistics by calculating confidence intervals	Quiz
Module 3	Hypothesis Testing	• Designing and Performing Hypothesis Tests • Improving the Customer Experience	• Develop and test hypotheses to assess the impact of changes on an entire population or estimate differences between populations • Quantify the evidence in favor of or against your hypothesis in order to make managerial decisions	Quiz
Module 4	Single Variable Linear Regression	• The Regression Line • Forecasting • Interpreting the Regression Output • Performing Regression Analysis • Forecasting Home Video Units	• Identify the best fit line for a data set and interpret its equation • Analyze the relationship between two variables and develop forecasts for values outside the data set • Perform a regression analysis using Excel and interpret the output	Quiz
Module 5	Multiple Regression	• The Multiple Regression Equation • Adapting Concepts from Single Regression • Performing Multiple Regression Analysis • New Concepts in Multiple Regression • The Caesars Staffing Problem	• Estimate the relative predictive power of different combinations of variables by performing and interpreting a multiple variable regression analysis using Excel • Expand the range of your analysis by using dummy and lagged variables	Quiz

Add extra content

- Confidence Interval vs. p-value
- F-Test for variance
- Chi-square test of independent
- One-Way ANOVA
- Logistic regression

My progress in HBS class

Business Analytics

	Introduction to Business Analytics with Prof. Jan Hammond	Describing and Summarizing Data	Sampling and Estimation	Hypothesis Testing	Single Variable Linear Regression	Multiple Regression
% COMPLETE	100%	100%	100%	100%	100%	100%
QUIZ SCORE		95% Due: 03/25/2020 01:00 PM ET Taken:03/19/2020 08:06 AM ET	95% Due: 04/08/2020 01:00 PM ET Taken:04/04/2020 03:04 PM ET	95% Due: 04/15/2020 01:00 PM ET Taken:04/05/2020 06:52 AM ET	95% Due: 04/29/2020 01:00 PM ET Taken:04/14/2020 07:14 AM ET	95% Due: 05/13/2020 01:00 PM ET Taken:04/28/2020 10:28 PM ET

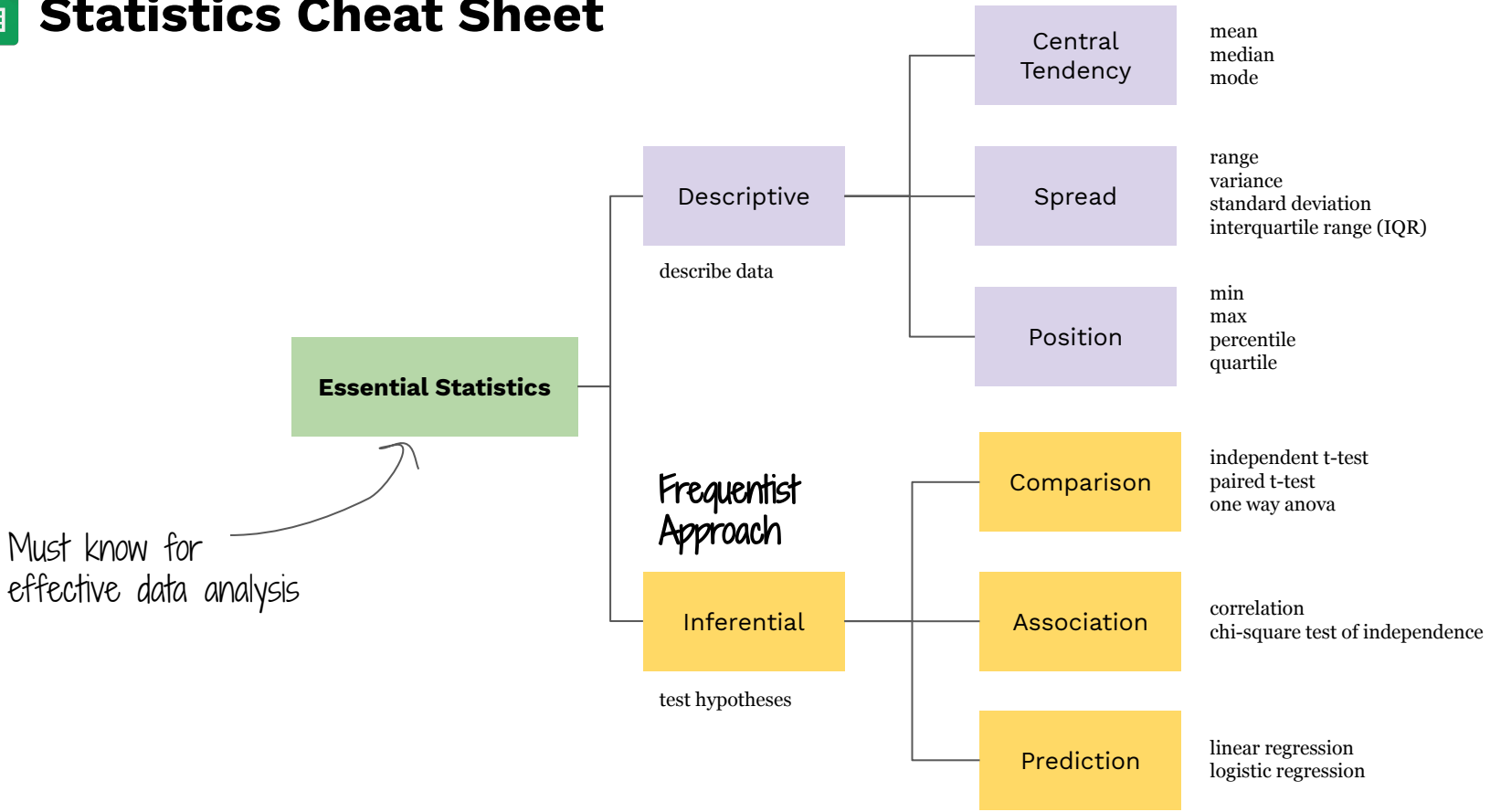
Our preferred tool to learn statistics

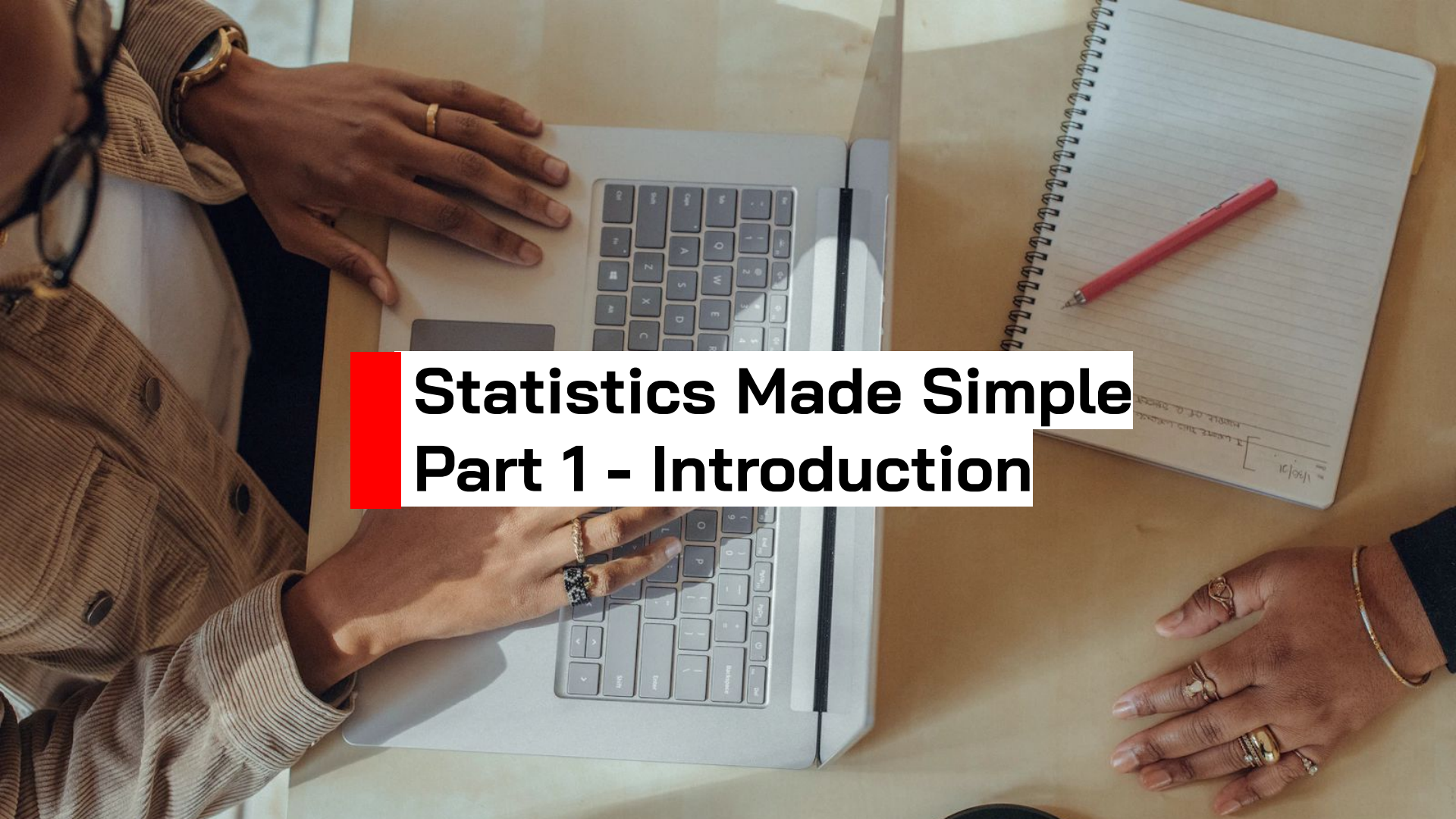
Inspired by Harvard

Harvard Business School Online

Spreadsheets is the best tool to learn statistics

Statistics Cheat Sheet




Statistics Made Simple
Part 1 - Introduction

What is statistics?

sta·tis·tics

/stəˈtistiks/

noun

the practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

sta·tis·tics

/stəˈtistiks/

noun

the practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

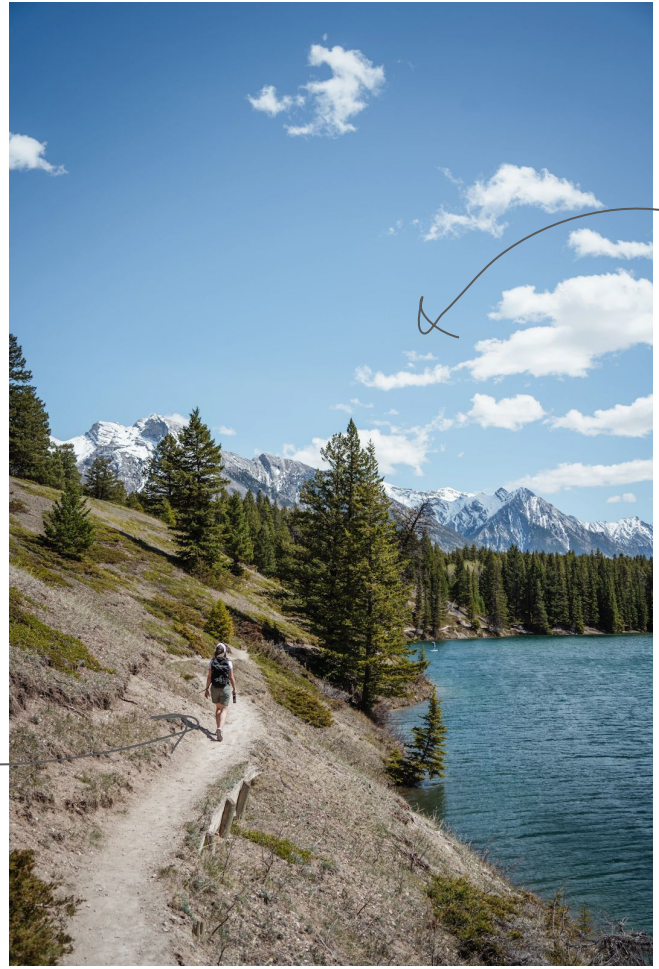
sta·tis·tics

/stəˈtistiks/

noun

the practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

Truth



Us

ที่มา https://www.facebook.com/fanmoph/posts/6229098400449796

VACCINES COMPARED

The Oxford University-AstraZeneca

TECHNOLOGY: Viral Vector (Genetically modified virus)

When injected, the vaccine instructs human cells to produce the SARS-CoV-2 spike protein — the immune system's main target in coronaviruses.

EFFICIENCY: 62-90%

PROCESS: Passed all three trials

MAJOR BUYERS: EU (400 million doses), US (300 million doses), UK (100 million doses)

THAILAND: 26 million doses

PRICE: US$4-5 per dose

DOSED REQUIRED: 2

Pfizer-BioNTech

TECHNOLOGY: mRNA

The new mRNA technology tricks the body into making the viral protein itself which, in turn, triggers an immune response

EFFICIENCY: 95%

PROCESS: Passed all three trials

MAJOR BUYERS: EU countries (200 million doses), US (100 million doses)

PRICE: US$20 per dose

DOSED REQUIRED: 2

Sinovac

TECHNOLOGY: Inactivated vaccine

Using the dead Covid-19 virus itself to trigger an immune response

EFFICIENCY: 50-70% (varies in tested countries)

PROCESS: Phase 3 trials

MAJOR BUYERS: Indonesia (40 million doses), Philippines (25 million doses)

THAILAND: 2 million doses

PRICE: US$5 per dose

DOSED REQUIRED: 2

Sputnik V (by Russia's Gamaleya Institute)

TECHNOLOGY: Adenoviral vector-based platform

The technology delivers the genetic instructions for SARS-CoV-2 antigens directly into patients' cells, triggering an immune response

EFFICIENCY: 91.4%

PROCESS: Phase 3 trials ongoing

MAJOR BUYERS: Brazil (10 million doses), Argentina (10 million doses) Bolivia (2.6 million doses), India (contracted to locally produce and distribute 100 million doses)

PRICE: US$10 per dose

DOSED REQUIRED: 2

Moderna

TECHNOLOGY: mRNA

A new type of vaccine which uses messenger RNA, which contains instructions for human cells to make proteins that mimic part of the coronavirus, to trigger an immune response.

EFFICIENCY: 95%

PROCESS: Passed all three trials

MAJOR BUYERS: EU (160 million doses), US (100 million doses), Canada (40 million doses)

PRICE: US$33 per dose

DOSED REQUIRED: 2

Johnson & Johnson

TECHNOLOGY: Uses a cold virus to deliver genetic material from the coronavirus into the body to prompt an immune response.

EFFICIENCY: Expected to be released by the end of January

PROCESS: Phase 3 clinical trials ongoing

MAJOR BUYERS: EU (160 million doses), US (100 million doses), Canada (40 million doses)

PRICE: Estimated US$10 per dose

DOSED REQUIRED: 1

Source: World Health Organization (WHO), agencies, respective companies

BANGKOK POST GRAPHICS

Vaccination overview

From Our World in Data · Last updated: 4 days ago

Thailand

Doses given	Fully vaccinated	% of population fully vaccinated
7.22M	1.97M	2.8%

Worldwide

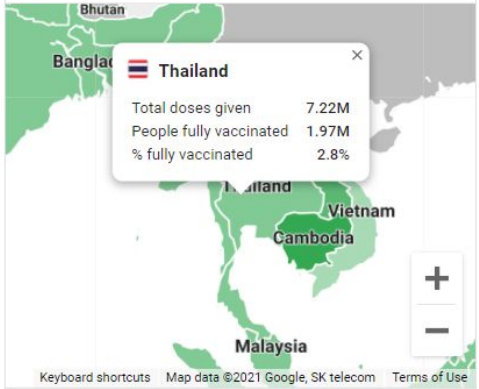
Doses given	Fully vaccinated	% of population fully vaccinated
2.5B	748M	9.6%

More locations and statistics

This data shows the total number of doses given in each location. Since some vaccines require more than one dose, the number of fully vaccinated people is likely to be lower. '+' shows data reported yesterday · About this data

Map of vaccinations

From Our World in Data · Last updated: 4 days ago



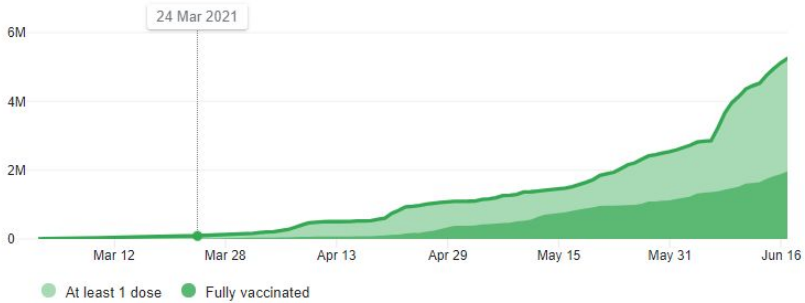
No data · Partial data · 0% · 2% · 5% · 10% · 20% · 30% · >40%

% of people fully vaccinated · About this data

Vaccinations

From Our World in Data · Last updated: 3 days ago

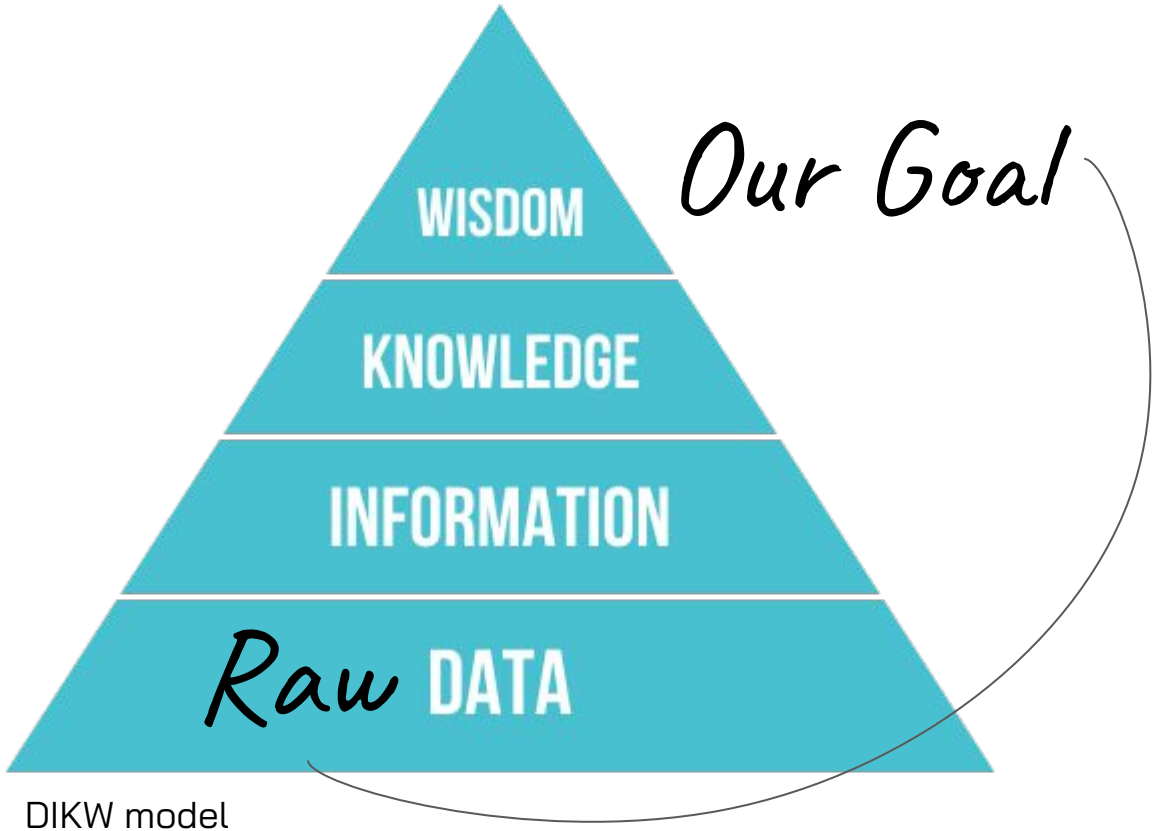
Total · Thailand · All time



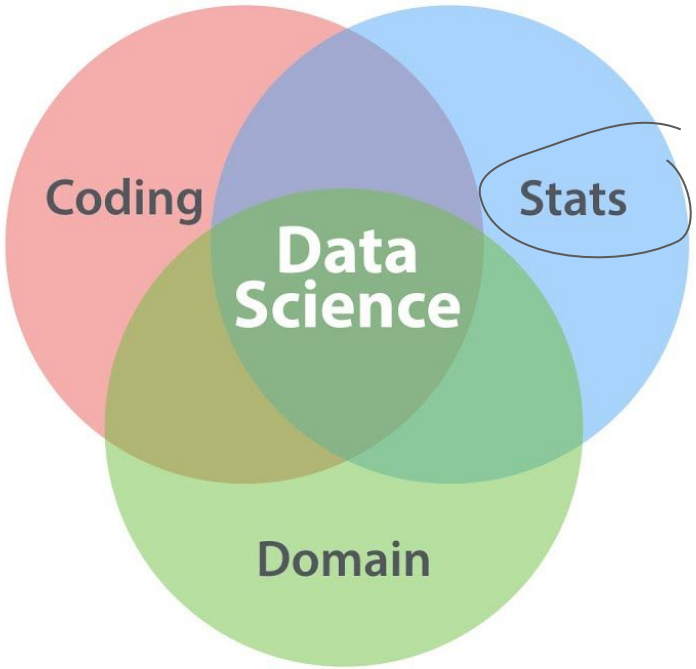
This data shows how many people have received at least one dose of a vaccine. People who are fully vaccinated may have received more than one dose. · About this data

More vaccine statistics

Statistics vs. Data Science



DIKW model

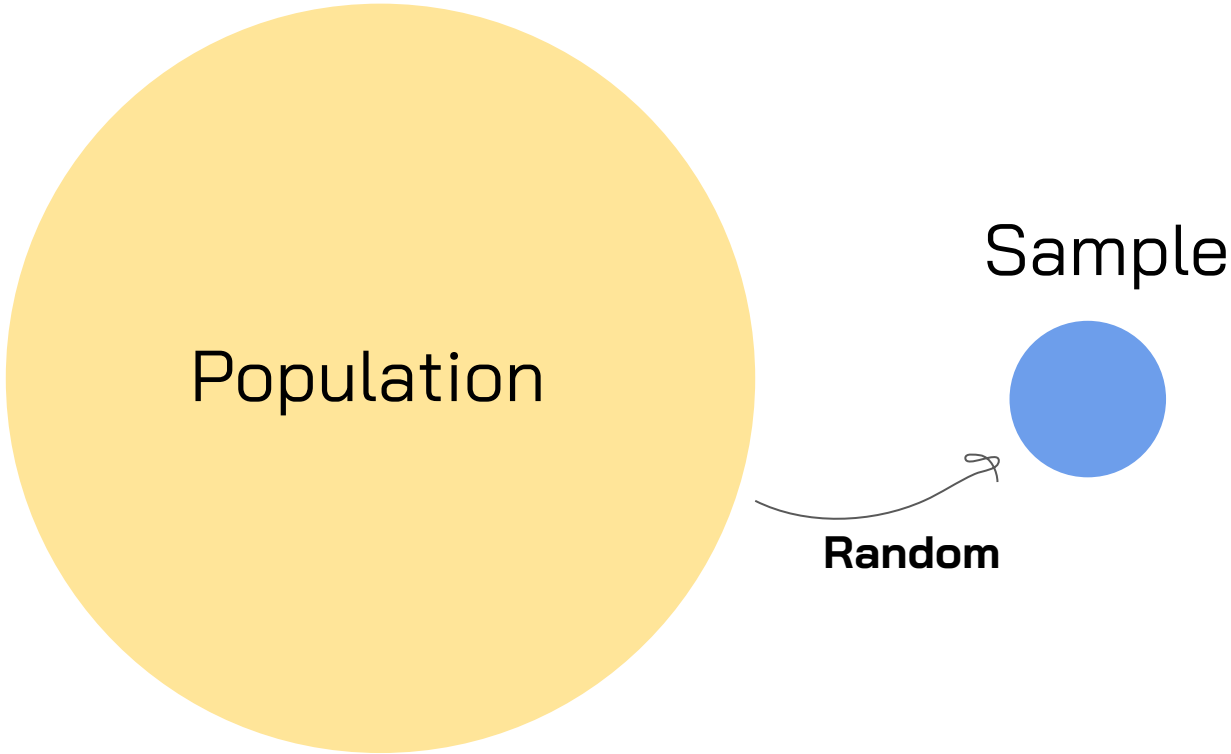


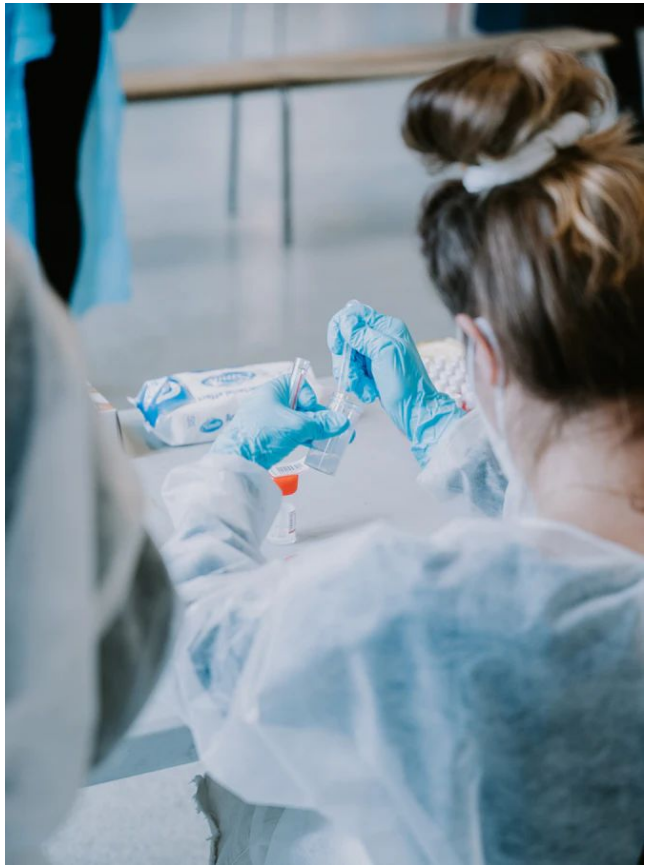
ที่มา Drew Conway

Data Science = {
Coding,
Statistics,
Domain Expertise
}

Descriptive vs. Inferential Statistics

Descriptive = Summarise
Inference = Generalize


Population
Sample
Random



Vaccine testing is a kind of **statistical test** (inference)

Because we cannot test vaccine with the whole population, we sample a few.

Case Study - Market Research

How to measure brand shares

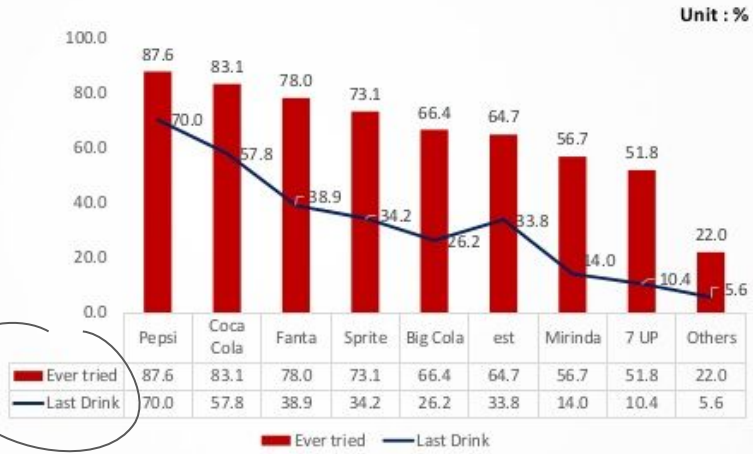
- Fact (e.g. real sales data)
- Claimed (e.g. research, survey)

Claimed

7. Soft drink's brands ever tried & last drink

87.6% ever tried Pepsi, and 70% drink Pepsi in the last month.

Although Coca Cola is the top-aware brand, however, Pepsi is the top brand receiving highest ever tried score and last drink score

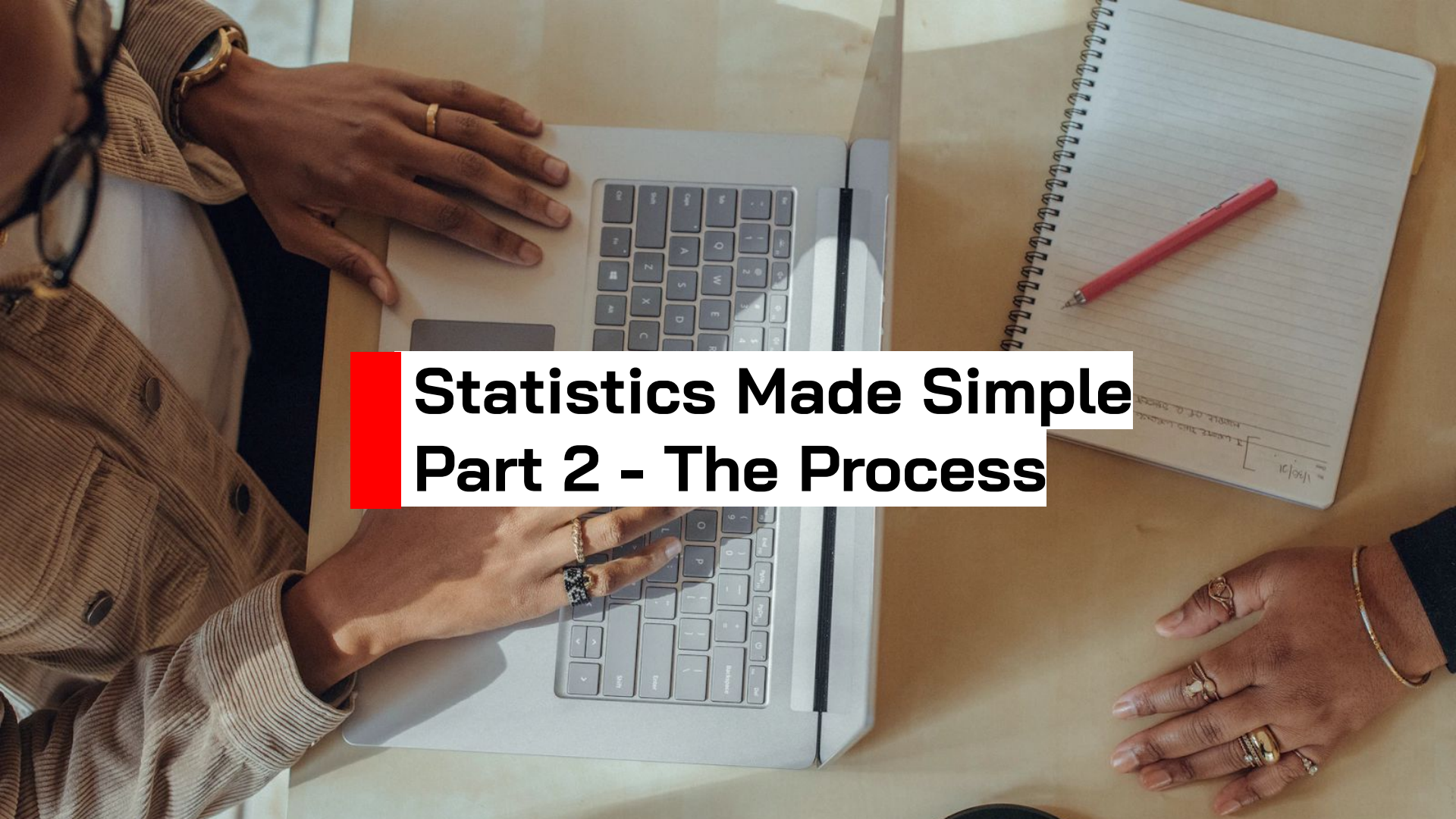


Q7. Among the list brands of soft drink, which one do you ever tried? [MA] n = 450
Q8. Among the list brands of soft drink, which one did you last drink last month? [MA] n = 450

ที่มา https://www.slideshare.net/WS-Thailand-Market-Research/soft-drink-marketstudiesthailand

AIDA Model

ที่มา https://thimpress.com/what-is-the-aida-model/


Statistics Made Simple
Part 2 - The Process


Ask good questions
Data Collection
Data Preparation
Data Analysis
Tell Your Story

It all starts with a **good** question

If you ask the **wrong** question, you are almost guaranteed to have the **wrong** answer.

ที่มา https://www.sportingnews.com/us/other-sports/news/takeru-kobayashi-hot-dog-champion/sbrtil9sbccv1ggve4zcg0wtp

25.8 in 12 minutes

50 in 12 minutes

Good question =>

Good action =>

Win the game


Ask good questions
Data Collection
Data Preparation
Data Analysis
Tell Your Story

Sources of Data

- Business
- Survey
- Internet/ Social Media/ UGC
- Mobile Phone (sensor, IoT)
- Public Datasets

Tesco/ Dunnhumby - The very first company who uses data analytics and science with commercially success

ID	Amount	Store	Customer

History [edit]

In 1993, Terry Leahy asked the Tesco marketing team to investigate the potential of loyalty cards. In the past Tesco had run Green Shield Stamps as a promotional tool which rewarded people for visits and spend but gained no customer information. The initial team, led by Grant Harrison, researched programmes across the world and developed a proposal which showed that a loyalty card could be very effective. The key change since the days of Green Shield Stamps was the ability to track individual customer behaviour cost-effectively using a magnetic stripe card.

In 1994, Harrison attended a conference where Clive Humby from marketing firm dunnhumby was speaking. Dunnhumby was already working with clients such as Cable & Wireless and BMW, and Harrison approached them to help with the loyalty card project.[2] Successful trials throughout 1994 led to the Tesco board asking Harrison and Humby to present to the annual Board strategy session.

The first response from the board came from Tesco's then chairman Lord MacLaurin, who said, "What scares me about this is that you know more about my customers after three months than I know after 30 years."[3]

Strategic Data Acquisition


Ask good questions
Data Collection
Data Preparation
Data Analysis
Tell Your Story

Common Tasks

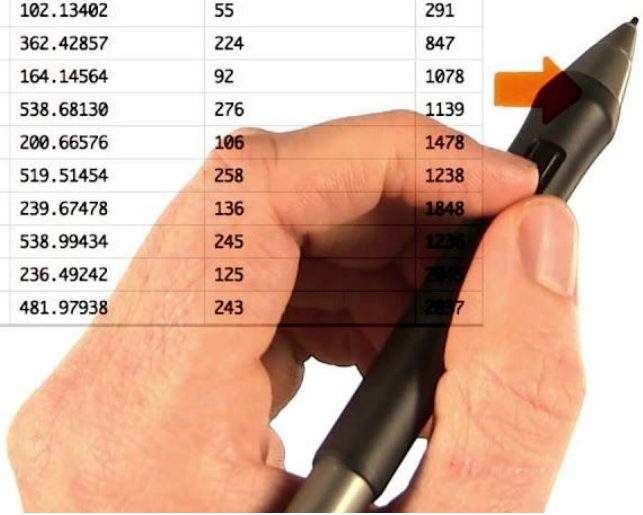
- Clean
- Transform
- Create new features
- Check outliers

ID	Color	Weight	Broken	Class
1	Black	80	Yes	1
2	Yellow	100	No	2
3	Yellow	120	Yes	2
4	Blue	90	No	2
5	Blue	85	No	2
6	**?**	60	No	1
7	Yellow	100	**?**	2
8	**?**	40	**?**	1

Reshape Data

Long

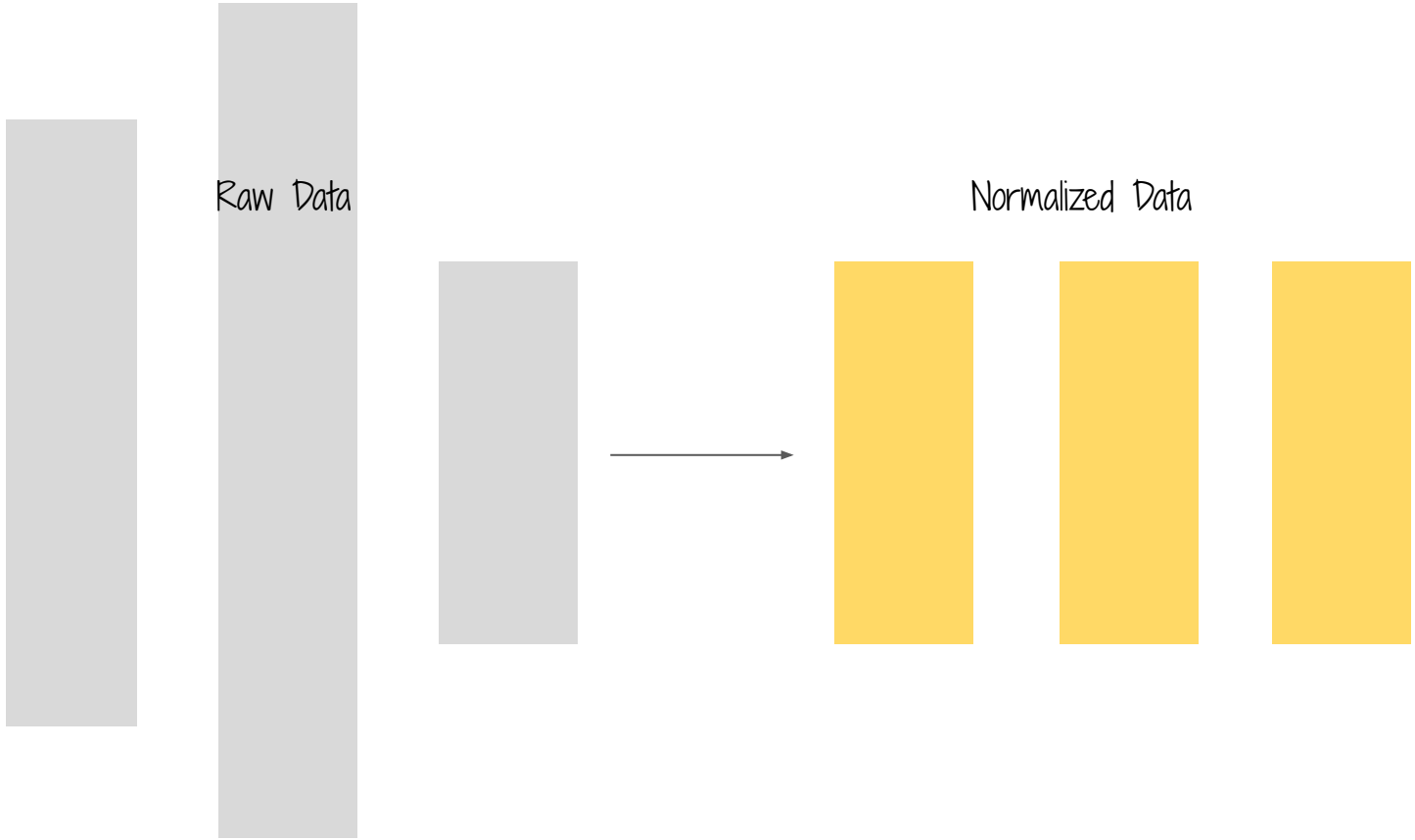
	age	gender	mean_friend_count	median_friend_count	n
1	13	female	259.16062	148	193
2	13	male	102.13402	55	291
3	14	female	362.42857	224	847
4	14	male	164.14564	92	1078
5	15	female	538.68130	276	1139
6	15	male	200.66576	106	1478
7	16	female	519.51454	258	1238
8	16	male	239.67478	136	1848
9	17	female	538.99434	245	1236
10	17	male	236.49242	125	2649
11	18	female	481.97938	243	2827

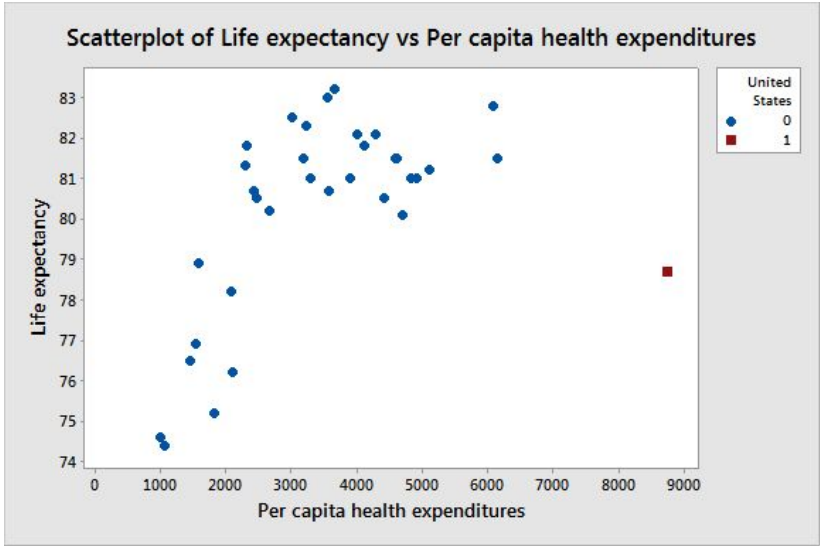


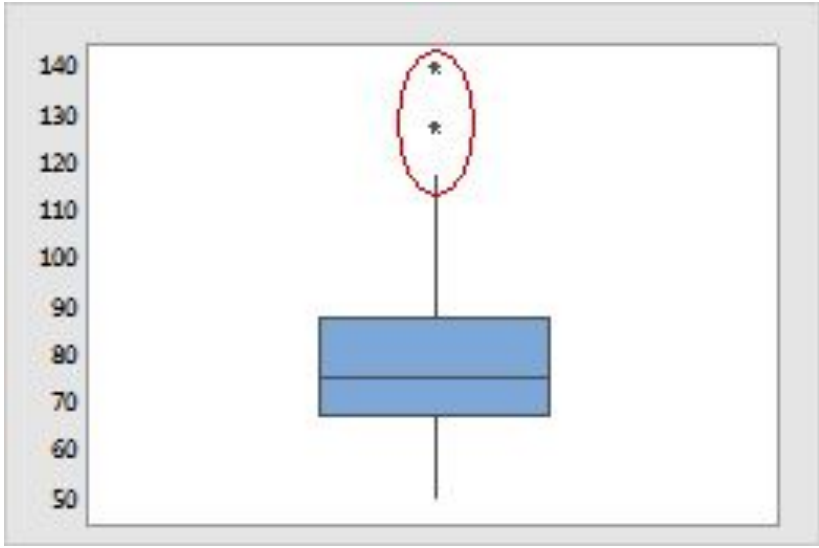
Wide

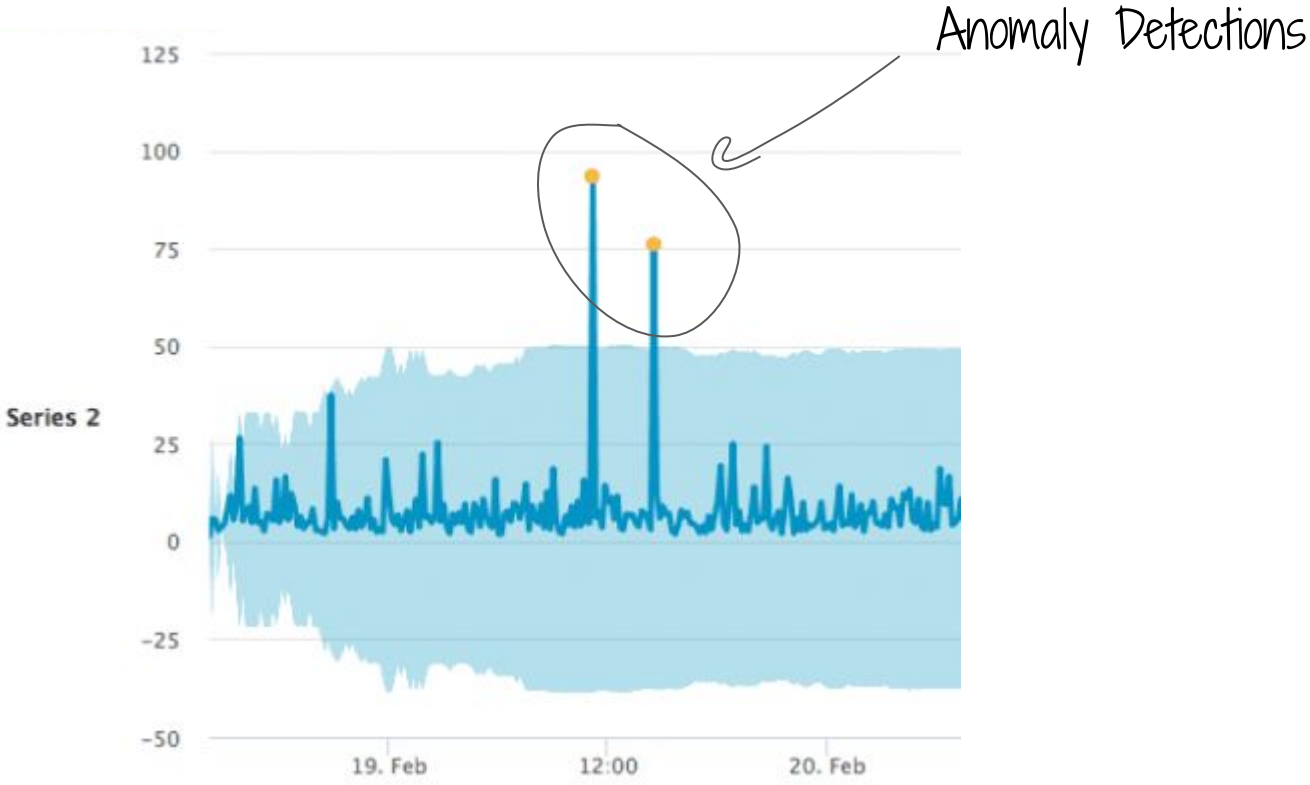
age	male	female
13		
14		
15		
⋮		

ที่มา https://www.youtube.com/watch?v=zlaeISxRESQ


Raw Data
Normalized Data


Scatterplot of Life expectancy vs Per capita health expenditures
Life expectancy
Per capita health expenditures
United States
0
1
83
82
81
80
79
78
77
76
75
74
0
1000
2000
3000
4000
5000
6000
7000
8000
9000


140
130
120
110
100
90
80
70
60
50



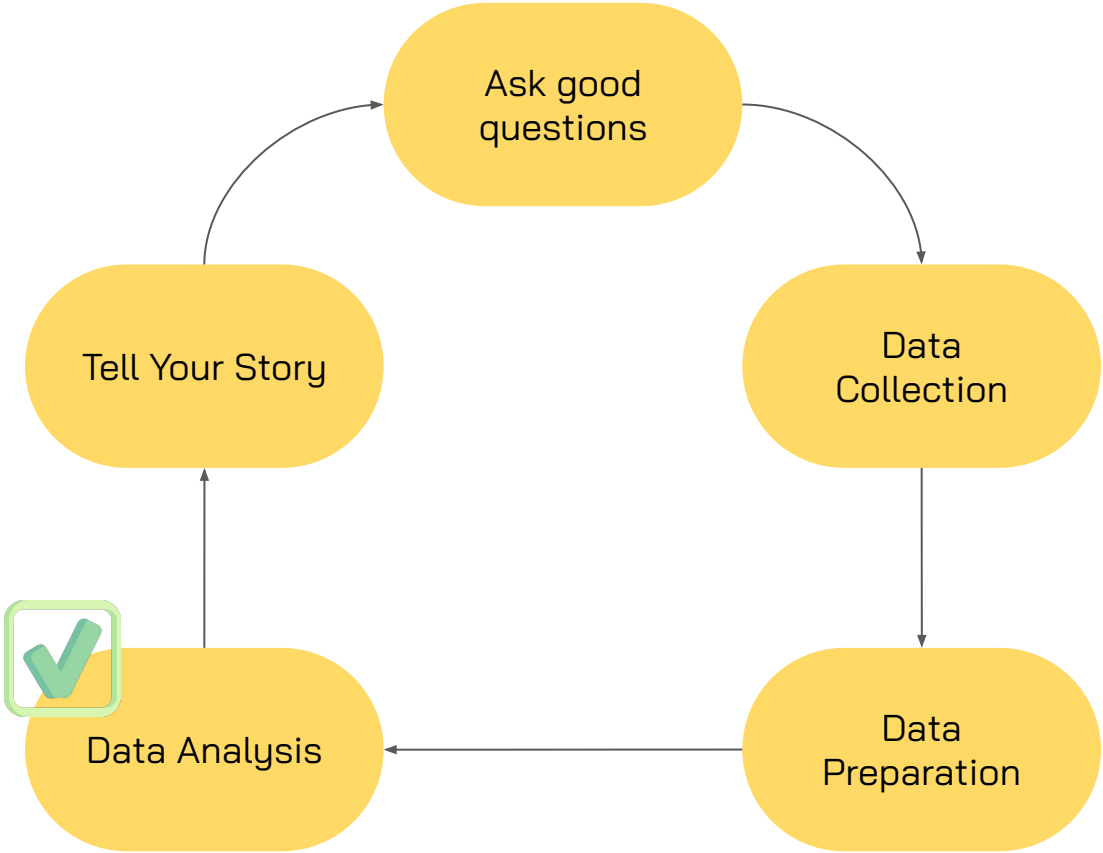
ที่มา https://docs.splunk.com/Documentation/MLApp/5.2.1/User/DNOlegacyassist

What is meant by anomaly detection?

Anomaly detection is the identification of rare events, items, or observations which are suspicious because they differ significantly from standard behaviors or patterns. **Anomalies** in data are also called standard deviations, outliers, noise, novelties, and exceptions.

avinetworks.com › glossary › anomaly-detection

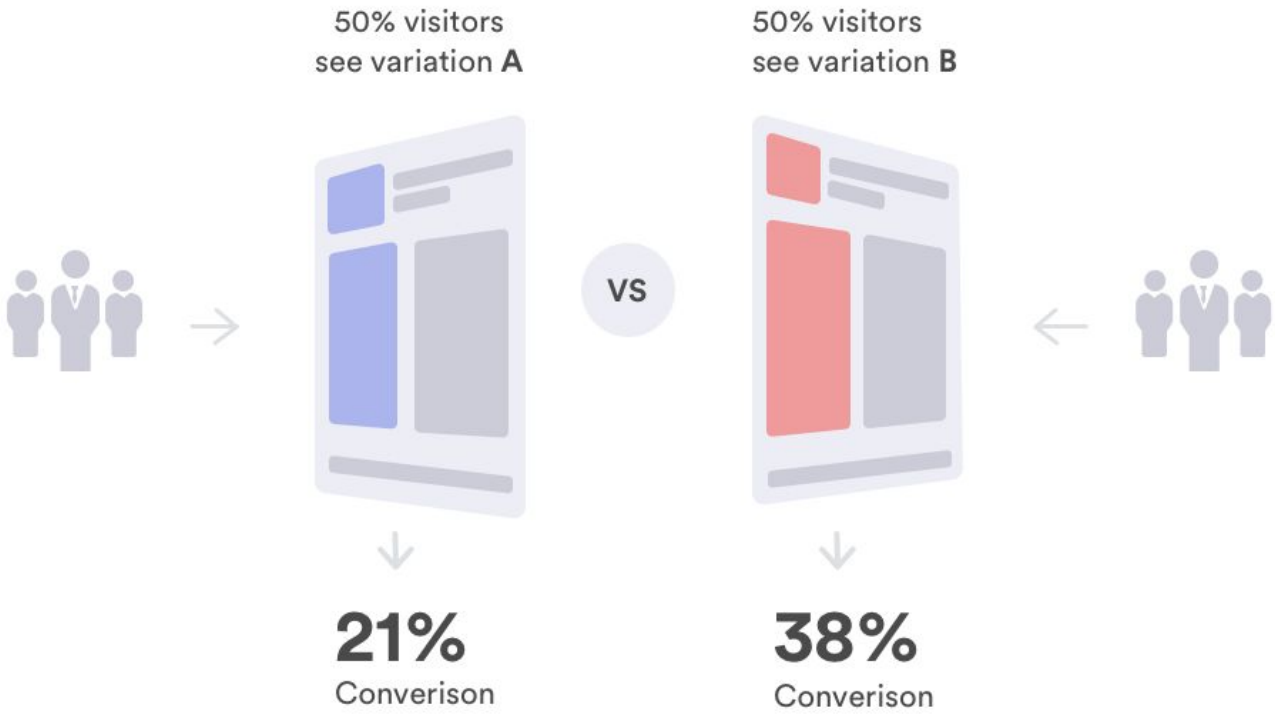
What is Anomaly Detection? Definition & FAQs | Avi Networks


Ask good questions
Data Collection
Data Preparation
Data Analysis
Tell Your Story

Common Techniques

- Descriptive statistics
- Inferential statistics
  - AB Test
  - Correlation and Regression
  - Hypothesis Testing
- Machine learning
- Mathematical modeling

A/B Testing

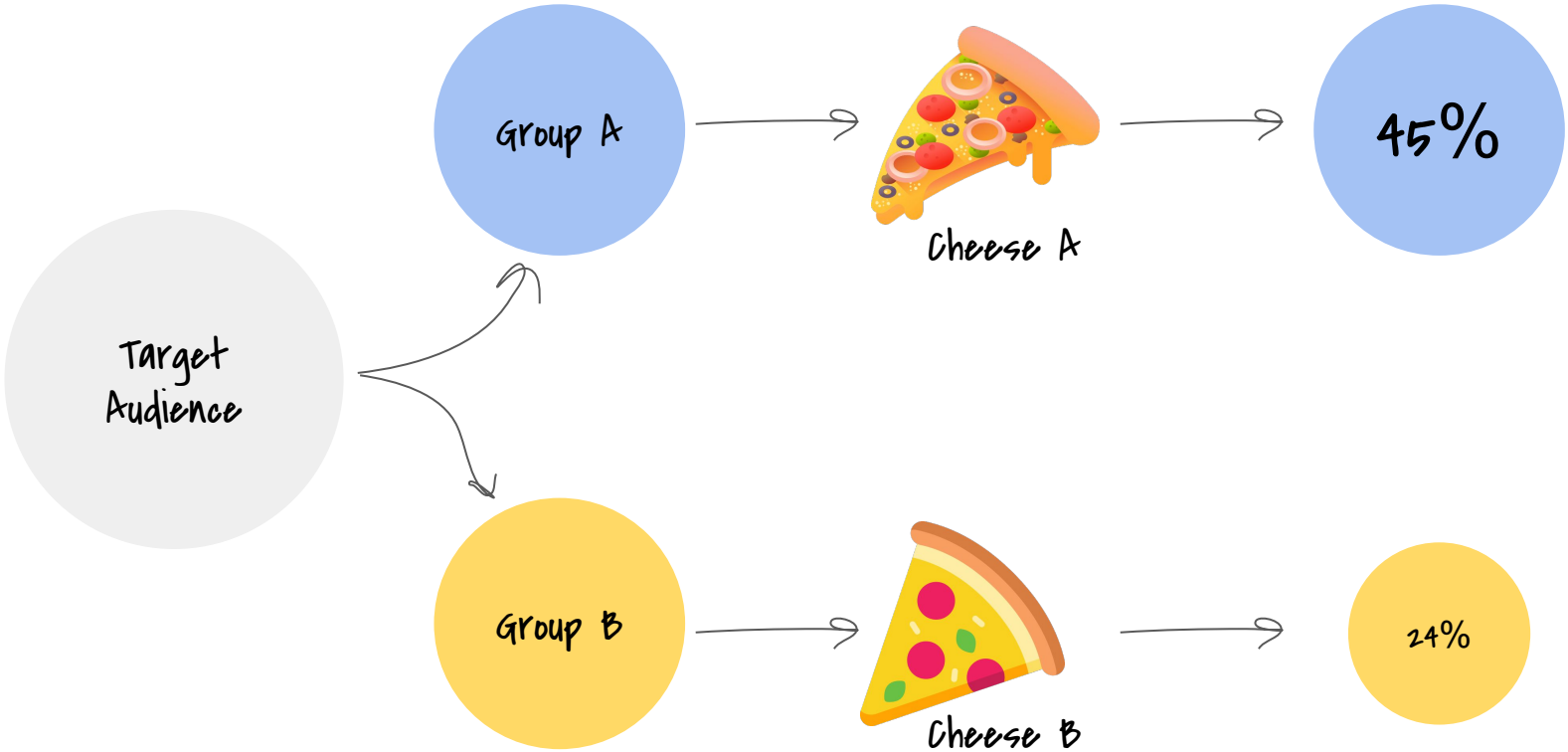

50% visitors
see variation A
50% visitors
see variation B
VS
21%
Converison
38%
Converison

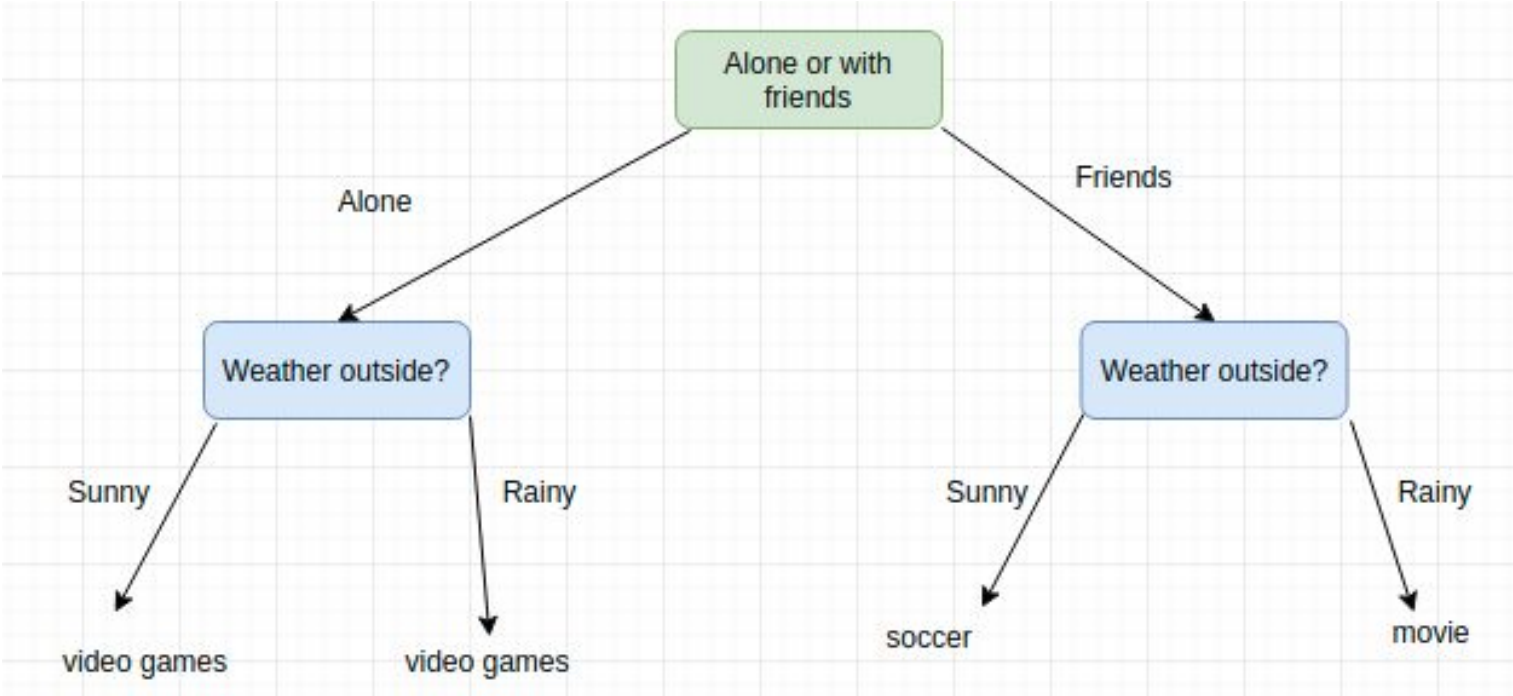
ORIGINAL DESIGN

PERSON DESIGN

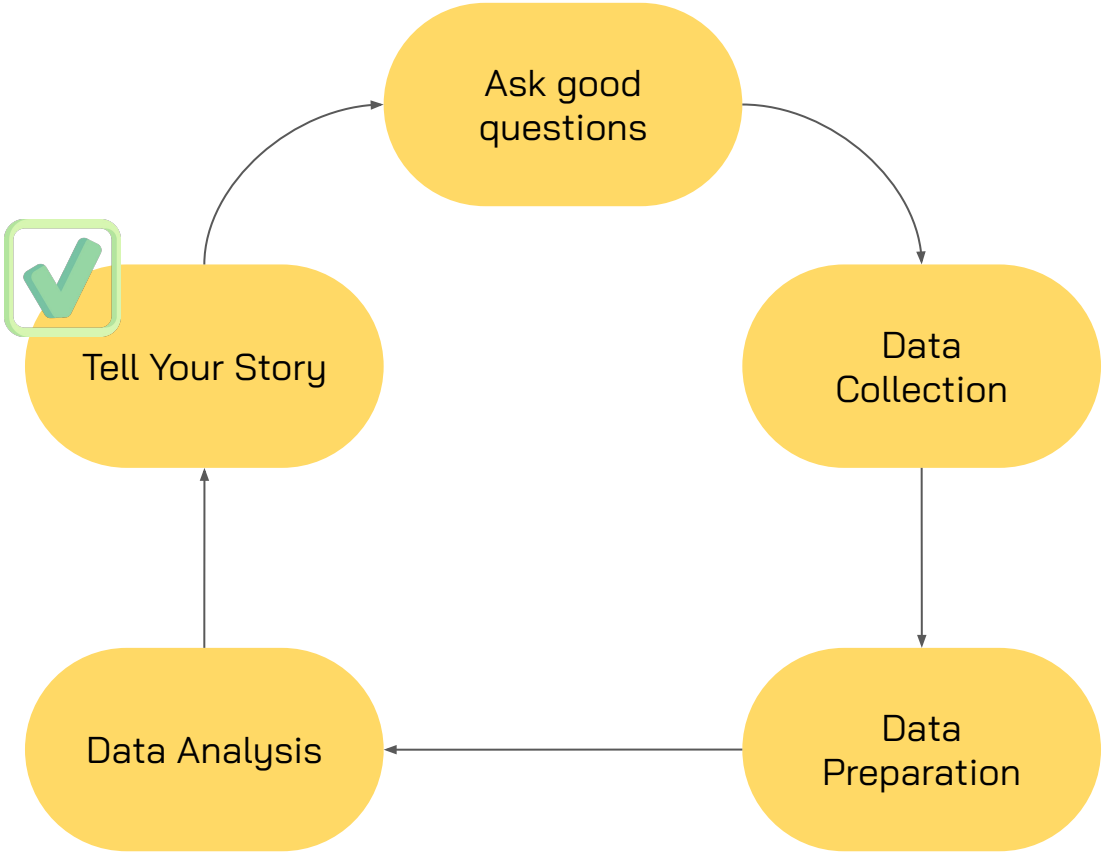
102.5%

ที่มา https://www.shopify.com/blog/6-product-photography-tests


Target Audience
Group A
Group B
Cheese A
Cheese B
45%
24%



ที่มา https://towardsdatascience.com/a-guide-to-decision-trees-for-machine-learning-and-data-science-fe2607241956


Ask good questions
Data Collection
Data Preparation
Data Analysis
Tell Your Story

1. Know your audience
2. Use image/ charts
3. Story has three acts

ที่มา หนังสือ Storytelling with Data (Wiley)

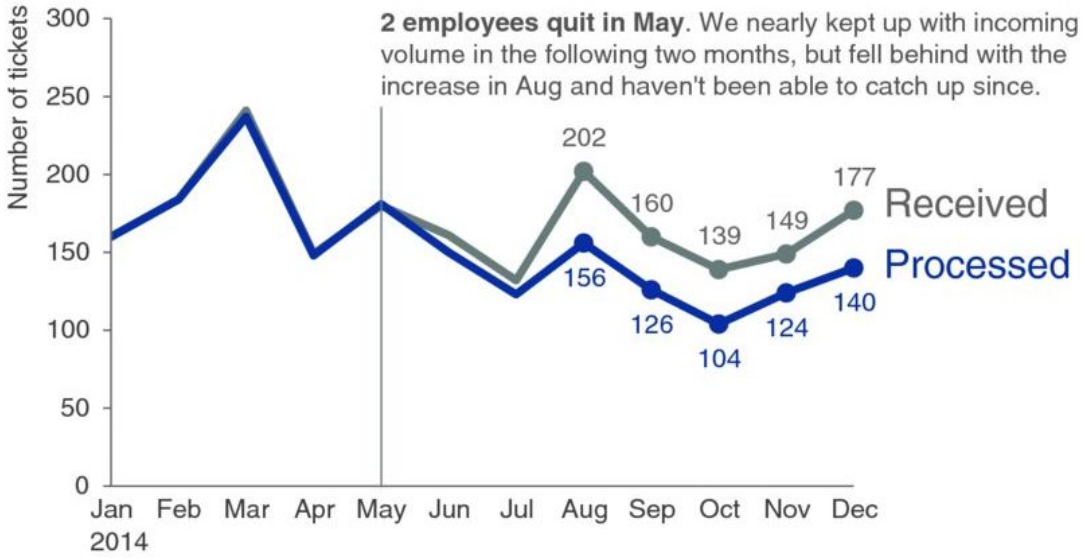

Story has **three acts**
Beginning > Middle > Ending


Story has **three acts**
Background > Problem > Solution

Please approve the hire of 2 FTEs

to backfill those who quit in the past year

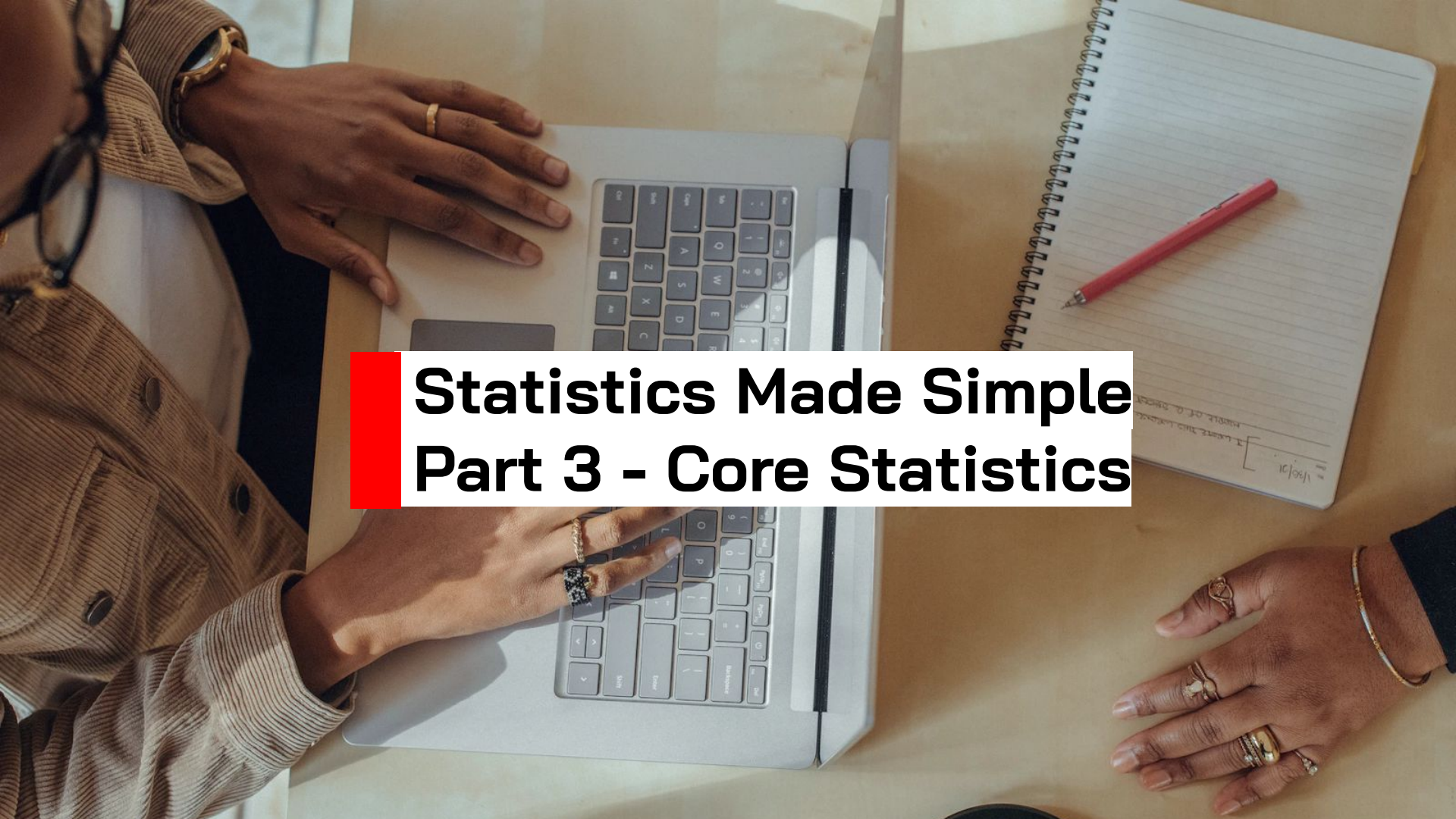
Ticket volume over time



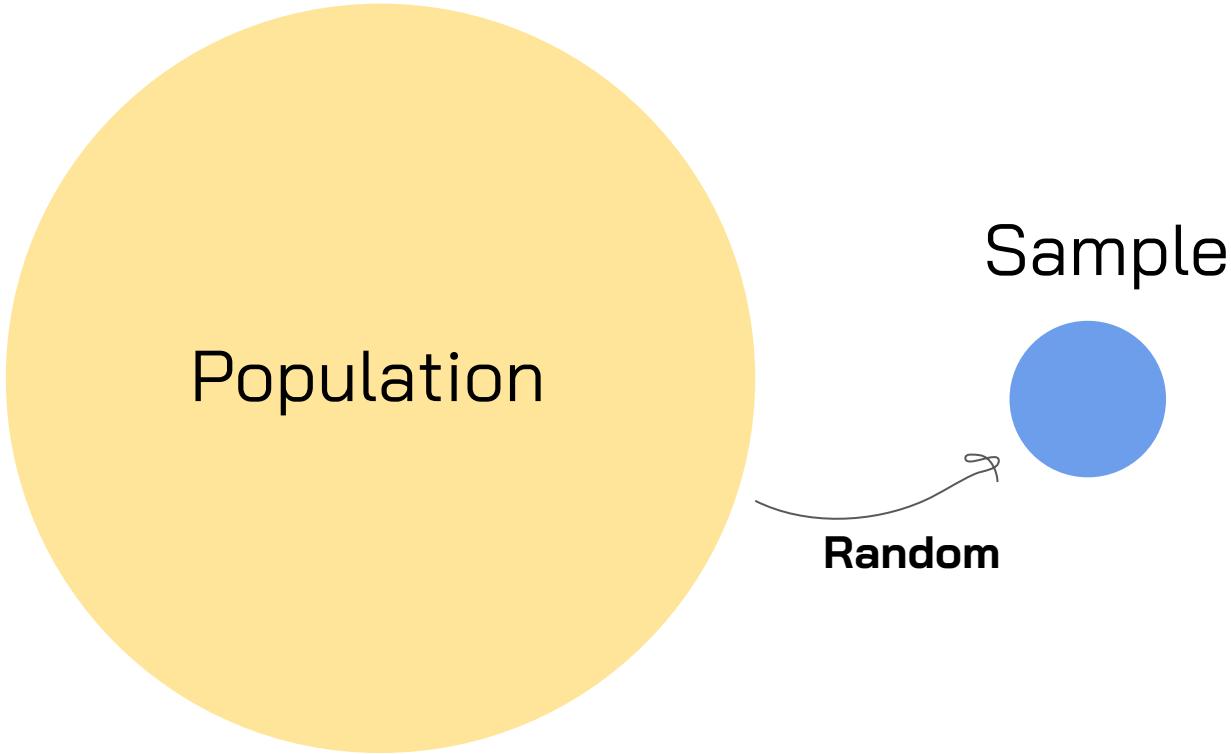
Data source: XYZ Dashboard, as of 12/31/2014 I A detailed analysis on tickets processed per person and time to resolve issues was undertaken to inform this request and can be provided if needed.

ที่มา หนังสือ Storytelling with Data (Wiley)

Good Data Analysis will lead to **Actions**


Statistics Made Simple
Part 3 - Core Statistics

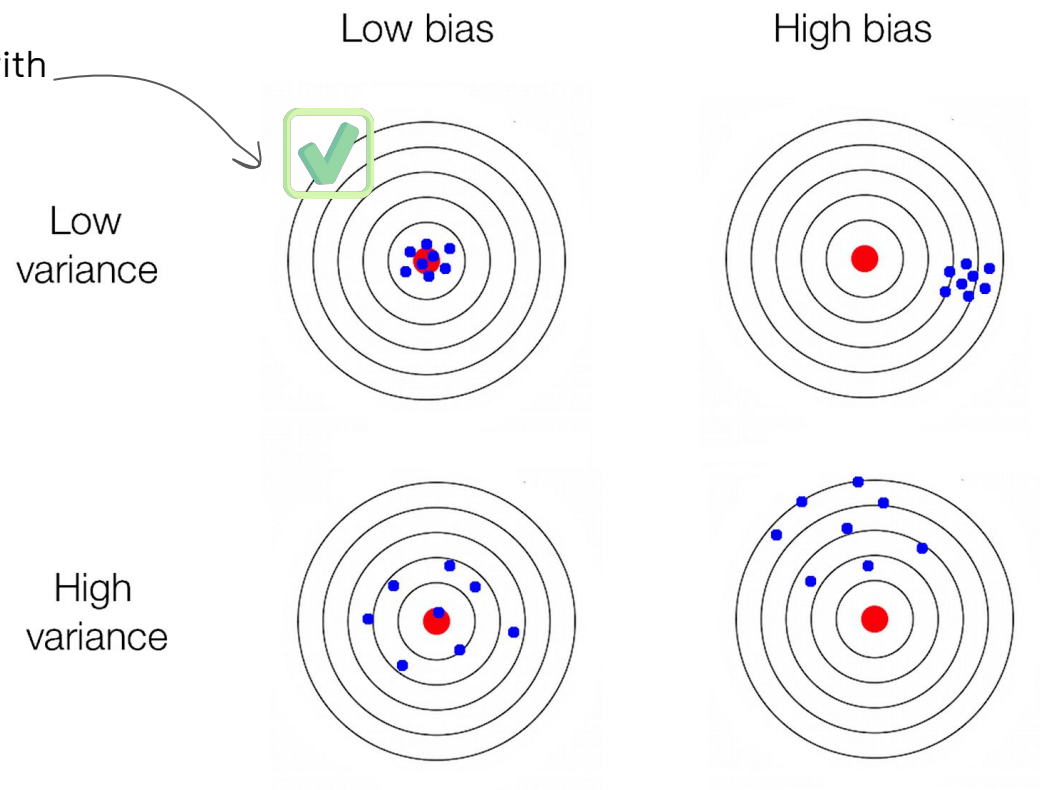
Random Sampling


Population
Sample
Random

Random sampling **reduce bias** and make our sample **more representative**

More accurate result

Random sampling with large sample size



ที่มา https://medium.com/@aymantidy/bias-variance-decomposition-the-story-behind-86a20160c3c


Sample
Population

Measures of Central Tendency

Central Tendency

- Mean
- Median
- Mode

16, 42, 50, 88, 120

Mean = (16+42+50+88+120)/ 5 = **63.2**

16, 42, 50, 88, 120

Median = 50

16, 42, 50, 60, 88, 120

Median = (50+60) / 2 = **55**

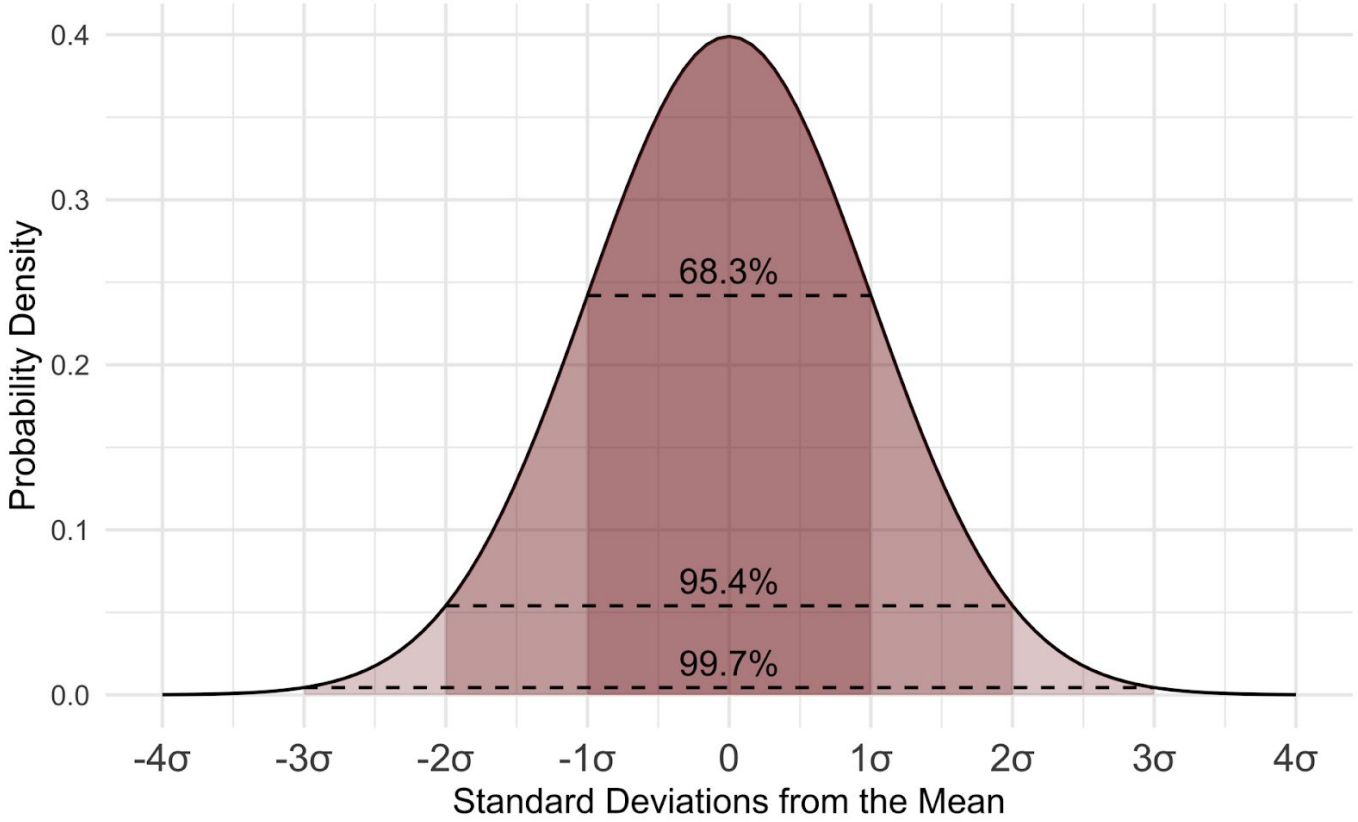
16, 42, 50, 50, 50, 88, 120

Mode = 50

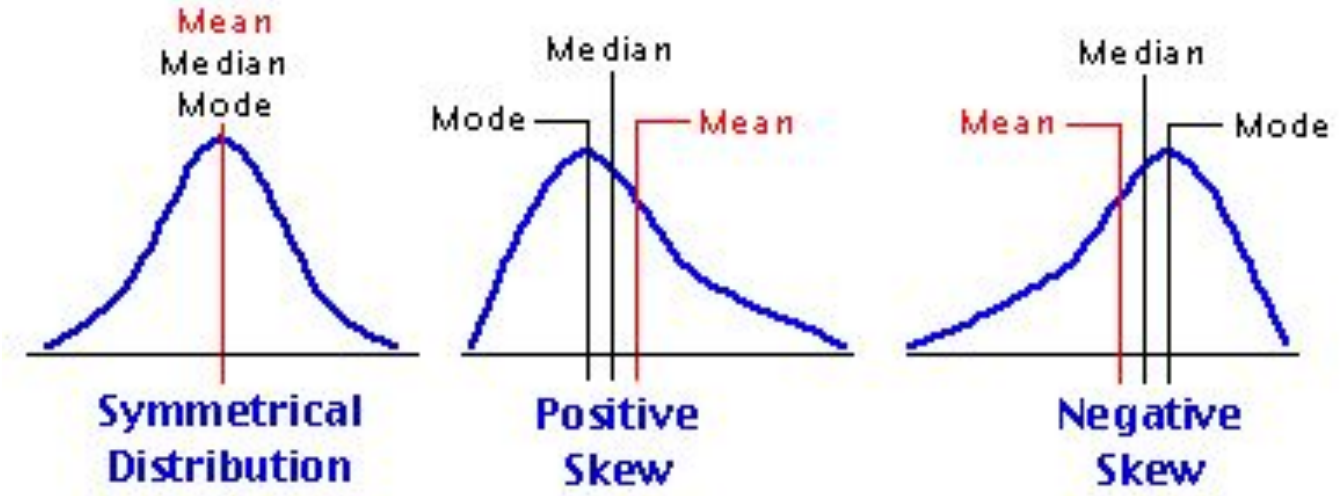
16, 42, 50, 50, 88, 120, 120

Mode = 50, 120

Bimodal (Multimodal)



ที่มา https://en.wikipedia.org/wiki/Statistics



ที่มา http://analystnotes.com/cfa-study-notes-symmetry-and-skewness-in-return-distributions.html

Use **median** instead of mean if the data is skewed

Measures of Spread

Spread or Variability

- Range
- Variance
- Standard Deviation

16, 42, 50, 88, 120

Range = 120 - 16 = 104

Variance Formula

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

16, 42, 50, 88, 120

Variance = 1673.2

	A	B	C	D	E	F	G
1	X	X_Bar	Diff	Diff^2		Sum(Diff^2) / (n-1)	
2	16	63.2	-47.2	2227.84		1673.2	
3	42	63.2	-21.2	449.44			
4	50	63.2	-13.2	174.24			
5	88	63.2	24.8	615.04			
6	120	63.2	56.8	3226.24			

SD Formula

$$s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N - 1}}$$

16, 42, 50, 88, 120

SD = SQRT(Variance) = 40.9

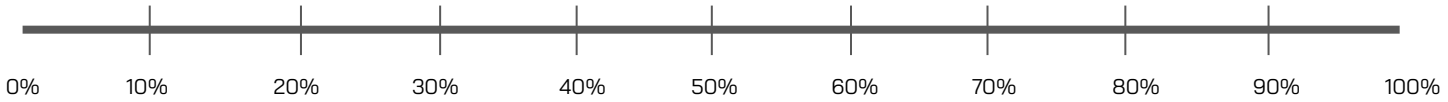
	A	B	C	D	E	F	G
1	X	X_Bar	Diff	Diff^2		Sum(Diff^2) / (n-1)	
2	16	63.2	-47.2	2227.84		1673.2	
3	42	63.2	-21.2	449.44			
4	50	63.2	-13.2	174.24		SQRT(1673.2)	
5	88	63.2	24.8	615.04		40.90477	
6	120	63.2	56.8	3226.24			

Measures of Position

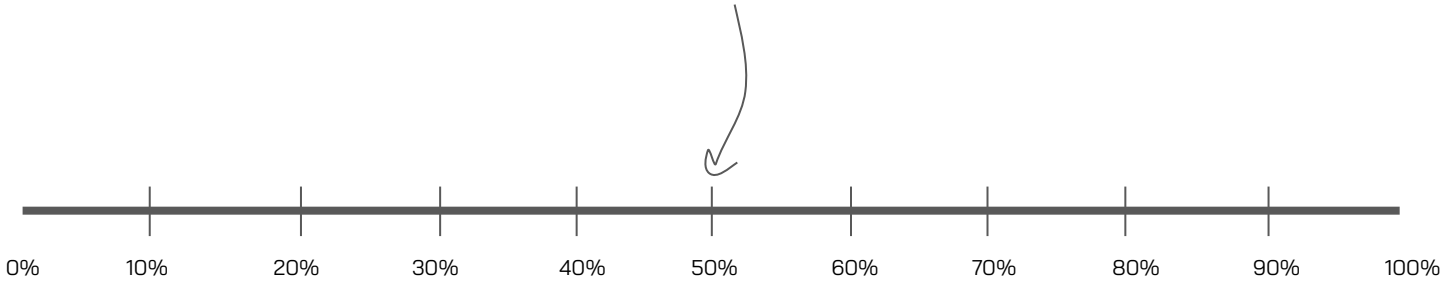
Positions

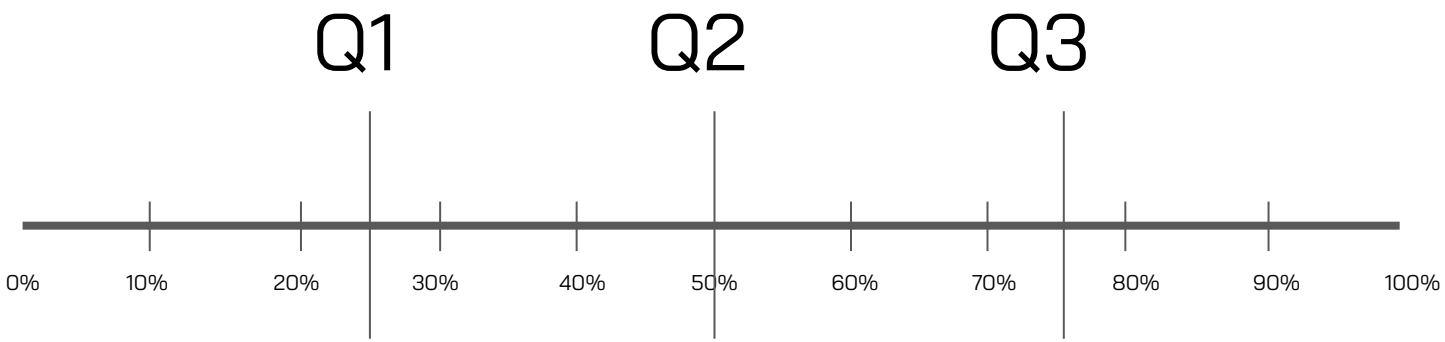
- Minimum, Maximum
- Percentile
- Quartile

Min Max



Median

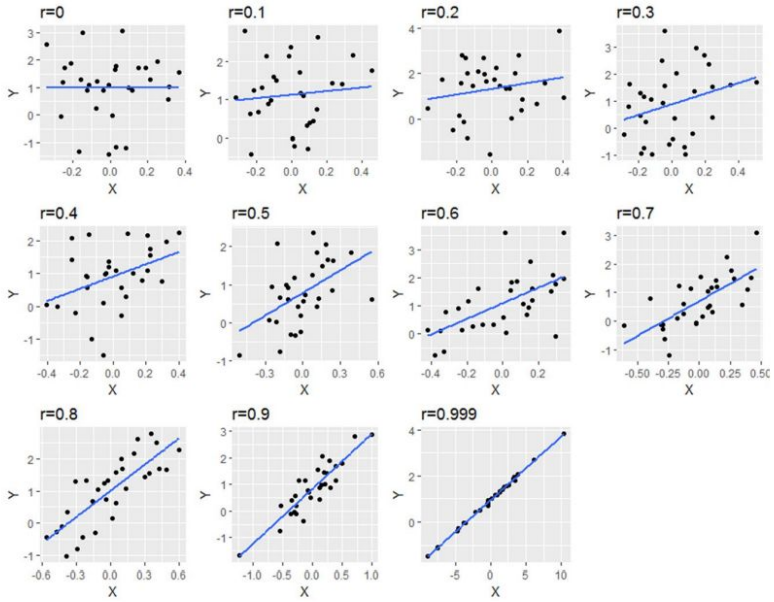



Q1
Q2
Q3
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%

Measures of Relationship

Relationships

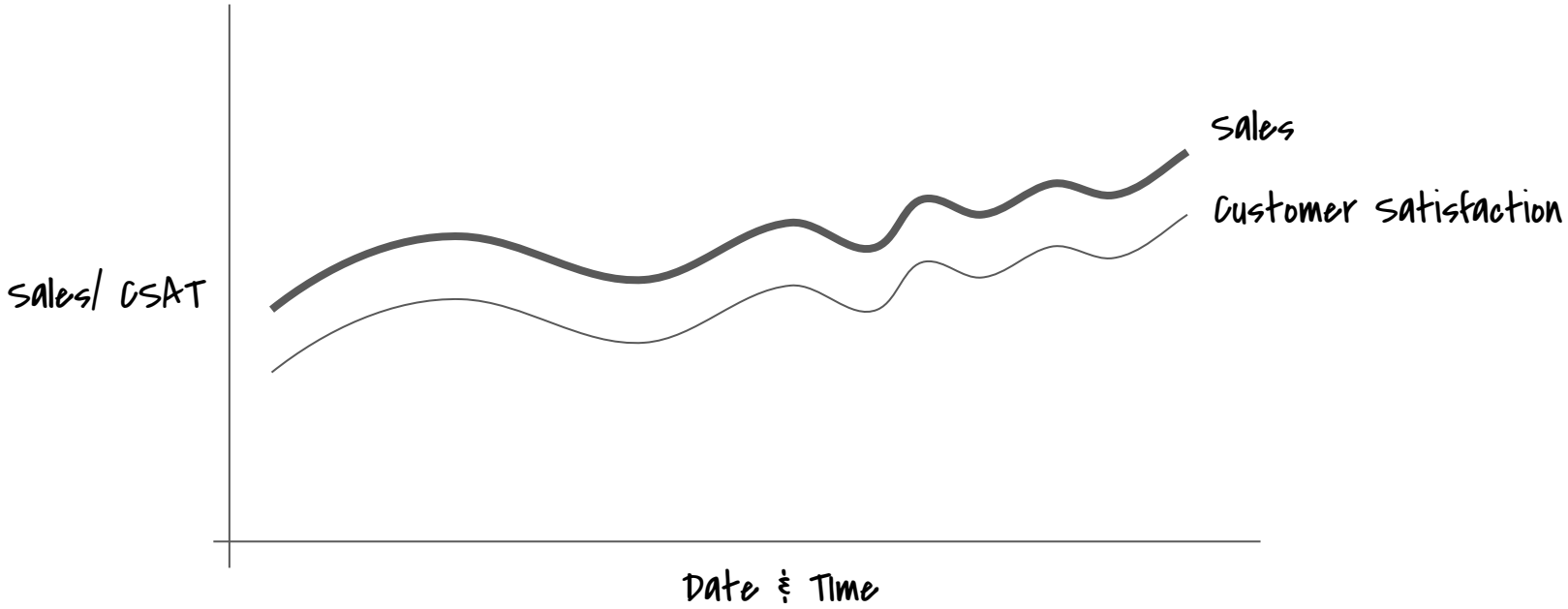
- Correlation
- Cross Tabulation



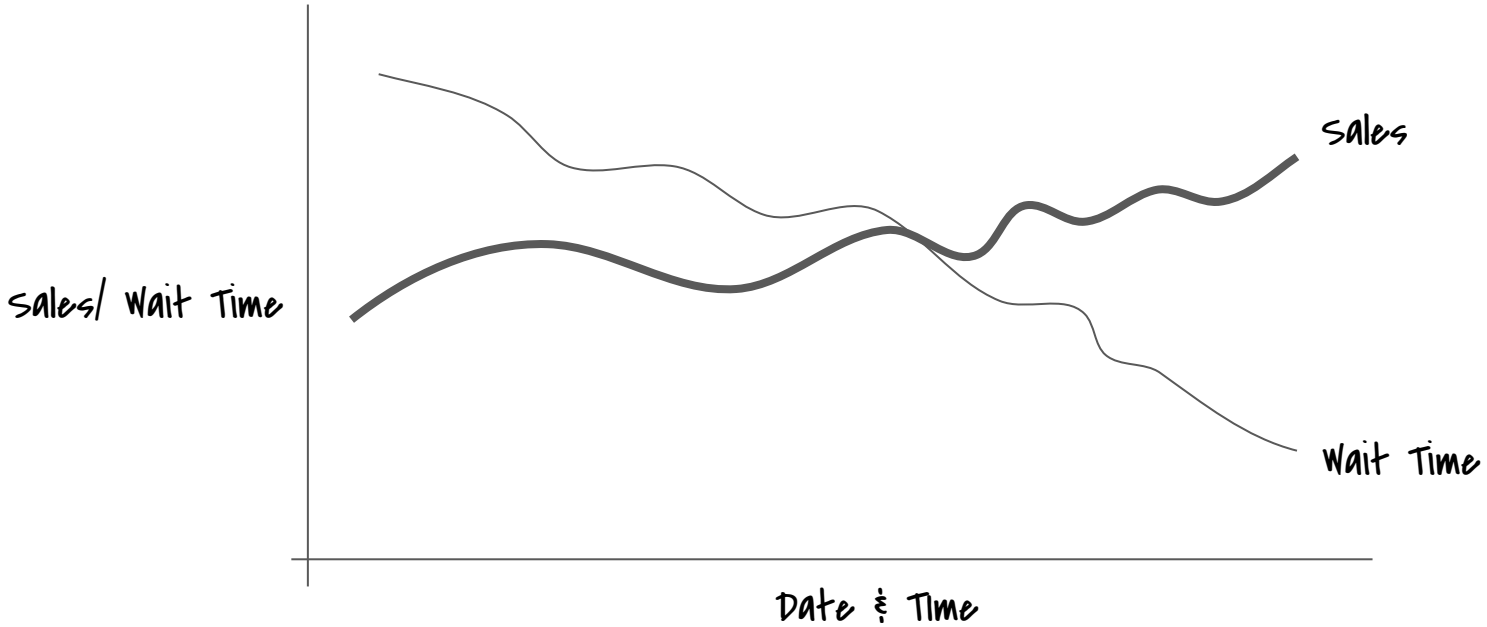
Correlation

- [-1, +1]
- 0 means no relationship

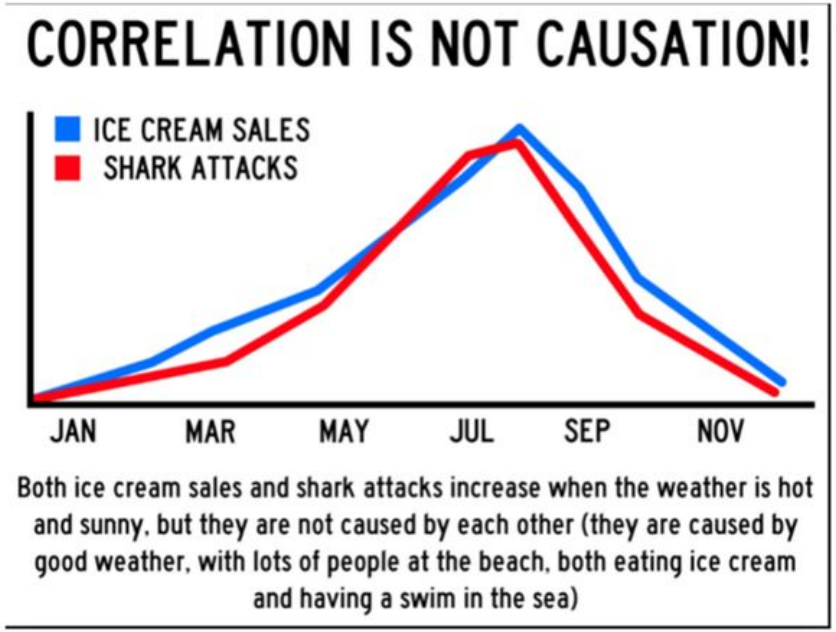
Positive



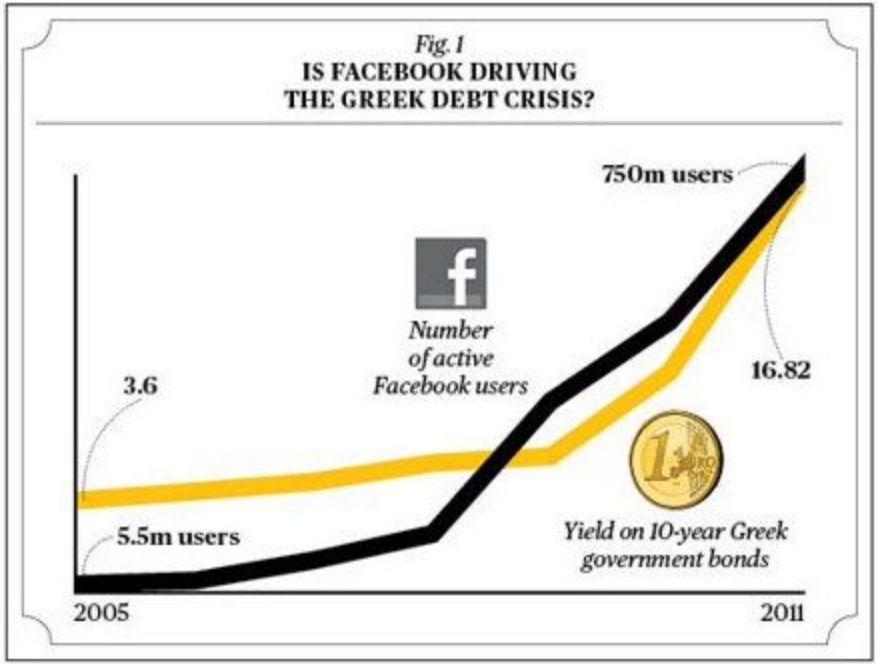
Negative



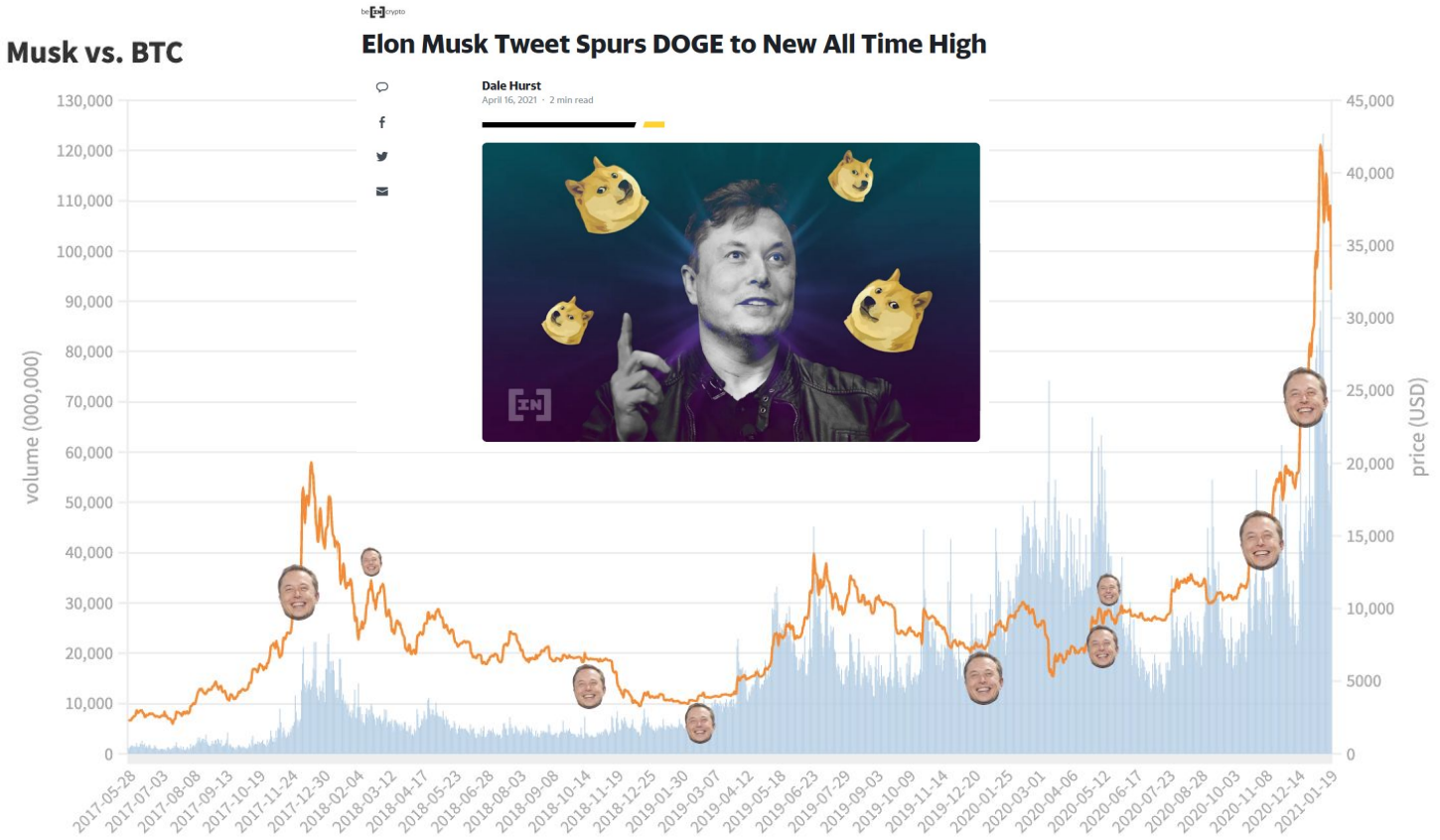
Correlation *does not* imply **Causation**



ที่มา https://www.kaggle.com/general/187094



ที่มา https://michaelnielsen.org/ddi/if-correlation-doesnt-imply-causation-then-what-does/



ที่มา https://finance.yahoo.com/news/elon-musk-tweet-spurs-doge-140000208.html

	Like Scary Movies	
	Yes	No
Girls	32	38
Men	30	12
Total	62	50

Cross Tabulation

To summarise two qualitative variables

Gender x Like Scary Movies

	Like Scary Movies	
	Yes	No
Girls	32	38
Men	30	12
Total	62	50

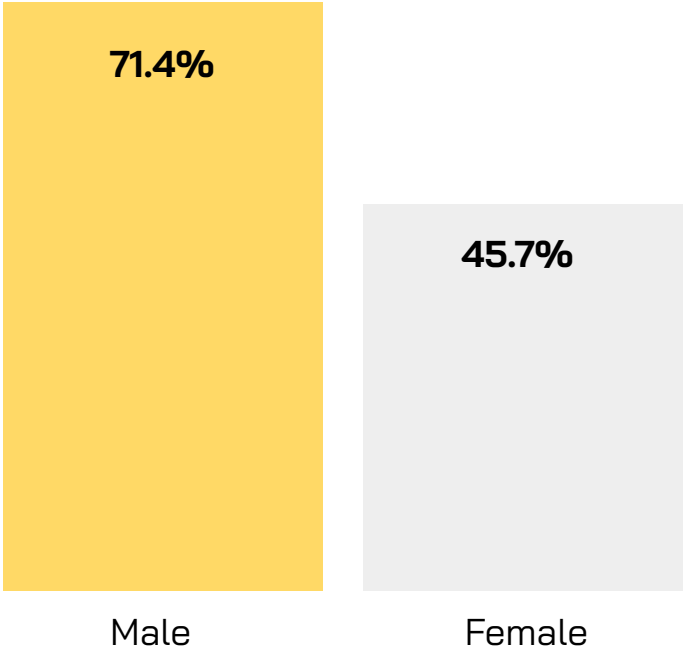
% Girls who like scary movies

= 32 / (32 + 38) = **45.7%**

	Like Scary Movies	
	Yes	No
Girls	32	38
Men	30	12
Total	62	50

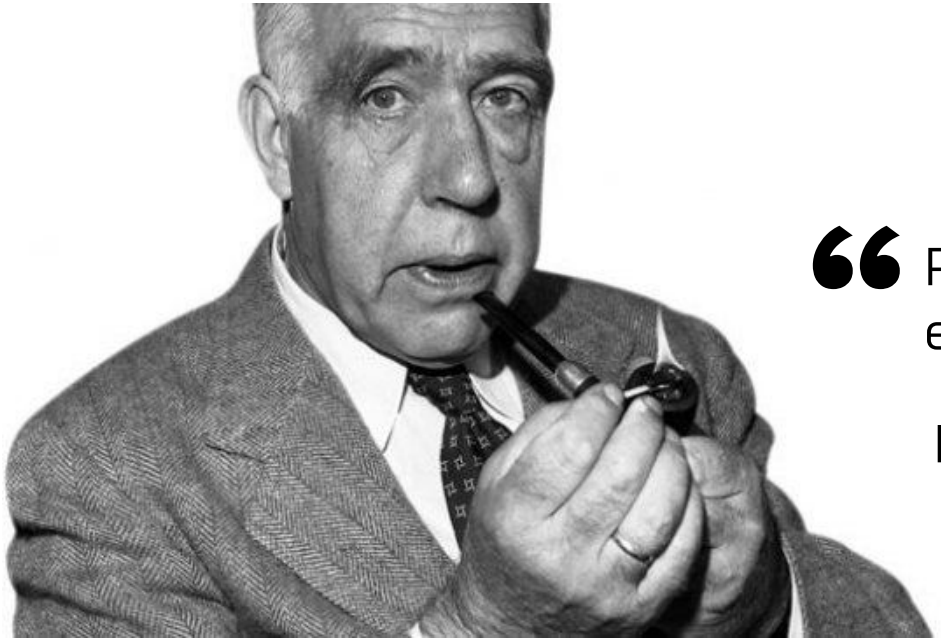
% Men who like scary movies

= 30 / (30 + 12) = **71.4%**



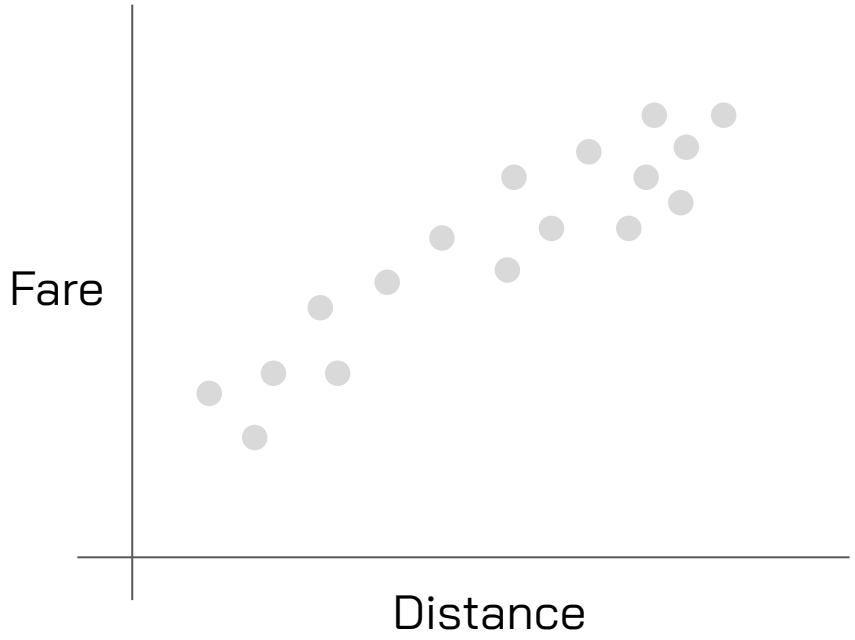
Male seems to like scary movies more than female **significantly**

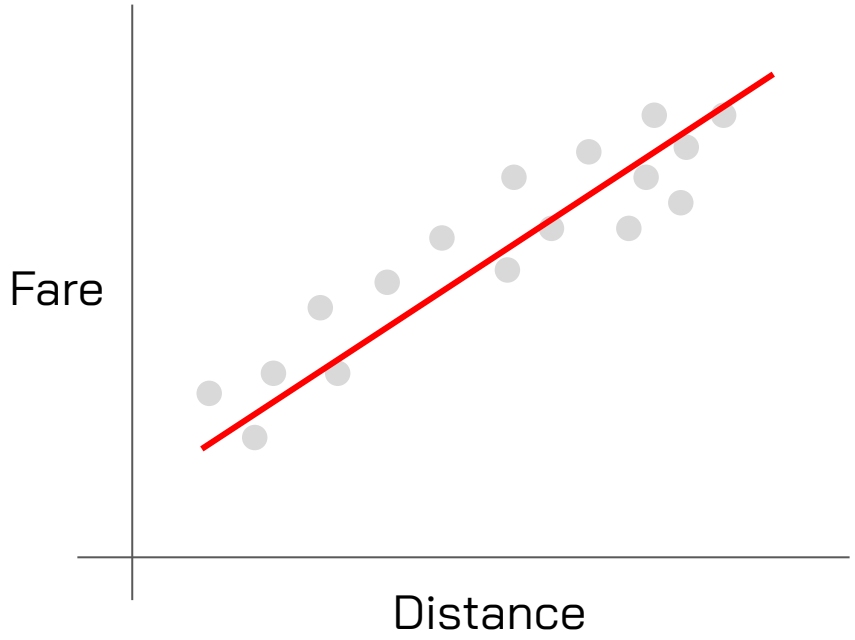
Prediction Techniques



> Prediction is very difficult,
> especially if it’s about the future.
>
> Niels Bohr

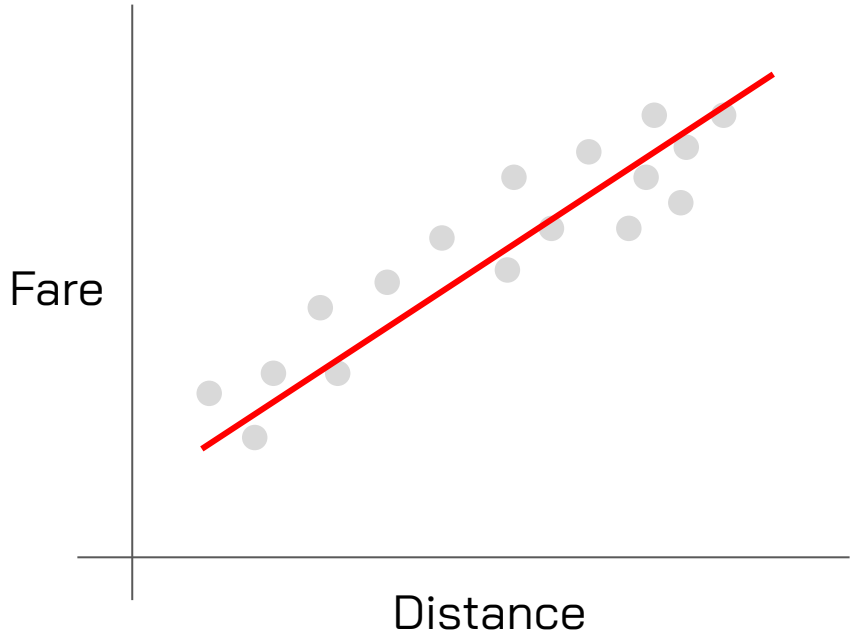
Taxi Fare **Prediction**


Fare
Distance



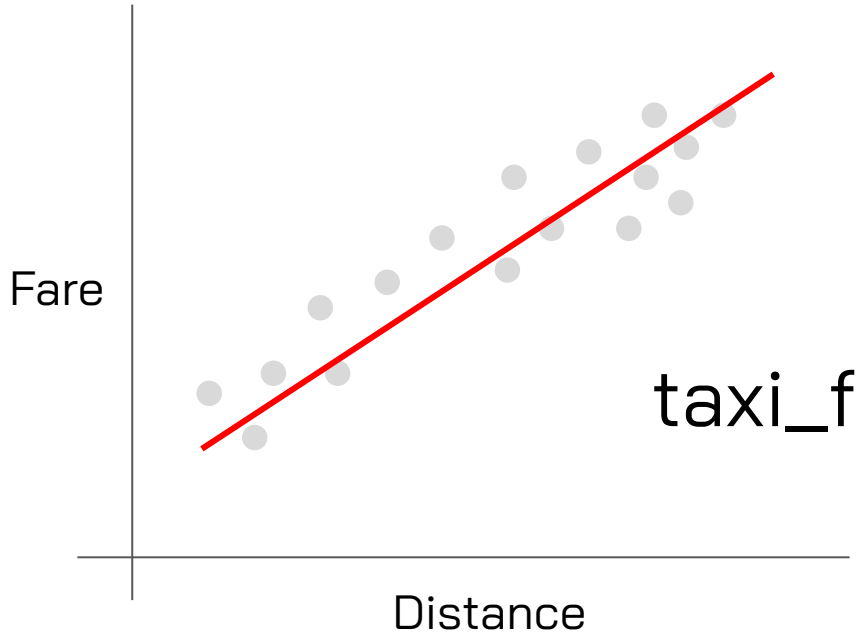
Linear Regression

is the **best fitted line** through our data

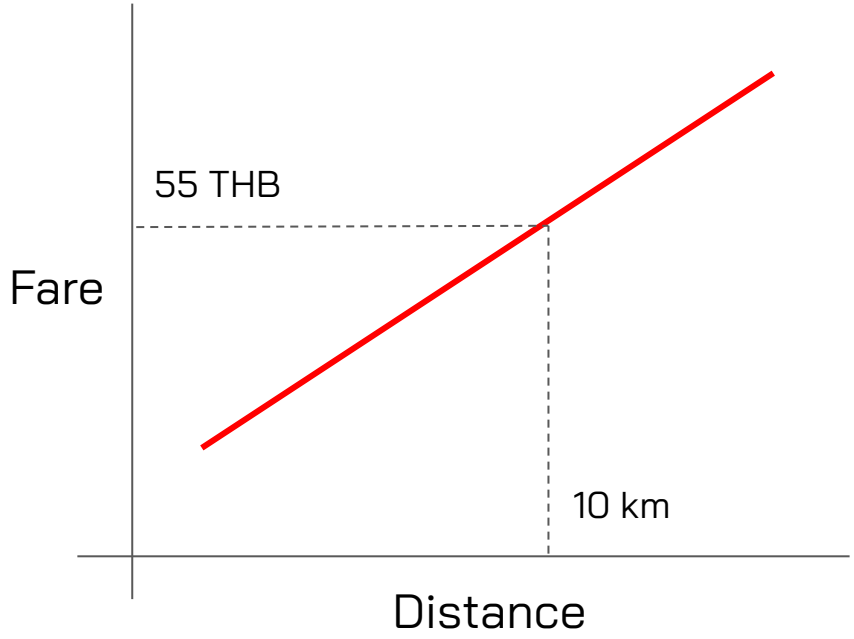


$$y = \mathbf{b0} + \mathbf{b1}^{*}x1$$

b0 = intercept
b1 = slope



taxi_fare = **35** + **2***distance



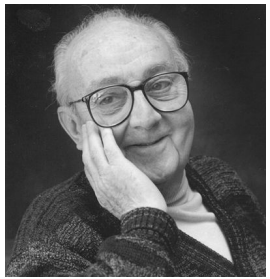
Linear Regression

can be used to **predict taxi fare**

taxi_fare = **35** + **2***distance

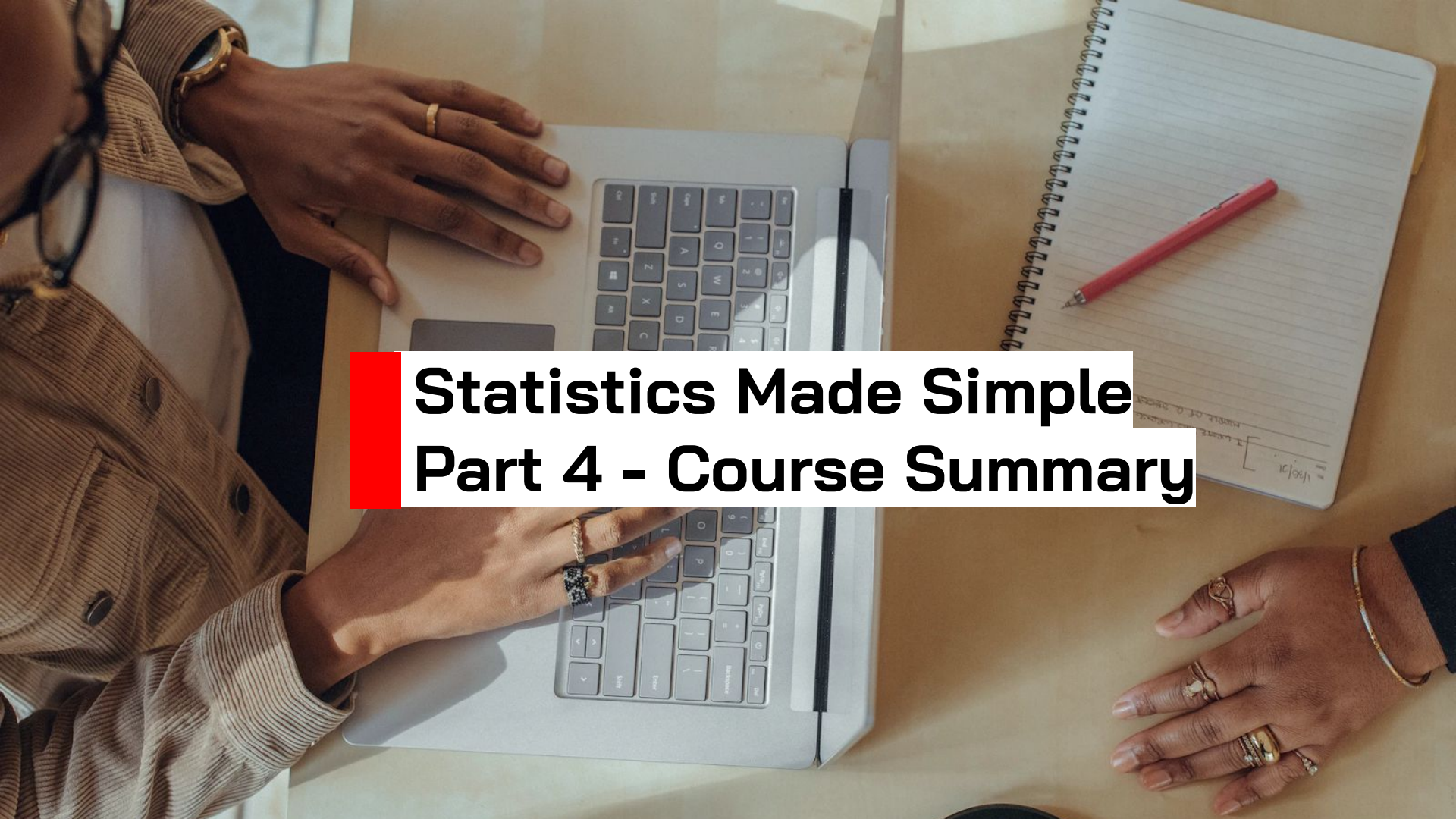
taxi_fare = 35 + 2*10 = 55

We use models to help make **educated decisions**



All models are wrong, but some are useful :)

George Box



Statistics Made Simple
Part 4 - Course Summary

- เข้าใจว่าสถิติสำคัญกับชีวิตอย่างไร
- เข้าใจหลักการสถิติพื้นฐาน
- สามารถวิเคราะห์ข้อมูลเบื้องต้นด้วย Spreadsheets


Statistics Made Simple
Data Science Bootcamp

Appendix

Data Science Bootcamp

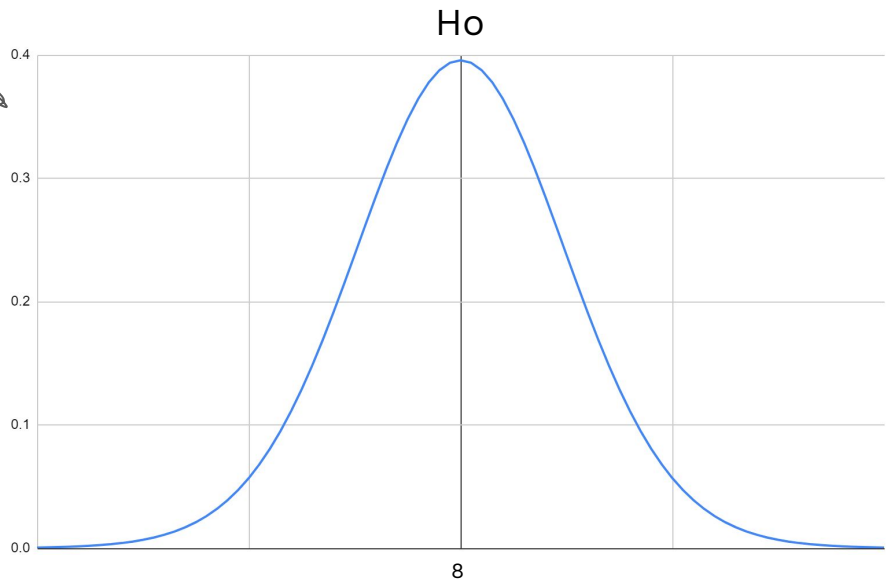
Hypothesis Testing

Ho: null hypothesis
Ha: alternative hypothesis

We'll **reject** Ho if p-value < alpha
alpha is normally set to .05

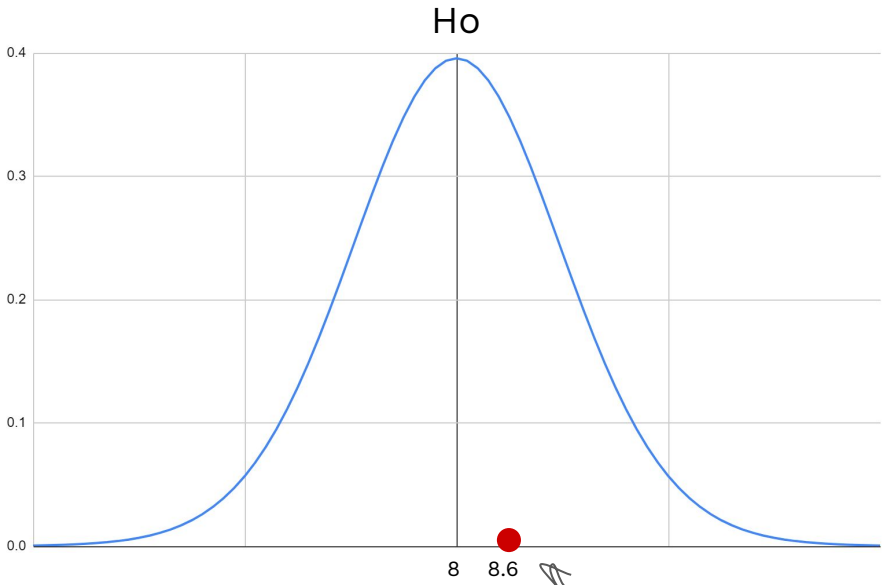
What is p-value

Ho: คนไทยนอนเฉลี่ย 8 ชั่วโมงต่อวัน (mean = 8)
Ha: คนไทยนอนเฉลี่ยไม่เท่ากับ 8 ชั่วโมงต่อวัน (mean != 8)



What is p-value

Ho: คนไทยนอนเฉลี่ย 8 ชั่วโมงต่อวัน
Ha: คนไทยนอนเฉลี่ยไม่เท่ากับ 8 ชั่วโมงต่อวัน

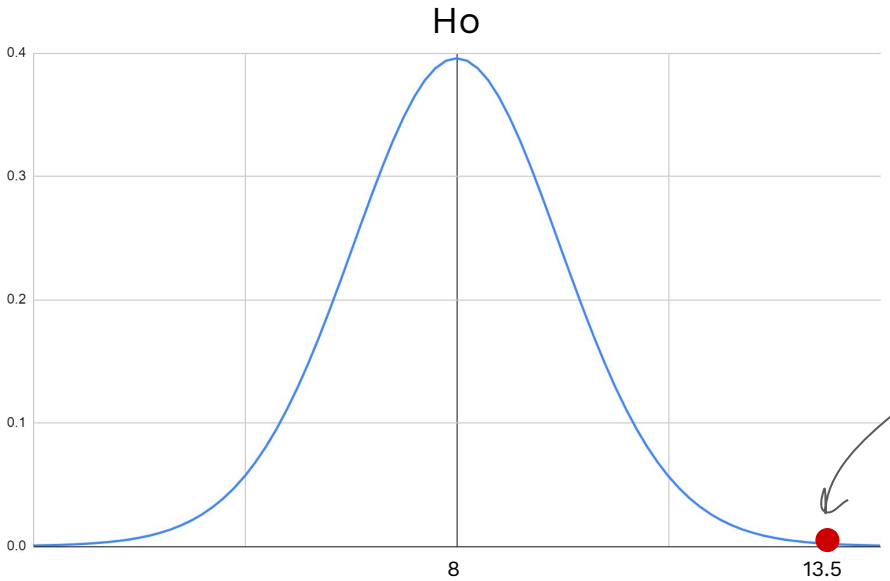


ถ้าคนไทยนอนเฉลี่ยวันละ 8 ชั่วโมงจริง (ho is true)

โอกาสที่เราจะเห็น sample mean = 8.6 จะ**สูงมาก**

What is p-value

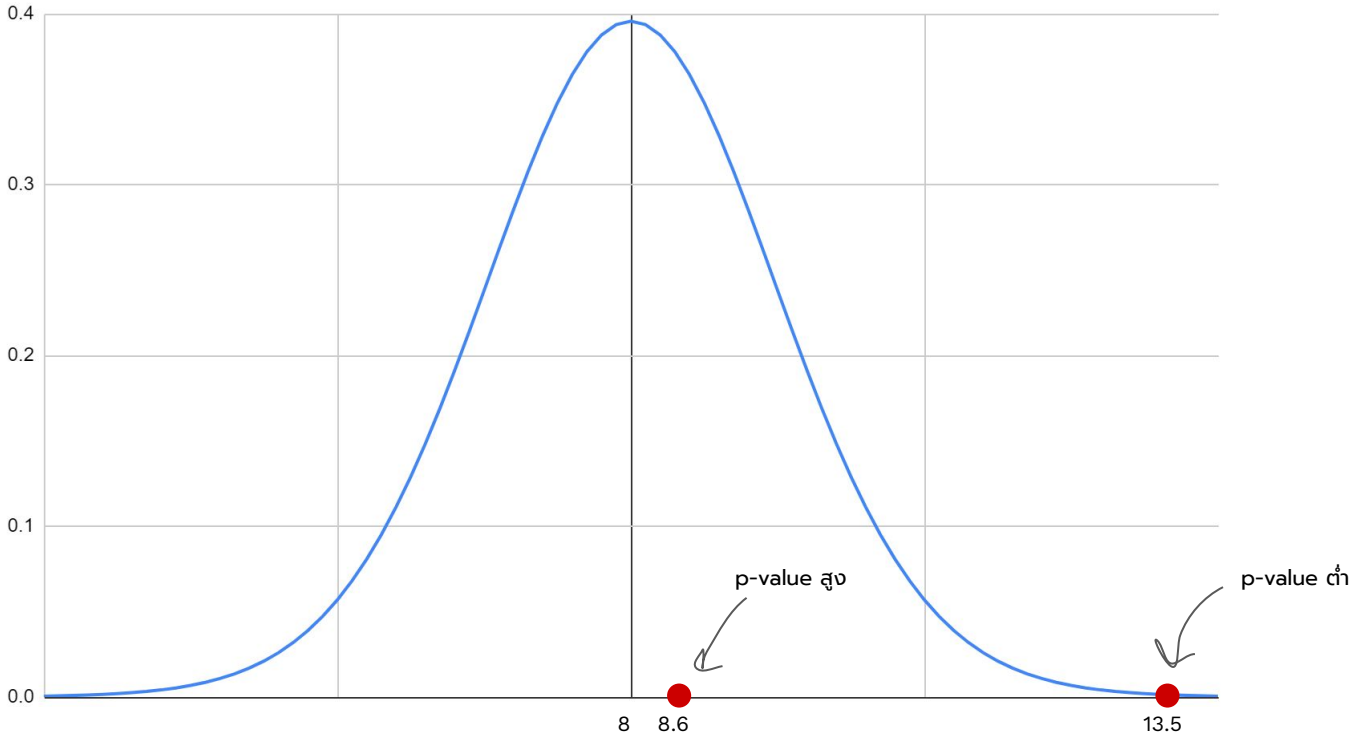
Ho: คนไทยนอนเฉลี่ย 8 ชั่วโมงต่อวัน
Ha: คนไทยนอนเฉลี่ยไม่เท่ากับ 8 ชั่วโมงต่อวัน



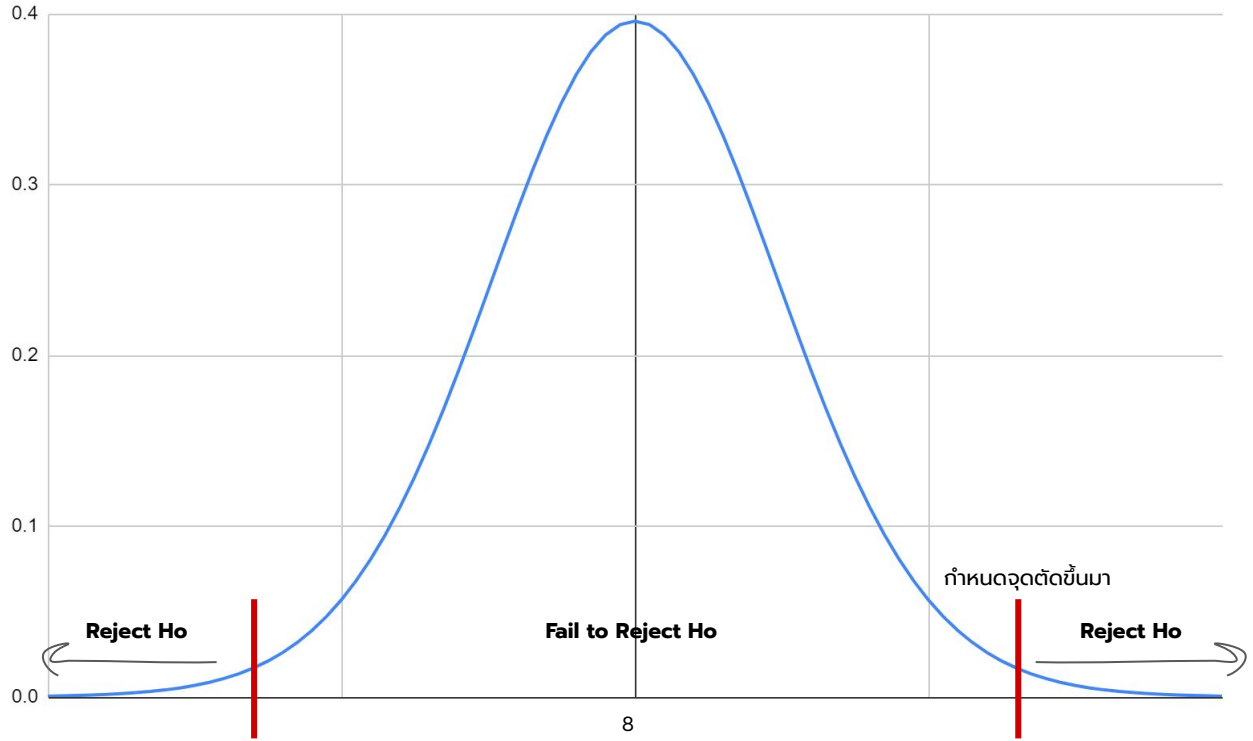
ถ้าคนไทยนอนเฉลี่ยวันละ 8 ชั่วโมงจริง (ho is true)

โอกาสที่เราจะเห็น sample mean = 13.5 จะ**ต่ำมาก**

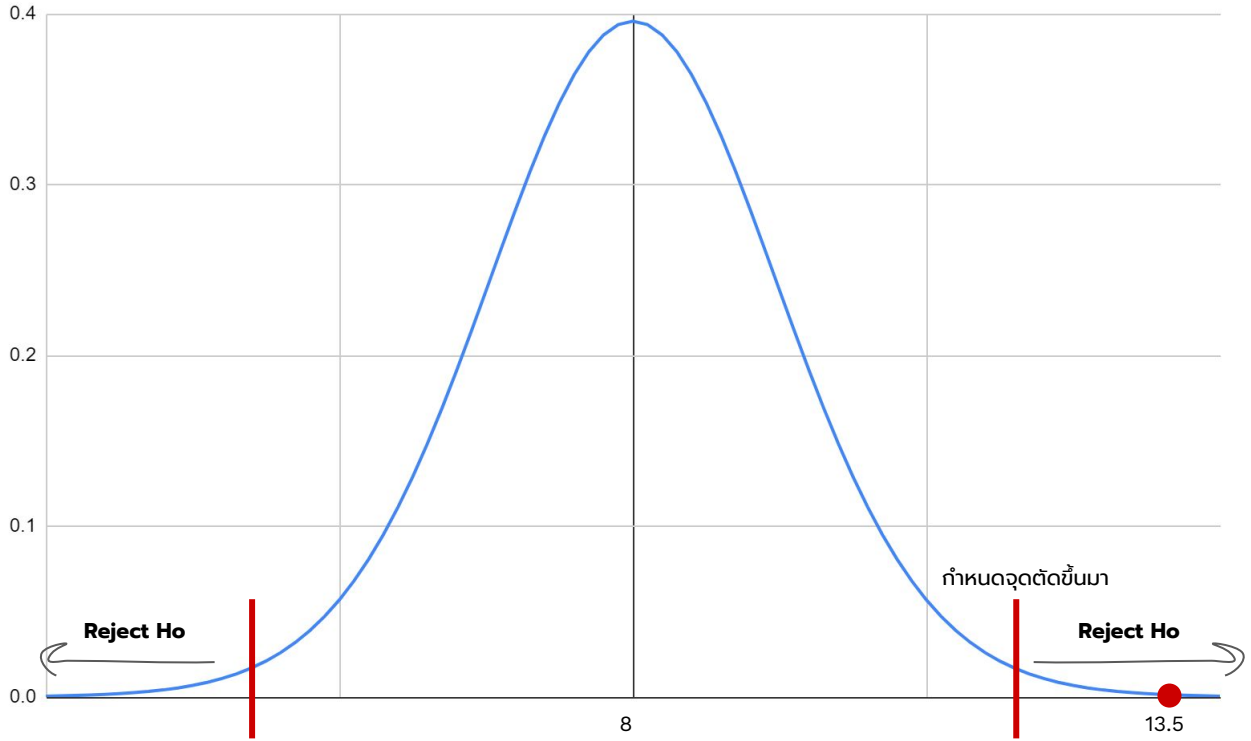
What is p-value



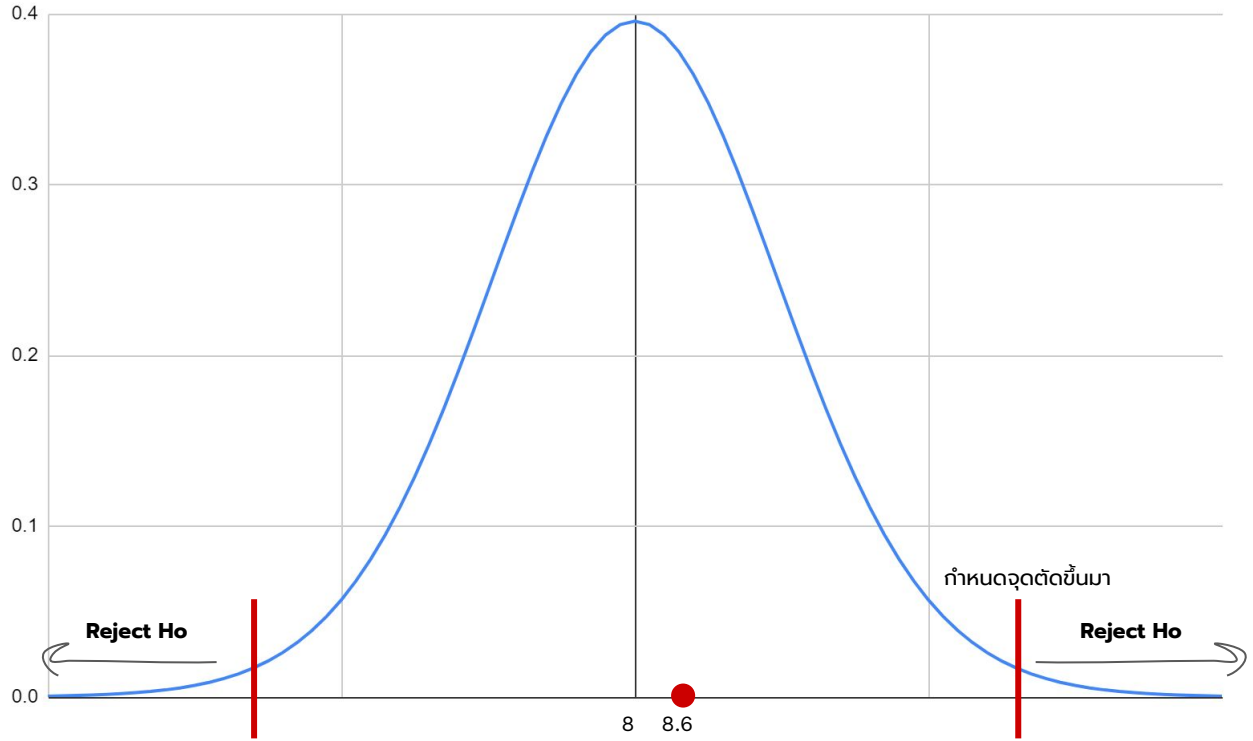
What is p-value



What is p-value

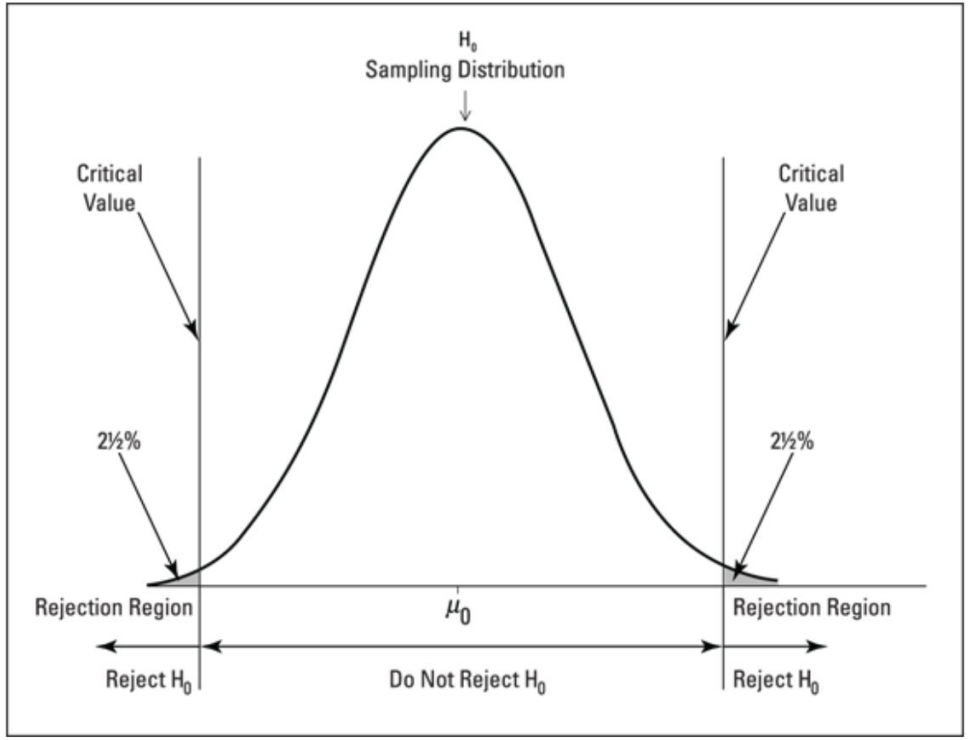


What is p-value



What is p-value

That’s all you need to know for Hypothesis Testing



Exponential Smoothing

How To Be a Smoothie, Exponentially

Exponential smoothing is similar to a moving average. It's a technique for forecasting based on prior data. In contrast with the moving average, which works just with a sequence of actual values, exponential smoothing takes its previous prediction into account.

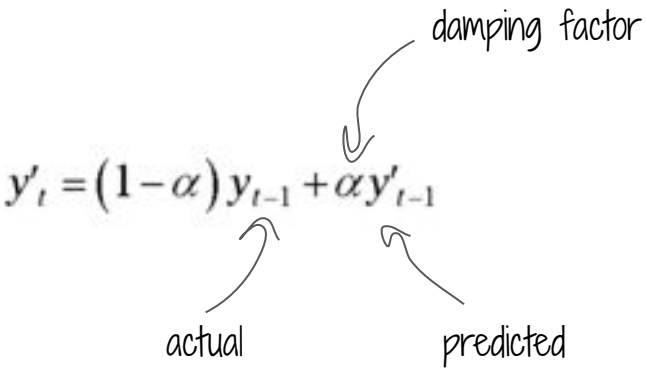
Exponential smoothing operates according to a *damping factor* — a number between zero and one. With α representing the damping factor, the formula is

$$y'_t = (1-\alpha)\,y_{t-1} + \alpha y'_{t-1}$$

In terms of sales figures from the preceding example, y'_t represents the predicted sales at a time: t. If t is the current quarter, $t\text{-}1$ is the immediately preceding quarter. So y_{t-1} is the preceding quarter's actual sales and y'_{t-1} is the preceding quarter's predicted sales. The sequence of predictions begins with the first predicted value as the observed value from the immediately preceding quarter.

A larger damping factor gives more weight to the preceding quarter's prediction. A smaller damping factor gives greater weight to the preceding quarter's actual value. A damping factor of 0.5 weighs each one equally.

Exponential Smoothing



Bootcamp Live 06
Business Analytics and Statistics

Website: https://datarockie.com