QBE DATA SCIENCE WORKSHOP

Data analysis

How data is leveraged in expanding business

Data Engineering

Here is a workflow / checklist of this to look out for and fix

Data Wrangling Exercises

Python Pandas

ANACONDA

jupyter

	Excel	Python	MySQL
type	spreadsheet	programming language	RDBMS
data source	flat files; reports	API feed; unstructured data (images, videos, documents)	database; data dump; logs
output	report; dashboard	export to flat file; database; API	updated data; dashboard
objective	simple calculations & visualization	data analytics	data storage and manipulation
scaling	software; worksheets	CPU; memory; libraries	disk space; optimization

Table of Contents

POINTS FOR DISCUSSION:

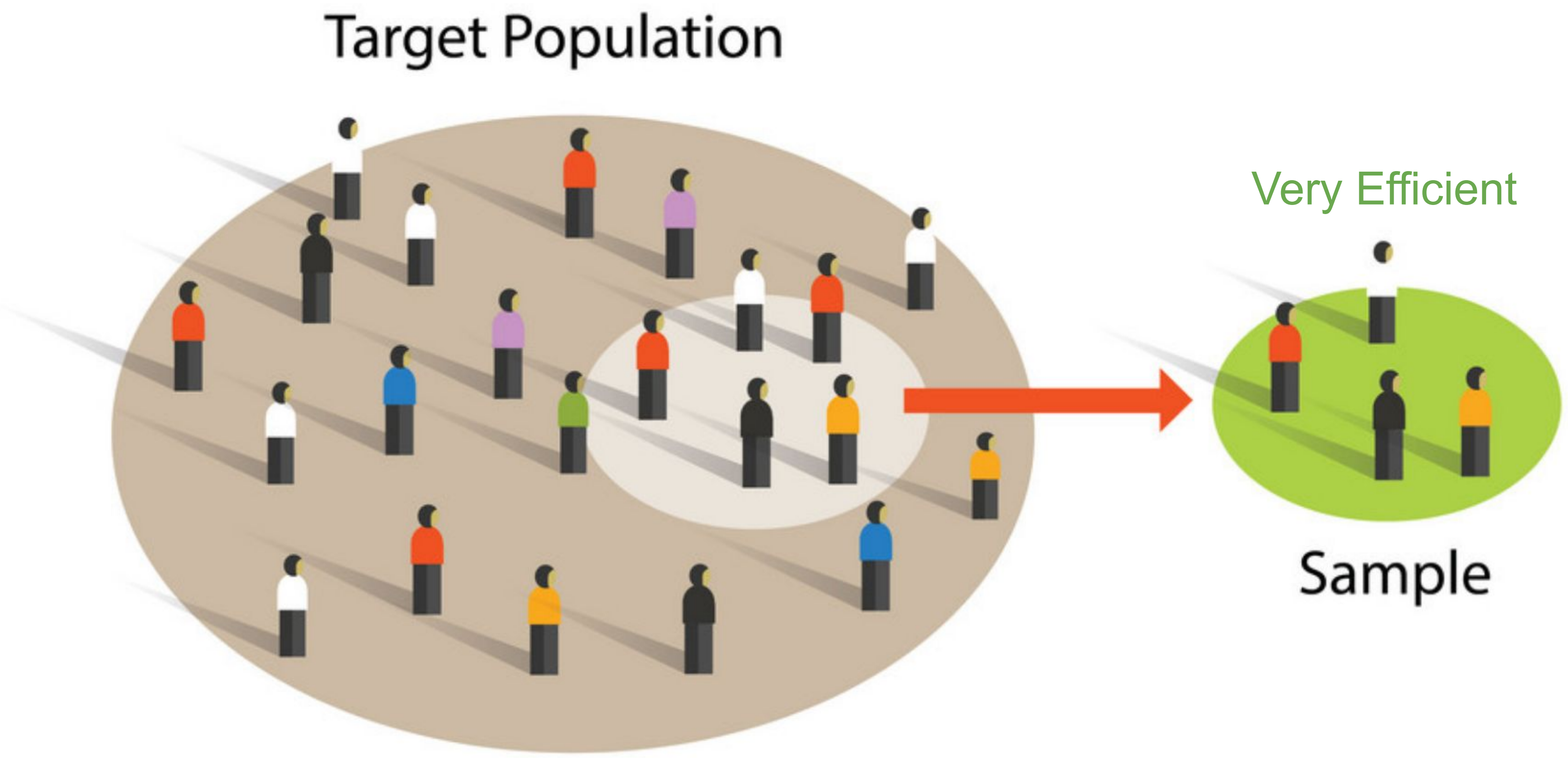
data are *samples* taken from the *population*


Target Population
Sample

but don't get me wrong, *population* is also *data*, so why play only with the *sample*?

analysis on the population as a whole costs *money* and *time*

Very Efficient



More Money, More Time

the goal, therefore, is to *estimate* the *population* *using the samples only*.

but how?

well to *properly estimate* the population,
we need *the sample to be a representative* of it.

Sample and Population

in

Halo-Halo

https://www.spot.ph/newsfeatures/5-facts-about-halo-halo-you-didn-t-know-adv-con hence, in order for the ***sample*** to be a ***representative of the population***, the sample must ***be random*** (or well-mixed).

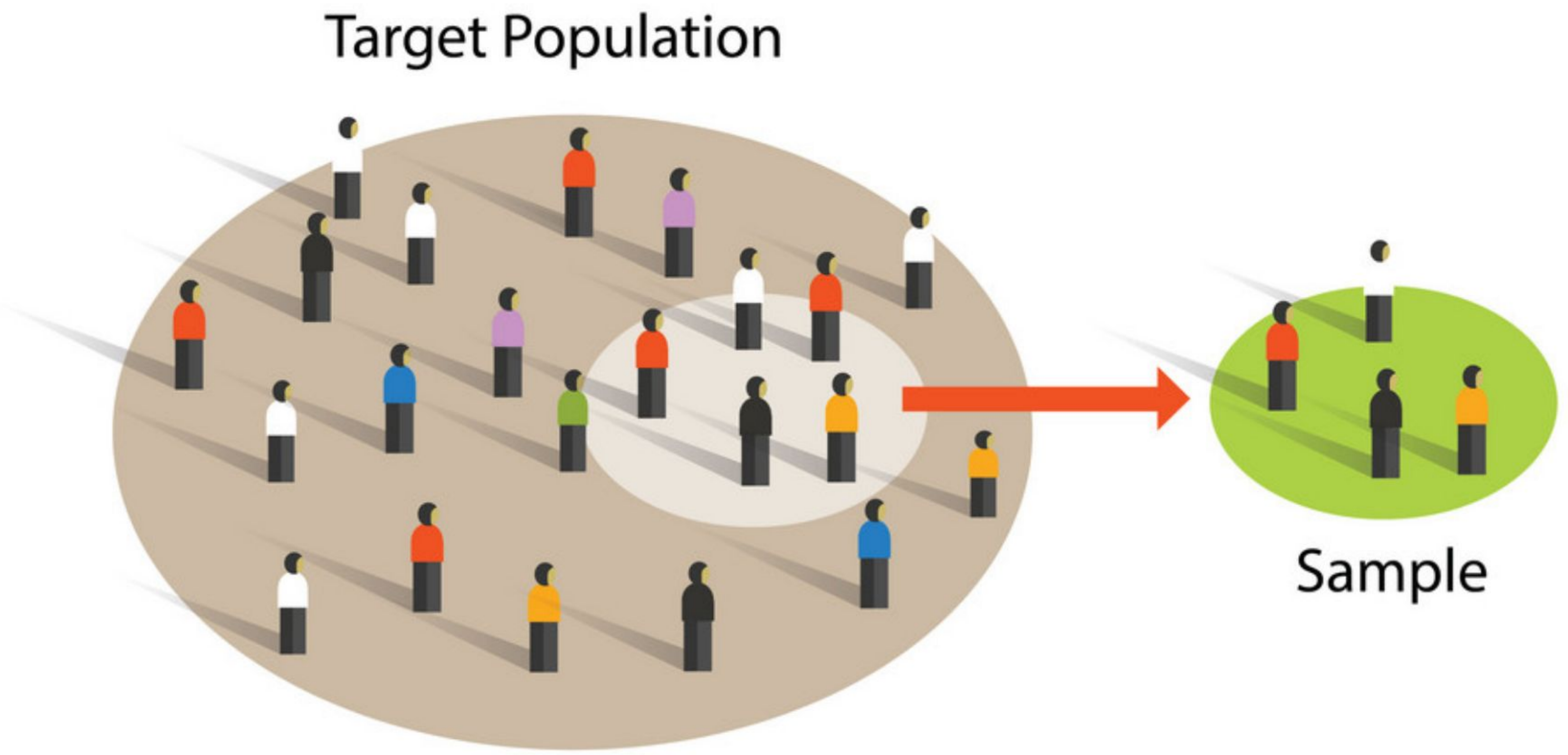
in summary, *data* is a *random sample* from the *population*.

are these representative of the target population?

Target Population

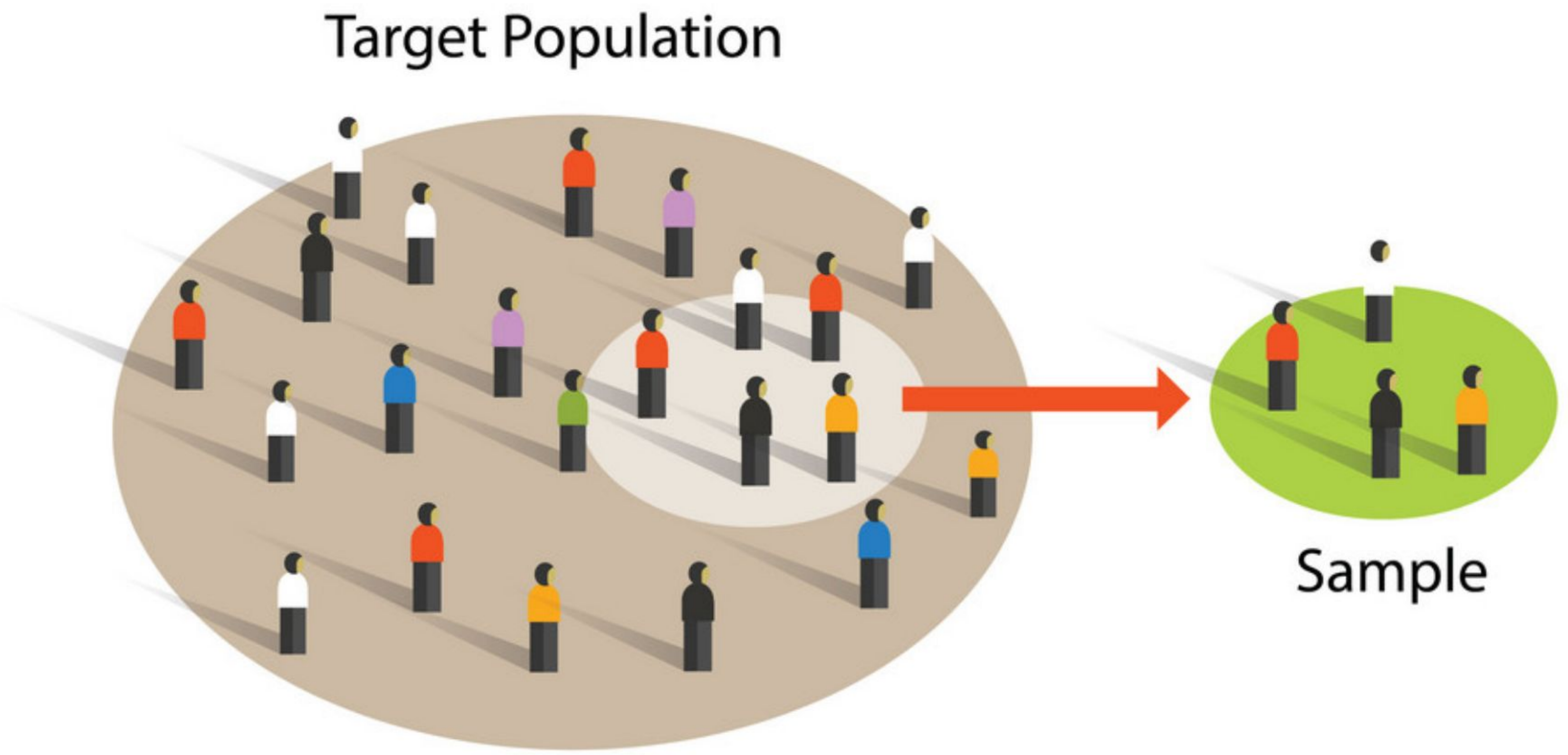
Colors:

- White
- Red
- Black
- Orange
- Purple
- Blue



Colors:

- White - 6
- Red - 5
- Black - 3
- Orange - 3
- Purple - 2
- Blue - 2



so how did we *select* this *sample*?

well, we need to *design* our *sampling procedure*.

there are two approaches:
non-probability and probability.

Sampling Design

- Probability vs Non-Probability

Probability

- ❑ It is defined as a quantitative measure of uncertainty state of information or event.
- ❑ It is an index with range from 0 to 1.
- ❑ It is approximated through proportion of number of events / total experiments:

Probability = 0 : certain the state will not happen.
Probability = 1 : the event will surely happen.
Probability = 0.5: we have maximum doubt about the state that it will happen

Non-Probability

- ❑ Odds of the event happening are not equal

Sampling Design

- Probability vs Non-Probability

  Given every member of this room

  Probability: every member of this population has a known and equal chance of being selected

  Non-Probability: names on the first page, or on the center has higher chance of being selected

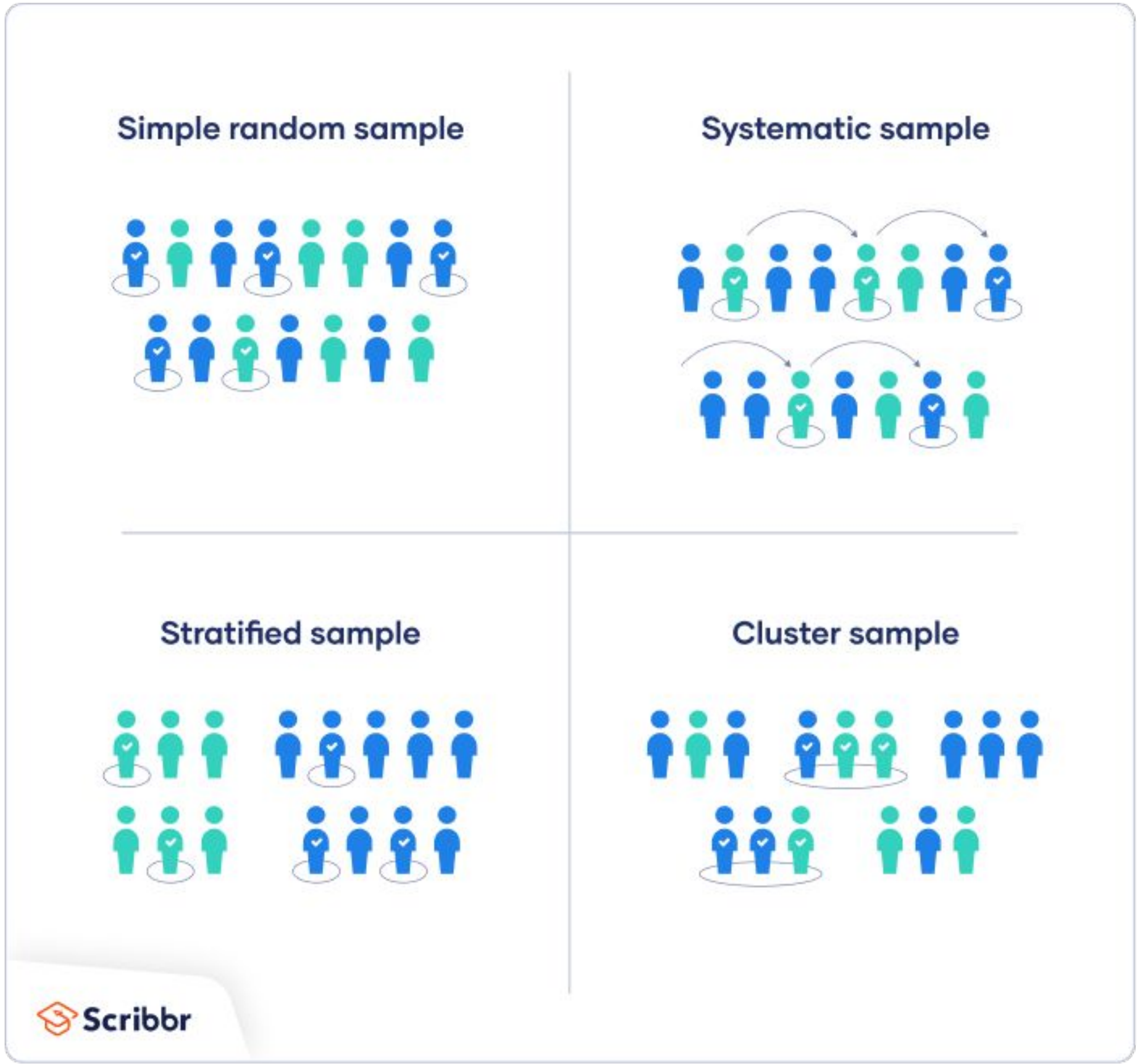
Sampling Design

- Non-Probability
    - Purposive Sampling
    - Snowball Sampling
    - Quota Sampling

- Probability
    - Simple Random Sampling
    - Systematic Sampling
    - Stratified Sampling

Sampling Design

- Probability
  - Simple Random Sampling (*let's do draw lots*)
  - Systematic Sampling (*I'll interview every 4th house in this street*)
  - Stratified Sampling *(I'll take sample for every year level or group)*

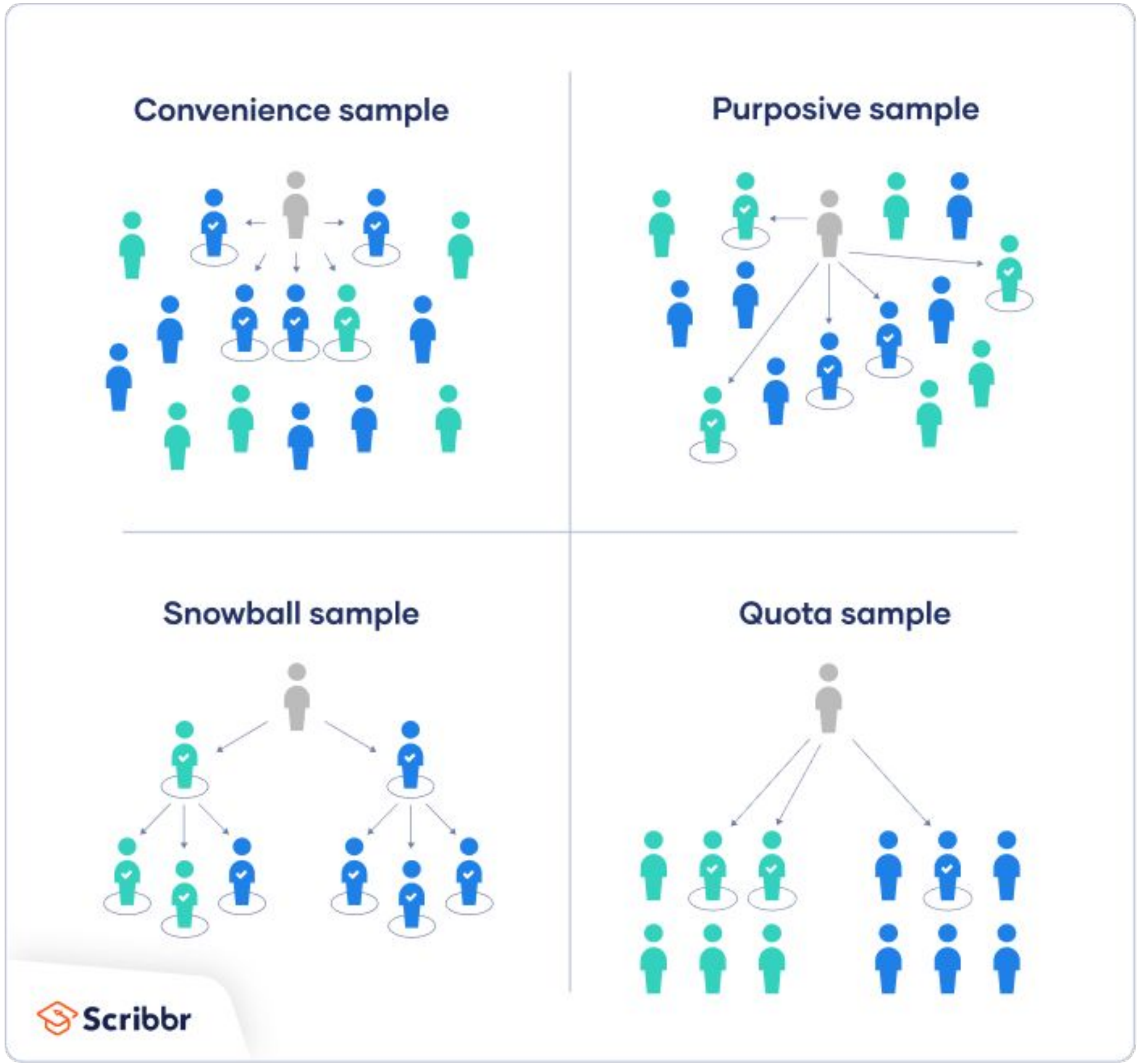


Probability Sampling: Pros and Cons

- Pros
  - creates samples that are highly representative of the population
  - minimize the risk of over or under representation -- ensuring your results are representative of the population
  - can use statistical means to validate your results (e.g. confidence intervals, margins of errors)
- Cons
  - tedious and time consuming, especially when creating larger samples

Sampling Design

- Non-Probability
  - Purposive Sampling
    (*hey, can I interview you?*)
  - Snowball Sampling
    (*do you know someone like you?*)
  - Quota Sampling
    (*I only need 30 people, can I interview the first 30 people in this group?*)



Non-Probability Sampling: Pros and Cons

- Pros
  - speed, cost-effectiveness, ease of availability
  - fine-tune the research by collecting results that only have vital insights

- Cons
  - preconceived notions that the researcher can influence the results (purposive)
  - involved high amount of ambiguity
  -

When to use Non-Probability Sampling

- use this to indicate if a particular trait or characteristic exists in a population
- when aim is conducting qualitative research, pilot studies, or exploratory research.
- have limited time to conduct research or have budget constraints.
- needs to observe whether a particular issue needs in-depth analysis
- **use it when you do not intend to generate results that will generalize the entire population**

https://www.questionpro.com/blog/non-probability-sampling/

Non-probability sampling	Probability sampling
Sample selection based on the subjective judgment of the researcher.	The sample is selected at random.
Not everyone has an equal chance to participate.	Everyone in the population has an equal chance of getting selected.
The researcher does not consider sampling bias.	Used when sampling bias has to be reduced.
Useful when the population has similar traits.	Useful when the population is diverse.
The sample does not accurately represent the population.	Used to create an accurate sample.
Finding respondents is easy.	Finding the right respondents is not easy.

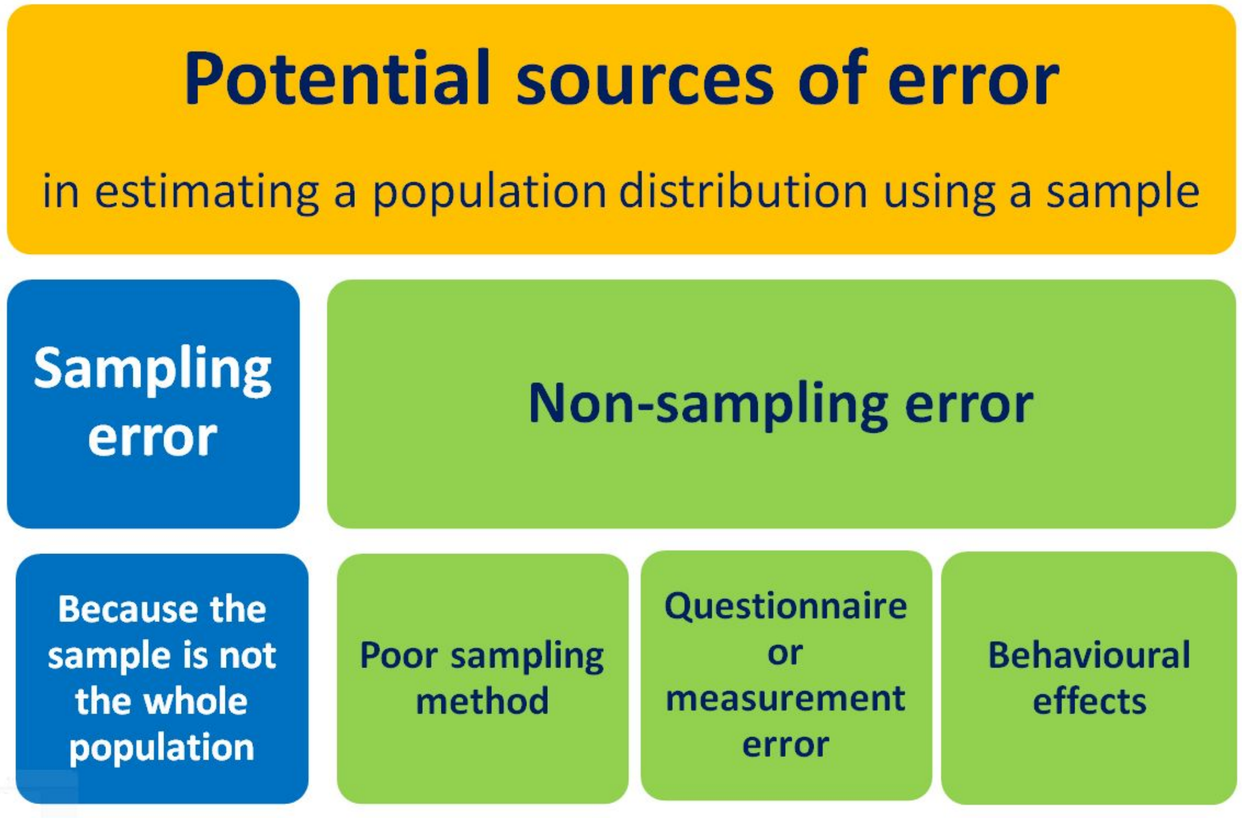
https://www.questionpro.com/blog/non-probability-sampling/

Sampling Design

- Non-Probability
  - Purposive Sampling (*hey, can I interview you?*)
  - Snowball Sampling (*do you know someone like you?*)
  - Quota Sampling (*I only need 30 people, can I interview the first 30 people in this group?*)

- Probability
  - Simple Random Sampling (*let's do draw lots*)
  - Systematic Sampling (*I'll interview every 4th house in this street*)
  - Stratified Sampling *(I'll take sample for every year level or group)*

Sampling vs Non-sampling Errors



Sampling Errors

Sampling errors are affected by factors such as the **size and design of the sample**, **population variability**, and **sampling fraction**.

Categories of Sampling Errors:

- Population Specification Error – Happens when the analysts do not understand who to survey. For example, for a survey of breakfast cereals, the population can be the mother, children, or the entire family.
- Selection Error – Occurs when the survey participation is self-selected by the respondents implying only those who are interested respond. Selection error can be reduced by encouraging participation.
- Sample Frame Error – Occurs when a sample is selected from the wrong population data.
- Non-Response Error – Occurs when a useful response is not obtained from the surveys. It may happen due to the inability to contact potential respondents or their refusal to respond.

https://corporatefinanceinstitute.com/resources/knowledge/other/non-sampling-error/

Non-sampling Errors

most common non-sampling errors include errors in **data entry, biased questions and decision-making, non-responses, false information,** and **inappropriate analysis**.

Mechanics of Non-Sampling Error

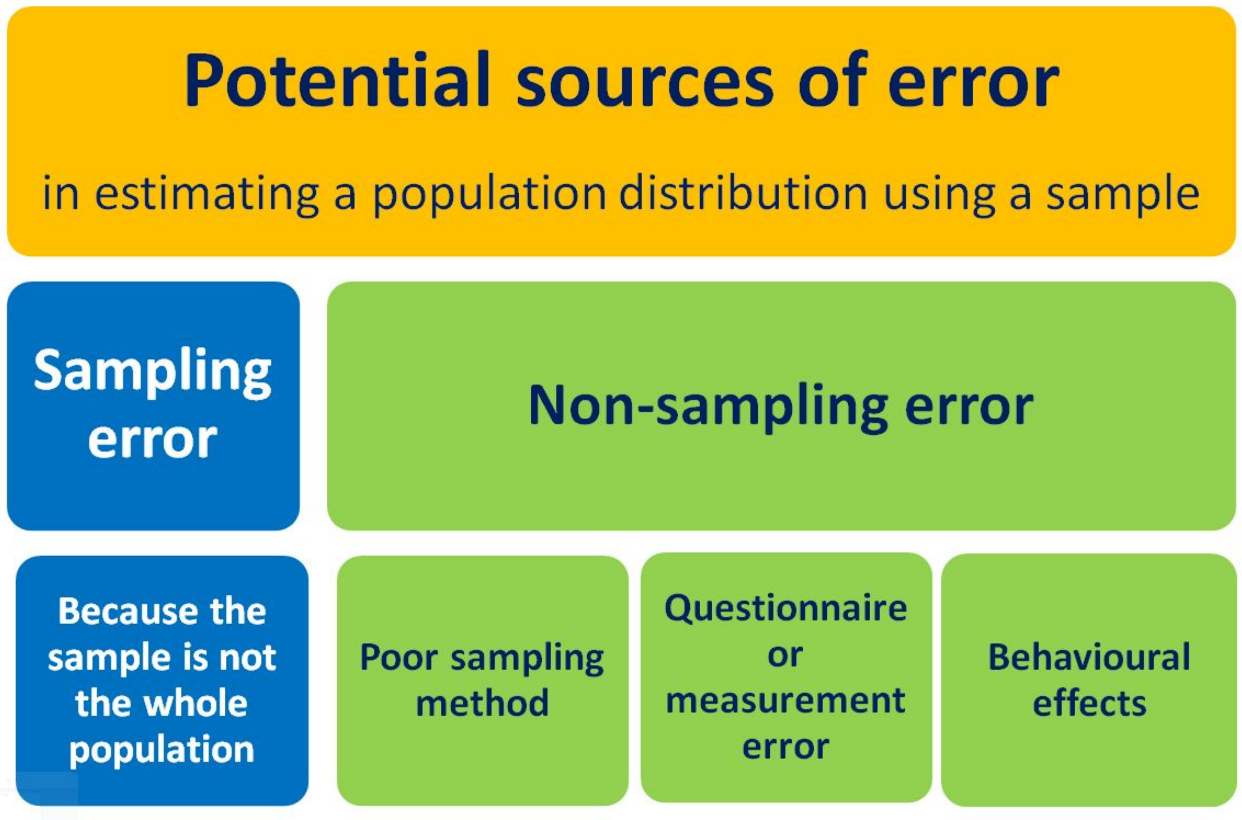
- **Random errors** -- Random errors are errors that cannot be accounted for and just happen. In statistical studies, it is believed that each random error offsets each other, generally speaking, so they are of little to no concern.
- **Systematic errors** -- Systematic errors affect the sample of the study and, as a result, will often create useless data. A systematic error is consistent and repeatable, so the creators of the study must take great care to mitigate such an error.

Non-sampling Errors

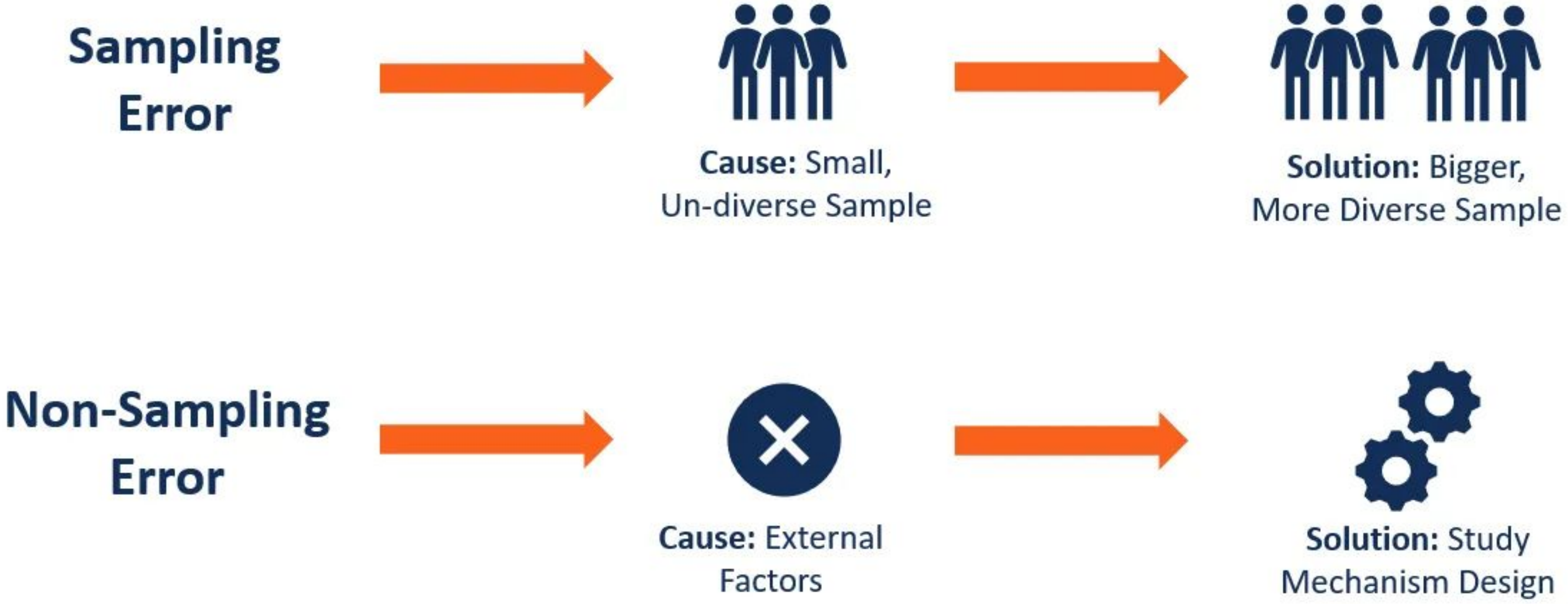
- **Non-response error** -- caused by the differences between the people who choose to participate compared to the people who do not participate in a given survey.
- **Measurement error** -- refers to all errors relating to the measurement of each sampling unit, as opposed to errors relating to how they were selected, often arises when there are confusing questions, low-quality data due to sampling fatigue (i.e., someone is tired of taking a survey), and low-quality measurement tools.
- **Interviewer error** -- occurs when the interviewer (or administrator) makes an error when recording a response. In qualitative research, an interviewer may lead a respondent to answer a certain way. In quantitative research, an interviewer may ask the question in a different way, which leads to a different end result.
- **Adjustment error** -- situation where the analysis of the data adjusts it in such a way that it is not entirely accurate. Forms of adjustment error include errors with weighting the data, data cleaning, and imputation.
- **Processing error** -- A processing error arises when there is a problem with processing the data that causes an error of some kind. An example will be if the data were entered incorrectly or if the data file is corrupt.

https://corporatefinanceinstitute.com/resources/knowledge/other/non-sampling-error/

Sampling vs Non-sampling Errors



Reducing Errors



https://corporatefinanceinstitute.com/resources/knowledge/other/non-sampling-error/

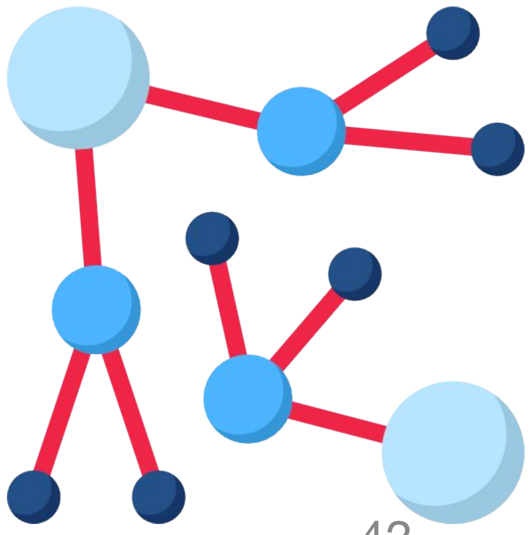
suppose we now have our representative sample data, how do we *get insights* from it?

well that’s where *basic statistics* comes in.

then?

Look for *patterns*, by doing *exploratory data analysis*.

How to look for patterns?



Descriptive Statistics and Data Visualization!

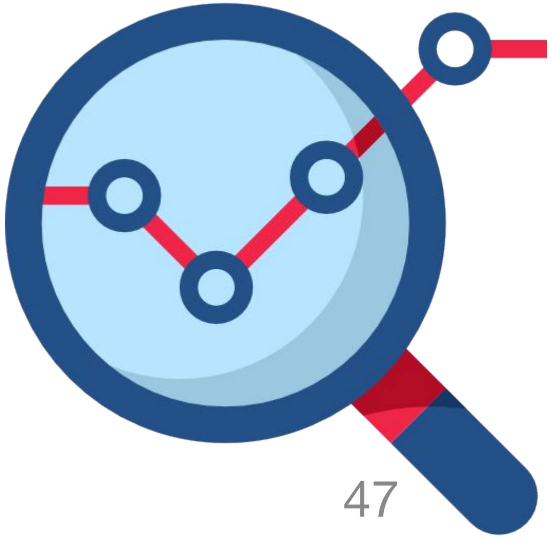
Identify your data

is it *univariate* or *multivariate*?

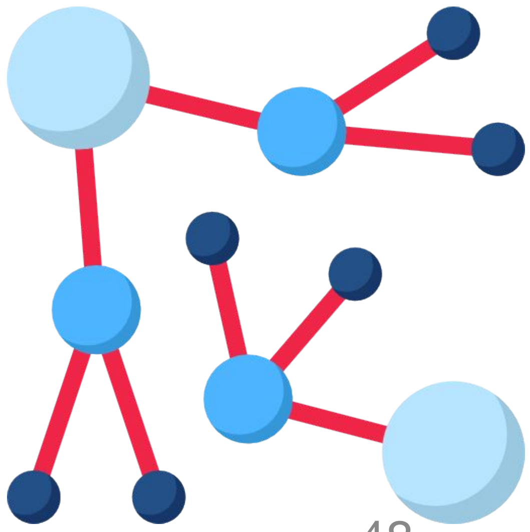
is it *categorical* or *continuous*?

then?

Look for *patterns*, by doing *exploratory data analysis*.



How to look for patterns?



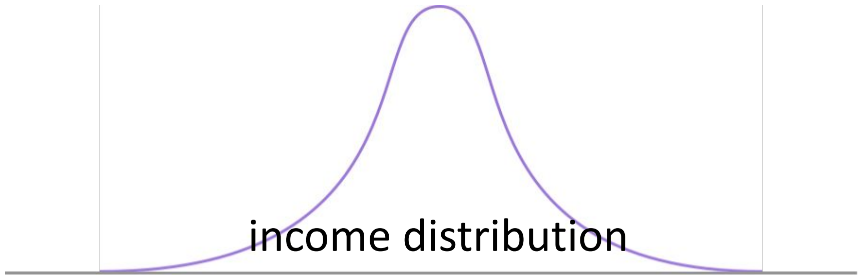
Descriptive Statistics and Data Visualization!

Descriptive Statistics

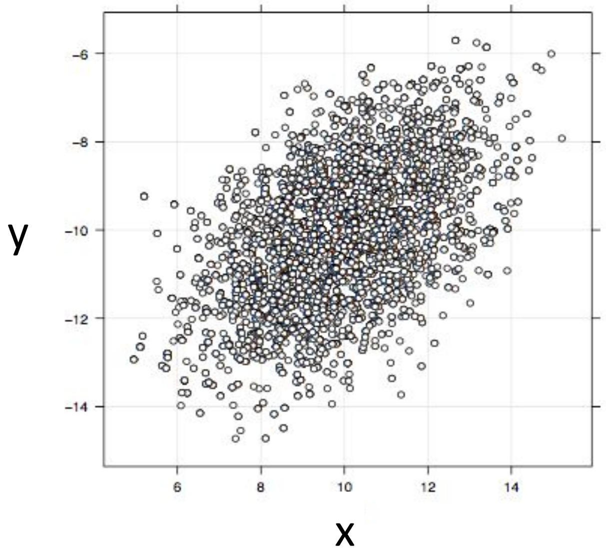
- Univariate
  - Mean
  - Median
  - Mode - categorical
  - Variance
  - Kurtosis
  - Skewness
- Multivariate
  - Pairwise Correlation

Data Visualization: Next Session

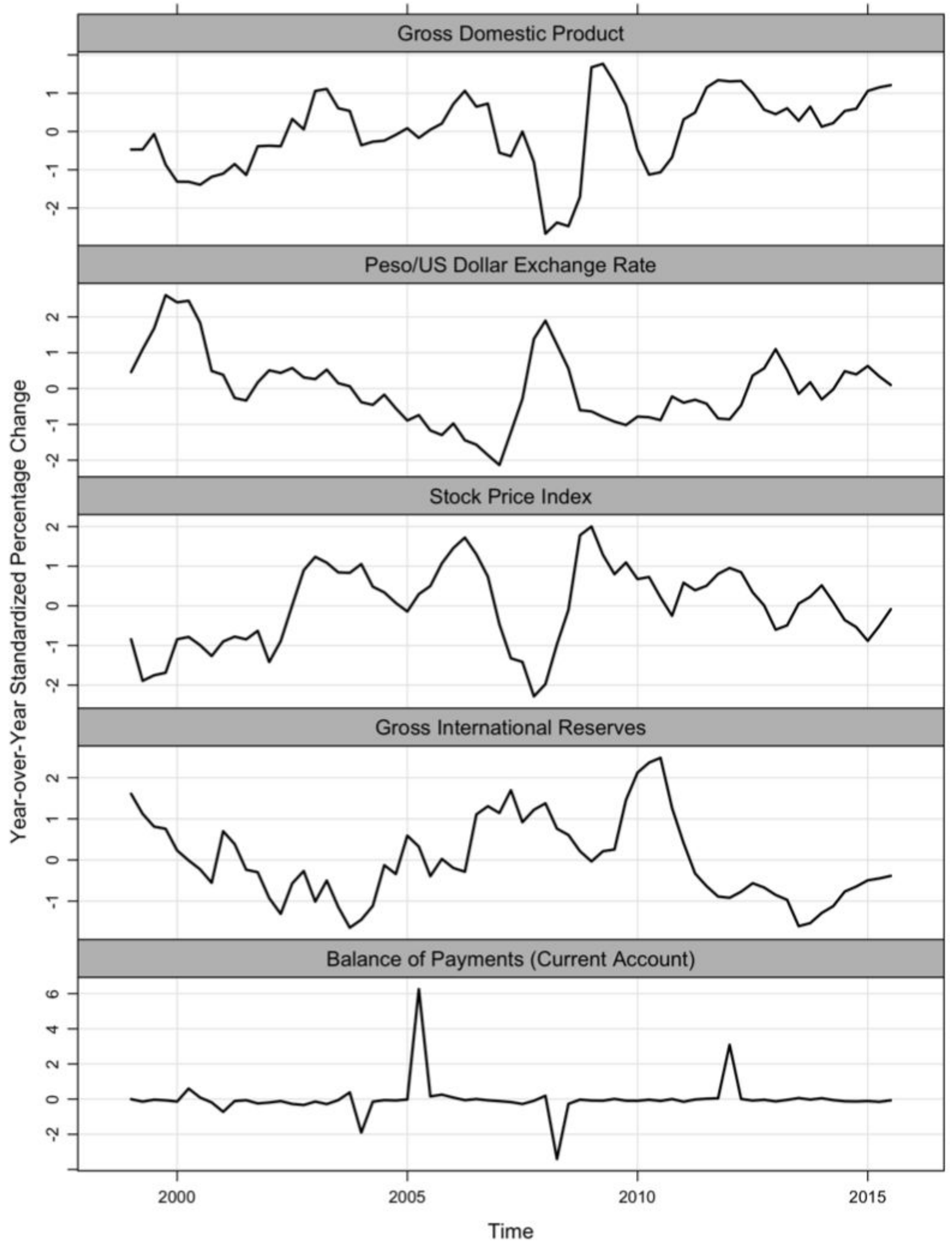
- Univariate
  - Plot the Histogram (Distribution)
  - Plot the Data if Time Series



- Multivariate
  - Plot X and Y
  - 3D



Time Series
(not covered)



Basic Statistics

- Mean
- Median
- Mode
- Variance
- Standard Deviation
- Kurtosis
- Skewness

Basic Statistics

- Mean (*the average value*)
- Median (*the middle value*)
- Mode (*the frequent value*)
- Variance (*the value for homogeneity or heterogeneity*)
- Standard Deviation (*standardized variance*)
- Kurtosis (*another measure for variance*)
- Skewness (*the location of concentration of points*)

Central Tendencies

- **Mean** -- the sum of a variable’s values divided by the total number of values
- **Median** -- the middle value of a variable
- **Mode** -- the value that occurs most often

Central Tendencies

The incomes of five randomly selected people in the United States are:

$10,000, $10,000, $45,000, $60,000, and $1,000,000

- **Mean** =
- **Median** =
- **Mode** =

Central Tendencies

The incomes of five randomly selected people in the United States are:

$10,000, $10,000, $45,000, $60,000, and $1,000,000

- **Mean** = (10,000 + 10,000 + 45,000 + 60,000 + 1,000,000) / 5 = $225,000
- **Median** =
- **Mode** =

Central Tendencies

The incomes of five randomly selected people in the United States are:

$10,000, $10,000, $45,000, $60,000, and $1,000,000

- **Mean** = (10,000 + 10,000 + 45,000 + 60,000 + 1,000,000) / 5 = $225,000
- **Median** = $45,000
- **Mode** =

Central Tendencies

The incomes of five randomly selected people in the United States are:

$10,000, $10,000, $45,000, $60,000, and $1,000,000

- **Mean** = (10,000 + 10,000 + 45,000 + 60,000 + 1,000,000) / 5 = $225,000
- **Median** = $45,000
- **Mode** = $10,000

Mean, Median and Mode

Employee	Salary (Annual)
John	420,000
Mike	510,000
Kate	630,000
Shane	450,000
Catty	550,000
Mathew	900,000
Sam	1,000,000

Mean, Median and Mode

Employee	Salary (Annual)
John	420,000
Mike	510,000
Kate	630,000
Shane	450,000
Catty	550,000
Mathew	900,000
Sam	1,000,000

Mean

Employee	Salary (Annual)
John	420,000
Mike	510,000
Kate	630,000
Shane	450,000
Catty	550,000
Mathew	900,000
Sam	1,000,000

Mean: (420,000 + 510,000 + 630,000 + . . . + 1,000,000) / 7 = **637,142.86**

Median: Sort the data first and take the middle

Employee	Salary (Annual)
John	420,000
Mike	510,000
Kate	630,000
Shane	450,000
Catty	550,000
Mathew	900,000
Sam	1,000,000

Median: 420000, 450000, 510000, **550000**, 630000, 900000, 1000000

Mode: Applicable for Categorical Variables Only

Employee	Salary (Annual)
John	420,000
Mike	510,000
Kate	630,000
Shane	450,000
Catty	550,000
Mathew	900,000
Sam	1,000,000

Mean, Median and Mode

Customer	Quantity of orders	Country
John	160	UK
Mike	1440	UK
Kate	1200	EIRE
Shane	10	France
Catty	20	UK
Mathew	30	France
Sam	60	Belgium

Mean

Customer	Quantity of orders	Country
John	160	UK
Mike	1440	UK
Kate	1200	EIRE
Shane	10	France
Catty	20	UK
Mathew	30	France
Sam	60	Belgium

Median

Customer	Quantity of orders	Country
John	160	UK
Mike	1440	UK
Kate	1200	EIRE
Shane	10	France
Catty	20	UK
Mathew	30	France
Sam	60	Belgium

Mode

Customer	Quantity of orders	Country
John	160	UK
Mike	1440	UK
Kate	1200	EIRE
Shane	10	France
Catty	20	UK
Mathew	30	France
Sam	60	Belgium

Mean

Customer	Quantity of orders	Country
John	160	UK
Mike	1440	UK
Kate	1200	EIRE
Shane	10	France
Catty	20	UK
Mathew	30	France
Sam	60	Belgium

Mean: (160 + 1440 + 1200 + . . . + 60) / 7 = **417.14**

Median: Sort the data first and take the middle

Customer	Quantity of orders
John	160
Mike	1440
Kate	1200
Shane	10
Catty	20
Mathew	30
Sam	60

Median: 10, 12, 30, **60**, 160, 1200, 1440

Mode: Applicable for Categorical Variables Only

Customer	Quantity of orders	Country
John	160	UK
Mike	1440	UK
Kate	1200	EIRE
Shane	10	France
Catty	20	UK
Mathew	30	France
Sam	60	Belgium

Mode: Applicable for Categorical Variables Only

Customer	Quantity of orders	Country
John	160	UK
Mike	1440	UK
Kate	1200	EIRE
Shane	10	France
Catty	20	UK
Mathew	30	France
Sam	60	Belgium

Number from UK = 3, Number from France = 2, Number of Belgium = 1, Number of EIRE = 1
Mode = **UK**

what are the *insights* from these statistics (mean, median, mode, etc.)?

how sure are we on our ***estimates***?

well, we can estimate that using the *variance*.

Measures of Dispersion

- **Range** -- difference between the smallest and largest values in the data
- **Variance** -- calculated by taking the average of the squared differences between each value and the mean
- **Standard Deviation** -- square root of the variance
- **Skew** -- measure of whether some values of a variable are extremely different from the majority of the values. Skew = 3 * (Mean – Median) / Standard Deviation

Measures of Dispersion

The incomes of five randomly selected people in the United States are:

$10,000, $10,000, $45,000, $60,000, and $1,000,000

- **Range** -- difference between the smallest and largest values in the data
- **Variance** --
- **Standard Deviation** --
- **Skew** --

Measures of Dispersion

The incomes of five randomly selected people in the United States are:

$10,000, $10,000, $45,000, $60,000, and $1,000,000

- **Range** -- 1,000,000 - 10,000 = 990,000
- **Variance** --
- **Standard Deviation** --
- **Skew** --

Measures of Dispersion

The incomes of five randomly selected people in the United States are:

$10,000, $10,000, $45,000, $60,000, and $1,000,000

- **Range** -- 1,000,000 - 10,000 = 990,000
- **Variance** -- calculated by taking the average of the squared differences between each value and the mean

  -- [(10,000 - 225,000)^2 + (10,000 - 225,000)^2 + (45,000 - 225,000)^2 + (60,000 - 225,000)^2 + (1,000,000 - 225,000)^2] / 5 = 150,540,000,000
- **Standard Deviation** --
- **Skew** --

Measures of Dispersion

The incomes of five randomly selected people in the United States are:

$10,000, $10,000, $45,000, $60,000, and $1,000,000

- **Range** -- 1,000,000 - 10,000 = 990,000
- **Variance** -- [(10,000 - 225,000)^2 + (10,000 - 225,000)^2 + (45,000 - 225,000)^2 + (60,000 - 225,000)^2 + (1,000,000 - 225,000)^2] / 5 = 150,540,000,000
- **Standard Deviation** -- square root of the variance

  -- Square Root (150,540,000,000) = 387,995
- **Skew** --

Measures of Dispersion

The incomes of five randomly selected people in the United States are:

$10,000, $10,000, $45,000, $60,000, and $1,000,000

- **Range** -- 1,000,000 - 10,000 = 990,000
- **Variance** -- [(10,000 - 225,000)^2 + (10,000 - 225,000)^2 + (45,000 - 225,000)^2 + (60,000 - 225,000)^2 + (1,000,000 - 225,000)^2] / 5 = 150,540,000,000
- **Standard Deviation** -- Square Root (150,540,000,000) = 387,995
- **Skew** -- Skew = 3 * (Mean – Median) / Standard Deviation

  -- Income is positively skewed

Degrees of Freedom

Standard deviation in a population is:

$$\sigma^2 = \frac{\sum (x - \mu)^2}{n}$$

The estimate of population standard deviation calculated from a random sample

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

Variance

Employee	Quantity of Orders
John	160
Mike	1440
Kate	1200
Shane	10
Catty	20
Mathew	30
Sam	60

Mean = **417.14**
Variance = (Salary - Mean)^2 / (n - 1)

(Salary - Mean)^2 = (160 - **417.14**)^2
+ (1440 - **417.14**)^2
+ (1200 - **417.14**)^2
:
+ (60 - **417.14**)^2
= 2326135.72

Variance = 2326135.72 / (7 - 1) = 2326135.72 / 6 = 387689.28

Standard Deviation

Employee	Quantity of Orders
John	160
Mike	1440
Kate	1200
Shane	10
Catty	20
Mathew	30
Sam	60

Mean = **417.14**
Variance = (Salary - Mean)^2 / (n - 1)

(Salary - Mean)^2 = (160 - **417.14**)^2
+ (1440 - **417.14**)^2
+ (1200 - **417.14**)^2
:
+ (60 - **417.14**)^2
= 2326135.72

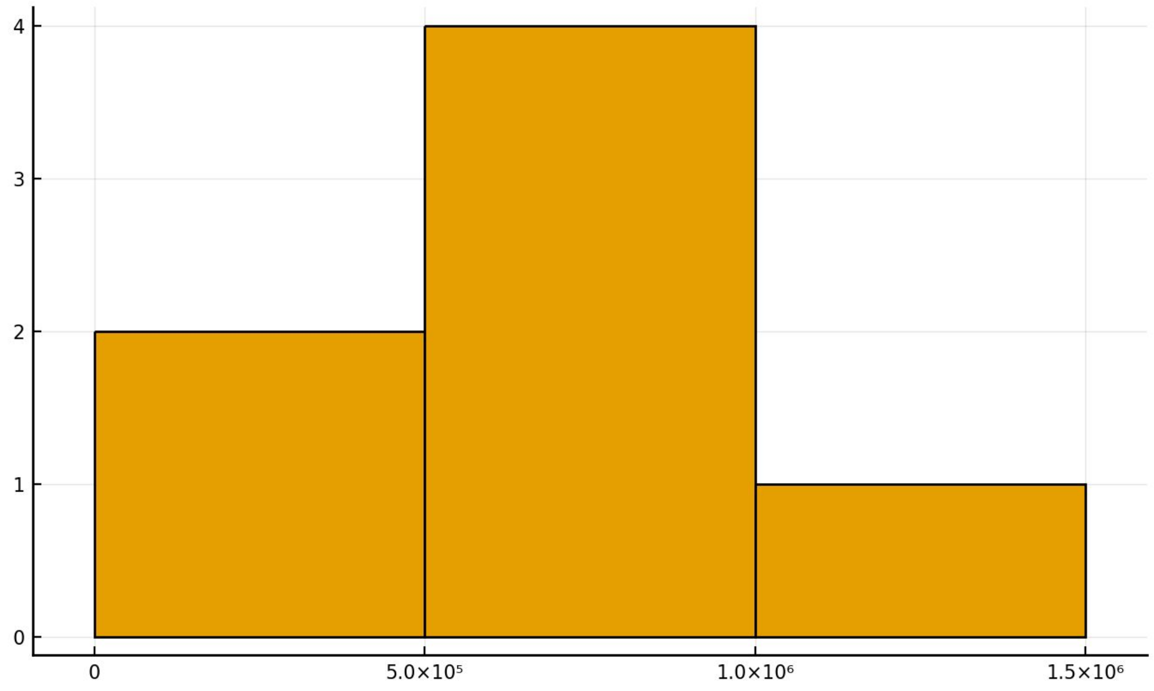
Variance = 2326135.72 / (7 - 1) = 2326135.72 / 6 = **387689.28**
Standard Deviation = Square Root of Variance = **622.65**

the ***larger the variance/standard deviation*** **the** ***larger the variability of the data*****, and vice-versa**

histogram

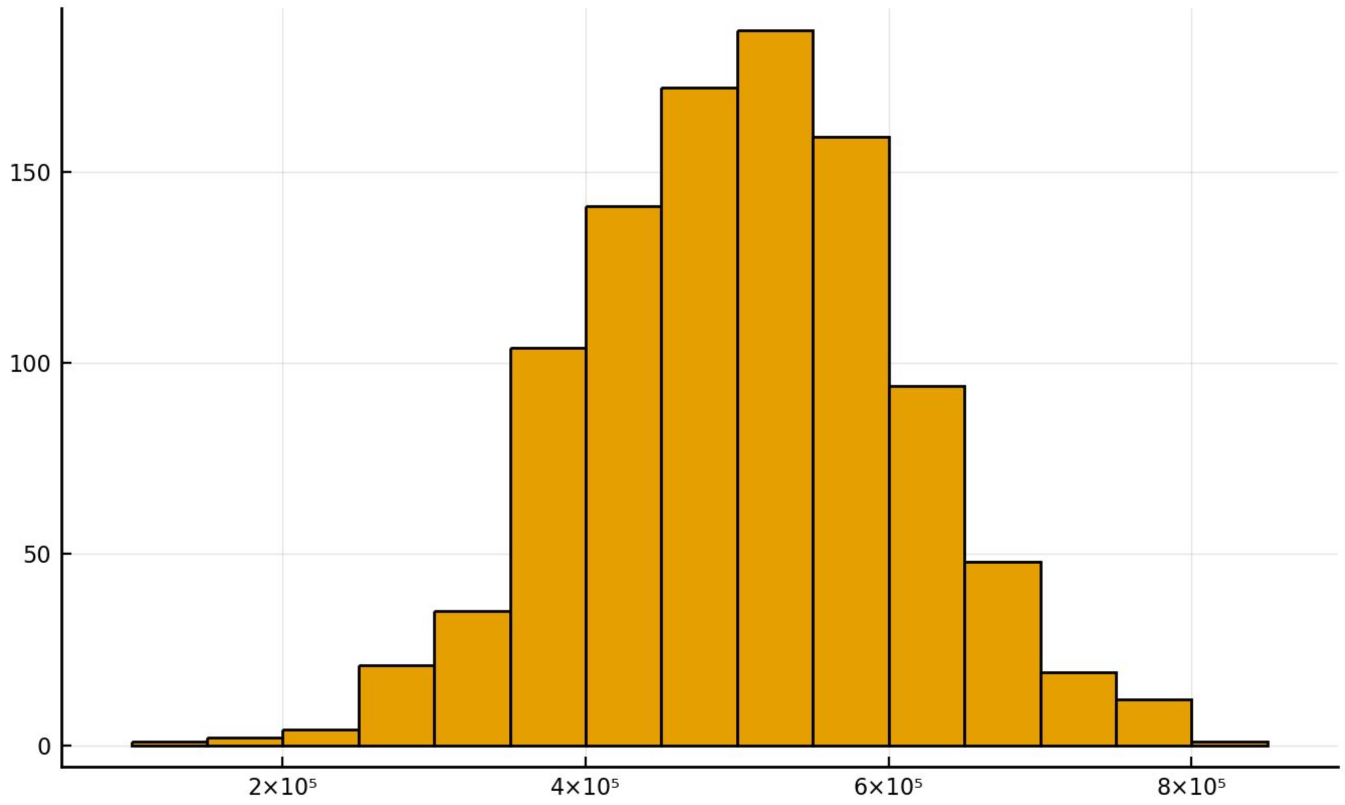
Histogram

Employee	Salary (Annual)
John	420,000
Mike	510,000
Kate	630,000
Shane	450,000
Catty	550,000
Mathew	900,000
Sam	1,000,000

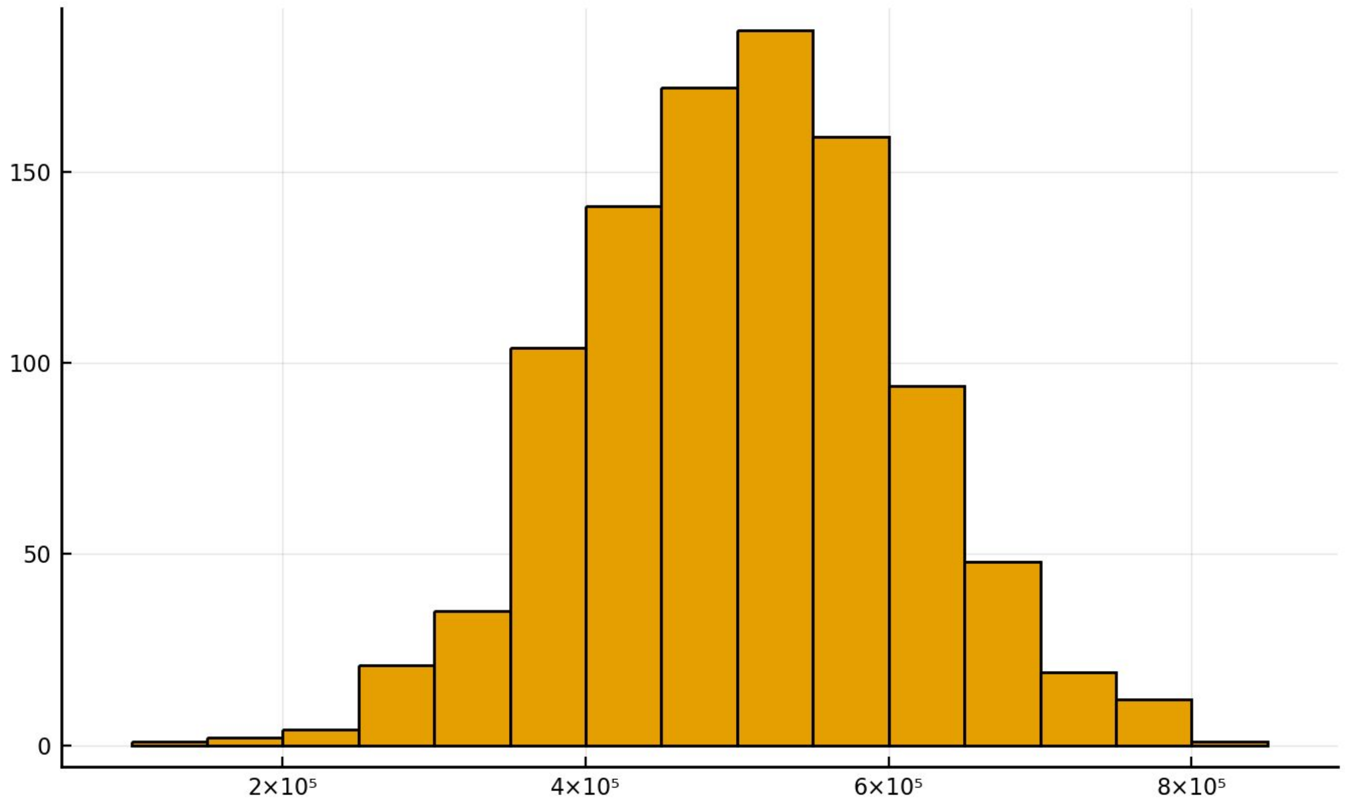


Histogram of Salaries of 1000 Employees

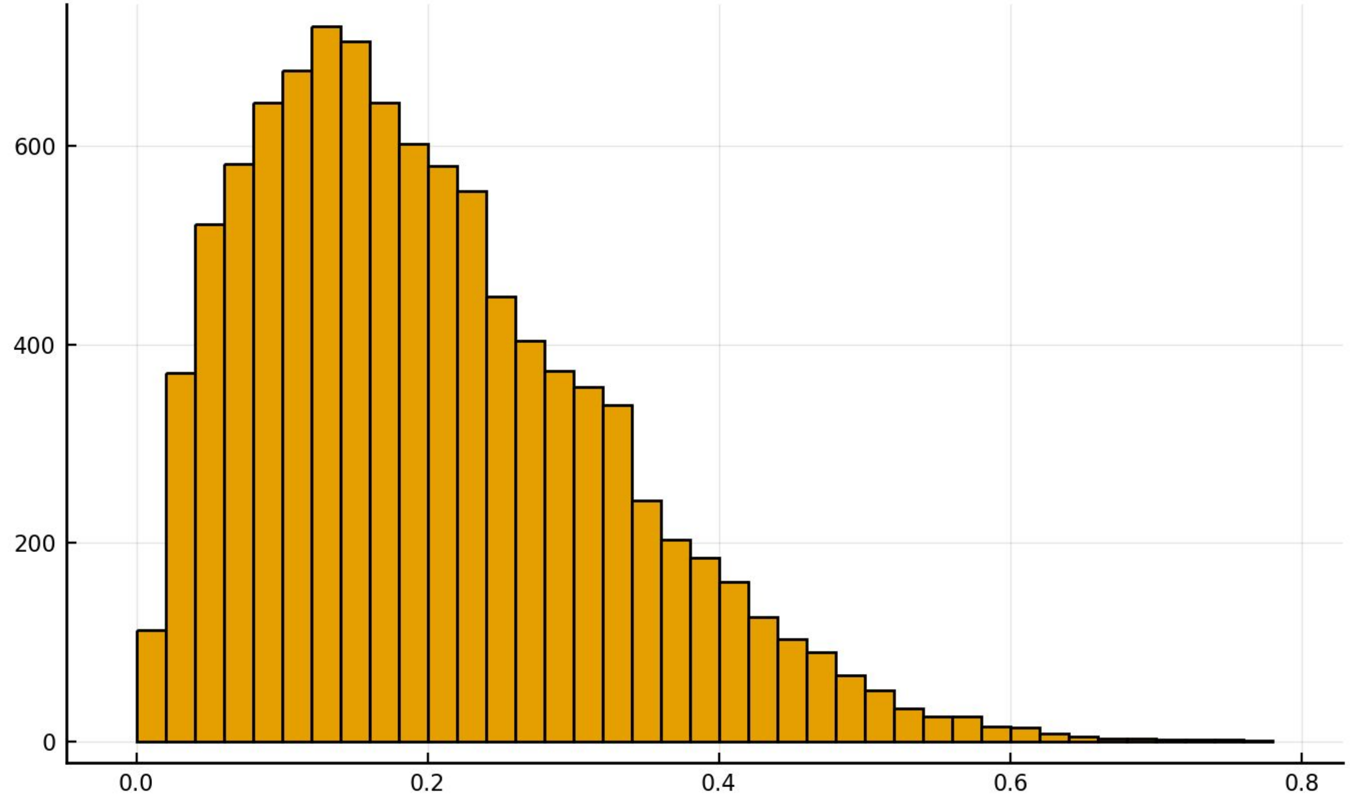
Employee	New Salary (Annual)
John	408,562.78
Mike	524,922.50
Kate	570,671.90
Shane	495,470.25
Catty	695,920.38
Mathew	637,457.63
Sam	625,171.31
:	:



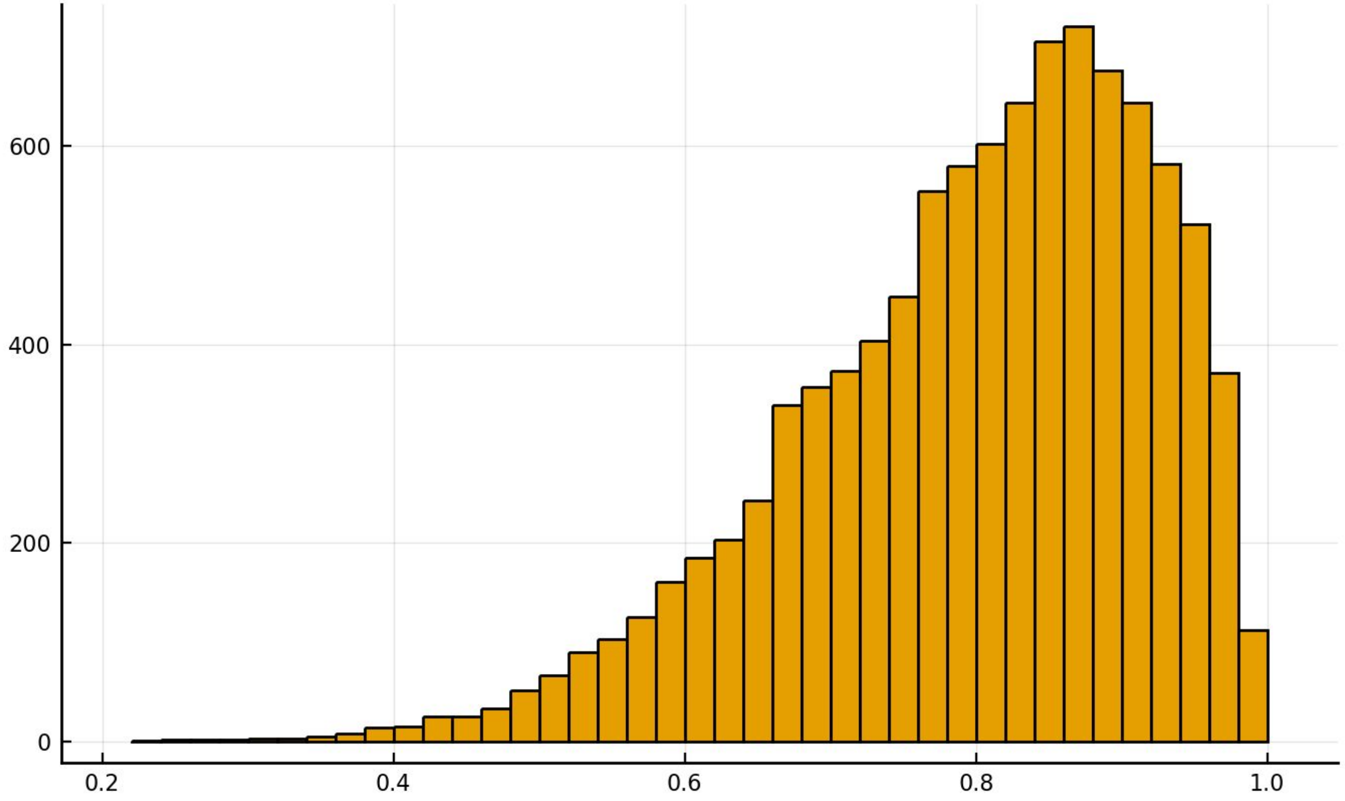
Bell Curve



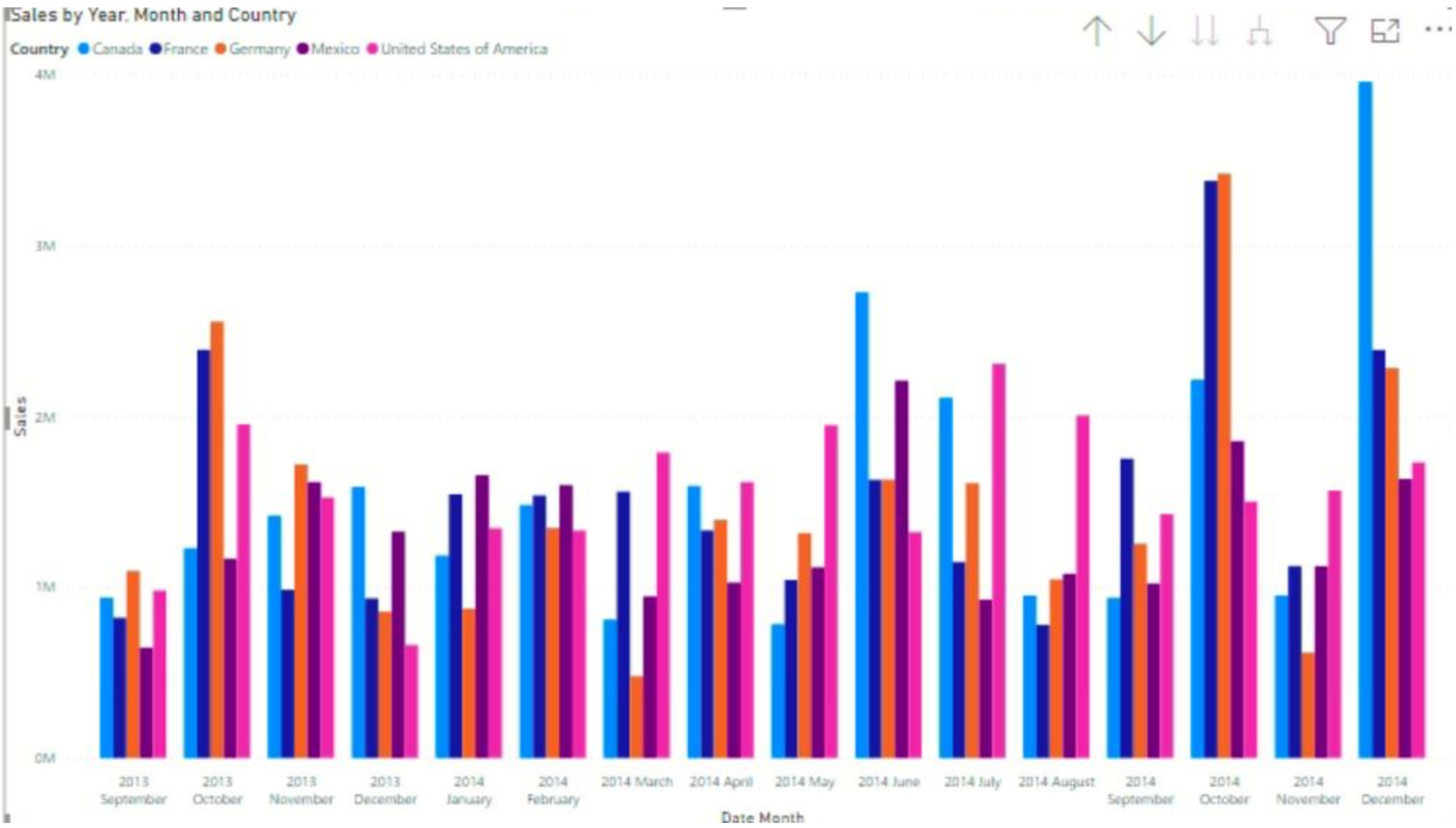
Skewed to the Right (Positively Skewed)



Skewed to the Left (Negatively Skewed)



Use Cases

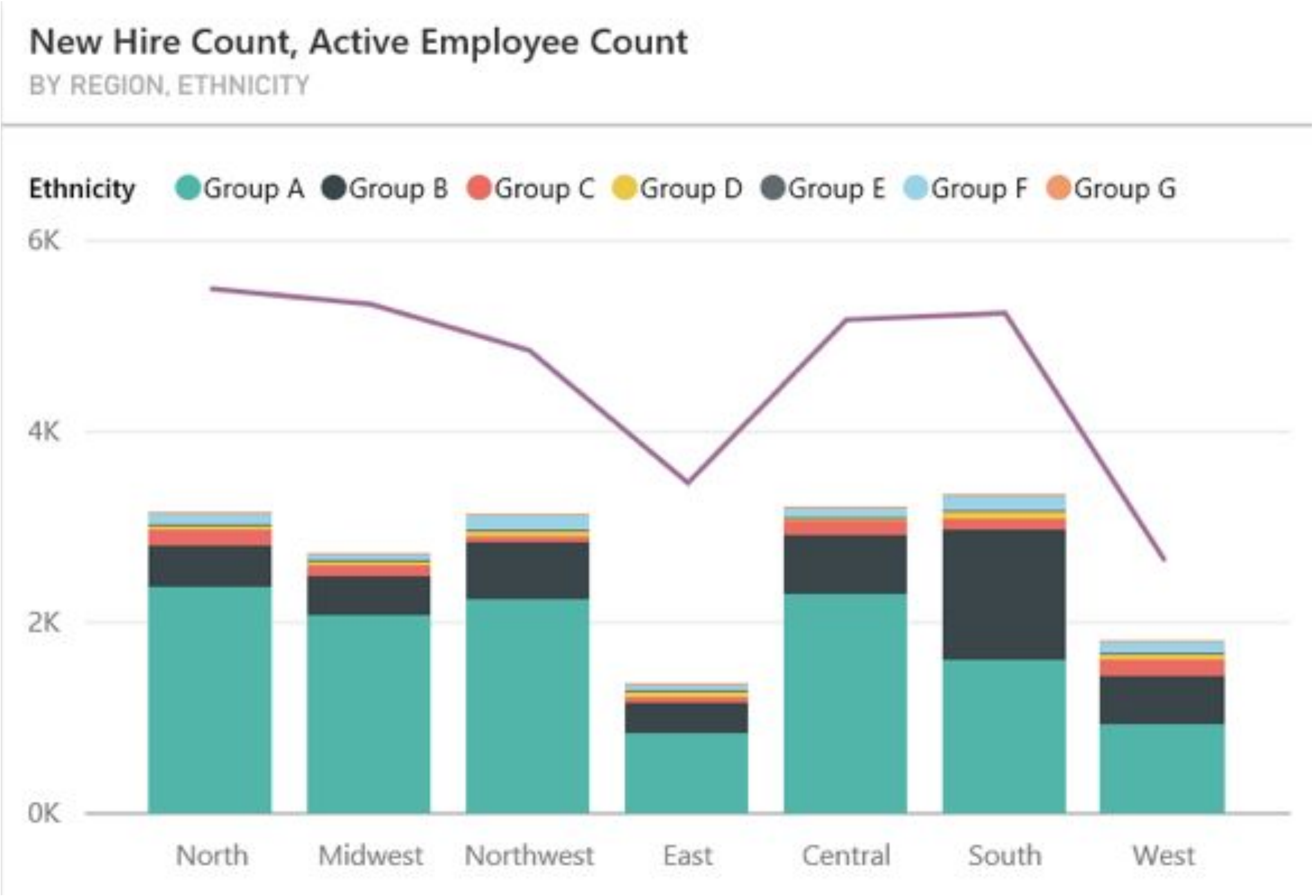


Sales by Year, Month and Country
Country Canada France Germany Mexico United States of America
4M
3M
2M
1M
0M
Sales
2013 September
2013 October
2013 November
2013 December
2014 January
2014 February
2014 March
2014 April
2014 May
2014 June
2014 July
2014 August
2014 September
2014 October
2014 November
2014 December
Date Month

Use Cases

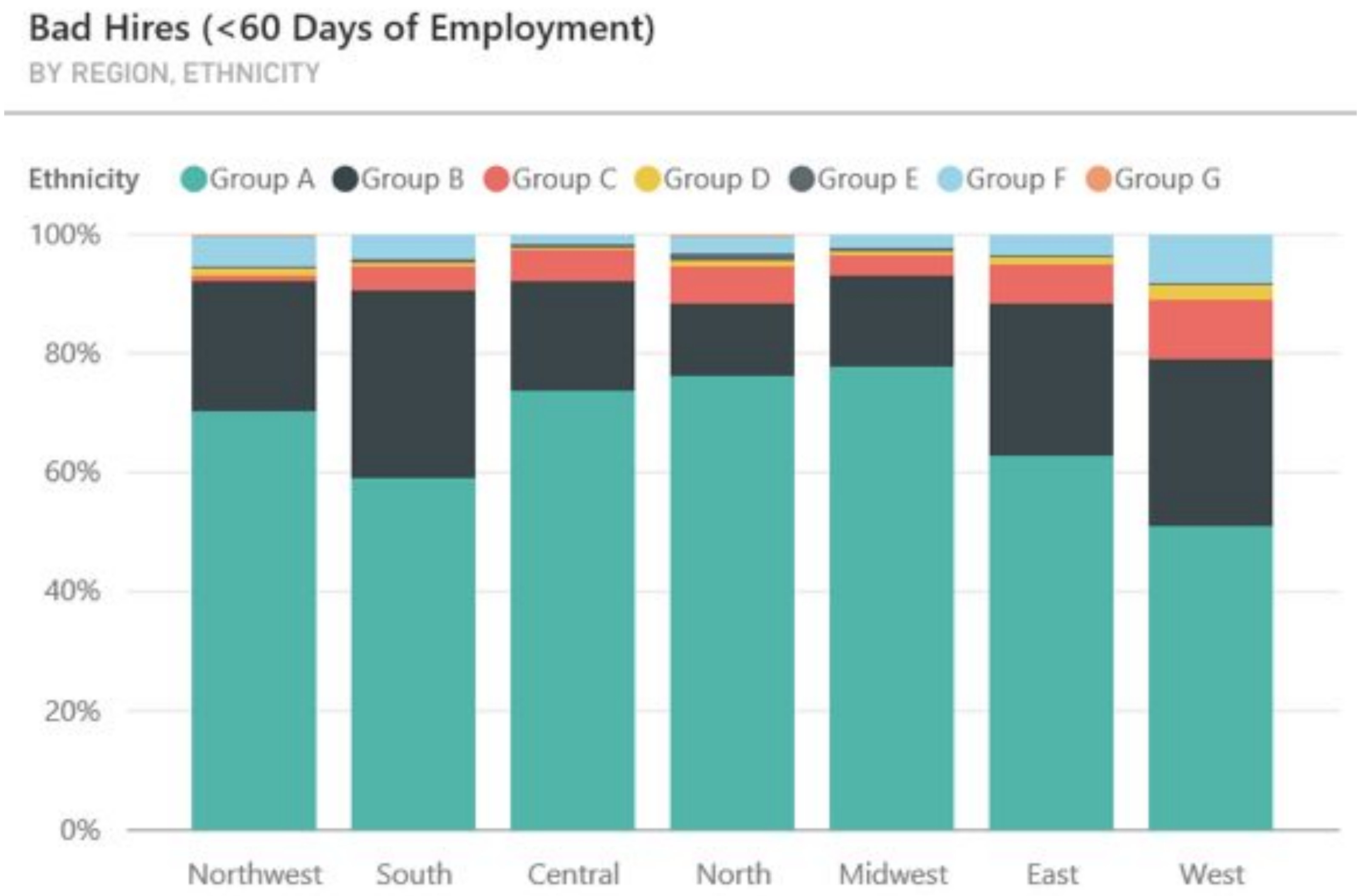
New Hire Count, New Hires Same Period Last Year, Actives YoY % Change
BY MONTH
New Hire Count
New Hires SPLY
Actives YoY % Change
2,500
2,000
1,500
1,000
500
0
10 %
8 %
6 %
4 %
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov

Use Cases



New Hire Count, Active Employee Count
BY REGION, ETHNICITY
Ethnicity
Group A
Group B
Group C
Group D
Group E
Group F
Group G
6K
4K
2K
0K
North
Midwest
Northwest
East
Central
South
West

Use Cases



Bad Hires (<60 Days of Employment)
BY REGION, ETHNICITY
Ethnicity
Group A
Group B
Group C
Group D
Group E
Group F
Group G
100%
80%
60%
40%
20%
0%
Northwest
South
Central
North
Midwest
East
West

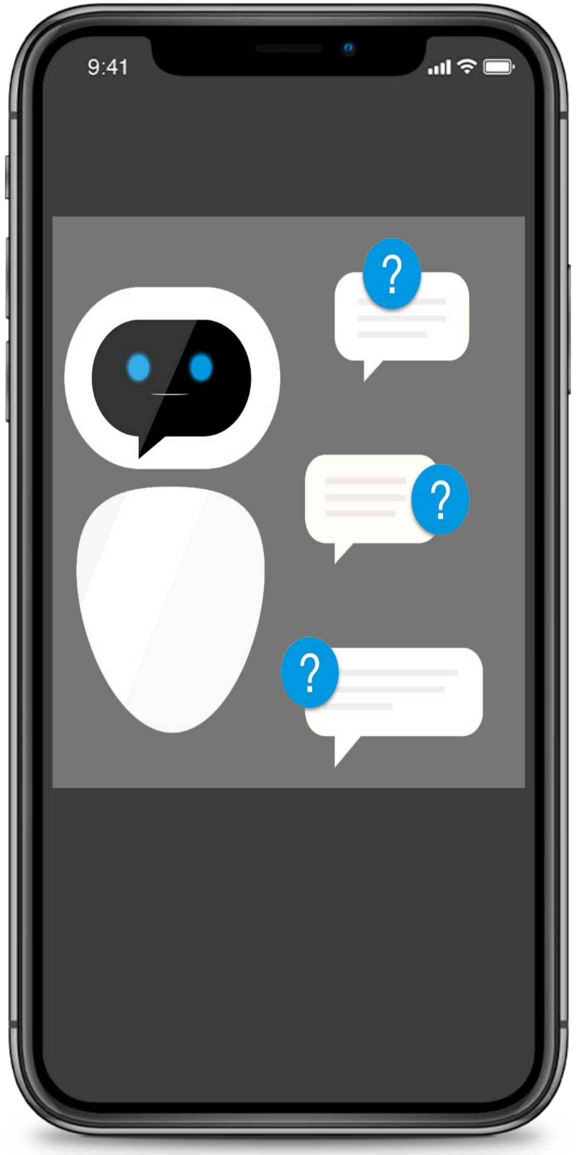
Measures of Dispersion

- **Range** -- difference between the smallest and largest values in the data
- **Variance** -- calculated by taking the average of the squared differences between each value and the mean
- **Standard Deviation** -- square root of the variance
- **Skew** -- measure of whether some values of a variable are extremely different from the majority of the values

Measure of Centrality

- **Mean** -- the sum of a variable's values divided by the total number of values
- **Median** -- the middle value of a variable
- **Mode** -- the value that occurs most often

Questions?



QBE DATA SCIENCE WORKSHOP

Data analysis

How data is leveraged in expanding business

Measures of Dispersion

- **Range** -- difference between the smallest and largest values in the data
- **Variance** -- calculated by taking the average of the squared differences between each value and the mean
- **Standard Deviation** -- square root of the variance
- **Skew** -- measure of whether some values of a variable are extremely different from the majority of the values

Measure of Centrality

- **Mean** -- the sum of a variable's values divided by the total number of values
- **Median** -- the middle value of a variable
- **Mode** -- the value that occurs most often

Table of Contents

POINTS FOR DISCUSSION:

Descriptive Statistics

Describes the data you have but can't be generalized beyond that.

Source: StatCast: What is a t-test?
https://www.youtube.com/watch?v=0Pd3dc1GcHc

Why not just look at the means?

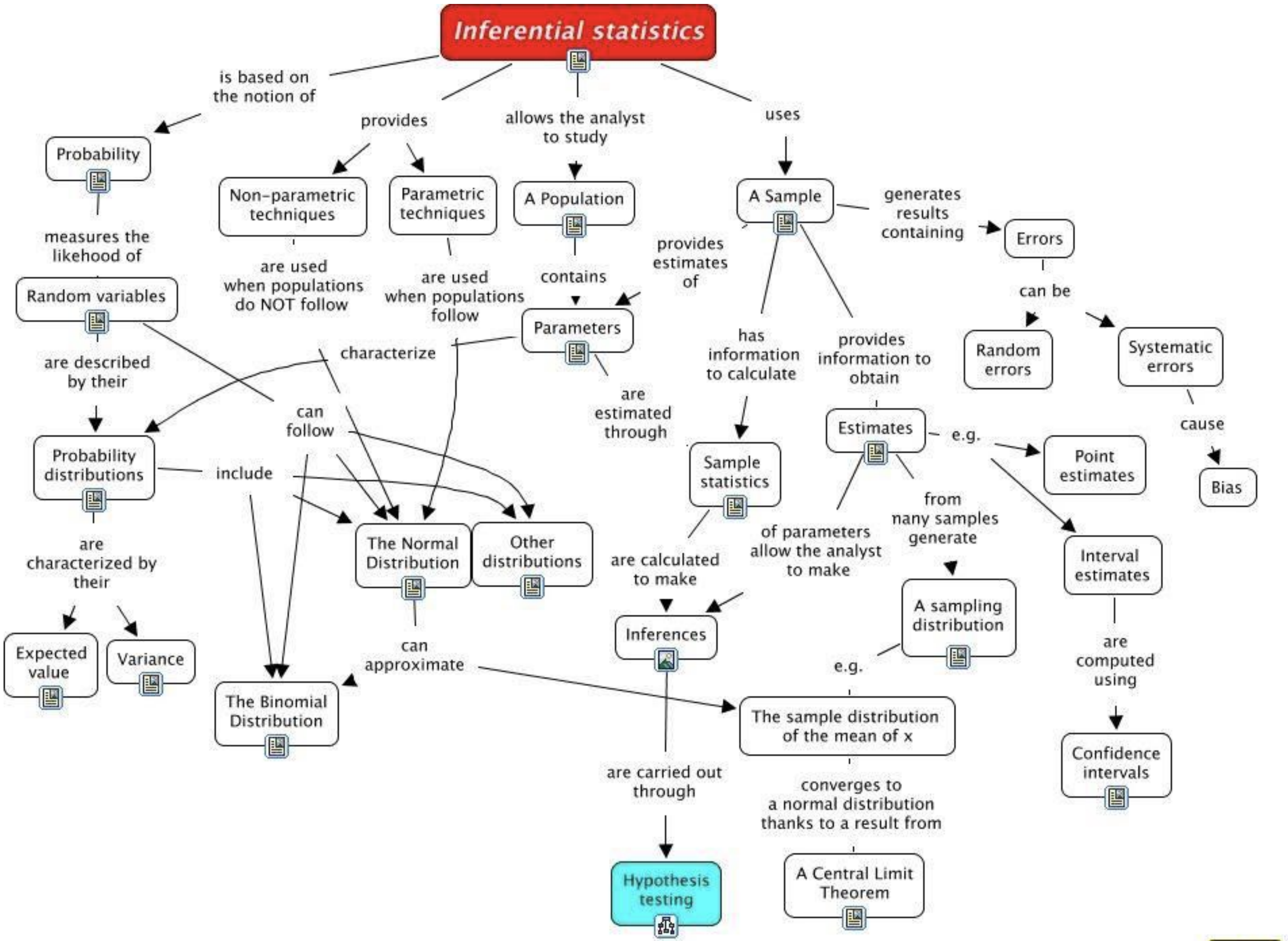
Looking at the means may show a difference, but we can't be sure if the difference is **reliable** (statistically significant).

Source: StatCast: What is a t-test?
https://www.youtube.com/watch?v=0Pd3dc1GcHc

Inferential Statistics

These are statistics, such as t-test, that allow us to make inferences about the population beyond our sample data.

Source: StatCast: What is a t-test?
https://www.youtube.com/watch?v=0Pd3dc1GcHc



Source: The Statisticians FB Group

What is a t-test?

used to compare two sets of samples (continuous variables)

A t-test is a statistic that checks if two means (averages) are **reliably** different from each other.

Source: StatCast: What is a t-test?
https://www.youtube.com/watch?v=0Pd3dc1GcHc

Assumptions of t-test

- Data points are independent
- Sample size is small ($n < 30$)
- Sample values are accurate
- Population variance is known

Source: Difference between t-test and z-test
https://keydifferences.com/difference-between-t-test-and-z-test.html

Assumptions of z-test

- Data points are independent
- Sample size is large (n > 30)
- Population variance is not known

Source: Difference between t-test and z-test
https://keydifferences.com/difference-between-t-test-and-z-test.html

BASIS FOR COMPARISON	T-TEST	Z-TEST
Meaning	T-test refers to a type of parametric test that is applied to identify, how the means of two sets of data differ from one another when variance is not given.	Z-test implies a hypothesis test which ascertains if the means of two datasets are different from each other when variance is given.
Based on	Student-t distribution	Normal distribution
Population variance	Unknown	Known
Sample Size	Small	Large

Source: Difference between t-test and z-test
https://keydifferences.com/difference-between-t-test-and-z-test.html

z-test statistic

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

x is the sample mean

σ is the population standard deviation

n is the sample size

μ is the population mean

Source: Difference between t-test and z-test
https://keydifferences.com/difference-between-t-test-and-z-test.html

t-test statistic

used to compare two sets of samples (continuous variables)

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

x is the sample mean

s is the sample standard deviation

n is the sample size

μ is the population mean

Source: Difference between t-test and z-test
https://keydifferences.com/difference-between-t-test-and-z-test.html

t-test statistic

$$t = \frac{\text{variance between groups}}{\text{variance within groups}}$$

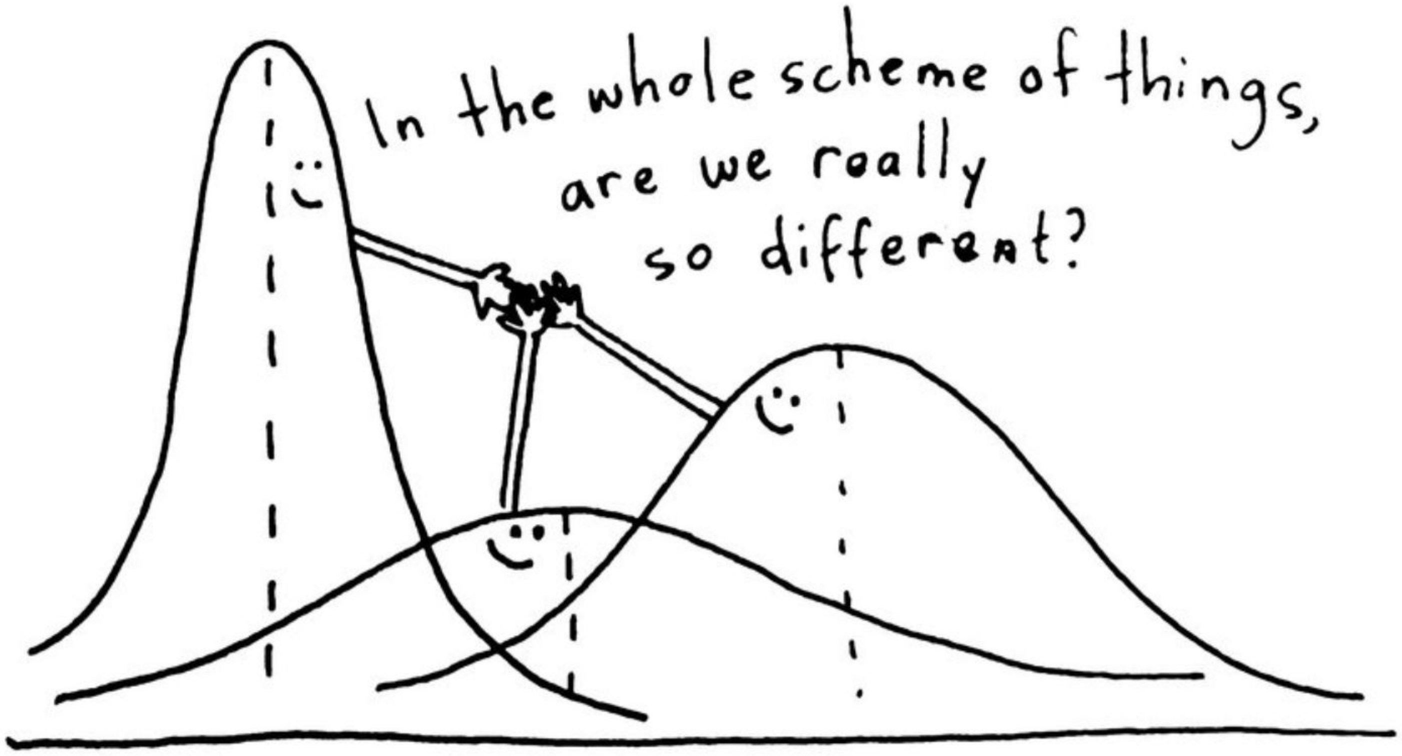
A big t-value = different groups

A small t-value = similar groups

Source: Difference between t-test and z-test
https://keydifferences.com/difference-between-t-test-and-z-test.html

Analysis of Variance (ANOVA)

used to compare multiple samples in a single test (generalized vs t-test)



Source: Questionpro

ANOVA

used to compare multiple samples in a single test (generalized vs t-test)

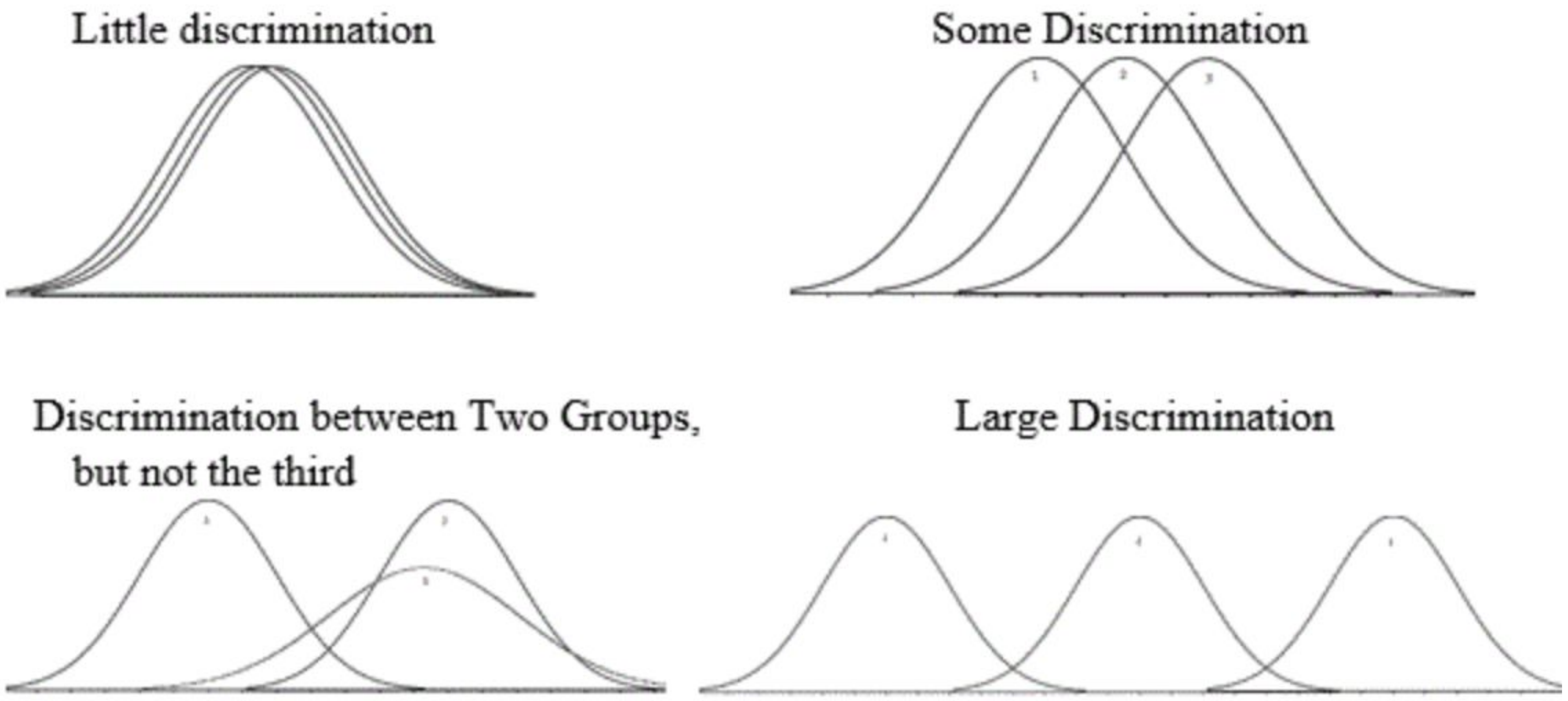
$H_o: \ \mu_1 = \mu_2 = \cdots = \mu_L$ *Null hypothesis*

$H_1: \ \mu_l \neq \mu_m$ *Alternate hypothesis*

Source: Analytics Vidhya

ANOVA

used to compare multiple samples in a single test (generalized vs t-test)



Source: PsychStat - Missouri State

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Source: http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf

What is a p-value?

Each t-value has a corresponding p-value.

What is a p-value?

The p-value is the probability that the pattern of data in the sample could be produced by random data (from the population).

What is a p-value?

if p = 0.05, there is 5% chance
there is no real difference

if p = 0.01, there is 1% chance

What is a p-value?

p-values are dependent on the sample size

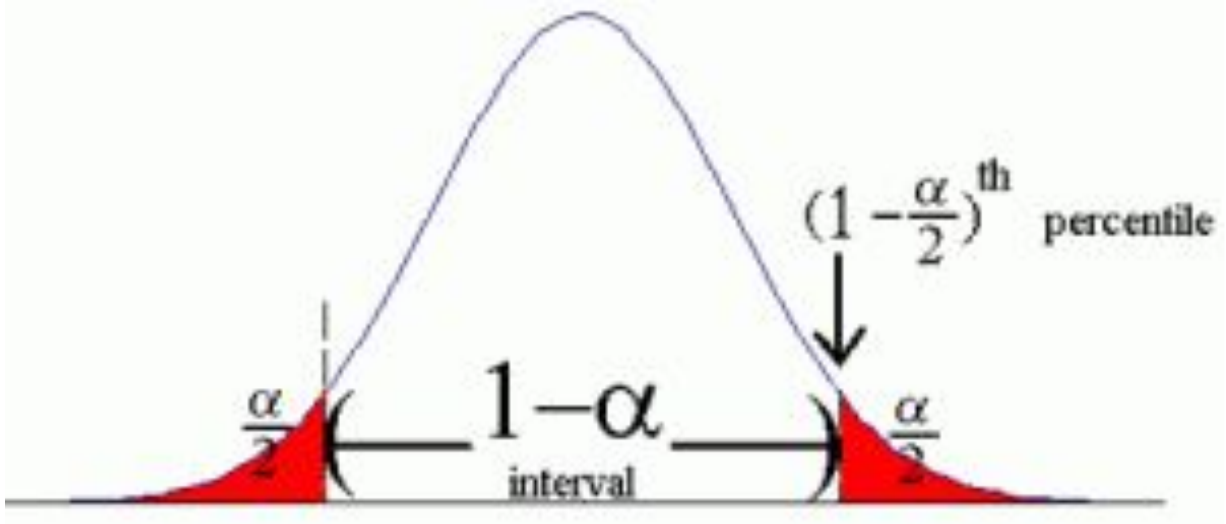
What is a p-value?

Most authors refer to

$p < 0.05$ as statistically significant

$p < 0.001$ as statistically highly significant

Significance Level: Alpha



Alpha levels (sometimes just called “significance levels”) are used in hypothesis tests;

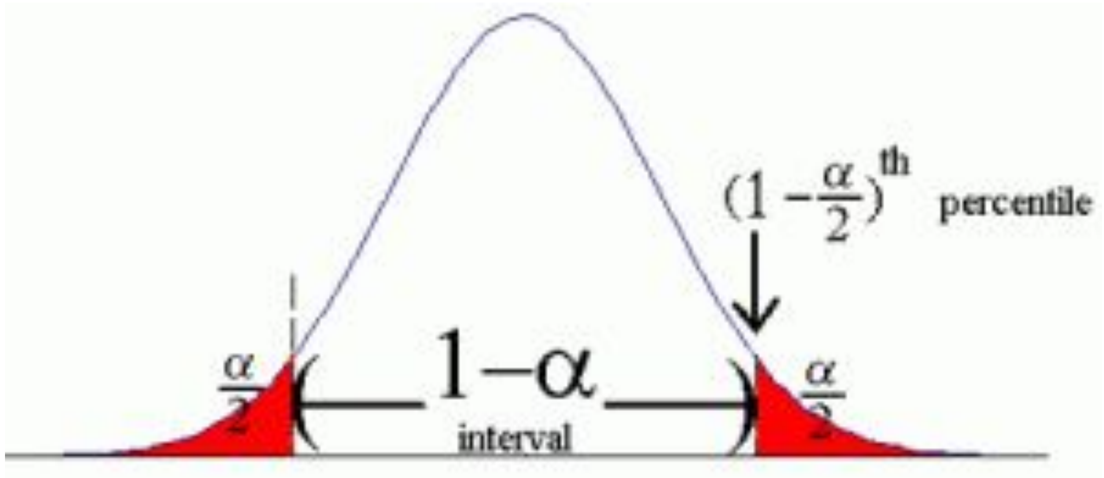
- it is the probability of making the wrong decision when the null hypothesis is true.

- **one-tailed test:** entire 5% of the alpha level in one tail (in either the left, or the right tail)
- **two-tailed test:** splits your alpha level in half (as in the image to the left).

Significance Level: Alpha

Let's say you're working with the standard alpha level of 0.5 (5%). A two tailed test will have half of this (2.5%) in each tail. Very simply, the hypothesis test might go like this:

1. A null hypothesis might state that the mean = *x*. You're testing if the mean is way above this *or* way below.
2. You run a t-test, which churns out a t-statistic.
3. If this test statistic falls in the top 2.5% or bottom 2.5% of its probability distribution (in this case, the t-distribution), you would reject the null hypothesis.

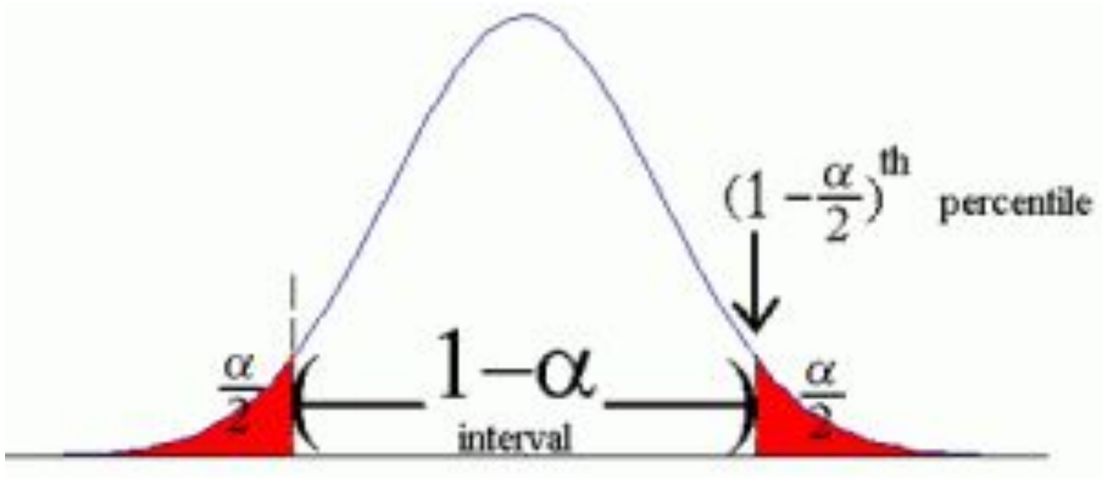


The "cut off" areas created by your alpha levels are called rejection regions. It's where you would reject the null hypothesis, if your test statistic happens to fall into one of those rejection areas. ***The terms "one tailed" and "two tailed" can more precisely be defined as referring to where your rejection regions are located.***

Significance Level: Alpha

The "cut off" areas created by your alpha levels are called rejection regions. It's where you would reject the null hypothesis, if your test statistic happens to fall into one of those rejection areas.

The terms "one tailed" and "two tailed" can more precisely be defined as referring to where your rejection regions are located.



Limitations

The results are only applied to the populations that resemble the sample tested

Limitations

The sample and the population should be roughly normal in distribution.

Limitations

Each group should have roughly the same number of data points.

Limitations

All data should be independent. Each point should not influence each other.

Limitations

And so on...

How to overcome limitations?

study data and apply appropriate statistics

Input Variable		Outcome variable					
		Nominal	Categorical (>2 Categories)	Ordinal	Quantitative Discrete	Quantitative Non-Normal	Quantitative Normal
	Nominal	X^2 or Fisher's	X^2	X^2-trend or Mann - Whitney	Mann-Whitney	Mann-Whitney or log-rank[a]	Student's t test
	Categorical (2>categories)	X^2	X^2	Kruskal-Wallis[b]	Kruskal-Wallis[b]	Kruskal-Wallis[b]	Analysis of variance[c]
	Ordinal (Ordered categories)	X^2-trend or Mann - Whitney	*	Spearman rank	Spearman rank	Spearman rank	Spearman rank or linear regression[d]
	Quantitative Discrete	Logistic regression	*	*	Spearman rank	Spearman rank	Spearman rank or linear regression[d]
	Quantitative non-Normal	Logistic regression	*	*	*	Plot data and Pearson or Spearman rank	Plot data and Pearson or Spearman rank and linear regression
	Quantitative Normal	Logistic regression	*	*	*	Linear regression[d]	Pearson and linear regression

My data is in categories (nominal) or ordered (ordinal)	non-parametric
My data has equal intervals (interval)	Parametric
My data represents a **normal distribution** of a population (bell curve shape / equal number above and below the mean)	Yes, normal distribution, fairly distributed → parametric Not a normal distribution / extreme scores → non-parametric
The variance (spread around the mean) of the 2 samples are **not** significantly different.	Not significantly different / SD is quite similar → parametric Significantly different. SD are quite different → non-parametric

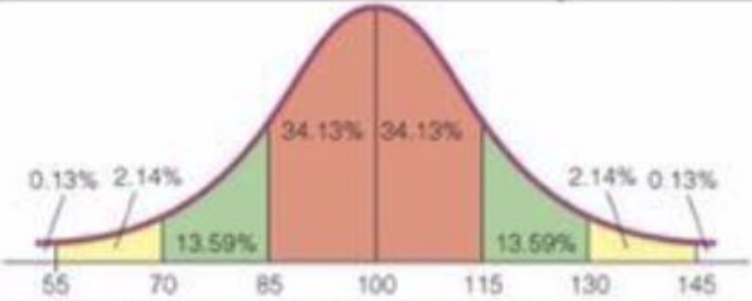


Figure 7.22 Normal distribution of IQ scores

You can **only** be asked to work out the **sign test** (non-parametric).
If you are not told otherwise, assume parametric.

Chi-Square Statistic

find if there are any significant association between the two categorical variables

- determines if a sample data matches a population
- tests to see whether distributions of categorical variables differ from each other

Source: Statistics How To
https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/chi-square/

Chi-Square Statistic

find if there are any significant association between the two categorical variables

- Example:
  - Null Hypothesis: Gender and voting preferences are independent
  - Alternative Hypothesis: Gender and voting preferences are not independent

$$\chi^2_c = \sum \frac{(O_i - E_i)^2}{E_i}$$

Source: Statistics How To
https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/chi-square/

	Testing difference (unrelated) Independent Groups	**Testing difference** (related) Repeated Measures / Matched Pairs	**Testing association or correlation**
Nominal Data in categories (males, females, football teams)	Chi-Squared test	Sign test	Chi-Squared test
Ordinal Ordered in some way via rank or rating scale. ('unsafe' data/subjective)	Mann-Whitney	Wilcoxon	Spearman's rho.
Interval Set measurements where each unit is the same (time, temperature, weight)	***Unrelated t-Test*** ***(parametric)***	***Related t-Test*** ***(parametric)***	***Pearson's r*** ***(parametric)***

Scenario

The HappyJoy Company sells chicken to customers.

Scenario

They are rewarding its best performing store with an all-expense paid trip to Boracay for all employees of the branch.

Scenario

However, the company didn't give any criteria of what *'best performing'* means and managers have to defend the performance of their stores.

Two-sample t-test

$$t = \frac{\overline{x}_a - \overline{x}_b}{\sqrt{s_a^2/N_a - s_b^2/N_b}}$$

$$t = \frac{157.28 - 149.11}{\sqrt{30.84^2/27 - 5.75^2/27}}$$

$$t = 1.354$$

The Two Managers

Store A mean: 157.28

Store A SD: 30.84

Store B mean: 149.11

Store B SD: 5.75

The Story of Two Managers

Store A	126	187	103	113	197	119	189
	191	161	174	205	166	115	121
	198	135	137	173	169	155	179
	167	184	120	184	110	151	175

Store B	141	148	148	157	141	144	146
	141	144	142	140	146	145	141
	158	153	154	151	154	152	152
	153	156	158	156	149	153	152

The Story of Two Managers

Store A	Week 1	126	187	103	113	197	119	189
	Week 2	191	161	174	205	166	115	121
	Week 3	198	135	137	173	169	155	179
	Week 4	167	184	120	184	110	151	175

Store B	Week 1	141	148	148	157	141	144	146
	Week 2	141	144	142	140	146	145	141
	Week 3	158	153	154	151	154	152	152
	Week 4	153	156	158	156	149	153	152

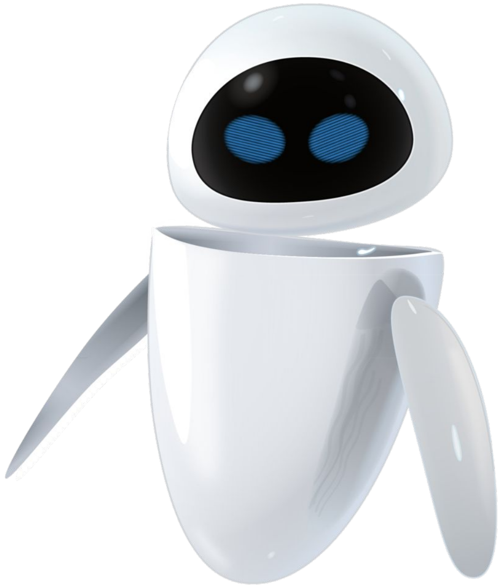
The Third Manager

{for ANOVA}

The Two Managers

I have the best performing store.
I have the highest sales average of happy chicken last month at ***157*** happy chicken per day.

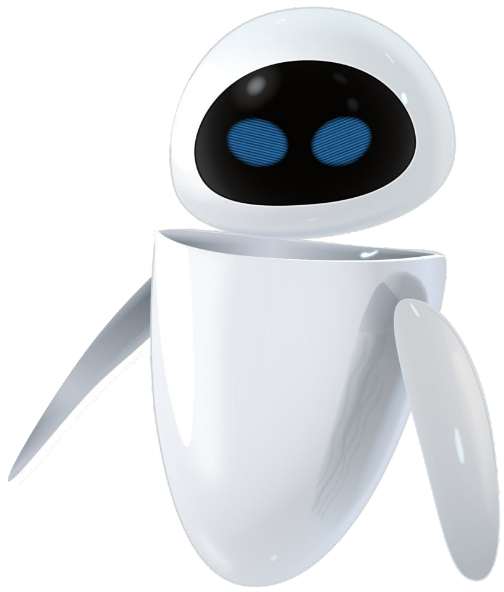
Your standard deviation is 30 while I only have 6. So I have a more stable store than you have.



The Two Managers

Store A mean: 157.28 per day
Store A SD: 30.84

Store B mean: 149.11 per day
Store B SD: 5.75



Perform t-test

Two-sample t-test

Use this equation:

$$t = \frac{\overline{x_a} - \overline{x_b}}{\sqrt{s_a^2/N_a - s_b^2/N_b}}$$

x are sample means

s are standard deviations

N are sample sizes

Hypothesis Testing

- State the hypothesis
- Pick level of significance
- Collect Data
- Calculate test statistic
- Decision
- Conclusion

Hypothesis Testing

- State the hypotheses:

  Null hypothesis, Ho:

  “Store A and Store B has the same average chicken sold for the last month”

  Alternative hypothesis, Ha:

  “They have different averages”

Hypothesis Testing

- State the hypotheses:

Null hypothesis, Ho: $\overline{x_a} = \overline{x_b}$ $t < t_{crit}$

Alternative hypothesis, Ha: $\overline{x_a} \neq \overline{x_b}$ $t > t_{crit}$

Hypothesis Testing

- Pick level of significance:

$$\alpha = 0.05$$

or

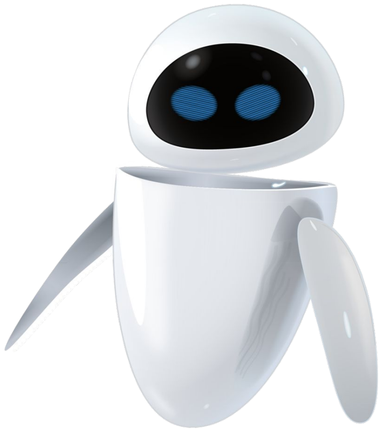
$$\alpha = 0.001$$

Hypothesis Testing

- Collect Data:

Store A mean: 157.28 per day
Store A SD: 30.84

Store B mean: 149.11 per day
Store B SD: 5.75



t-test statistic

used to compare two sets of samples (continuous variables)

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

x is the sample mean

s is the sample standard deviation

n is the sample size

μ is the population mean

Source: Difference between t-test and z-test
https://keydifferences.com/difference-between-t-test-and-z-test.html

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Source: http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf

Hypothesis Testing

- Calculate the statistic:

$$t = \frac{\overline{x_a} - \overline{x_b}}{\sqrt{s_a^2/N_a - s_b^2/N_b}}$$

$$t = \frac{157.28 - 149.11}{\sqrt{30.84^2/27 - 5.75^2/27}}$$

$$t = 1.354$$

Hypothesis Testing

- Calculate the statistic:

$$t = 1.354$$

$$t_{crit} = 2.052$$

Hypothesis Testing

- Decision:

$$t = 1.354$$

$$t_{crit} = 2.052$$

$$t < t_{crit}$$

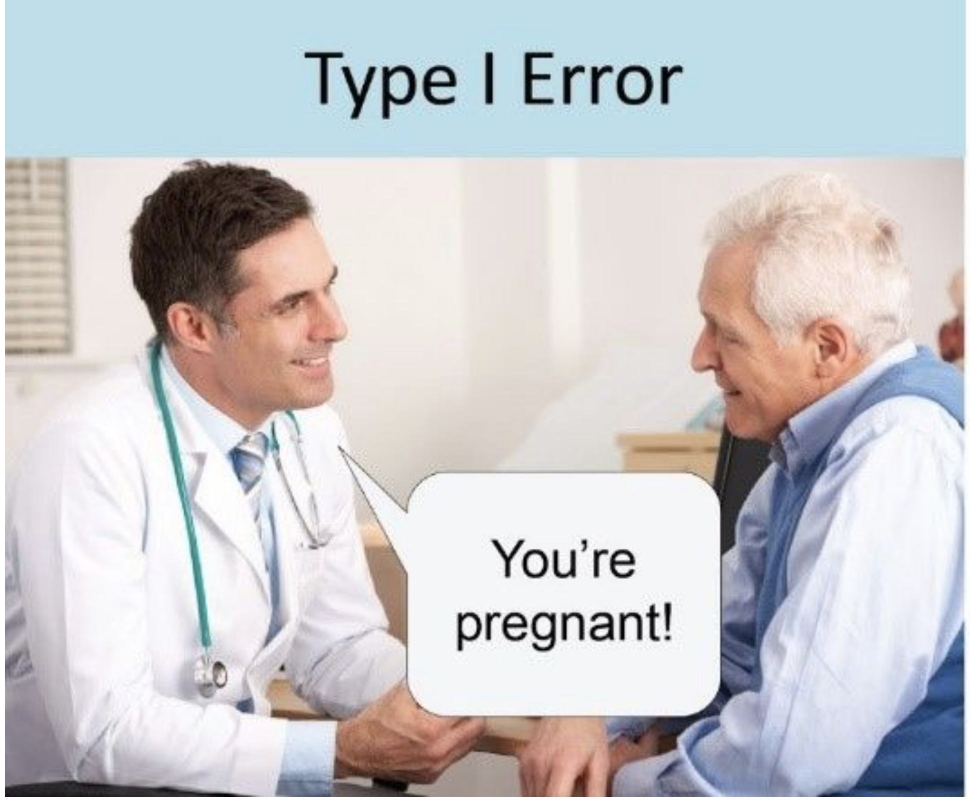
The t-score is less than the critical t score so we don’t reject the null hypothesis

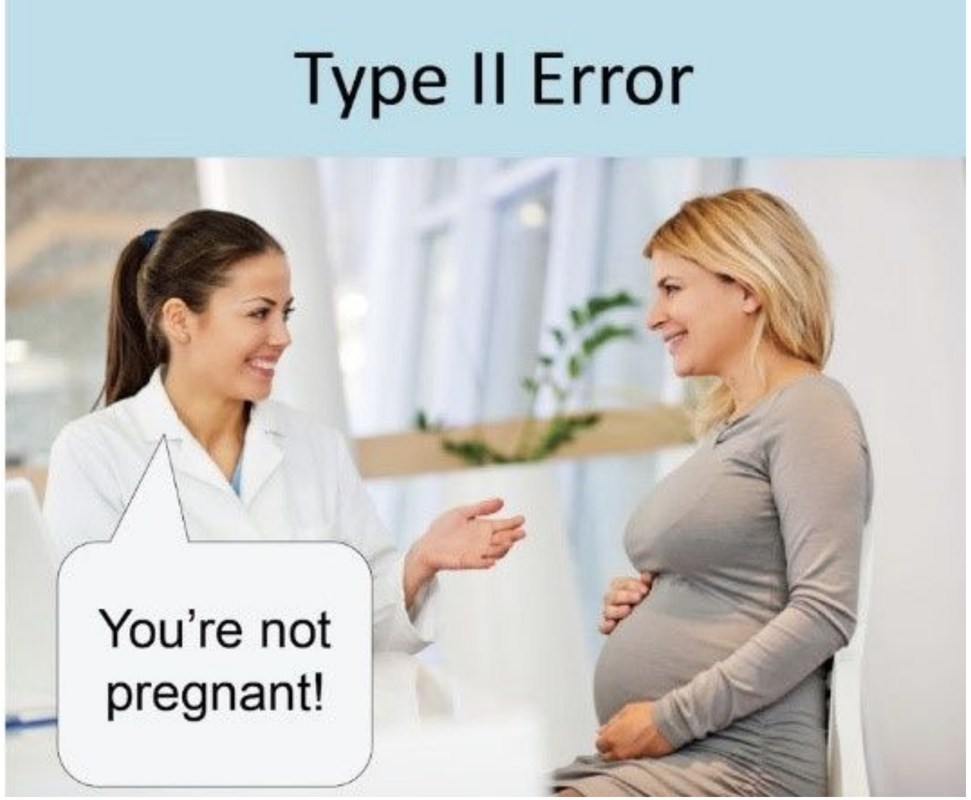
Hypothesis Testing

- Conclusion:

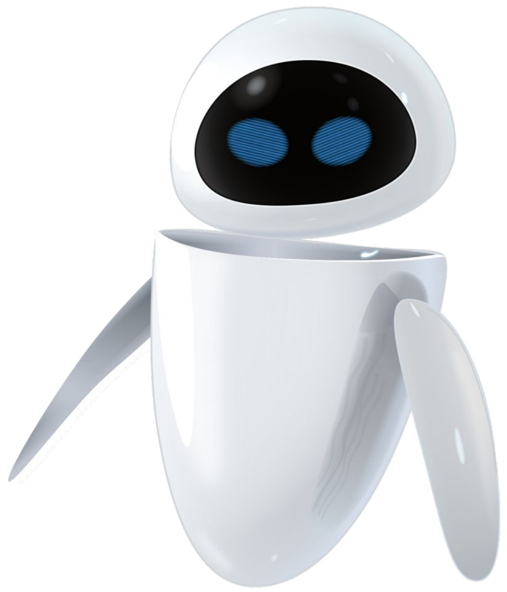
  “We have enough evidence to not reject the null hypothesis therefore Store A and Store B has no statistically significant difference between their average happychicken sales.”

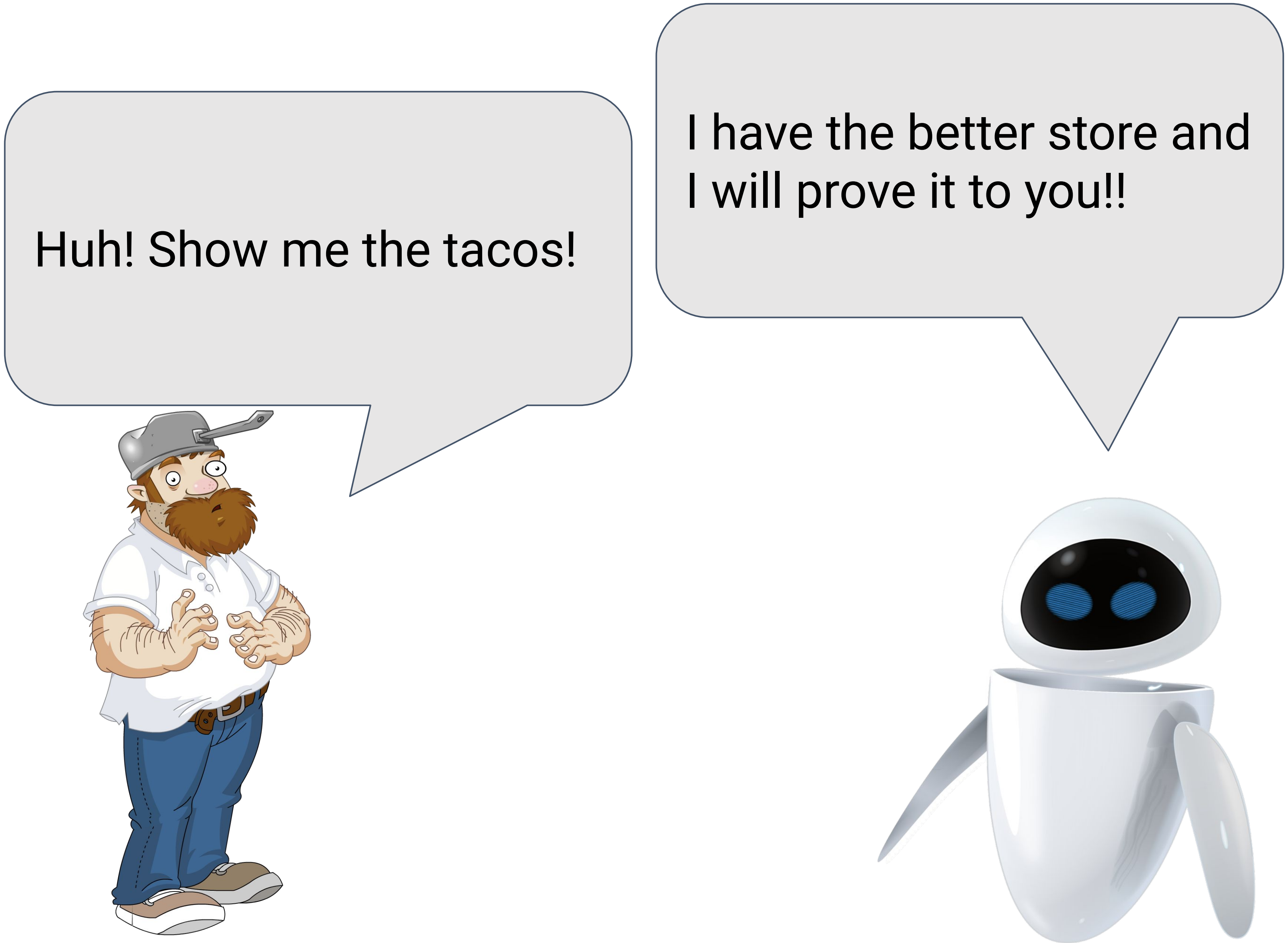
Types of Errors





Does this mean they both win (or lose)?




Huh! Show me the tacos!
I have the better store and I will prove it to you!!

One-sample t-test

National daily average of happy chicken sales:

146 chicken per day

t-test statistic

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

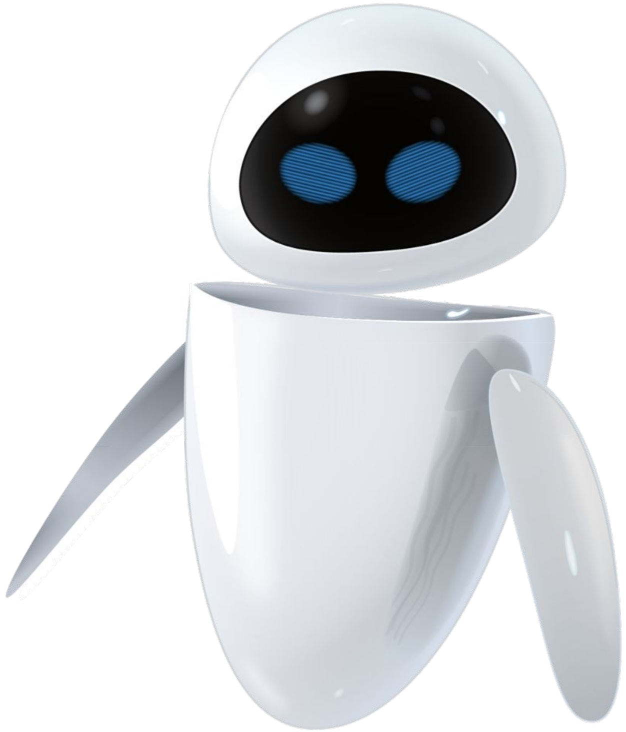
x is the sample mean

s is the sample standard deviation

n is the sample size

μ is the population mean

Let’s help Eve win the competition!!



Apply the steps of hypothesis testing to Eve’s problem

Hypothesis Testing

- State the hypothesis
- Pick level of significance
- Collect Data
- Calculate test statistic
- Decision
- Conclusion

t-test statistic

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

x is the sample mean (149.11)

s is the sample standard deviation (5.75)

n is the sample size (27)

μ is the population mean (146)

t-test statistic

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

$$t = 2.809$$

$$t > t_{crit}$$

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Source: http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf

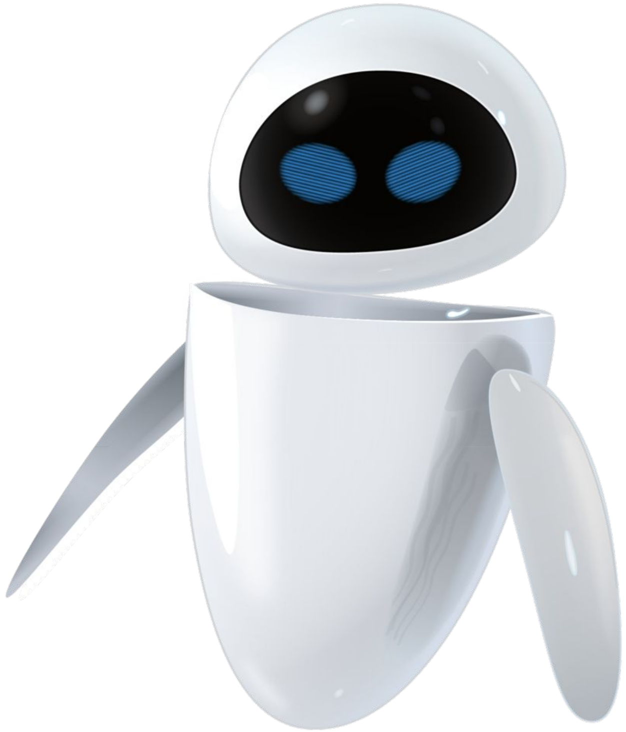
t-test statistic

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

$$t = 2.809$$

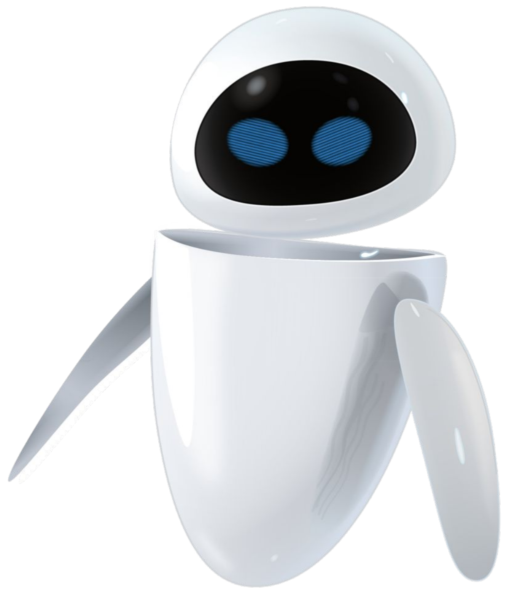
$$t > t_{crit}$$

I win! Thanks for the help!



I’m not done yet! I had promos on the 2nd week of operations. I’m sure we improved after that!

Really? We did some promotion too. Let’s see which promo was more successful.



Hypothesis Testing

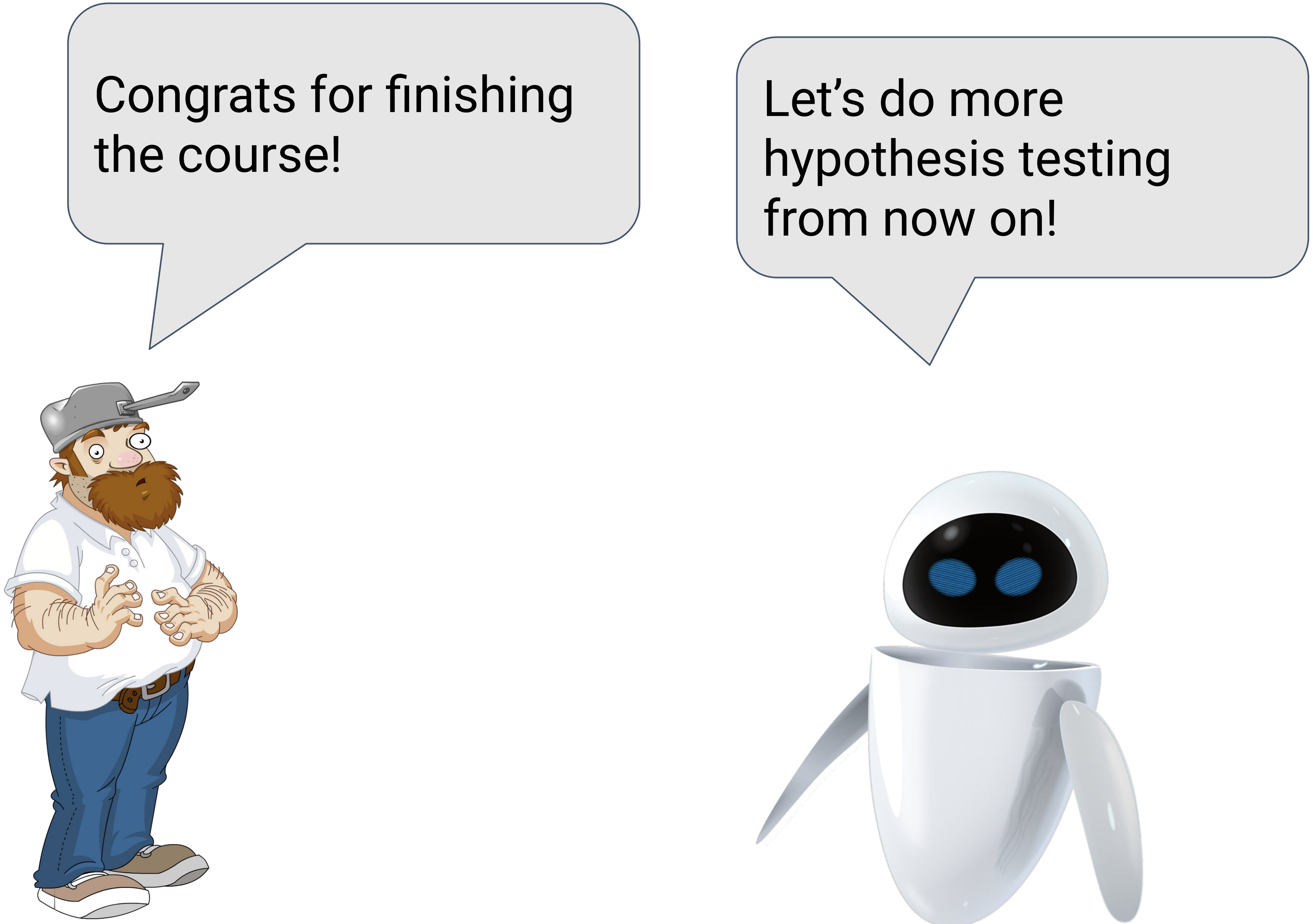
- State the hypothesis
- Pick level of significance
- Collect Data
- Calculate test statistic
- Decision
- Conclusion

Perform another hypothesis testing!

Think of a management dilemma and solve it using hypothesis testing

Hypothesis Testing

- State the hypothesis
- Pick level of significance
- Collect Data
- Calculate test statistic
- Decision
- Conclusion


Congrats for finishing the course!
Let’s do more hypothesis testing from now on!