

# Introduction to statistics

Pirommas Techitnutsarut, Ph.D.  
Data Scientist



12 July 2023



# Agenda

- 01** Overview
- 02** Illusions in statistics
- 03** Descriptive Statistics
- 04** Inferential Statistics





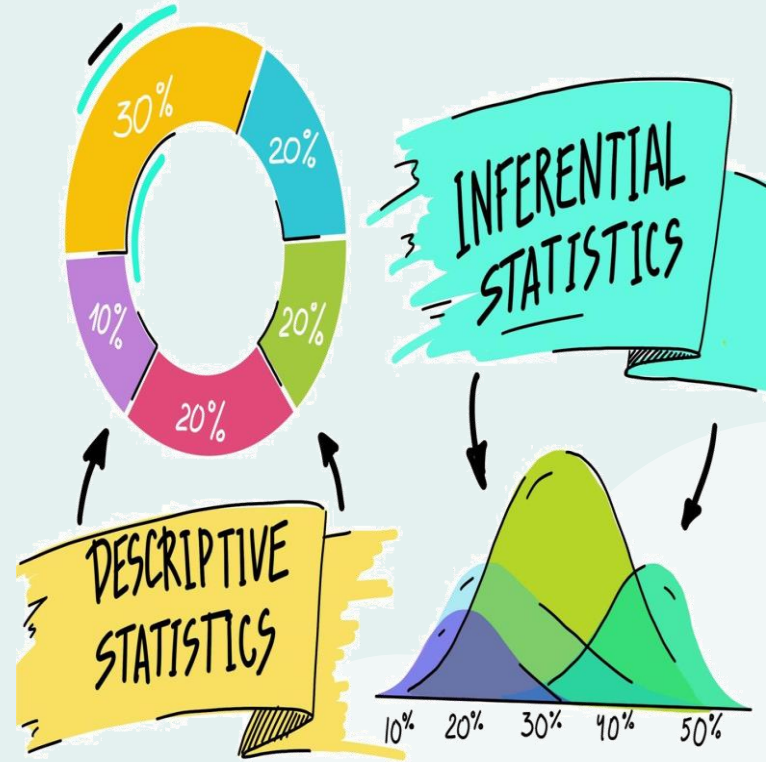
01

# Overview

.....

# What is STATISTICS?

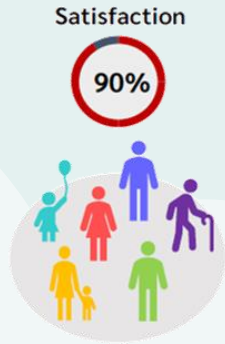
STATISTICS is the science of collecting, organizing, and interpreting numerical facts, which we call data.



# Type of Statistics

## Descriptive statistics

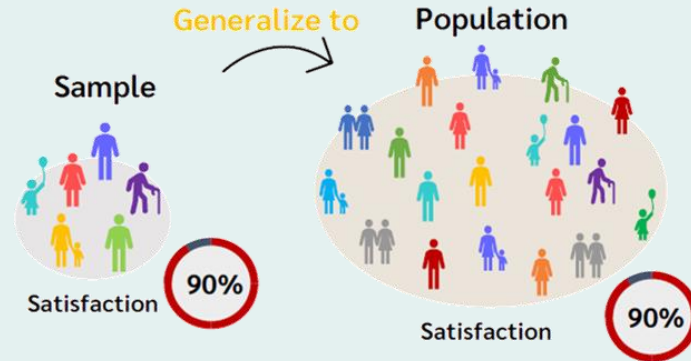
Collecting, Summarizing, and Presenting data



i.e. 90% satisfaction of all customers

## Inferential statistics

Drawing conclusions about a population based only on sample data



i.e. 90% satisfaction of a sample of 6 customers  
-> 90% satisfaction of all customers

# Data analysis process

## Data collection and preparation

Collect data

Prepare codebook

Set up structure of data

Enter data

Screen data for errors

## Exploration of data

Descriptive Statistics

Graphs

## Analysis

Explore relationship between variables

Compare groups

# Data matrix

- Data matrix
  - Case/Observation (แถว, กรณี, บุคคล)
  - Variable (คอลัมน์, ตัวแปร)
- Case/Observation คือบุคคล, สัตว์หรือสิ่งของในการศึกษา
- Variable คือลักษณะที่น่าสนใจ เช่น น้ำหนัก ส่วนสูง
- ข้อมูลดิบที่เป็นลักษณะนี้สามารถนำมาวิเคราะห์ได้ง่าย



A diagram illustrating a data matrix. A horizontal double-headed arrow above the table is labeled "variable". A vertical double-headed arrow to the left of the table is labeled "cases".

ลำดับ	จังหวัด	รายได้
1	กรุงเทพมหานคร	12000
2	กรุงเทพมหานคร	24000
3	กรุงเทพมหานคร	28000
4	นนทบุรี	42000
5	นนทบุรี	32000
6	สมุทรปราการ	19000
7	สมุทรปราการ	17500
8	สมุทรปราการ	20000
9	นนทบุรี	35000
10	กรุงเทพมหานคร	50000

# Example

Name	Dept	Project	1/7/2013	1/14/2013	1/21/2013	1/28/2013	2/4/2013	02/11/2013
Tom	IT	Budget	10	15	17	35	27	18
Dick	IT	Budget	15	18	21	18	28	13
Harry	Acct	Budget	7	12	5	33	14	9



Name	Dept	Project	Date	Hours
Tom	IT	Budget	02/11/13	18
Tom	IT	Budget	1/7/2013	10
Tom	IT	Budget	1/14/2013	15
Tom	IT	Budget	1/21/2013	17
Tom	IT	Budget	1/28/2013	35
Tom	IT	Budget	2/4/2013	27
Dick	IT	Budget	02/11/13	13
Dick	IT	Budget	1/7/2013	15
Dick	IT	Budget	1/14/2013	18
Dick	IT	Budget	1/21/2013	21
Dick	IT	Budget	1/28/2013	18
Dick	IT	Budget	2/4/2013	28
Harry	Acct	Budget	02/11/13	9
Harry	Acct	Budget	1/7/2013	7
Harry	Acct	Budget	1/14/2013	12
Harry	Acct	Budget	1/21/2013	5
Harry	Acct	Budget	1/28/2013	33
Harry	Acct	Budget	2/4/2013	14





# Types of Variables/Data

## Qualitative

### Nominal

- Counts by category
- Binary (Yes/No)
- No meaning between categories



Marital status, Type of car owned

### Ordinal

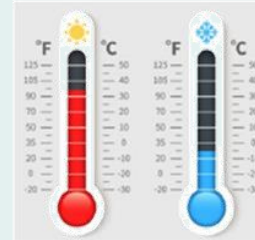
- Rank
- Scales
- Space between ranks is subjective



Service quality rating, Student letter grades

### Interval

- Zero is just another value - doesn't mean "absence of"



Temperature in Fahrenheit, Standardized exam score

## Quantitative

### Ratio

- Zero means "absence of"



Height, Age, Weekly Food Spending

# Meaningful Zero

ข้อมูลที่เป็น **Ratio** คือ ข้อมูลที่มีค่าเท่ากับศูนย์จะมีความหมายเท่ากับศูนย์จริงหรือไม่มีปริมาณจริง (**Meaningful Zero**)

- ส่วนสูง เท่ากับศูนย์ คือ ไม่มีความสูงเลย
- น้ำหนัก เท่ากับศูนย์ คือ ไม่มีน้ำหนักเลย
- ยอดขายเท่ากับศูนย์ คือ ไม่มียอดขายเลย

แต่ข้อมูลประเภทอื่นๆเช่น **Nominal Ordinal หรือ Interval** ข้อมูลที่มีค่าเท่ากับศูนย์จะมีได้หมายความว่าสิ่งนั้นจะเป็นศูนย์หรือไม่มีค่าจริง

- **Nominal Data:** การกรอกข้อมูลกำหนดให้เพศชายแทนด้วย 0 ดังนั้น ศูนย์ ในที่นี้มิได้หมายความว่าไม่มีเพศ
- **Ordinal Data:** การมีขนาดไซส์เบอร์ 0 นั้นไม่ได้มีความหมายว่าไม่มีขนาด
- **Interval Data:** อุณหภูมิ เท่ากับ ศูนย์ มิได้หมายความว่า ไม่มีอุณหภูมิ

# Time out to think



- แบบสำรวจความพึงพอใจสอบถามว่า “คุณชอบทานไอศกรีมรสอะไร”
- รสชาติของไอศกรีมเป็น *nominal*
- แต่สมมุติว่านักวิจัยนำข้อมูลเข้าระบบแล้วใช้ตัวเลขแทนรสชาติของไอศกรีม เช่น 1 = vanilla, 2 = chocolate, 3 = strawberry
- ข้อมูลรสชาติไอศกรีมจะเปลี่ยนจาก *nominal* เป็น *ordinal* หรือไม่?



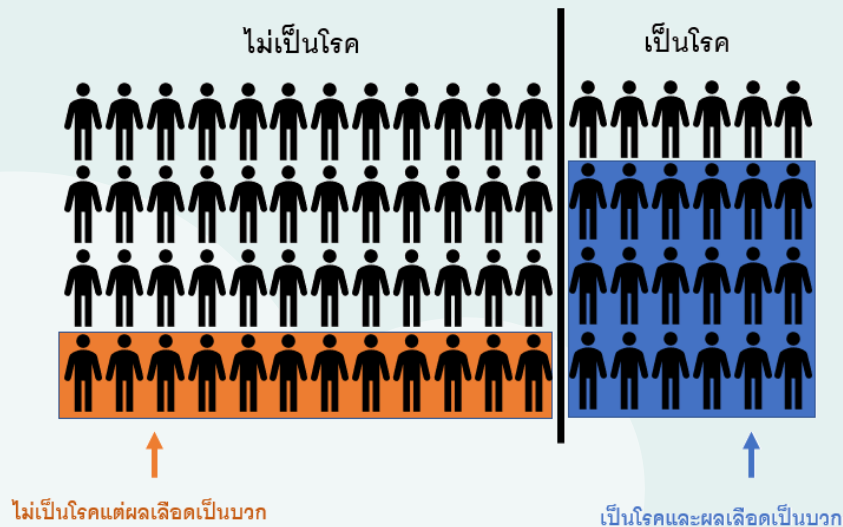
**02**

# Illusions in statistics



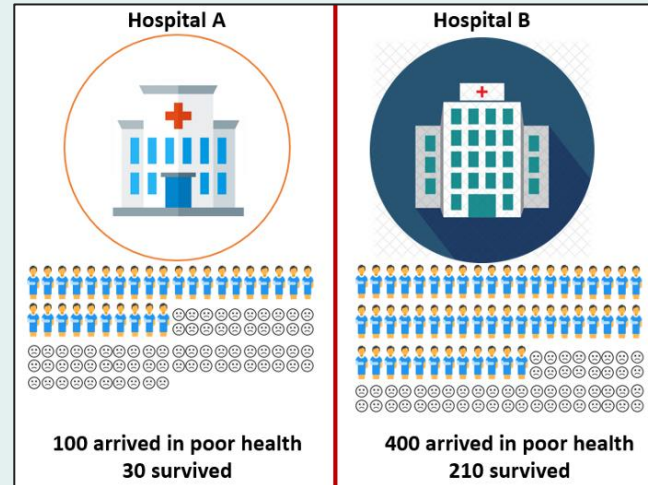
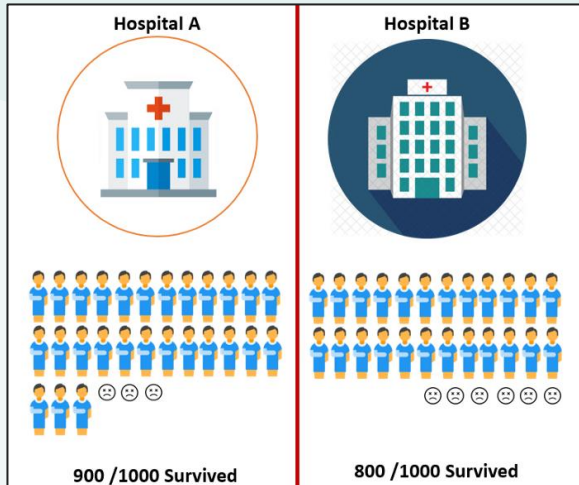
# BAYESIAN TRAP

Bayesian Trap - เป็นการตั้งข้อสรุปผิดพลาดที่เกิดจากการที่ไม่ได้คำนึงถึงจำนวนประชากร



# SIMPSON'S PARADOX

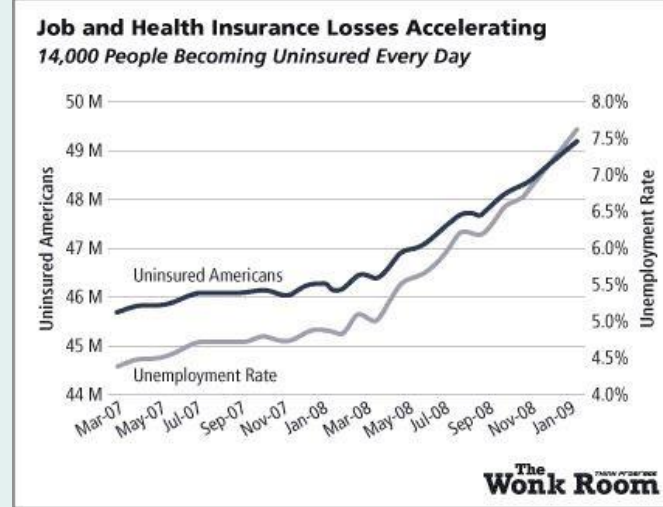
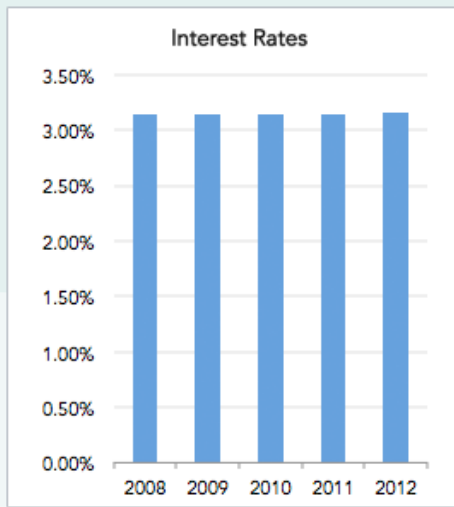
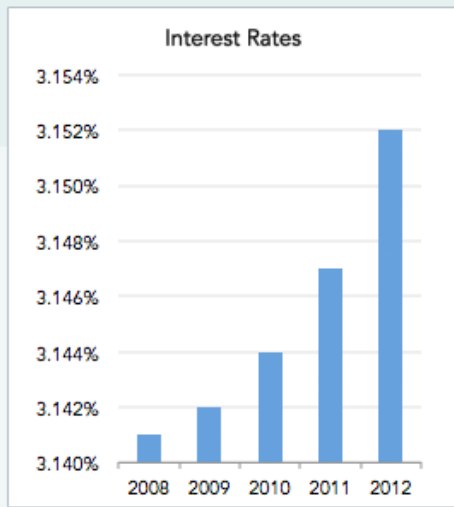
**Simpson's paradox** is a phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.



# LABELS ON Y-AXIS

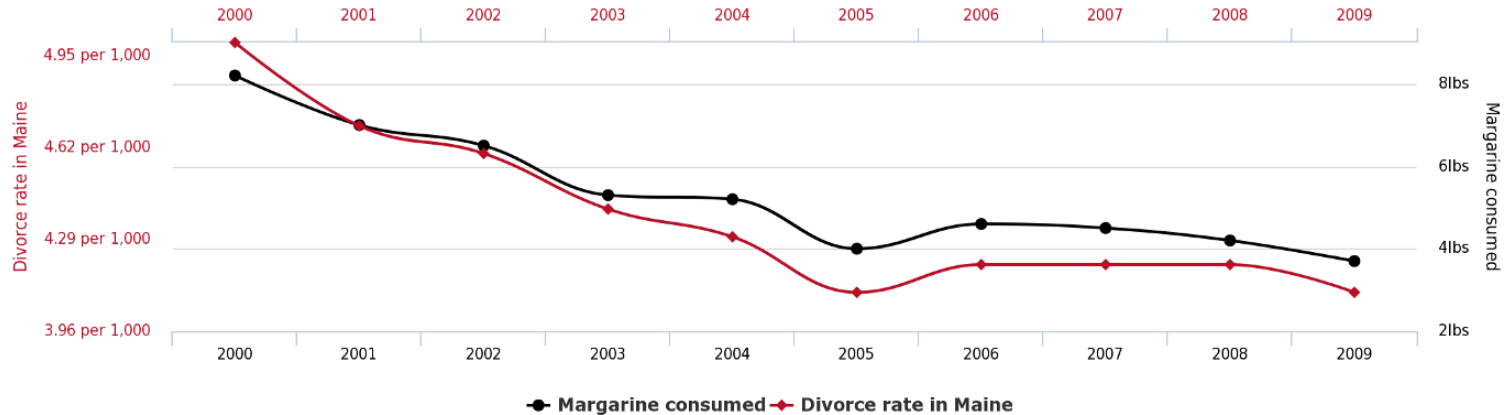
ไม่ควรเลื่อนแกนในกรณีข้อมูลเป็นแบบที่มี Meaningful Zero

## Same Data, Different Y-Axis



# CORRELATION & CAUSATION

## Divorce rate in Maine correlates with Per capita consumption of margarine



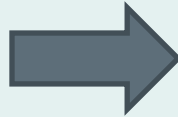


# REGRESSION TO THE MEAN

A statistical phenomenon that can make natural variation in repeated data look like real change

## Example: Israeli Air Force Cadet

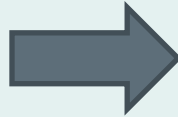
If a cadet has a good flight, praise



performance drops



If a cadet has a bad flight, criticize



performance improves



Conclusion: Criticism is best for improvement.

This is wrong because having a good flight may also depends on luck and the pilot simply are not lucky on the second flight

**03**

# Descriptive Statistics

.....



# Methods in Descriptive Statistics

## Tabular Method

Raw Data

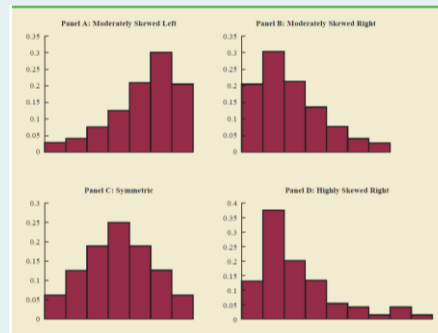
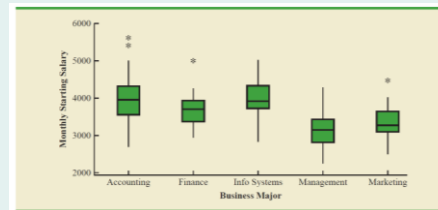
Restaurant	Quality Rating	Meal Price (\$)
1	Good	18
2	Very Good	22
3	Good	28
4	Excellent	38
5	Very Good	33
6	Good	28
7	Very Good	19
8	Very Good	11
9	Very Good	23
10	Good	13
.	.	.
.	.	.



Pivoted  
Table

Quality Rating	Meal Price				Total
	\$10–19	\$20–29	\$30–39	\$40–49	
Good	42	40	2	0	84
Very Good	34	64	46	6	150
Excellent	2	14	28	22	66
Total	78	118	76	28	300

## Graphical Method



# Why Statistics? Part I – Descriptive

ตารางด้านล่างเหมาะจะใช้ตอบคำถามใดมากกว่ากัน

Emp_Id	satisfaction_level	last_evaluation	Department	salary
IND02438	38%	53%	sales	low
IND28133	80%	86%	sales	medium
IND07164	11%	88%	sales	medium
IND30478	72%	87%	sales	low
IND24003	37%	52%	sales	low
IND08609	41%	50%	sales	low
IND14345	10%	77%	sales	low
IND16300	92%	85%	sales	low
IND27336	89%	100%	sales	low
IND41409	42%	53%	sales	low
IND01460	45%	54%	sales	low
IND07665	11%	81%	sales	low
IND13556	84%	92%	sales	low
IND20559	41%	55%	sales	low

1. พนักงานแต่ละคนได้รับเงินเดือนเหมาะสมกับการประเมินหรือไม่

คำถามรายคน

2. แผนกไหนที่มีผลการประเมินต่ำสุด

คำถามรายกลุ่ม

# Why Statistics? Part I – Descriptive

สามารถทำการจัดระเบียบข้อมูลเพื่อการวิเคราะห์ภาพรวมได้ดีขึ้น

Row Labels	Average of satisfaction_level	Average of last_evaluation
accounting	49%	72%
hr	49%	69%
IT	50%	73%
management	49%	75%
marketing	52%	71%
product_mng	53%	72%
RandD	51%	73%
sales	51%	71%
support	51%	73%
technical	50%	73%
<b>Grand Total</b>	<b>51%</b>	<b>72%</b>

**GROUP BY - Department**  
**Aggregate Function - Average**

มีสอง concept ใหญ่ๆ ในการสรุปข้อมูลในลักษณะนี้

1. Group By – แบ่งกลุ่มข้อมูลแยกโดยอะไร
2. Aggregate – ใช้วิธีไหนในการสรุปตัวเลข

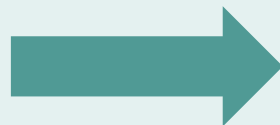


# Quantitative

ข้อมูลเชิงปริมาณ

# Quantitative

ลำดับ	จังหวัด	รายได้
1	กรุงเทพมหานคร	12000
2	กรุงเทพมหานคร	24000
3	กรุงเทพมหานคร	28000
4	นนทบุรี	42000
5	นนทบุรี	32000
6	สมุทรปราการ	19000
7	สมุทรปราการ	17500
8	สมุทรปราการ	20000
9	นนทบุรี	35000
10	กรุงเทพมหานคร	50000



ค่ากลาง

ค่าสูงสุด - ต่ำสุด

การกระจายตัว

เปรียบเทียบตำแหน่ง

# Measures of Central Tendency

## การวัดแนวโน้มเข้าสู่ส่วนกลาง (Measures of central tendency)

เป็นระเบียบวิธีทางสถิติในการหาค่าเพียงค่าเดียวที่จะใช้เป็นตัวแทนของข้อมูลทั้งหมด ค่าที่หาได้นี้จะทำให้สามารถทราบถึงลักษณะของข้อมูลทั้งหมดที่เก็บรวบรวมมาได้ ค่าที่หาได้นี้จะเป็ค่ากลาง ๆ เรียกว่า ค่ากลาง

การวัดแนวโน้มเข้าสู่ส่วนกลางมีอยู่หลายวิธีด้วยกัน ที่นิยมกันมาก ได้แก่

1

ค่าเฉลี่ยเลขคณิต (Mean)

2

มัธยฐาน (Median)

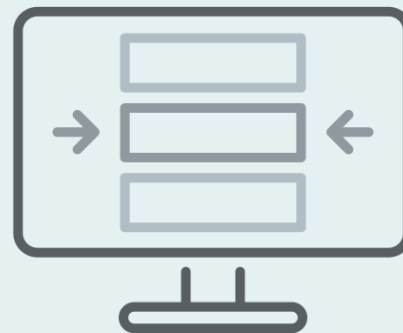
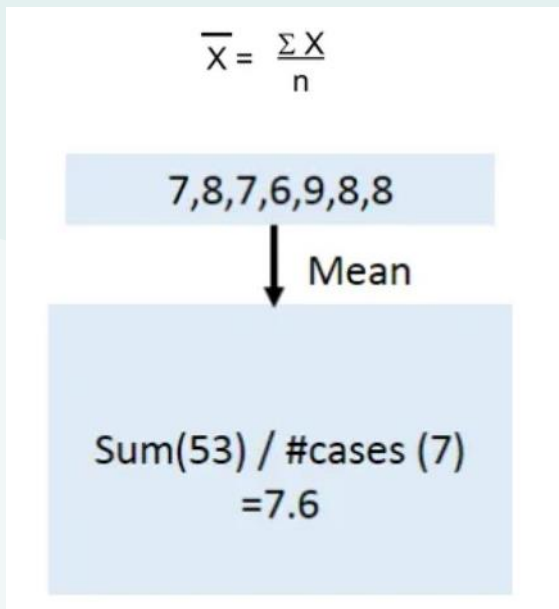
3

ฐานนิยม (Mode)



# Mean

ค่าเฉลี่ยเลขคณิต หาได้จากนำข้อมูลมารวมกัน แล้วหารด้วยจำนวนข้อมูลทั้งหมด



# Median

ค่ามัธยฐานคือค่าในตำแหน่งที่แบ่งข้อมูลออกเป็นสองส่วนเท่า ๆ กัน คือ มากกว่ามัธยฐาน 50% น้อยกว่ามัธยฐาน 50% หรือคือค่าในตำแหน่งกึ่งกลางของการแจกแจง ดังนั้นค่ามัธยฐานก็คือ ค่าของข้อมูล ณ ตำแหน่งที่  $(n+1)/2$  เมื่อเรียงลำดับข้อมูลแล้ว

## Example

7, 8, 7, 6, 9, 8, 8



~~6, 7, 7, 8, 8, 8, 9~~



The median is 8.

7, 8, 7, 6, 9, 8, 8, 7



~~6, 7, 7, 7, 8, 8, 8, 9~~



The median is  
 $(7+8)/2 = 7.5$ .

# Mode

ฐานนิยมคือ ข้อมูลที่มีการซ้ำซ้อนกันมากที่สุดในชุดข้อมูลนั้น ๆ

## Example

7, 8, 7, 6, 9, 8, 8



The mode is 8.

7, 8, 7, 6, 9, 8, 8, 7



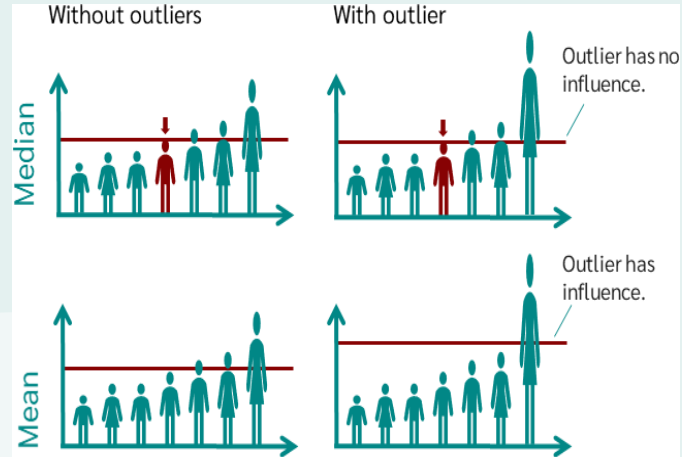
The modes are 7 and 8.

Here you have two modes.

# Mean vs Median

Q: When to use Mean & Median?

A: Depends on your application. Simple rule of thumb



1. If the distribution is mostly symmetrical and there are **no outliers**, use mean.
2. If the distribution is either skewed or there are **outliers present**, use median.

# Measure of Dispersion/variability

การพิจารณาหรือสรุปลักษณะของข้อมูลโดยใช้ค่ากลางหรือค่าเฉลี่ยเพียงอย่างเดียว อาจทำให้ไม่ทราบถึงลักษณะของข้อมูลได้ชัดเจนเนื่องจากข้อมูลที่มีค่ากลางเท่ากันแต่ลักษณะของข้อมูลอาจจะต่างกัน นั่นคือมีการกระจายของข้อมูลไม่เหมือนกัน

ดังนั้นในการเปรียบเทียบข้อมูลหลายๆชุด ควรจะพิจารณา ค่ากลาง และ การกระจาย ของข้อมูลควบคู่กันไป  
การวัดการกระจายที่นิยมใช้ในการศึกษา ได้แก่

1

พิสัย (Range)

3

สัมประสิทธิ์ความแปรผัน

2

ส่วนเบี่ยงเบนมาตรฐาน

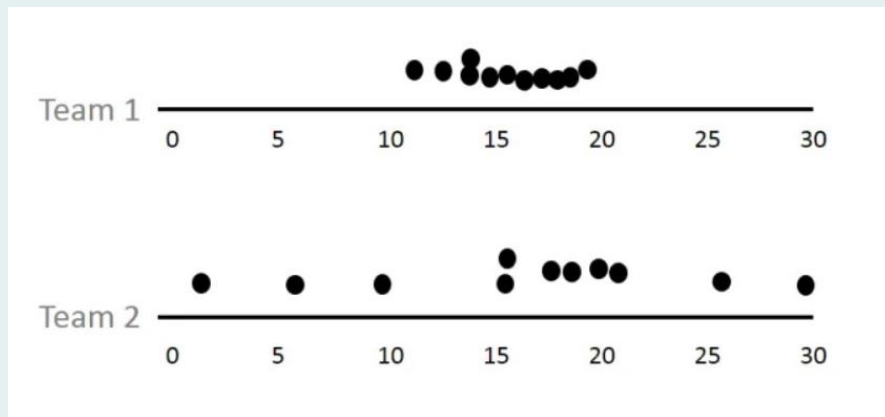
4

ส่วนเบี่ยงเบนควอไทล์

# Example

ข้อมูลใดมีการกระจายมากกว่ากัน ?

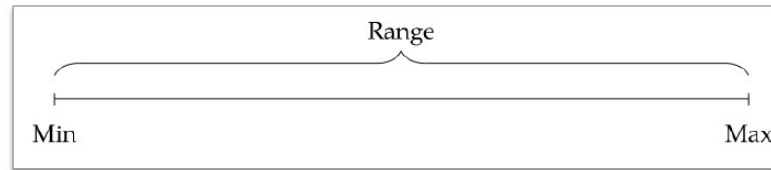
Mode = 14.1, Median = 15 และ Mean = 15



# Range

การหาค่าพิสัย (Range) ของข้อมูลหาได้โดยนำ **ข้อมูลที่มีค่าสูงที่สุด** ลบกับ **ข้อมูลที่มีค่าต่ำที่สุด** เพื่อให้ได้ค่าที่เป็นช่วงของการกระจาย ซึ่งสามารถบอกถึงความกว้างของข้อมูลชุดนั้นๆ

$$\text{พิสัย (R)} = X_{\max} - X_{\min}$$



# Standard Deviation : SD

ค่าส่วนเบี่ยงเบนมาตรฐานเป็นค่าที่ใช้วัดการกระจายของข้อมูลได้ดีกว่า เพราะไม่ได้ขึ้นอยู่กับ **ค่าสูงสุดและต่ำสุด** ของข้อมูลในกลุ่มเท่านั้น

- ถ้าข้อมูลมีค่าใกล้เคียงกับค่าเฉลี่ย ค่าเบี่ยงเบนมาตรฐานจะน้อย
- ถ้าข้อมูลมีค่าแตกต่างไปจากค่าเฉลี่ยมาก ค่าเบี่ยงเบนมาตรฐานจะมีค่ามาก

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

Population

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

Sample



# Variance: Var

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

Population

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

Sample

# Example

หาค่าการกระจายของข้อมูลกลุ่มตัวอย่างต่อไปนี้

$x$
0
24.1
5.6
14.1
17.2
8.7
19.2
14.1
27.7
15
19.3

ดังนั้น

จำนวนข้อมูล

$$n = 11$$

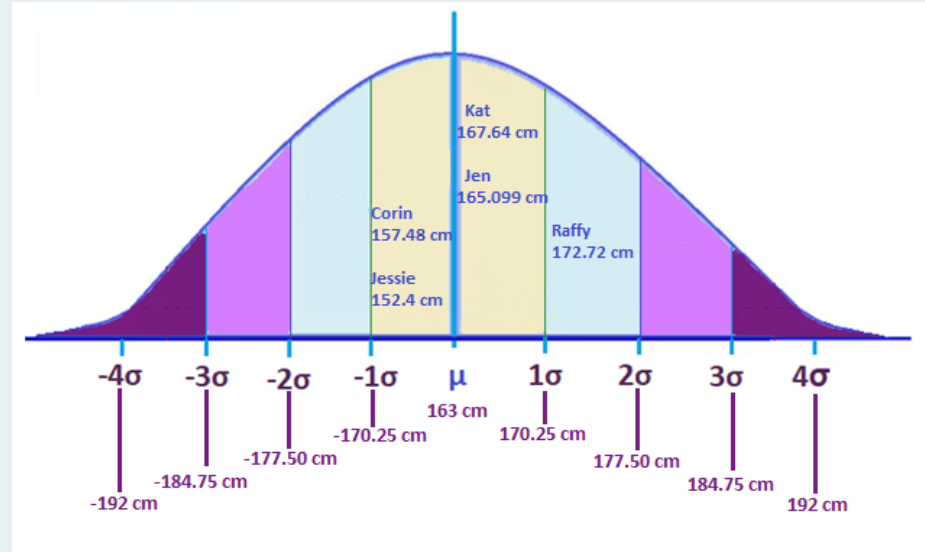
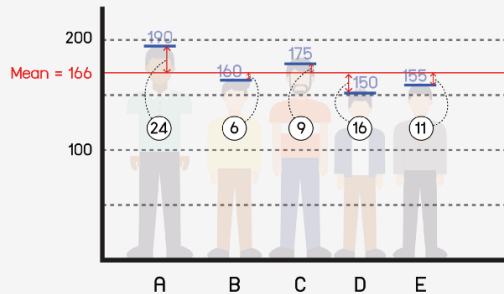
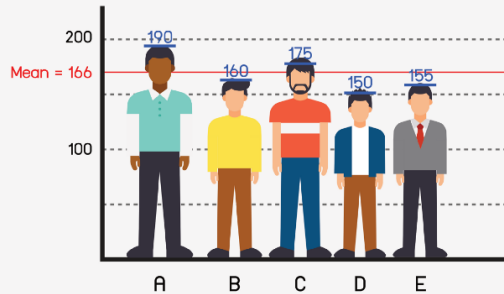
$$n - 1 = 10$$

ค่าเฉลี่ย **Mean = 15**

$$\text{ผลรวมของ } (x - \bar{x})^2 = 639.74$$

$$\text{ส่วนเบี่ยงเบนมาตรฐาน S.D.} = \sqrt{639.74/10} = 8.00$$

# Example



S.D. = 7.25 cm  
mean = 163 cm

# สัมประสิทธิ์ความแปรผัน

$x$	$x - \bar{x}$	$(x - \bar{x})^2$
0	-15	225
24.1	9.1	82.81
5.6	-9.4	88.36
14.1	-0.9	0.81
17.2	2.2	4.84
8.7	-6.3	39.69
19.2	4.2	17.64
14.1	-0.9	0.81
27.7	12.7	161.29
15	0	0
19.3	4.3	18.49

$$\text{สัมประสิทธิ์ความแปรผัน} = \frac{S.D.}{\bar{x}}$$

ค่าเฉลี่ย Mean = 15

$$\text{ส่วนเบี่ยงเบนมาตรฐาน S.D.} = \sqrt{639.74/10} = 8.00$$

$$\text{สัมประสิทธิ์ความแปรผัน} = 8/15 = 0.53$$

ใช้เปรียบเทียบระหว่างแต่ละกลุ่มตัวอย่างได้

เช่น น้ำหนัก vs ส่วนสูง

# การวัดตำแหน่งการเปรียบเทียบ

- เป็นการบอกให้ทราบว่า ค่าที่ได้มานั้นมีตำแหน่งอยู่ที่ใดหรือส่วนใดของค่าทั้งหมด
- เป็นการแสดงให้เห็น **ความสัมพันธ์ระหว่างค่าที่ได้กับข้อมูลทั้งหมด**

เช่น ครูผู้สอนต้องการแสดงให้เห็นว่าส่วนสูงของนักเรียน ก. มีความสัมพันธ์กับส่วนสูงของเพื่อนในชั้นอย่างไร จึงต้องใช้การวัดตำแหน่งการเปรียบเทียบ ได้แก่

1

ควอไทล์ (Quartiles)

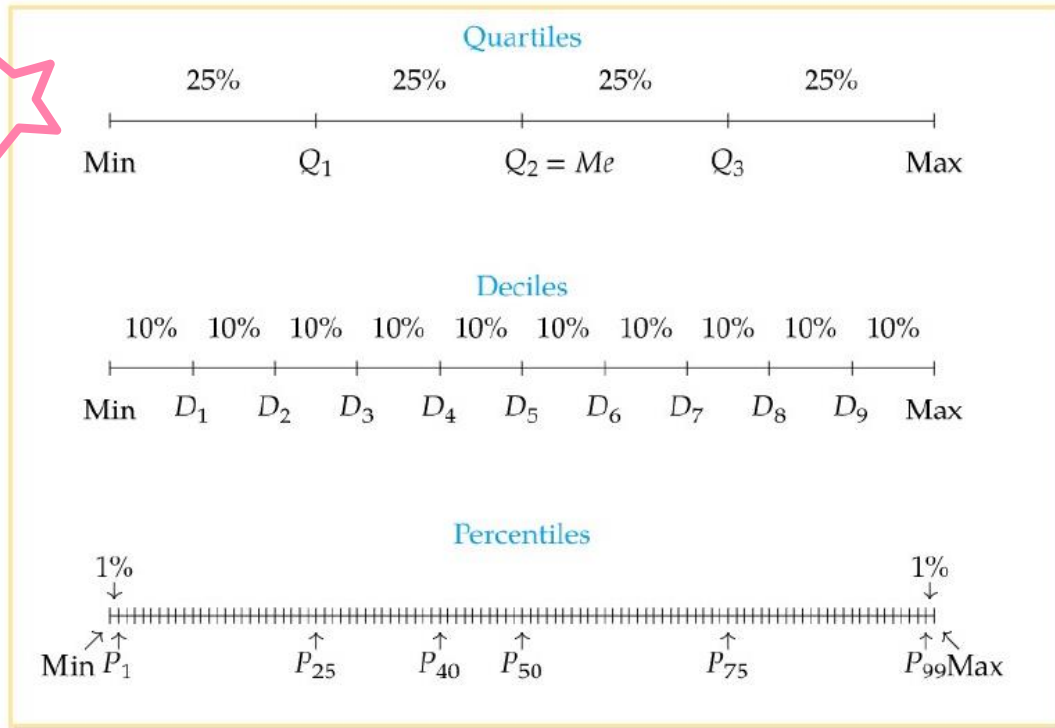
2

เดไซล์ (Deciles)

3

เปอร์เซ็นต์ไทล์ (Percentile)

# การวัดตำแหน่งการเปรียบเทียบ

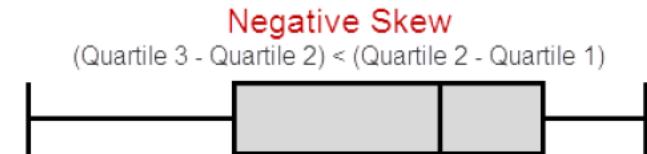
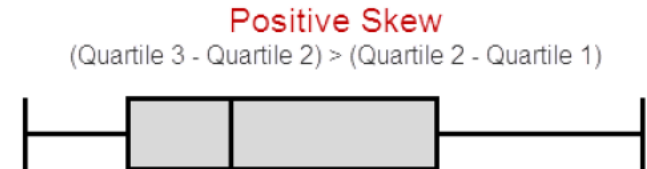
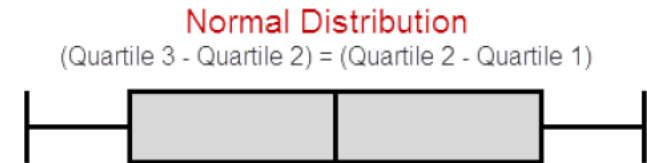
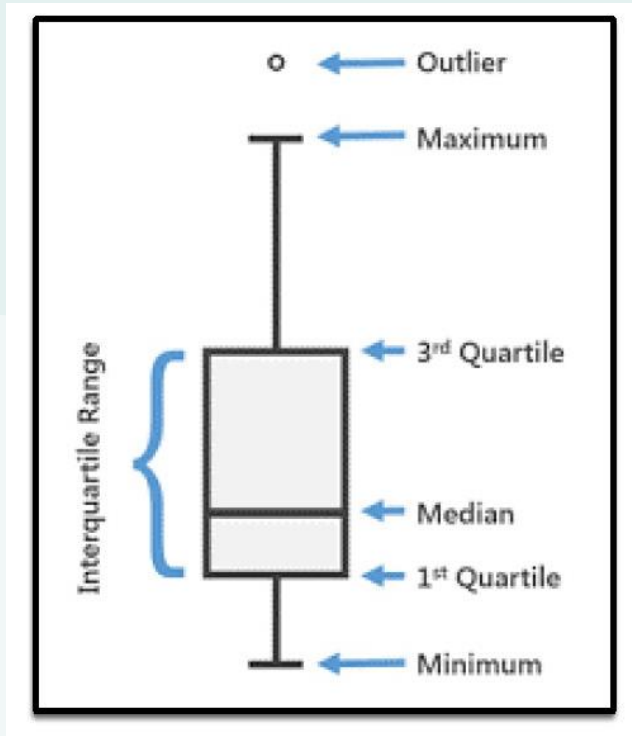


# Quartiles

ควอไทล์เป็นการแบ่งข้อมูลออกเป็น 4 ส่วนเท่าๆ กันส่วนละ 25% โดยเรียงลำดับข้อมูลจากน้อยไปมาก  
ดังนั้น

- ค่าควอไทล์ 1 (Q1 หมายถึงค่าของข้อมูลที่มีจำนวนข้อมูลที่มีค่าต่ำกว่า Q1 อยู่ 25%
- ค่าควอไทล์ 2 (Q2 หมายถึงค่าของข้อมูลที่มีจำนวนข้อมูลที่มีค่าต่ำกว่า Q2 อยู่ 50% และมีจำนวนข้อมูลที่มีค่ามากกว่า Q2 อยู่ 50%
- ค่าควอไทล์ 3 (Q3 หมายถึงค่าของข้อมูลที่มีจำนวนข้อมูลที่มีค่าต่ำกว่า Q3 อยู่ 75% และมีจำนวนข้อมูลที่มีค่ามากกว่า Q3 อยู่ 25%

# Quartiles & Box plot

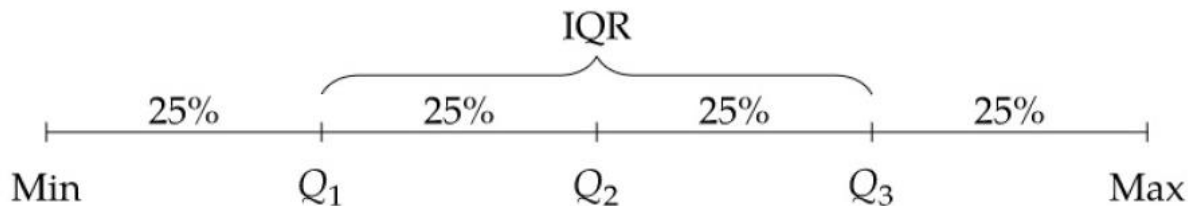




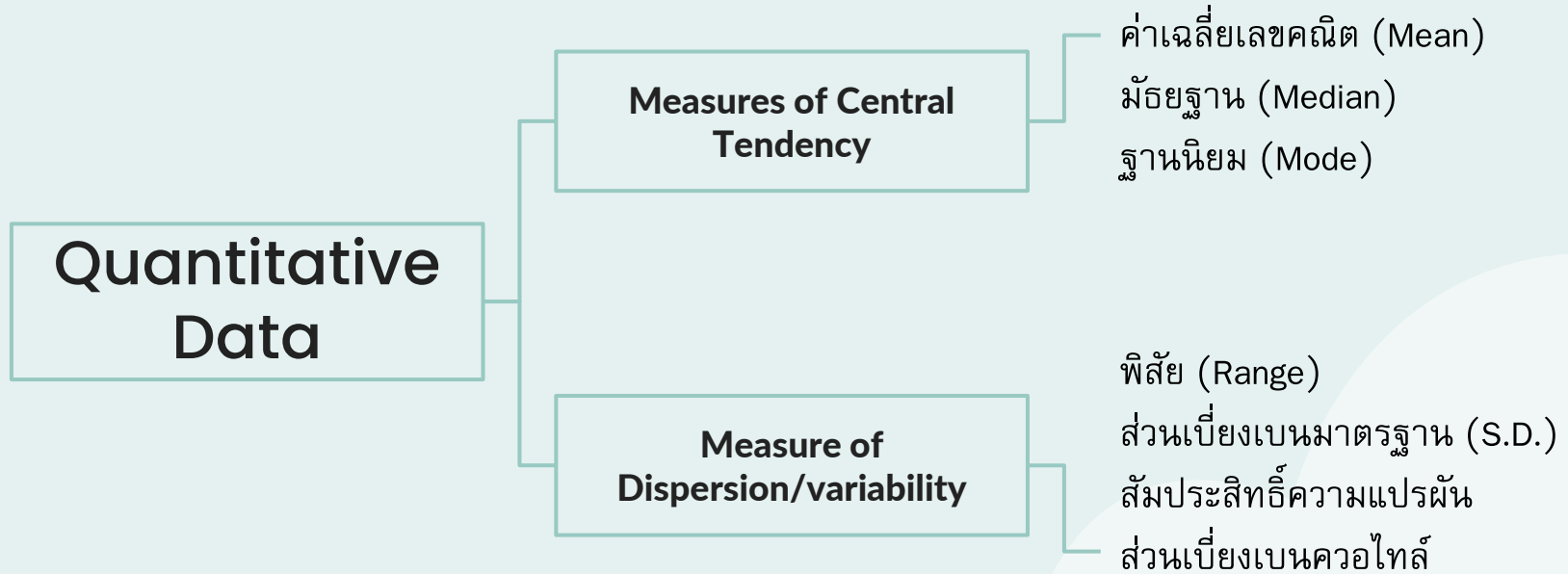
# Interquartile Range (IQR)

**Definition - Sample interquartile range.** The *sample interquartile range* of a variable  $X$  is the difference between the third and the first sample quartiles.

$$\text{IQR} = Q_3 - Q_1$$



# Summary of Quantitative Data

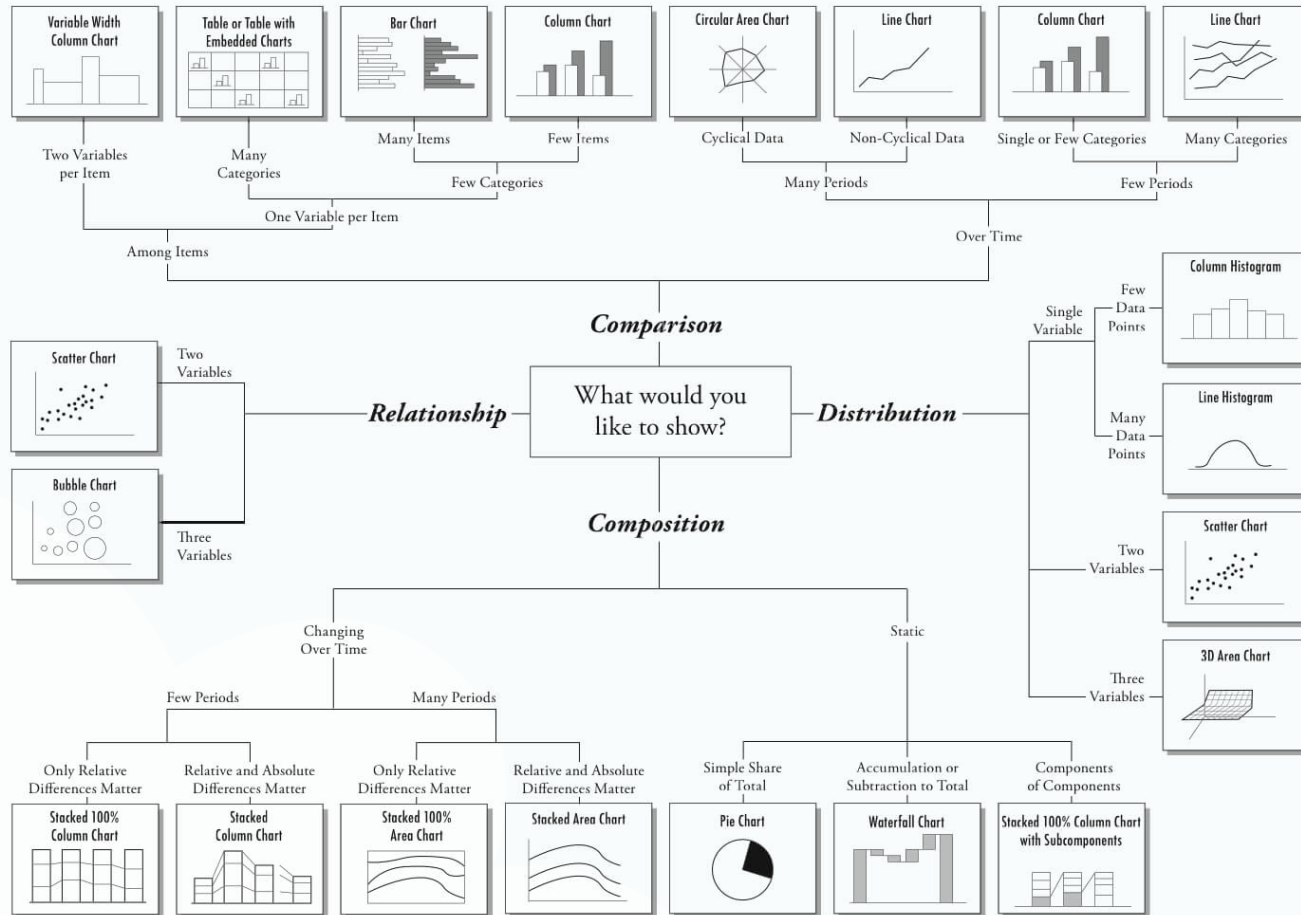




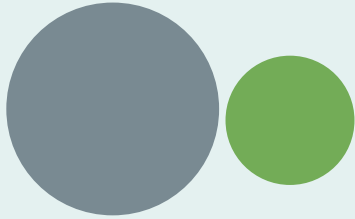
# Graphical Approach

# Chart Suggestions—A Thought-Starter

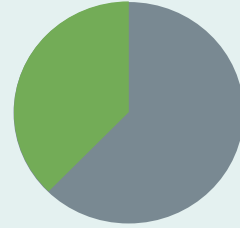
www.ExtremePresentation.com  
© 2009 A. Abela — a.v.abela@gmail.com



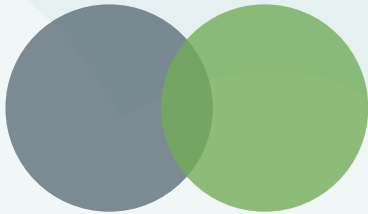
# What do you want to tell from your data?



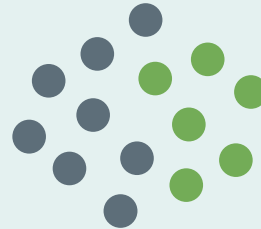
Comparison



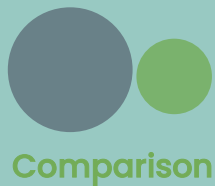
Composition



Relationship



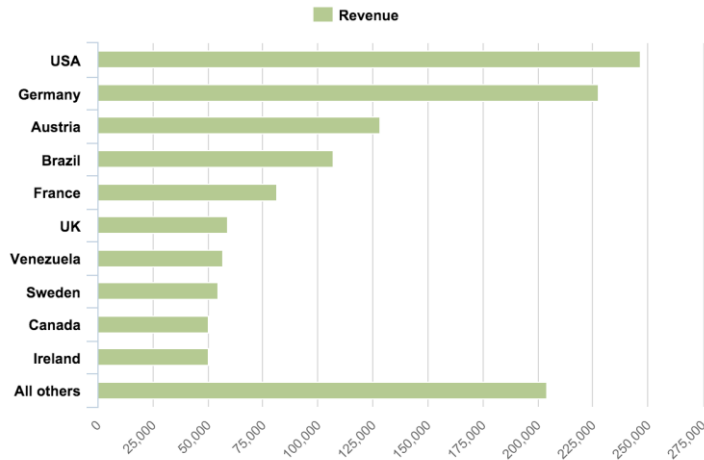
Distribution



Comparison

# 1. Bar Chart / Column Chart

- เปรียบเทียบข้อมูลในแต่ละหมวดหมู่



<https://eazybi.com/blog/data-visualization-and-chart-types#:~:text=Bar%20charts%20are%20good%20for,never%20for%20comparisons%20or%20distributions.>



Comparison



Relationship



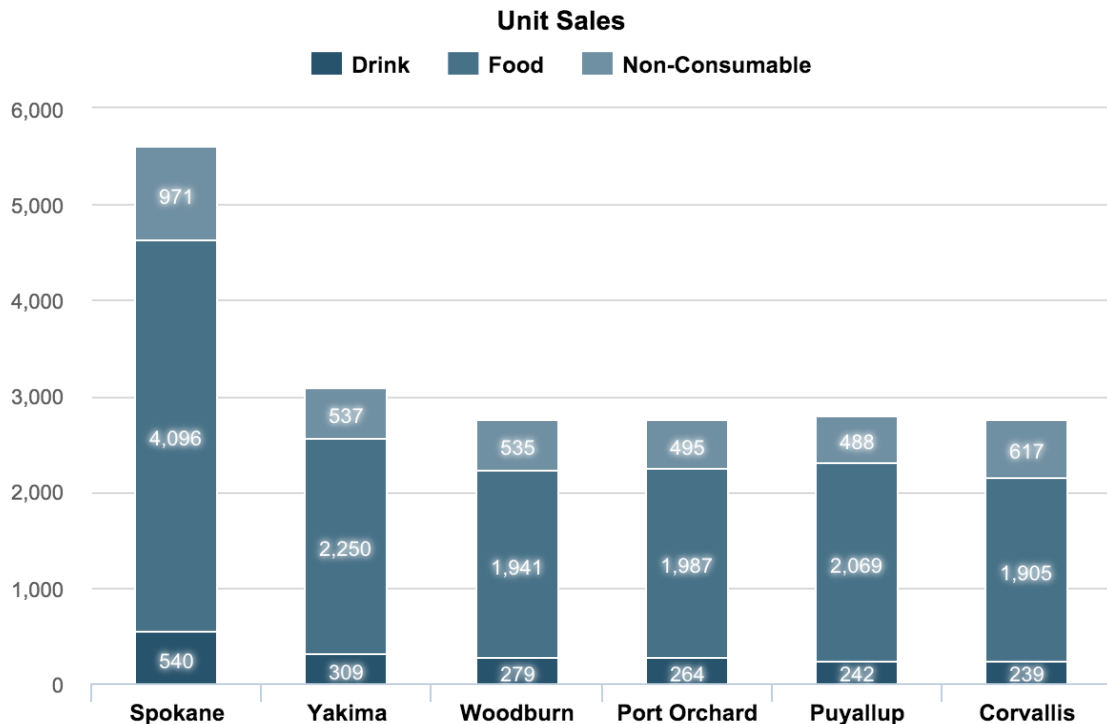
Composition



Distribution

## 2. Stacked Column Chart

- เปรียบเทียบข้อมูลในแต่ละหมวดหมู่ รวมถึงองค์ประกอบ



<https://eazybi.com/blog/data-visualization-and-chart-types#:~:text=Bar%20charts%20are%20good%20for,never%20for%20comparisons%20or%20distributions.>



Comparison



Relationship



Composition



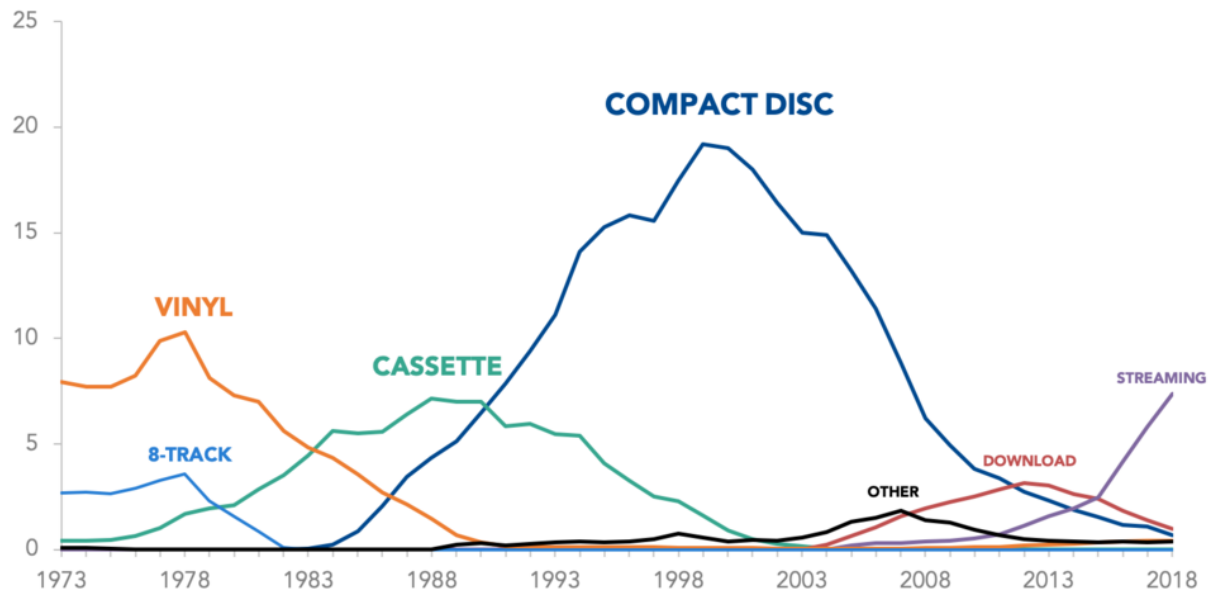
Distribution

## 4. Line Chart

- เปรียบเทียบแนวโน้มตามช่วงเวลา

### US music sales by format (inflation-adjusted)

IN BILLIONS (USD)



SOURCE: Recording Industry Association of America

<http://www.storytellingwithdata.com/blog/2020/4/9/what-is-an-area-graph>





Comparison



Relationship



Composition



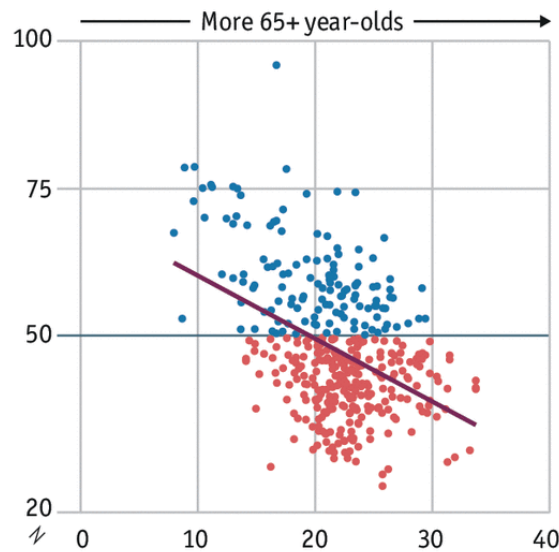
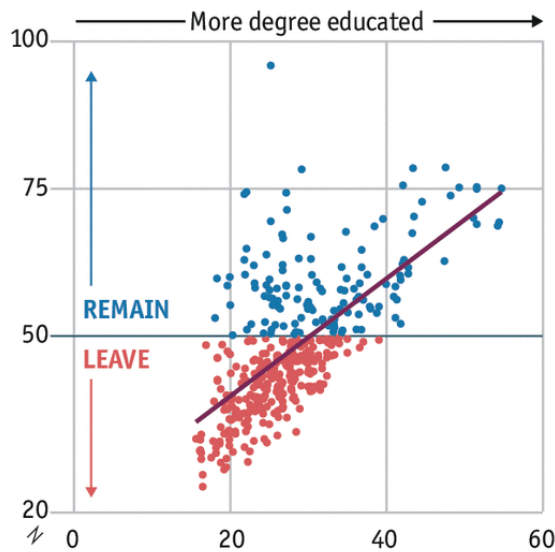
Distribution

# 1. Scatter Chart

- ความสัมพันธ์ระหว่างตัวแปร

## EU referendum results by demographics

Remain vote % by counting area



Sources: BBC; 2011 Census, UK Data Service

Economist.com



Comparison



Relationship



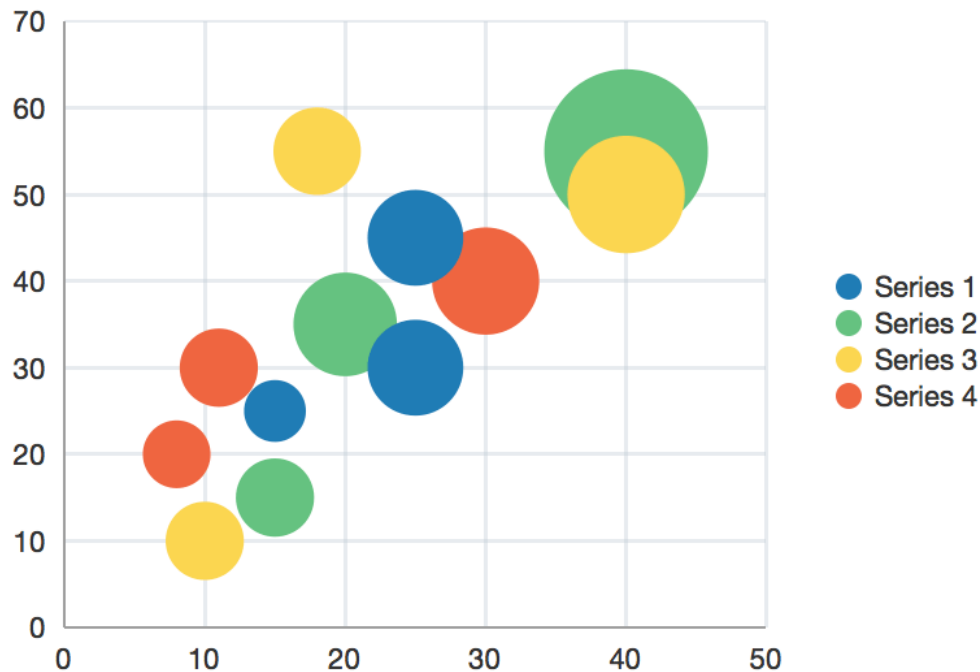
Composition



Distribution

## 2. Bubble Chart

- ความสัมพันธ์ระหว่างตัวแปร >2 ตัวแปร





Comparison



Relationship



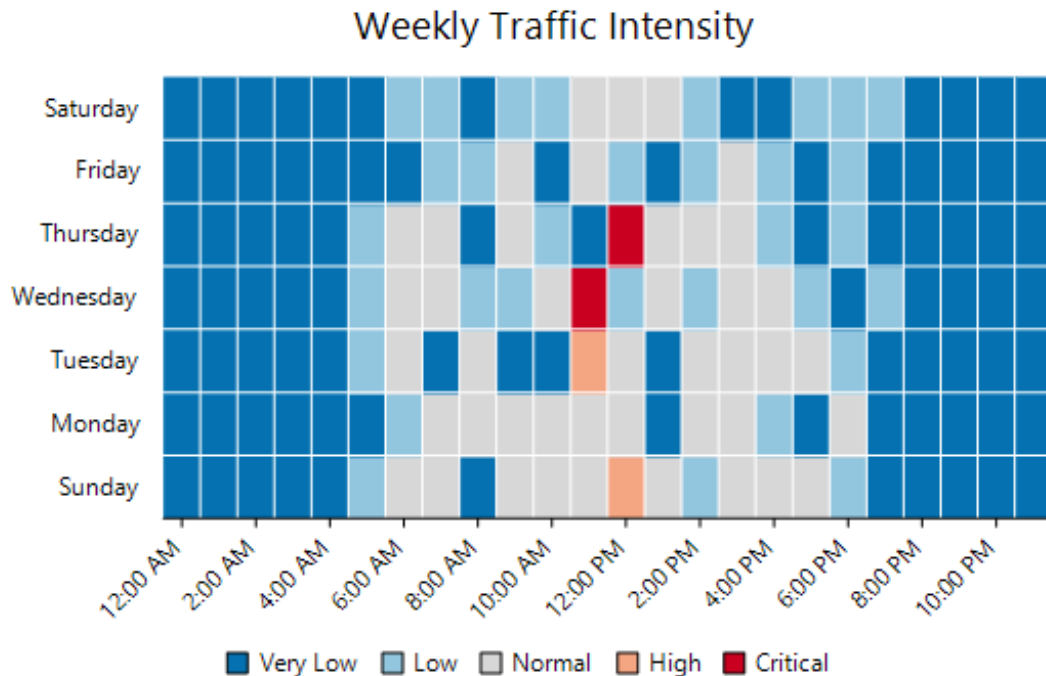
Composition



Distribution

### 3. Heatmap Chart

- หารูปแบบความสัมพันธ์ของข้อมูล การกระจุกหรือการกระจายตัว





Comparison



Relationship



Composition



Distribution

## 4. Crosstab Chart

- ตารางเพื่อแสดงความสัมพันธ์ระหว่างตัวแปร

Column % Column Comparisons	Under 25	25 to 39	40 or more	Male		Female	
				Under \$45,000	\$45,000 or more	Under \$45,000	\$45,000 or more
Coca-Cola	53% c	55% c	35%	45%	46%	55%	38%
Diet Coke	6%	13%	13%	9%	7%	11%	15%
Coke Zero	16%	19%	20%	21%	15%	18%	23%
Pepsi	6%	7%	10%	15%	10%	0%	8%
Diet Pepsi	1%	0%	5%	3%	0%	3%	5%
Pepsi Max	17% b	6%	15%	6%	20%	13%	10%
Dislike all cola	1%	0%	1%	0%	1%	0%	1%
Don't care	0%	0%	2%	0%	1%	0%	1%
NET	100% -	100% -	100% -	100% -	100% -	100% -	100% -
Column n	83	69	175	33	114	38	105
Column Names	A	B	C	A	B	A	B

<https://www.displayr.com/what-is-a-crosstab/>



Comparison



Relationship



Composition

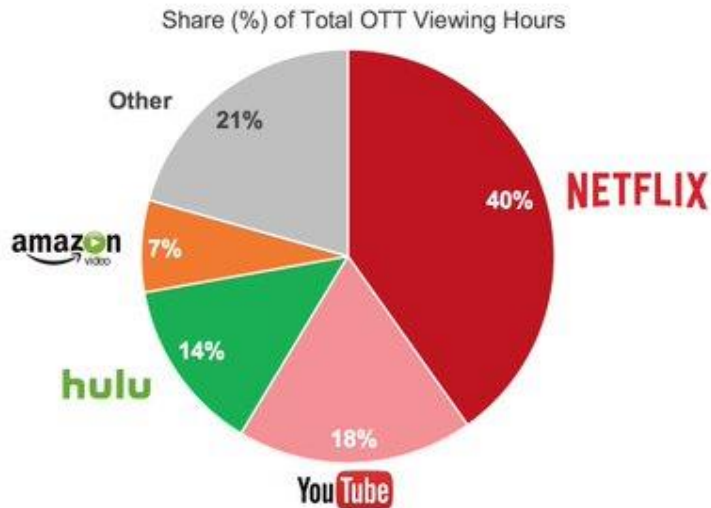


Distribution

# 1. Pie Chart

- แสดงองค์ประกอบของทั้งหมด
- รวมเป็น 1 หรือ 100%

The four major OTT streaming services account for nearly 80% of viewing time for OTT households



## 2. Tree Map Chart

- แสดงองค์ประกอบและลำดับชั้นของข้อมูล



Comparison



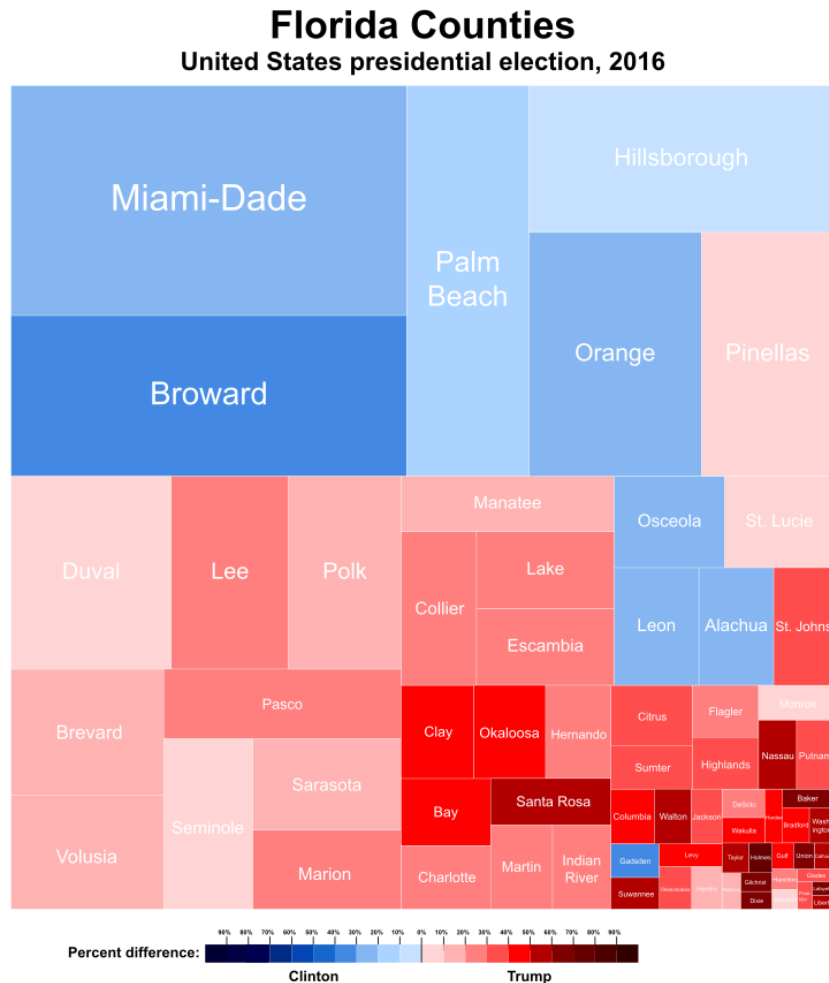
Relationship



Composition



Distribution

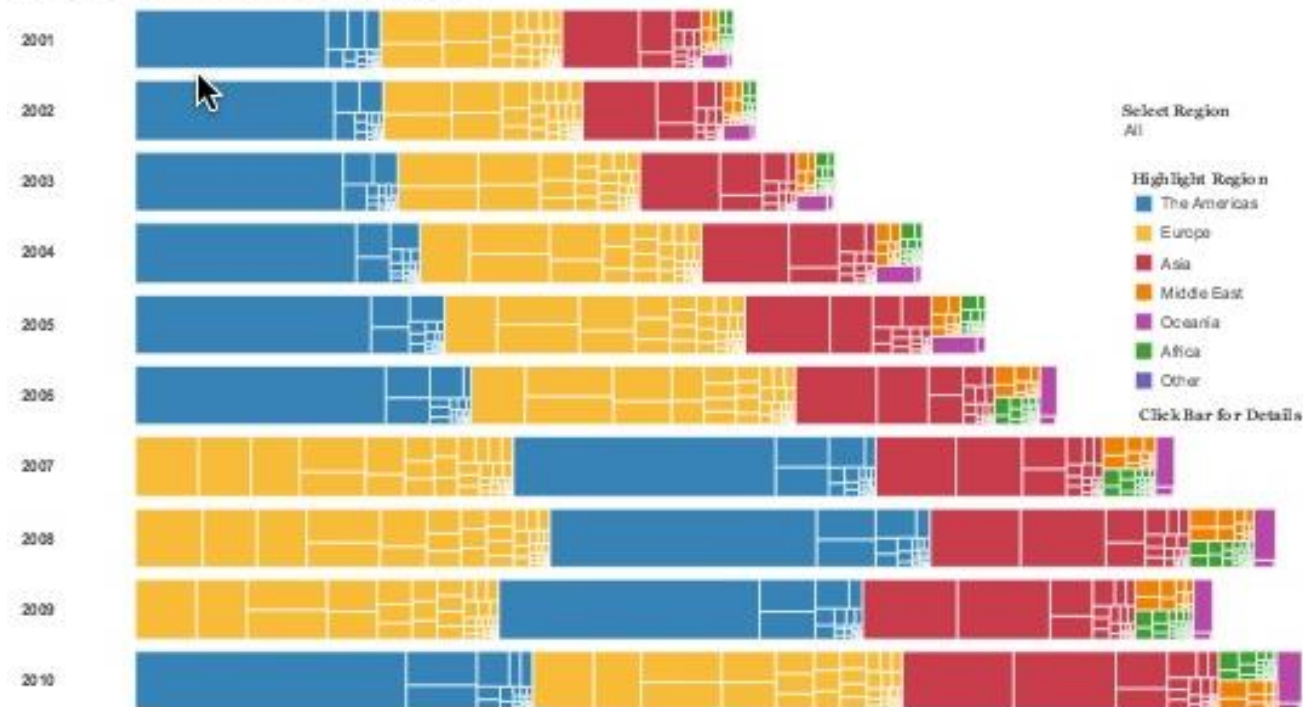


[https://en.wikipedia.org/wiki/Treemapping#/media/File:United\\_States\\_presidential\\_election\\_in\\_Florida,\\_2016.svg](https://en.wikipedia.org/wiki/Treemapping#/media/File:United_States_presidential_election_in_Florida,_2016.svg)

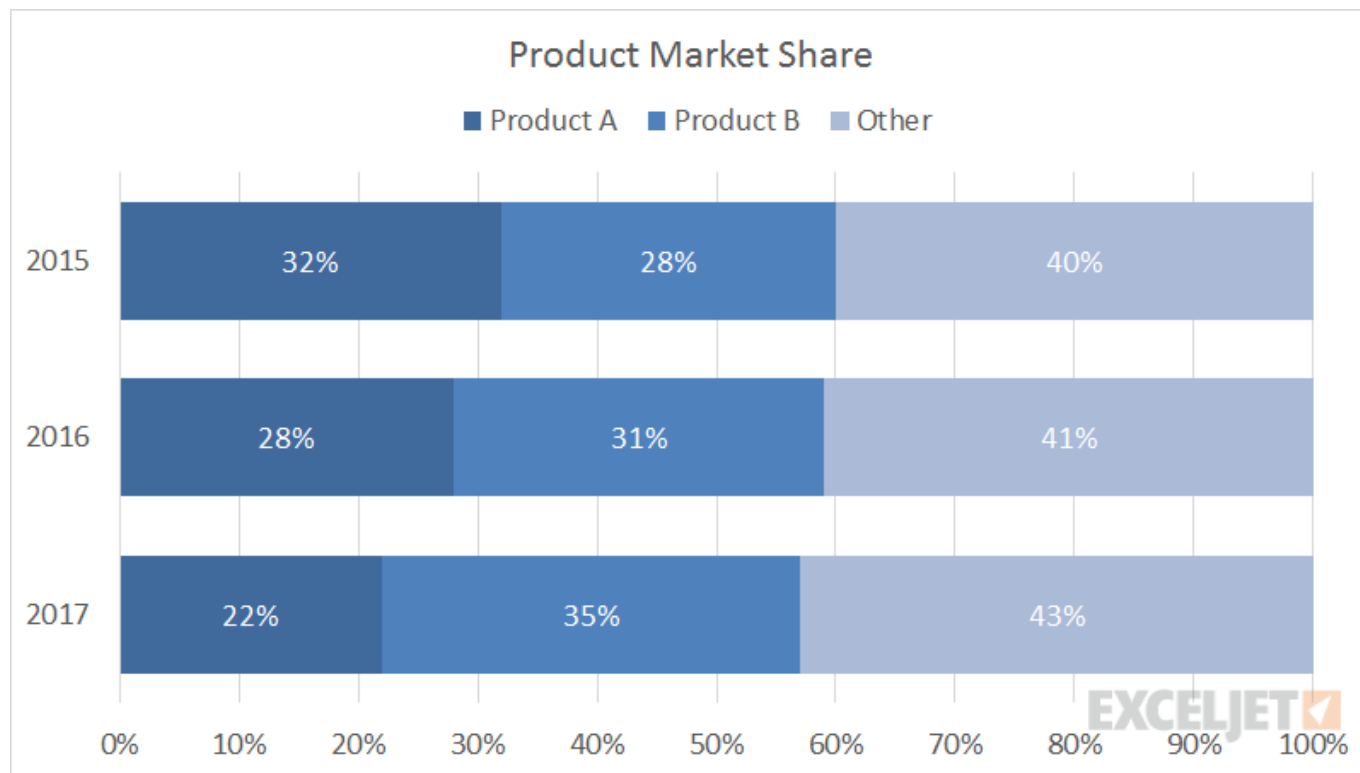
# Mixed Chart

- Treemap + Bar

## World GDP Through Time



### 3. 100% Stacked Bar Chart







Comparison



Relationship



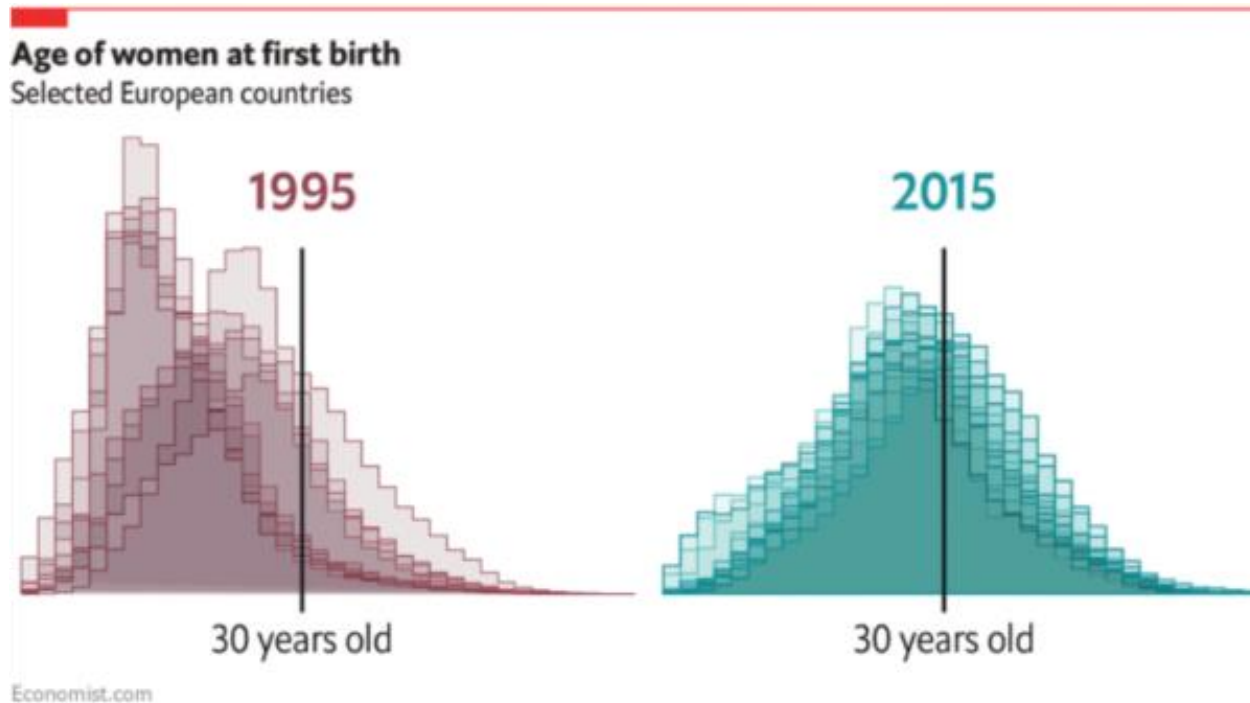
Composition



Distribution

# 1. Histogram Chart

- แสดงการกระจายตัวของข้อมูล



# 1. Histogram Chart



Comparison



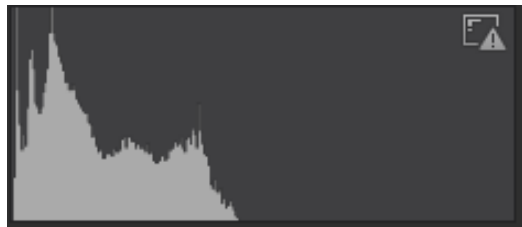
Relationship



Composition



Distribution



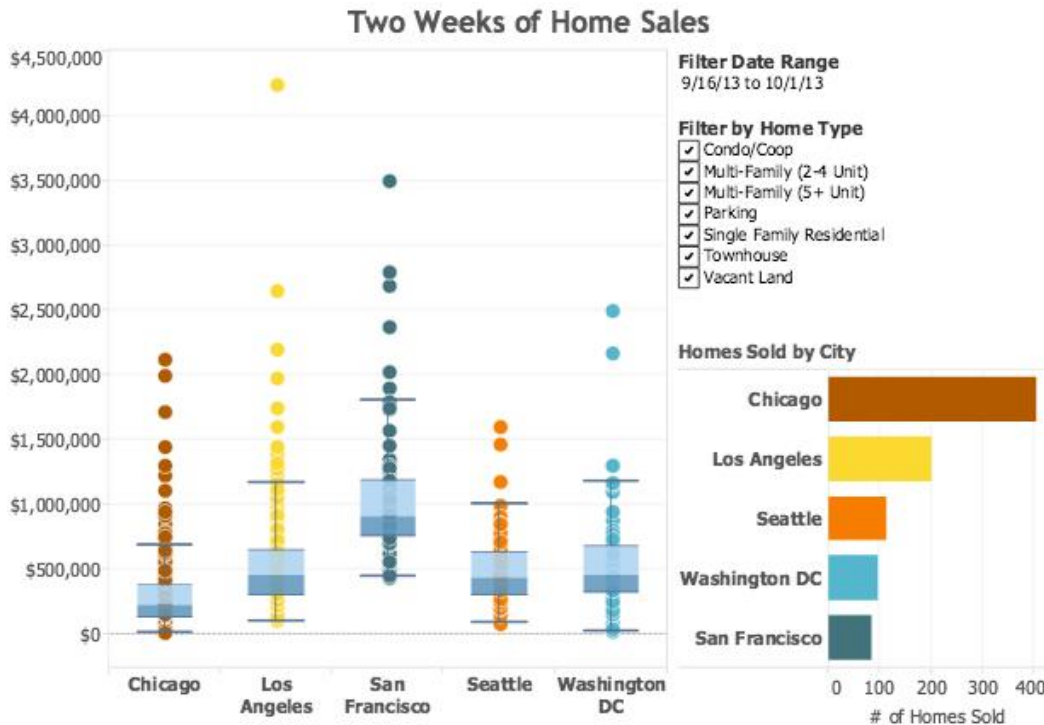
A



B

## 2. Box Chart

- แสดงการกระจายตัวของข้อมูล



<https://imgconvert.csdnimg.cn/aHR0cHM6Ly91cGxvYWQtaW1hZ2VzLmppYW5zaHUuaW8vdXBsb2FkX2ltYWdlcy8xMDEzNjA1NC03MWEyNzRiNGYwNDM1NzA3LmpwZz9pbWFnZU1vZ3lyL2F1dG8tb3JpZW50L3N0cmIwGltYWdlVmlldzlvMi93LzUyNC9mb3JtYXQvd2VicA?x-oss-process=image/format,png>

# Excel Workshop



Statistics

# EXCEL for Statistics

- เปิดไฟล์ Excel “workshopexcel.xlsx”
- คำนวณค่าสถิติของคอลัมน์ Income:

Mean :

Standard Deviation (SD) :

Q1 :

Median :

Q3 :

IQR :

Minimum :

Maximum :

Count :



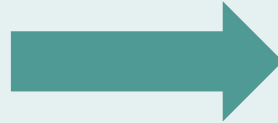


# Qualitative

ข้อมูลเชิงคุณภาพ

# Qualitative

ลำดับ	จังหวัด	รายได้
1	กรุงเทพมหานคร	12000
2	กรุงเทพมหานคร	24000
3	กรุงเทพมหานคร	28000
4	นนทบุรี	42000
5	นนทบุรี	32000
6	สมุทรปราการ	19000
7	สมุทรปราการ	17500
8	สมุทรปราการ	20000
9	นนทบุรี	35000
10	กรุงเทพมหานคร	50000



แต่ละจังหวัดมีกี่คน

จังหวัดไหนมีประชากร  
มากที่สุด / น้อยที่สุด

# Analysis relationship of data variables



# Relationship

- หมวดหมี vs ตัวเลข
  - Correlation Analysis
- หมวดหมี vs หมวดหมี
  - Crosstab / Contingency Table
- ตัวเลข vs ตัวเลข
  - Correlation Analysis



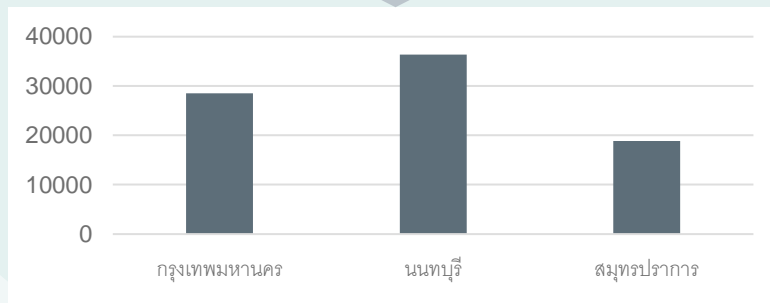
# Group and Aggregate

- Group ตัวแปรประเภทหมวดหมู่
- Aggregate ตัวแปรประเภทตัวเลข โดยใช้ ค่ากลาง เช่น ค่าเฉลี่ย มัธยฐาน หรือ การกระจายตัว เช่น S.D.

ลำดับ	จังหวัด	รายได้
1	กรุงเทพมหานคร	12000
2	กรุงเทพมหานคร	24000
3	กรุงเทพมหานคร	28000
4	นนทบุรี	42000
5	นนทบุรี	32000
6	สมุทรปราการ	19000
7	สมุทรปราการ	17500
8	สมุทรปราการ	20000
9	นนทบุรี	35000
10	กรุงเทพมหานคร	50000



จังหวัด	รายได้เฉลี่ย	S.D.
กรุงเทพมหานคร	28500	15864.01
นนทบุรี	36333.33	5131.60
สมุทรปราการ	18833.33	1258.30
Grand Total	27950	11889.42



# Excel Workshop



Pivot Table

FileHomeInsertDrawPage LayoutFormulasDataReviewViewDeveloperHelp

PivotTable

Recommended PivotTables

Table

Tables

Illustrations

Get Add-ins

My Add-ins

Add-ins

Recommended Charts

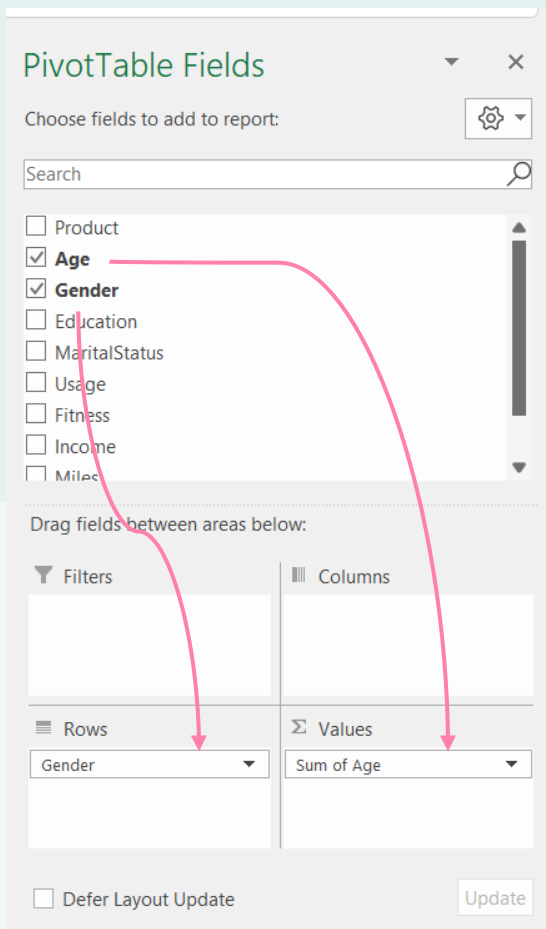
Charts

Charts

L16

	A	B	C	D	E	F	G	H	I
1	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
2	TM195	18	Male	14	Single	3	4	29562	112
3	TM195	19	Male	15	Single	2	3	31836	75
4	TM195	19	Female	14	Partnered	4	3	30699	66
5	TM195	19	Male	12	Single	3	3	32973	85
6	TM195	20	Male	13	Partnered	4	2	35247	47
7	TM195	20	Female	14	Partnered	3	3	32973	66
8	TM195	21	Female	14	Partnered	3	3	35247	75
9	TM195	21	Male	13	Single	3	3	32973	85
10	TM195	21	Male	15	Single	5	4	35247	141
11	TM195	21	Female	15	Partnered	2	3	37521	85
12	TM195	22	Male	14	Single	3	3	36384	85
13	TM195	22	Female	14	Partnered	3	2	35247	66
14	TM195	22	Female	16	Single	4	3	36384	75
15	TM195	22	Female	14	Single	3	3	35247	75
16	TM195	23	Male	16	Partnered	3	1	38658	47
17	TM195	23	Male	16	Partnered	3	3	40932	75
18	TM195	23	Female	14	Single	2	3	34110	103
19	TM195	23	Male	16	Partnered	4	3	39795	94
20	TM195	23	Female	16	Single	4	3	38658	113

1. ทำ sheet1 ให้เป็น Table
2. คลิก Insert -> Pivot Table
3. เลือก New Worksheet หรือ  
Existed Worksheet คลิก OK

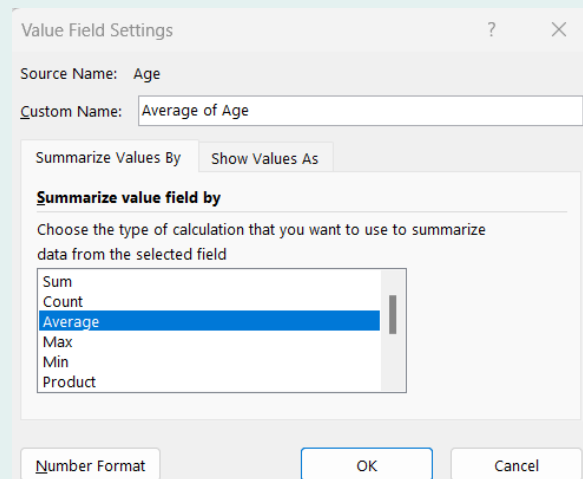
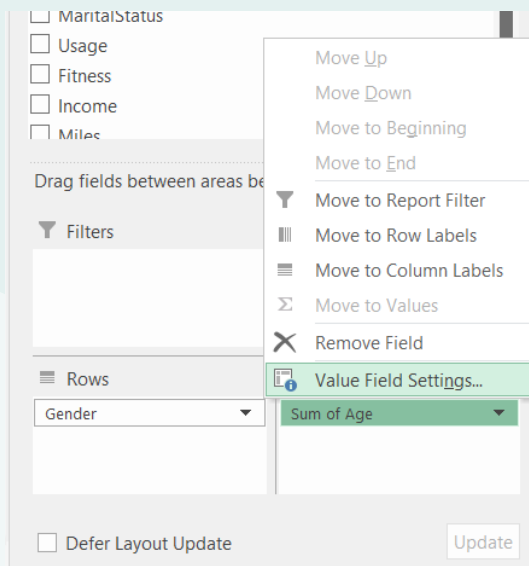


4. ลากตัวแปรหมวดหมู่ไปใส่ Rows

5. ลากตัวแปรตัวเลขไปใส่ Values

6. ตรง Values จะแสดงเป็น Sum of .... ถ้าต้องการเปลี่ยน

ให้คลิกขวา Value Field Settings แล้วเปลี่ยนได้



# Excel Workshop



Correlation Analysis

File Home **Insert** Draw Page Layout Formulas Data Review View Developer Help Table Design

PivotTable Recommended PivotTables Table Illustrations Get Add-ins My Add-ins Recommended Charts Maps PivotCharts

B1 Age

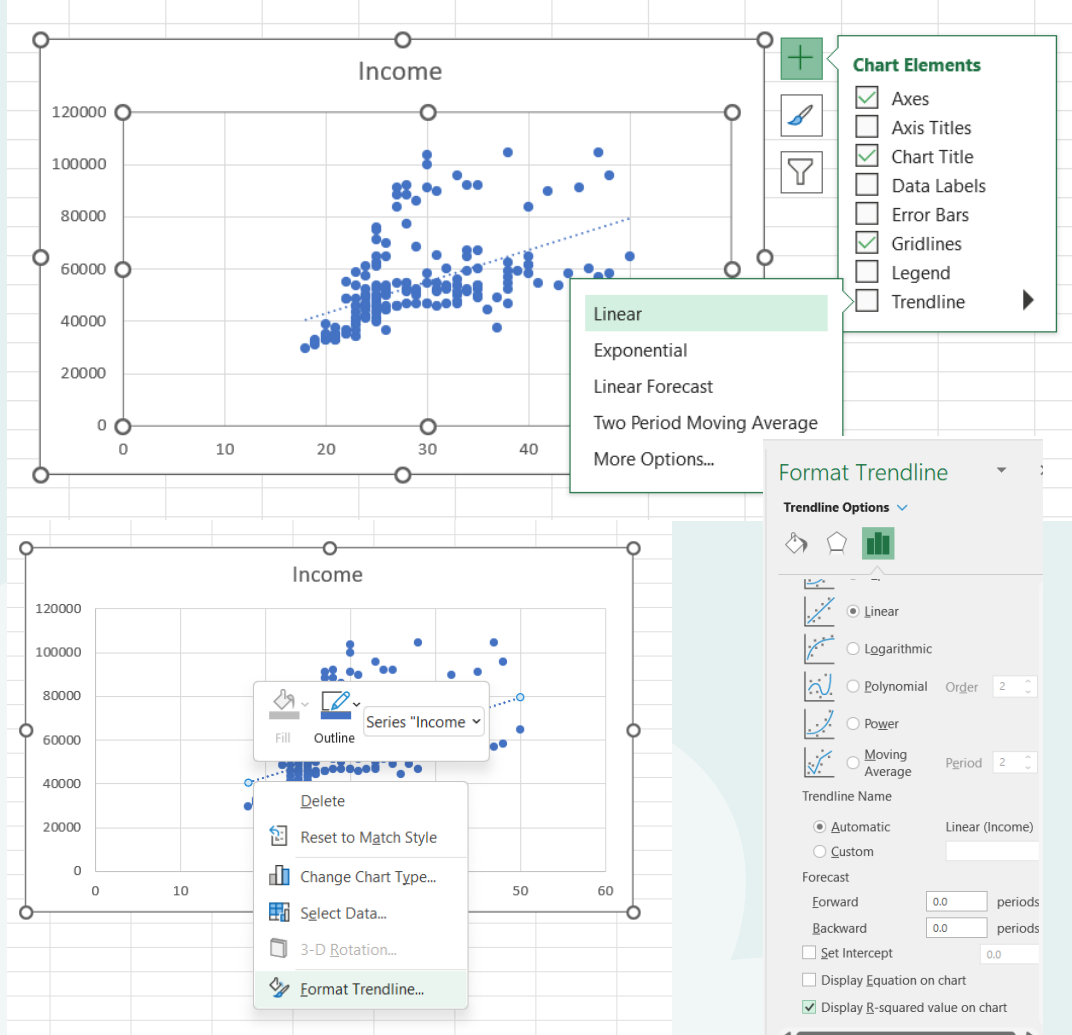
	A	B	C	D	E	F	G	H
	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income
2	TM195	18	Male	14	Single	3	4	295
3	TM195	19	Male	15	Single	2	3	318
4	TM195	19	Female	14	Partnered	4	3	306
5	TM195	19	Male	12	Single	3	3	329
6	TM195	20	Male	13	Partnered	4	2	352
7	TM195	20	Female	14	Partnered	3	3	329
8	TM195	21	Female	14	Partnered	3	3	352
9	TM195	21	Male	13	Single	3	3	329
10	TM195	21	Male	15	Single	5	4	352
11	TM195	21	Female	15	Partnered	2	3	37521
12	TM195	22	Male	14	Single	3	3	36384
13	TM195	22	Female	14	Partnered	3	2	35247
14	TM195	22	Female	16	Single	4	3	36384
15	TM195	22	Female	14	Single	3	3	35247

Scatter

Bubble

More Scatter Charts...

1. เลือกคอลัมน์ที่เป็นข้อมูลเชิงปริมาณ
2. คอลัมน์
2. คลิก Insert -> Insert Scatter Chart



3. เลือกกราฟที่โผล่มา คลิก + เลือก

Trendline -> Linear

4. จะมีเส้นตรงปรากฏขึ้นมา คลิกขวาที่

เส้นเลือก Format Trendline

5. เลือก Display R-squared value on  
chart จะได้ค่า R-squared



**04**

# Inferential Statistics

.....

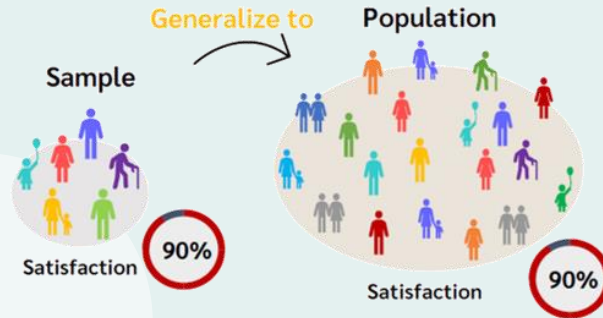


# Why Statistics? Part II – Inferential

It's not always possible to have all the data. Hence, we may try infer something about data we don't have using the data we have.

## Inferential statistics

Drawing conclusions about a population based only on sample data



i.e. 90% satisfaction of a sample of 6 customers  
-> 90% satisfaction of all customers

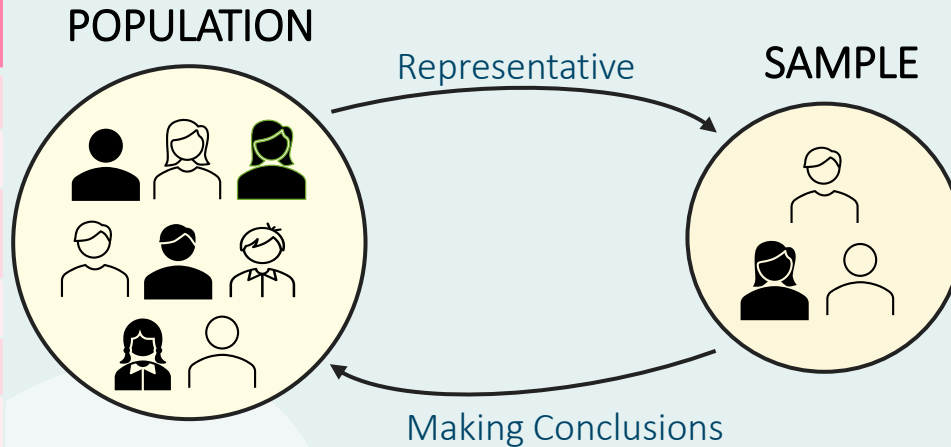
# PROBLEMS WITH POINT ESTIMATION



Different sampling results in different estimation –  
The smaller the sample size, the less reliable it is

# INFERENCE STATISTICS

Description	Population Parameter
Mean	$\mu$
Variance	$\sigma^2$
Standard Deviation	$\sigma$
Size	$N$
Correlation	$\rho$
Proportion	$p$



Description	Sample statistic
Mean	$\bar{x}$
Variance	$s^2$
Standard Deviation	$s$
Size	$n$
Correlation	$r$
Proportion	$\hat{p}$

# Inferential Statistical



```
graph TD; A[Inferential Statistical] --> B[Parameter estimation]; A --> C[Hypothesis testing];
```

## Parameter estimation

Using sample data to estimate the parameters of a distribution

---

How to reliably estimate the population mean or proportion?

e.g., Estimate the population mean weight using the sample mean weight

## Hypothesis testing

How to use a random sample to judge if it is evidence that supports or not the hypothesis

---

Is the population mean or proportion equal to what we believe it is?

e.g., Test the claim that the population mean weight is 120 lbs.

# PARAMETER ESTIMATION

## Point Estimation

PARAMETER	STATISTIC
Population Mean : $\mu$	Sample Mean : $\bar{x}$
Population Variance : $\sigma^2$	Sample Variance : $s^2$
Population Standard Deviation : $\sigma$	Sample Standard Deviation : $s$
Population Correlation : $\rho$	Sample Correlation : $r$
Population Proportion : $p$	Sample Proportion : $\hat{p}$

Quantitative Data  
(numerical)

Categorical Data  
(Yes/No)

# CONFIDENCE INTERVALS

Gives the range of “what the reasonable results could be?”

This is very helpful in the area of Risk Management so the organization can be prepared for the worse (yet probable) case possible.

## **Example:**

- รัฐบาลขายล็อตเตอรี่จะต้องเตรียมเงินไว้ให้ผู้ซื้อขึ้นรางวัลเท่าไร
- นักลงทุนต้องการประเมินว่าการลงทุนครั้งนี้มีความเสี่ยงที่จะเสียเงินไปเท่าไร
- บริษัทประกันคำนวณงบประมาณสำรอง

# SAMPLING FOR PROPORTION

Suppose we take repeated random sample of size n where each observation can be one of the only two possible outcomes

The two outcomes is often regarded “success” and “failure”. For example, in a poll the “success” could be the “yes” vote and a “failure” a “no” vote.

We have the estimate for the proportion,

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

$\hat{p}$  – Estimated proportion  
 $X$  – Number of “success”  
 $n$  – Sample size



# SAMPLING FOR PROPORTION

The sample proportion  $\hat{p}$  can be approximated by a normal distribution and the confidence interval is given by

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$$

$\hat{p}$  = Estimated proportion  
 $\hat{q} = 1 - \hat{p}$

Confidence	$z^*$
0.90	1.64
0.95	1.96
0.99	2.58

← Sometimes replaced with 2

- The term  $2 \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$  is called the margin of error which indicates how far of our estimate can be from the truth (within 95% chance). The bigger the sample size, the lower this margin will be
- This formula works well when  $n \times \hat{p} > 10$  and  $n \times \hat{q} > 10$

# Example: Laptop Case

A campus of 1,000 students. A survey sample 50 students and ask whether they have a personal laptop. Suppose 15 people answered with “yes”.

What would be the confident interval of the proportion of the 1,000 students that have a computer?

**Point Estimate:**

$$\hat{p} = \frac{15}{50}$$

**Assumption check:**

- ✓ The population size (1,000) > 10 times sample size (50)
- ✓ The terms  $n \times \hat{p} = 50 \times 0.3 = 15 > 10$  and  $n \times \hat{q} = 50 \times 0.7 > 10$

# Example: Laptop Case

95% Confidence Interval:

$$\begin{aligned}\hat{p} \pm 2 \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}} &= 0.3 \pm 2 \times \sqrt{\frac{0.3 \times 0.7}{50}} \\ &= 0.3 \pm 0.13\end{aligned}$$

**Conclusion:**

With 95% probability, we may estimate  
 $1,000 \times (0.3 \pm 0.13) = 300 \pm 130$  students to have a laptop.

# SAMPLING FOR MEAN

The sample mean  $\mu$  follows t-distribution with degree of freedom  $n$  and the 95% confidence interval is given by

$$\bar{x} \pm t_{n-1}^* \frac{s}{\sqrt{n}}$$

$\bar{x}$  = Sample average  
 $s$  = Sample standard deviation

Critical Values of  $t^*$  for 95% Confidence

n	$t_{n-1}^*$
20	2.09
100	1.98
500	1.96

Sample python command for  $n = 500$

```
from scipy.stats import t  
t.interval(0.95, 500-1, loc=0, scale=1)[1]
```

# Example: Pricing a Buffet

Suppose a restaurant owner wants to start a buffet service. He wishes to estimate the average price per customer. He does a survey with 10,000 customers and it turns out the sample's average cost is \$500 and the Sample's standard deviation is \$100.

What would be the 95% confidence interval for the average cost for the whole population?

For large sample size, we can use the normal distribution to estimate as

$$\mu = \bar{x} \pm 2 \cdot \frac{s}{\sqrt{n}} = 500 \pm 2 \times \frac{100}{\sqrt{10,000}} = 500 \pm 200$$


This would be a decent estimate but not exactly, to get the right kind of estimation, we need to use t-distribution.

# Example: Pricing a Buffet

Using t-distribution,

$$\begin{aligned}\mu &= \bar{x} \pm t_{n-1}^* \cdot \frac{s}{\sqrt{n}} = 500 \pm 1.96 \times \frac{100}{\sqrt{10,000}} \\ &= 500 \pm 196\end{aligned}$$

```
from scipy.stats import t  
t.interval(0.95, 1000-1, loc=0, scale=1)[1]  
  
1.9623414611334487
```



The difference between t-distribution's and the z-distribution's estimate are very significant. So, in many cases, using the number 2 instead of 1.96 would give us a good enough answer.

# Thank you

