

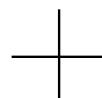
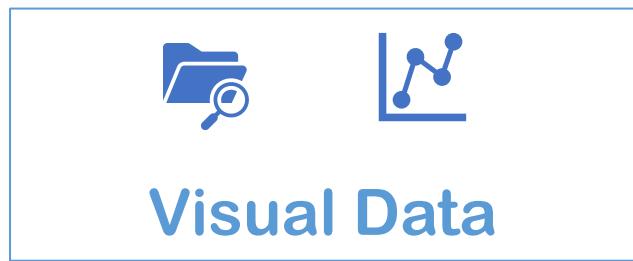
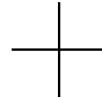


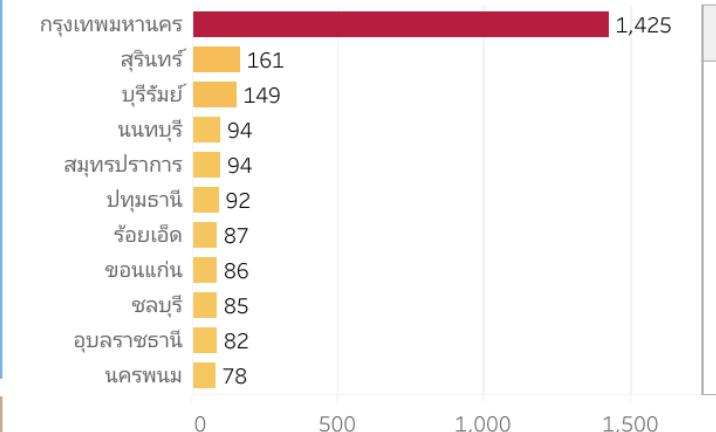
GBDi

# Introduction to Statistics

Khwansiri Sirimangkhala, Ph.D

# PROCESS





[Link](#)

## สถิติอุบัติเหตุทางถนน ในช่วงเทศกาลสงกรานต์



	2560	2561
ยอดรวมการเกิด อุบัติเหตุรุนแรง	3,690	3,724
ผู้บาดเจ็บ	3,808	3,897
ผู้เสียชีวิต	390	418
ดัชนีความรุนแรง*	10.57	11.2

\*severity index: SI  
คำนวณจากจำนวนเสียชีวิต<sup>ต่อ 100 ครั้งของอุบัติเหตุ</sup>

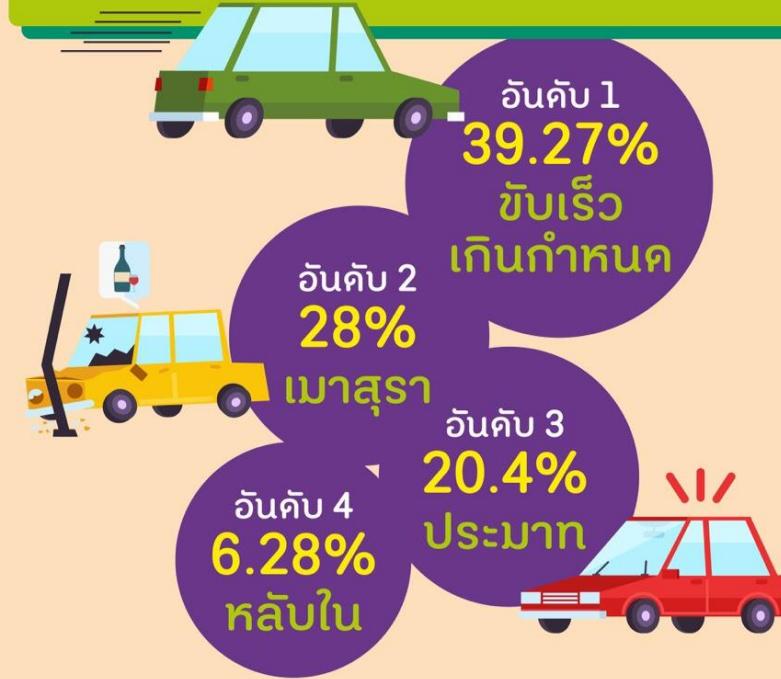


สสส.

### อุบัติเหตุทางถนน สามารถป้องกันได้ด้วย

- ผู้ขับขี่มีวินัยจราจร
- ไม่ดื่มทั้งก่อนและระหว่างขับ
- นอนหลับเพียงพอ
- เลิกพฤติกรรมเสี่ยง  
ภาค จี้ เปียด ตัดหน้า
- ไม่ประมาท
- สวมหมวกนิรภัย
- คาดเข็มขัดนิรภัยทุกครั้ง
- หมั่นตรวจสอบสภาพรถ
- และวางแผน  
การเดินทางทุกครั้ง

## สาเหตุการเกิดอุบัติเหตุและเสียชีวิต ช่วงเทศกาลสงกรานต์ 2561



ศูนย์วิชาการเพื่อความปลอดภัยทางถนน (ศวปก.)

ชี้ว่า ค่าเฉลี่ยการเสียชีวิตจากอุบัติเหตุทางถนน  
ในช่วง 10 ปีที่ผ่านมา เฉพาะวันที่ 13 เมษายน  
มีความสูงเสี่ยงถึง 754 คพ

ปัญหานำจากต้นน้ำ คือ  
งานรื้นเริงที่จัดโดยมีเครื่องดื่มแอลกอฮอล์



สสส.



# ความเสี่ยงบนท้องถนน มาจากไหน?

คน  
**48%**

นอนน้อย ดื่มเยอะ เมาค้าง  
ไม่ชำนาญเส้นทาง ไม่มีวินัยจราจร  
เร่งรีบ



กุญแจสิ่งแวดล้อม  
**38%**

จุดเสี่ยง จุดติดไฟแดง  
ไม่มีป้ายเตือน ไม่มีไฟส่องสว่าง  
ถนนขรุขระ

ยานพาหนะ  
**15%**

ชำรุด ขาดการซ่อมบำรุง  
ขาดอุปกรณ์ป้องกันความปลอดภัย ปรับแต่งรถผิด  
มาตรฐาน บรรทุกเกินขนาด



**รู้หรือไม่!!!** อัตราการตายเกิดกับรถจักรยานยนต์สูงสุด  
รองลงมาคือ รถระยะ ระยะรถเก๋ง

มีหลายปัจจัยที่เราทุกคนสามารถควบคุมได้ ก็ต่อเมื่อวินัย ใจดี และอุปกรณ์  
ความปลอดภัยให้อยู่ในสภาพที่พร้อมขับขี่ หากอยู่ในสภาพแวดล้อมที่ควบคุมไม่ได้  
จำเป็นต้องเพิ่มความระมัดระวังยิ่งกว่าเดิม

สสส.สนับสนุนการอิเคราะห์หาสาเหตุของการเกิดอุบัติเหตุ  
ช่วยให้ผู้ใช้รถ ใช้ถนน เพิ่มความระมัดระวังได้ตรงจุด!



# ปิดเทอมนี้



## เด็กและเยาวชนอย่างกำมะไร?

(ข้อมูลจากการสำรวจความคิดเห็นบนหน้าเฟซบุ๊ก สสส. ตั้งแต่วันที่ 18 - 22 กุมภาพันธ์ 2562)

ปิดเทอมใหญ่ เป็นช่วงเวลาว่างของเด็กเยาวชน ที่ผู้ปกครองต้องมองหากิจกรรม  
หรือสถานที่เสริมสร้างทักษะความรู้ให้แก่บุตรหลาน นอกเหนือจากห้องเรียน

เข้าค่ายวิชา  
เสริมสร้างประสบการณ์

**45%**



ทำงานพิเศษ

**24%**



ท่องเที่ยว

**22%**



เล่นกับเพื่อน

**4%**



เรียนพิเศษ

**3%**



เข้าค่ายวิชาการ

**2%**

สสส. และภาคีเครือข่าย สนับสนุนให้เกิดกิจกรรม ‘ปิดเทอมสร้างสรรค์’  
เพื่อพัฒนาทักษะชีวิตและสังคม ปลดปล่อยศักยภาพ ตลอดจนการเรียนรู้ไปกิจวัตร  
และเติบโตกิจกรรมในทุกมิติ

สามารถติดตามกิจกรรมและสร้างความต้องรับปิดเทอม ได้ที่ พพพ.ปิดเทอมสร้างสรรค์.com



# Pantip Analytics

**Pantip Analytics**  
มาดูประเด็นสนทนาในห้องท่องเที่ยว (Blueplanet) ตั้งแต่ ม.ค. 2564

ช่วงระยะเวลาที่สนใจ : 14 พ.ค. 2022 - 12 มิ.ย. 2022

ข้อมูลล่าสุด : 12 มิ.ย. 2022

**ประเภทกระทู้ที่ถูกพูดถึง**

หัวข้อ	จำนวน
ตอบคำถาม	956
สนทนา	143
รีวิว	129
ข่าว	8

**ประเด็นที่ถูกพูดถึง**

ประเด็น	จำนวน	% Δ
รถโดยสาร	54	28.6% ↑
เที่ยวต่างประเทศ	41	5.1% ↑
ท่องเที่ยว	39	-17.0% ↓
แผนการเดินทางและท่องเที่ยว	37	-2.6% ↓
รถโดยสารประจำทาง	35	34.6% ↑
แผนที่เดินทาง	31	10.7% ↑

% Δ : เปอร์เซ็นต์ความแตกต่างของความสำคัญเทียบกับช่วงเวลาก่อนหน้า

**จังหวัดยอดนิยม**

**จำนวนหัวข้อกระทู้สนทนา**

**รายละเอียดกระทู้สนทนา**

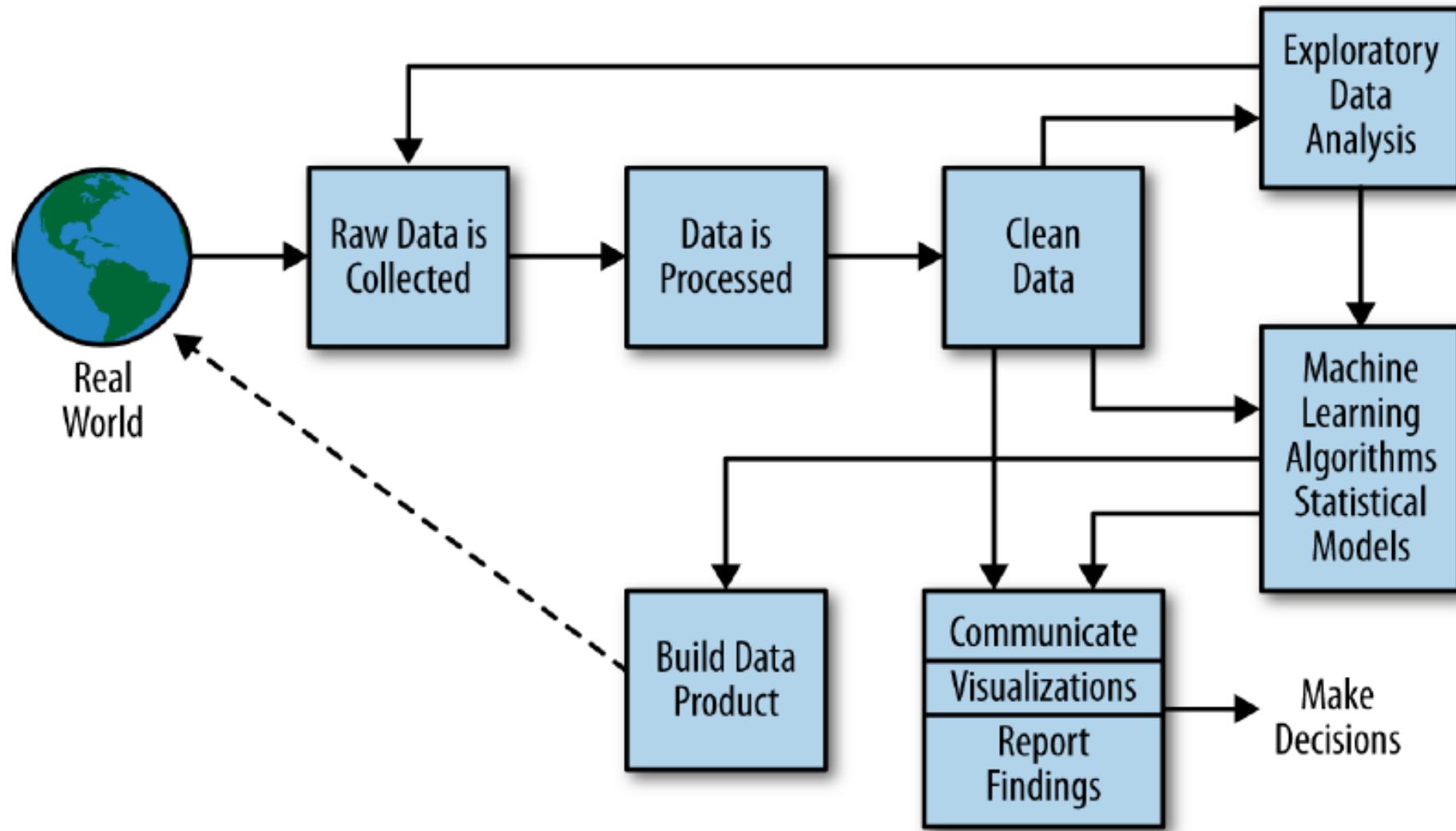
วันที่	กระทู้สนทนา	ประเภท	จังหวัด	ความเห็น	ยอดนิวต์	ยอดชม
12 มิ.ย. 2022	ขออนุญาต ก่อนแกะ-กลบ อีกครั้งที่ใบอนุฯ	ตอบคำถาม	ชลบุรี	0	0	23
11 มิ.ย. 2022	สถานที่น้ำตกสวยมาก	ตอบคำถาม	-	3	0	86
11 มิ.ย. 2022	อพากานการเดินทางจาก บริษัท หมู่บ้าน ไป น้ำตกตุ่นไน่สองกัน	ตอบคำถาม	-	2	0	126
11 มิ.ย. 2022	ขอถึงคุณเมืองใหม่ ท่องเที่ยว ชนเผ่าที่หายตัวไป	ตอบคำถาม	-	2	0	274
11 มิ.ย. 2022	มีอะไรดีๆ ให้ลอง ท่องเที่ยว ชุมชนที่หายตัวไป	รีวิว	-	10	0	116
11 มิ.ย. 2022	พอดีจะเดินทางไปกรุงเทพมหานครให้ล่อ (วางแผนเดินทาง)	ตอบคำถาม	นนทบุรี	2	0	164
11 มิ.ย. 2022	Power bank	ตอบคำถาม	-	1	0	130
11 มิ.ย. 2022	แนะนำที่เที่ยว บ้านโนนป่าไม่ จ.นนทบุรีไปเมืองไทยตอนนี้	ตอบคำถาม	-	2	0	144

1 - 1000 / 1596 < >

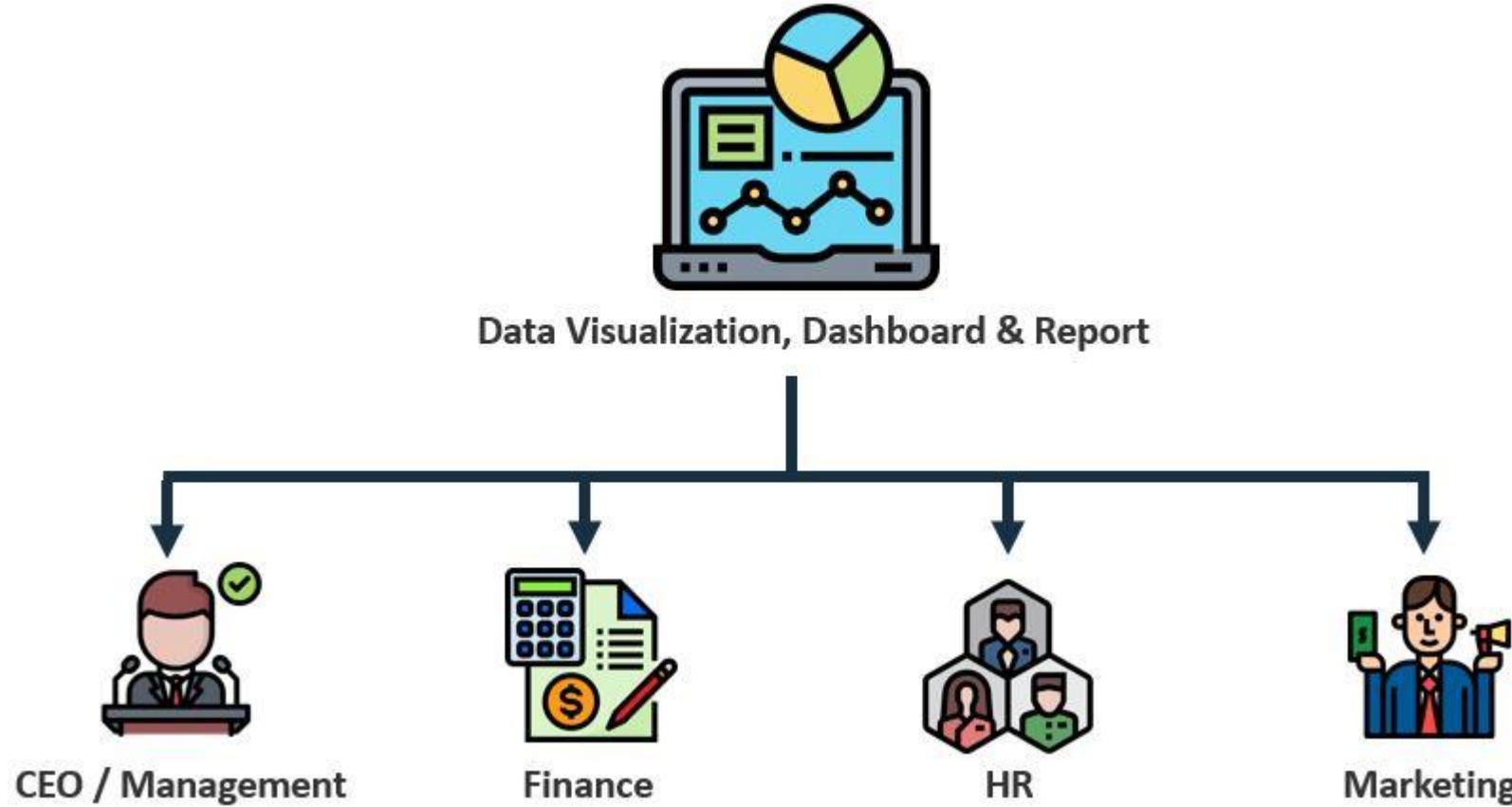
Google Data Studio



[Link](#)



# USEFUL



- Source: <https://techsauce.co/tech-and-biz/what-is-dashboard>

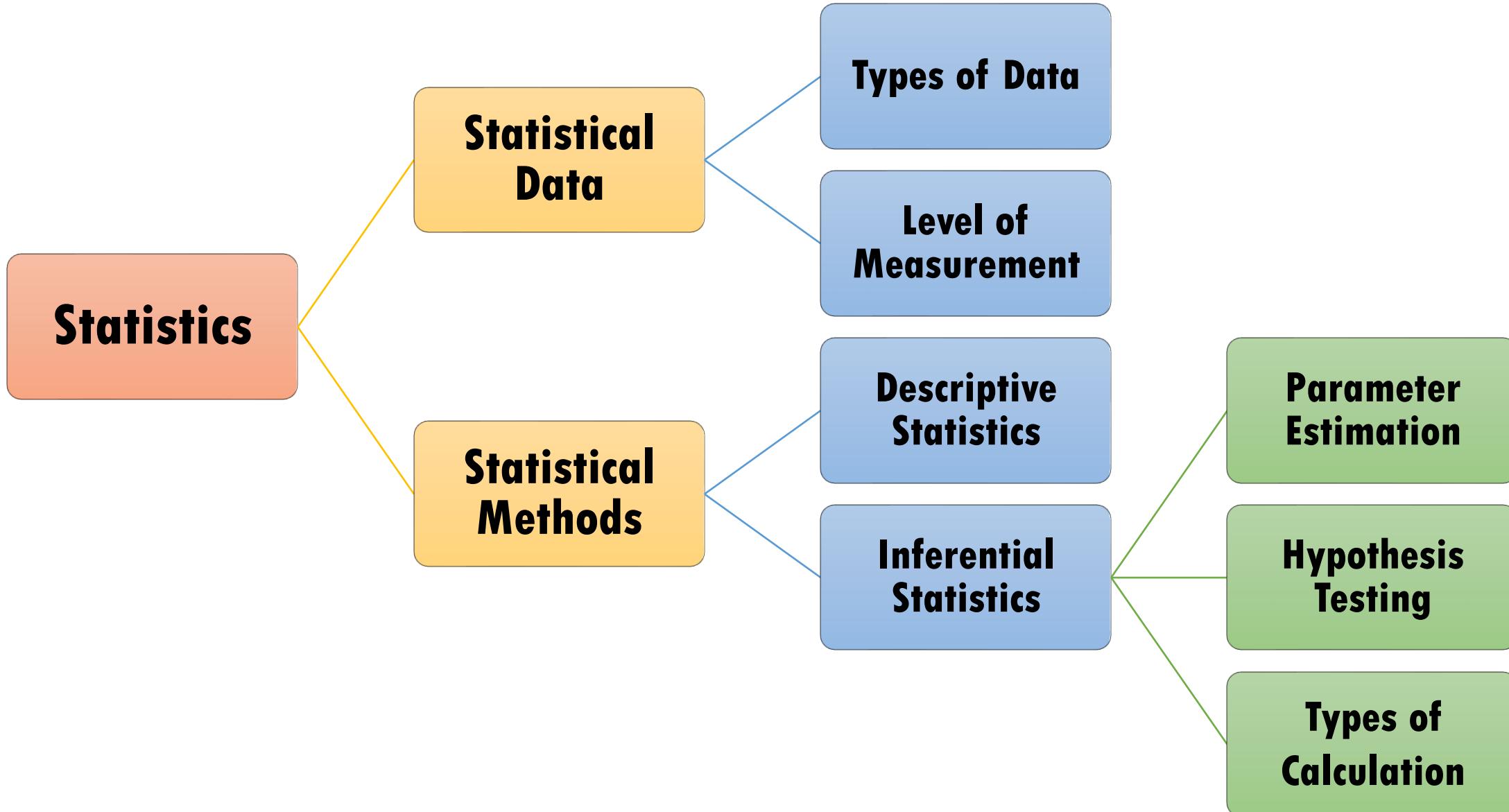
# STATISTICS

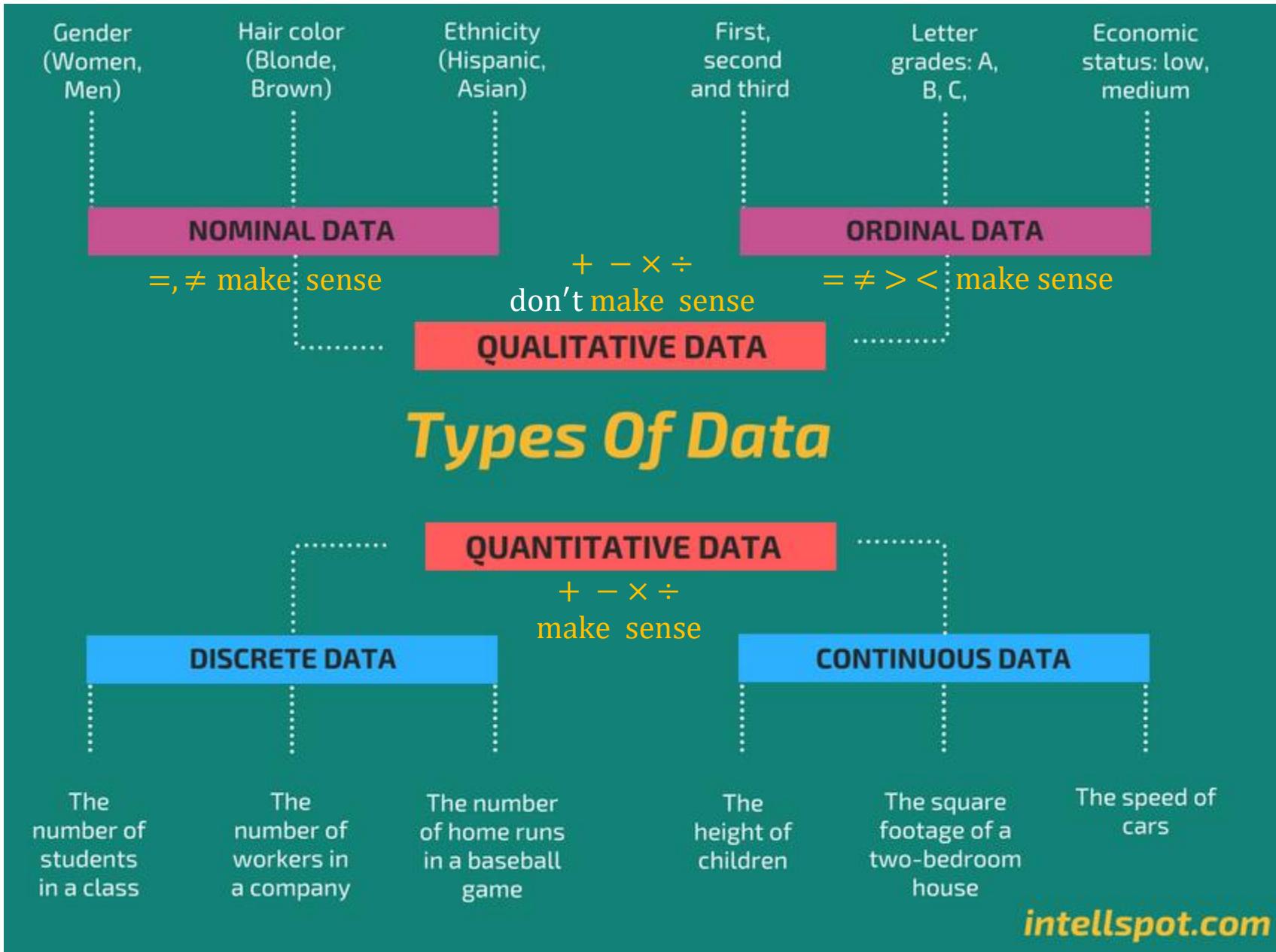
**STATISTICS** is the science of collecting, organizing, and interpreting numerical facts, which we call **data**.

## OBJECTIVE

- Understand basic **statistical methods** to **summarize data**
- Correctly apply **statistical tools** and make **valid conclusion**

# OVERVIEW





# Example 1: Qualitative or Quantitative?

A: Daily confirmed cases of coronavirus

B: Brand names of smartphone in a consumer survey

## Example 2: Discrete or Continuous?

A: Number of customers

B: Number of COVID-19 patients

C: Years of work experience

# Data

## Quantitative Data

### Tabular Methods

Frequency Distribution

Relative Frequency Distribution

Percent Frequency Distribution

Cumulative Frequency  
Distribution

Cumulative Relative Frequency  
Distribution

Cumulative Percent Frequency  
Distribution

Crosstabulation

### Graphical Methods

Dot Plot

Histogram

Ogive

Stem-and-Leaf Display

Scatter Diagram

## Categorical Data

### Tabular Methods

Frequency Distribution

Relative Frequency Distribution

Percent Frequency Distribution

Crosstabulation

### Graphical Methods

Bar Chart

Pie Chart

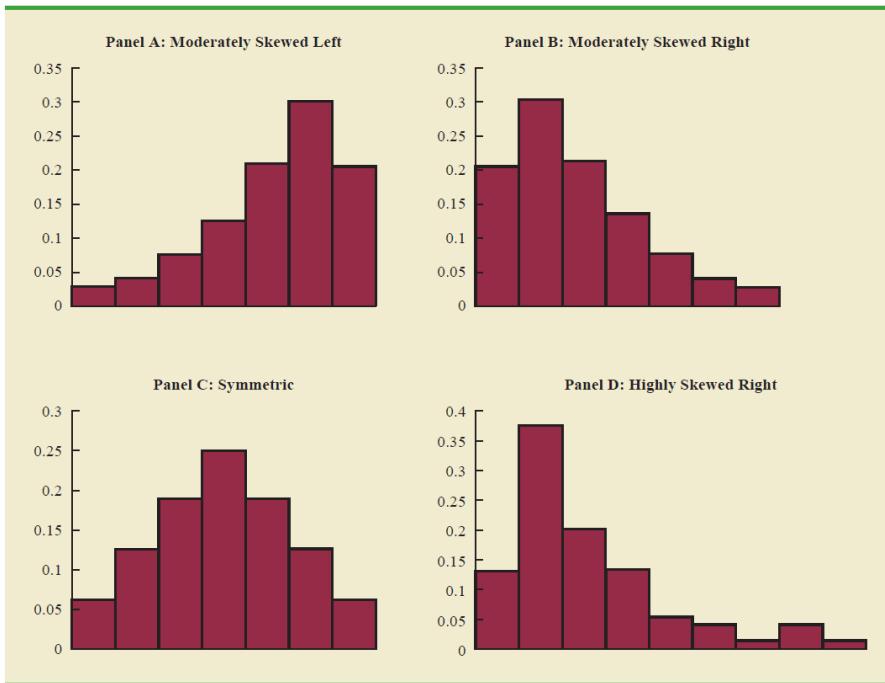
# TABULAR METHODS

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
.	.	.
.	.	.
.	.	.

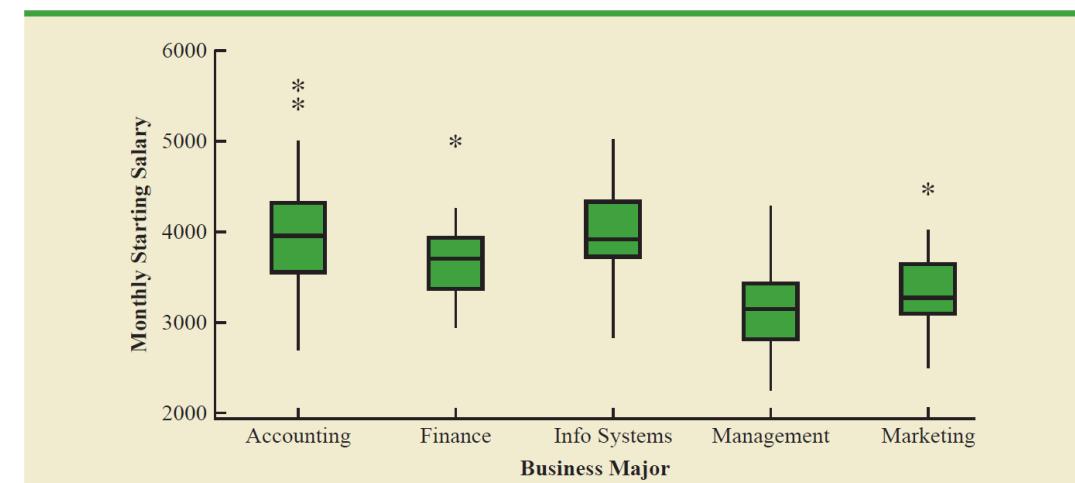
Quality Rating	Meal Price				Total
	\$10–19	\$20–29	\$30–39	\$40–49	
Good	42	40	2	0	84
Very Good	34	64	46	6	150
Excellent	2	14	28	22	66
Total	78	118	76	28	300

# GRAPHICAL METHODS

## Histogram

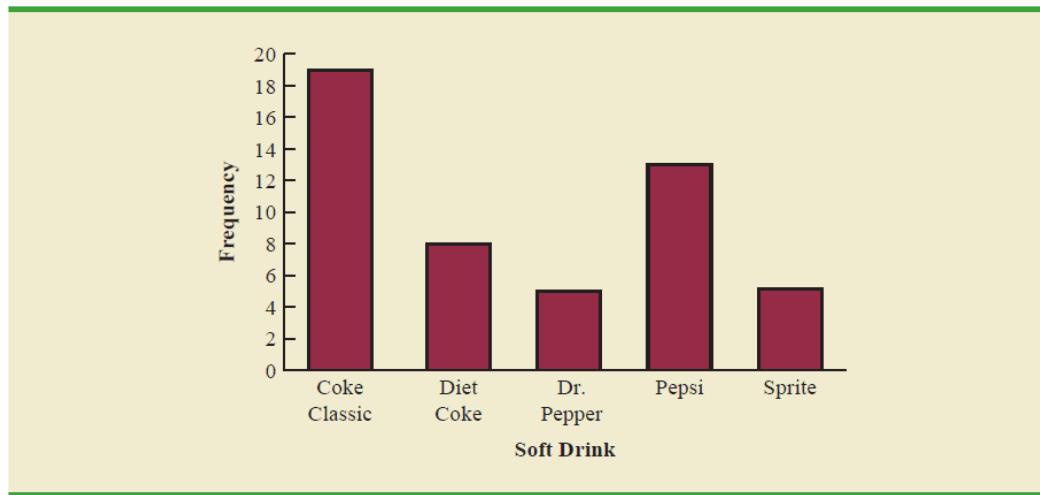


## Box Plot

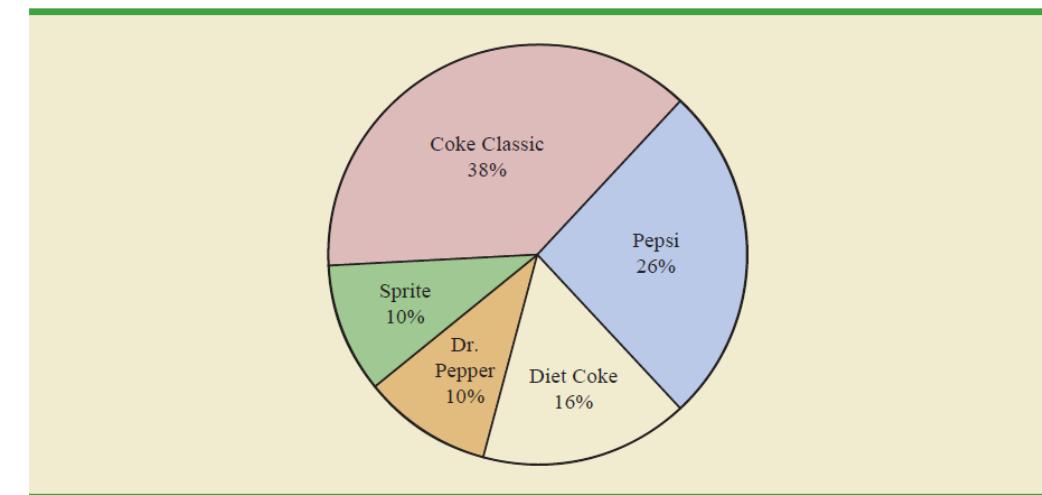


# GRAPHICAL METHODS

Bar Chart

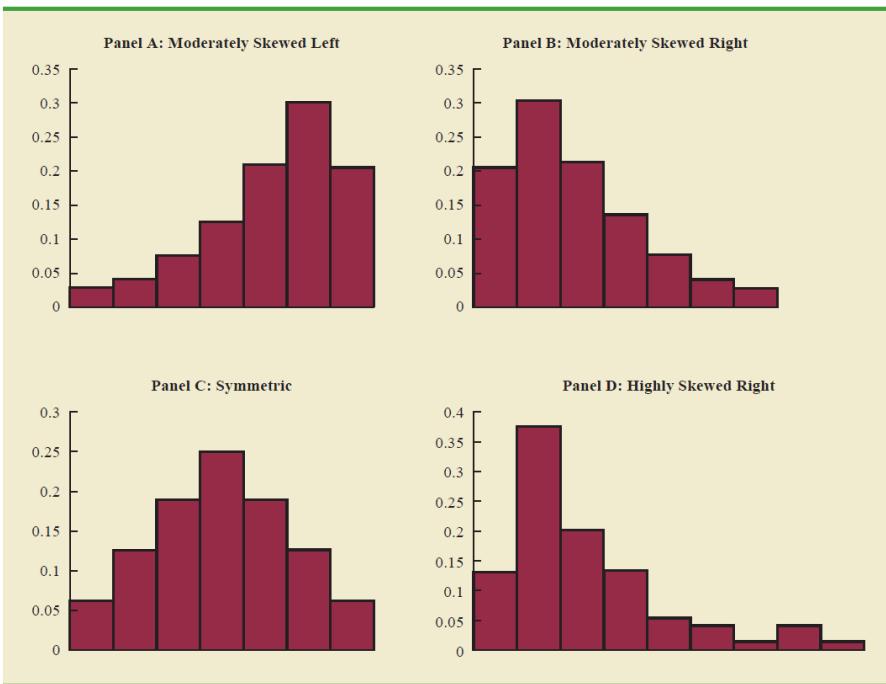


Pie Chart

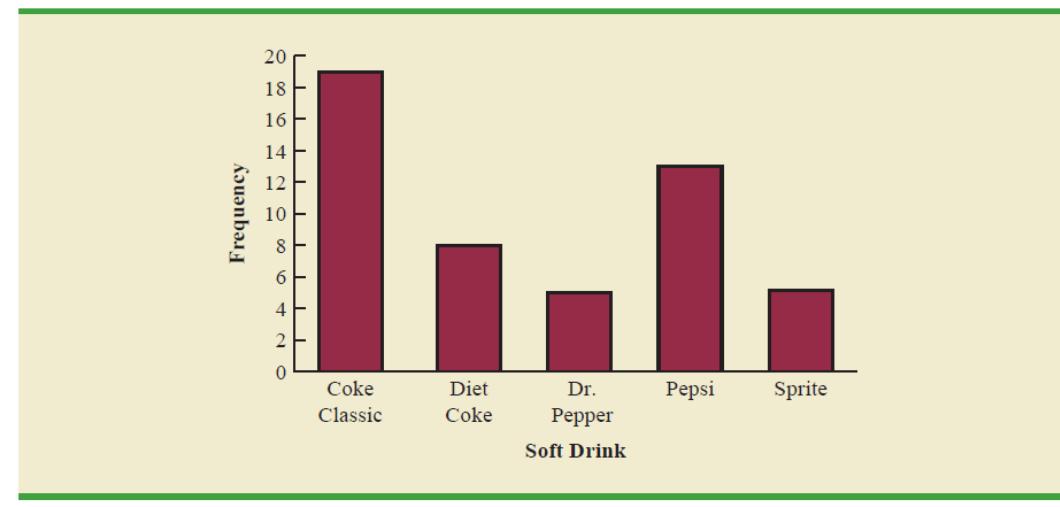


# GRAPHICAL METHODS

Histogram



Bar Chart



# LEVELS OF MEASUREMENT

## 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

## Ratio

- Zero means "absence of"



Height, Age, Weekly Food Spending

# MEANINGFUL ZERO

ข้อมูลที่เป็น Ratio คือ ข้อมูลที่มีค่าเท่ากับศูนย์จะมีความหมายเท่ากับศูนย์จริงหรือไม่มีปริมาณจริง

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

แต่ข้อมูลประเภทอื่นๆ เช่น Nominal Ordinal หรือ Interval ข้อมูลที่มีค่าเท่ากับศูนย์จะมิได้หมายความว่าสิ่งนั้นจะเป็นศูนย์หรือไม่มีค่าจริง

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

# LEVELS OF MEASUREMENT

	Nominal	Ordinal	Interval	Ratio
Rank	✗	✓	✓	✓
Measure	✗	✗	✓	✓
True Zero	✗	✗	✗	✓

# DATA TYPES & LEVELS OF MEASUREMENT

Data Type

Qualitative

Quantitative

(Discrete/Continuous)

Level of Measurement

Nominal

Ordinal

Interval

Ratio

# ILLUSIONS IN STATISTICS

ในบางครั้งข้อมูลที่เรามีอาจจะชี้นำเราสู่ข้อสรุปที่ไม่ถูกต้อง ซึ่งเราอาจจะต้องพิจารณาตัวบริบทของข้อมูล เพื่อประกอบการวิเคราะห์ของเรา

# 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.*

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	44%	4321	35%

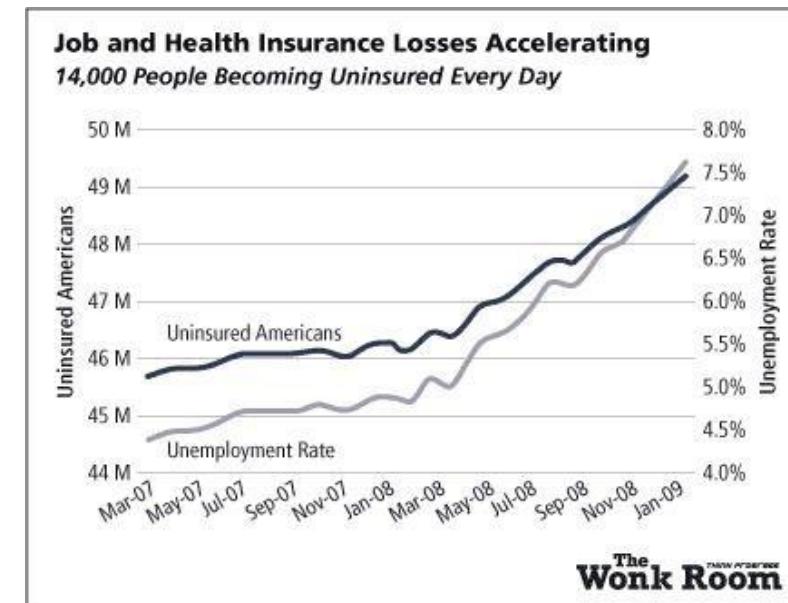
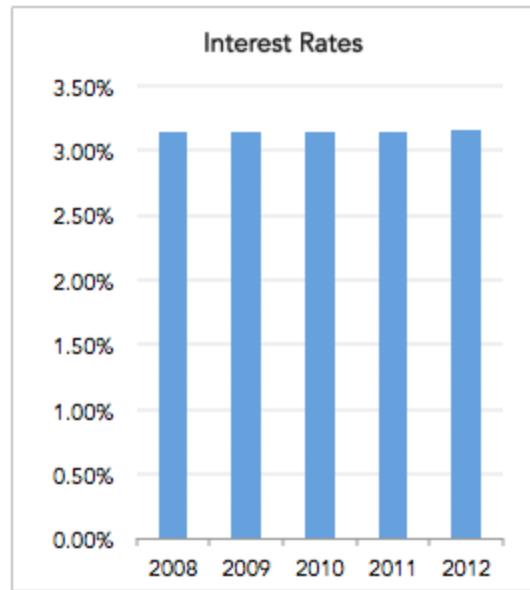
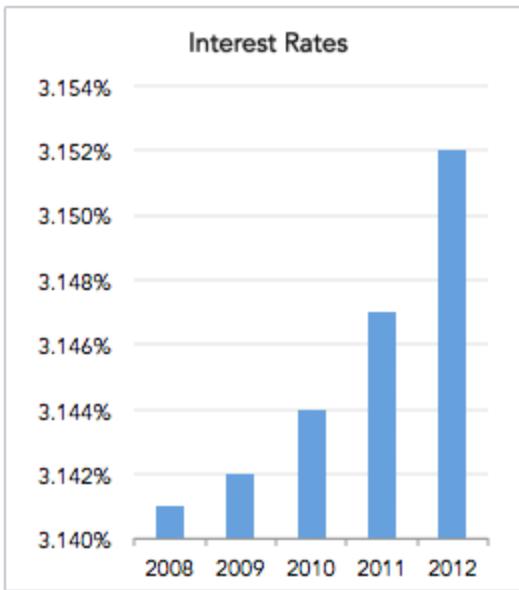
# SIMPSON'S PARADOX

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	44%	4321	35%

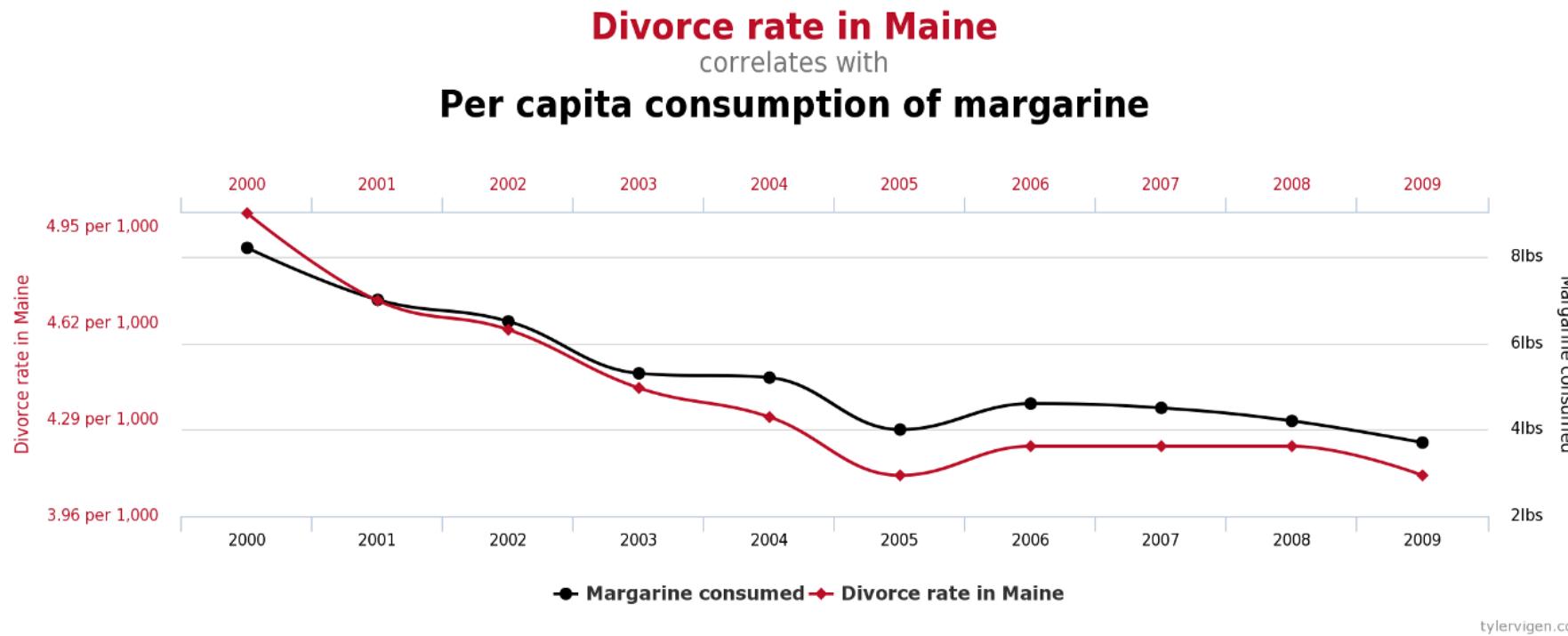
Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

# LABELS ON Y-AXIS

Same Data, Different Y-Axis



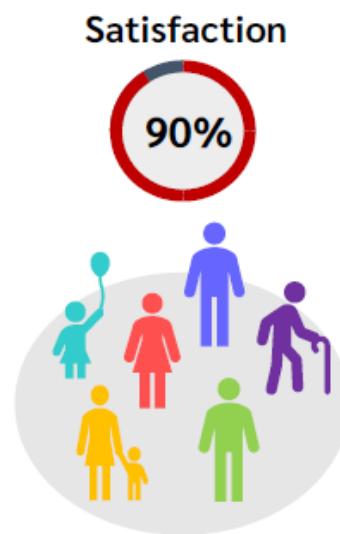
# CORRELATION & CAUSATION



# TYPES OF STATISTICS

## Descriptive statistics

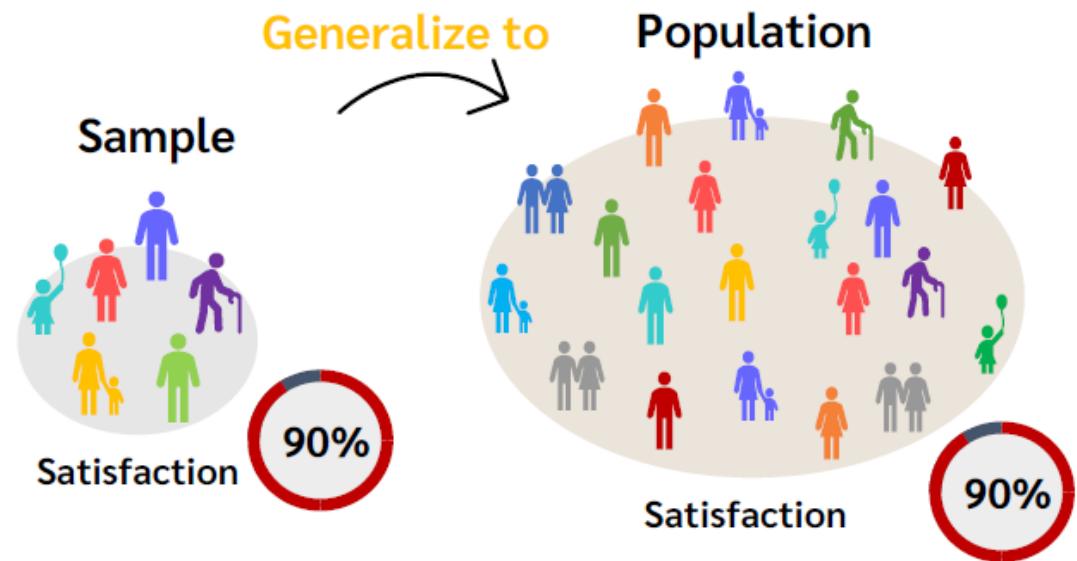
Collecting, Summarizing, and Presenting data



i.e. 90% satisfaction of all customers

## Inferential statistics

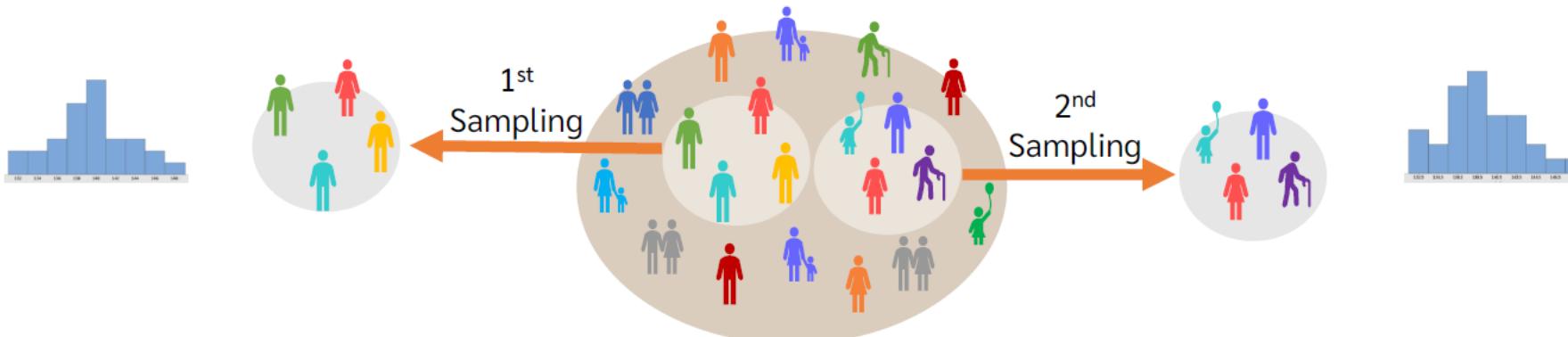
Drawing conclusions about a population based only on sample data



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

# SAMPLE & POPULATION

- **Parameter** = a number that describes an entire **population**.
- **Statistic** = a number that describes a **sample** of population.
  - The value of a statistic can change from sample to sample.



- A statistic is used to estimate an unknown parameter.

# DESCRIPTIVE STATISTICS

# Descriptive Statistics

Categorical or Qualitative data

1 variable

- Frequency Tables
- Pie Charts
- Bar Charts

2 variables

- Contingency Tables
- Grouped Pie Charts
- Grouped Bar Charts

Numerical or Quantitative data

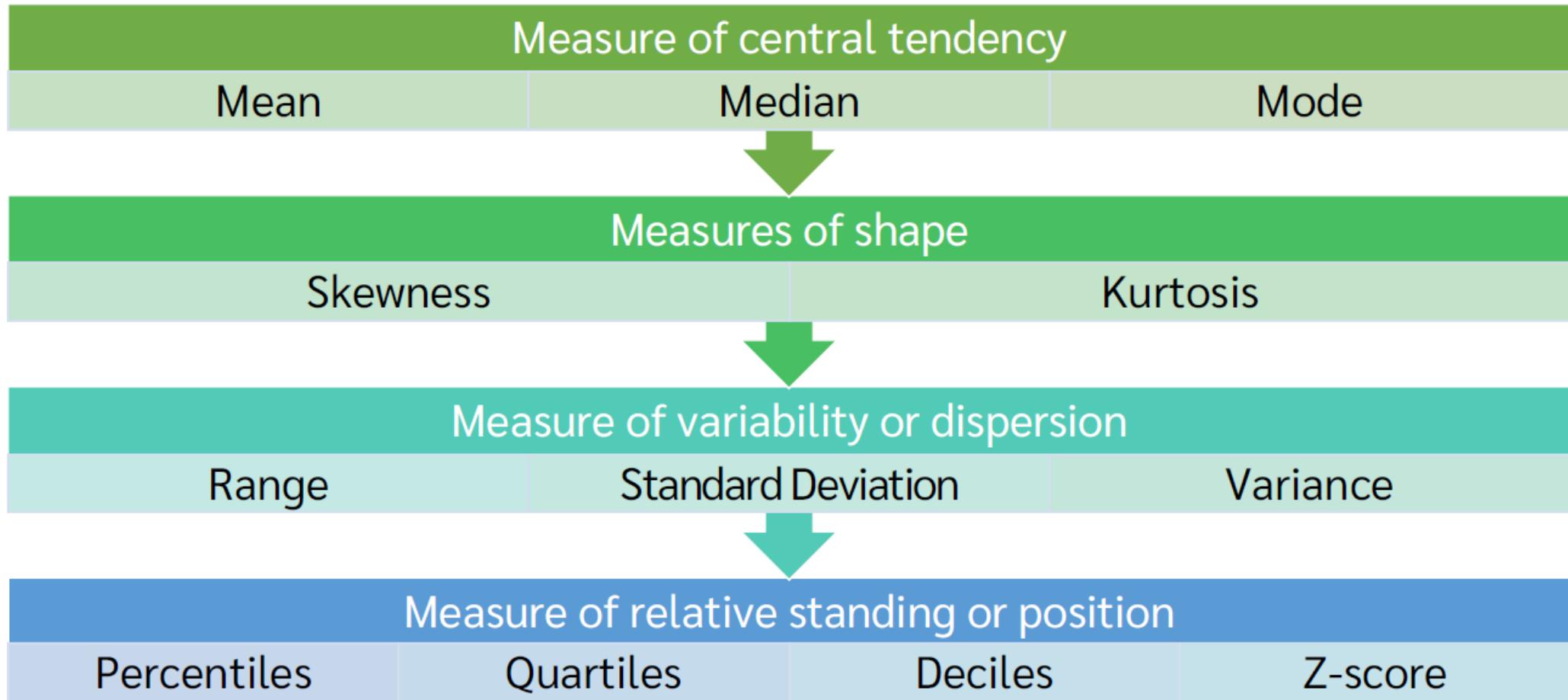
Display Graphically

- Frequency Tables
- Histogram
- Bar charts
- Line charts

Summary statistics

- Center: Mean, Median, Mode
- Spread: Range, Variance, Standard Deviation
- Shape: Skewness, Kurtosis
- Unique features: Outlier

# STATISTICS FOR NUMERICAL DATA



# CENTRAL TENDENCY

Consider 46, 54, 42, 46, 32

## MEAN

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{46 + 54 + 42 + 46 + 32}{5} = 44$$

## MODE

The mode is the value that occurs with greatest frequency.

For above data, mode is 46.

# CENTRAL TENDENCY

## MEDIAN

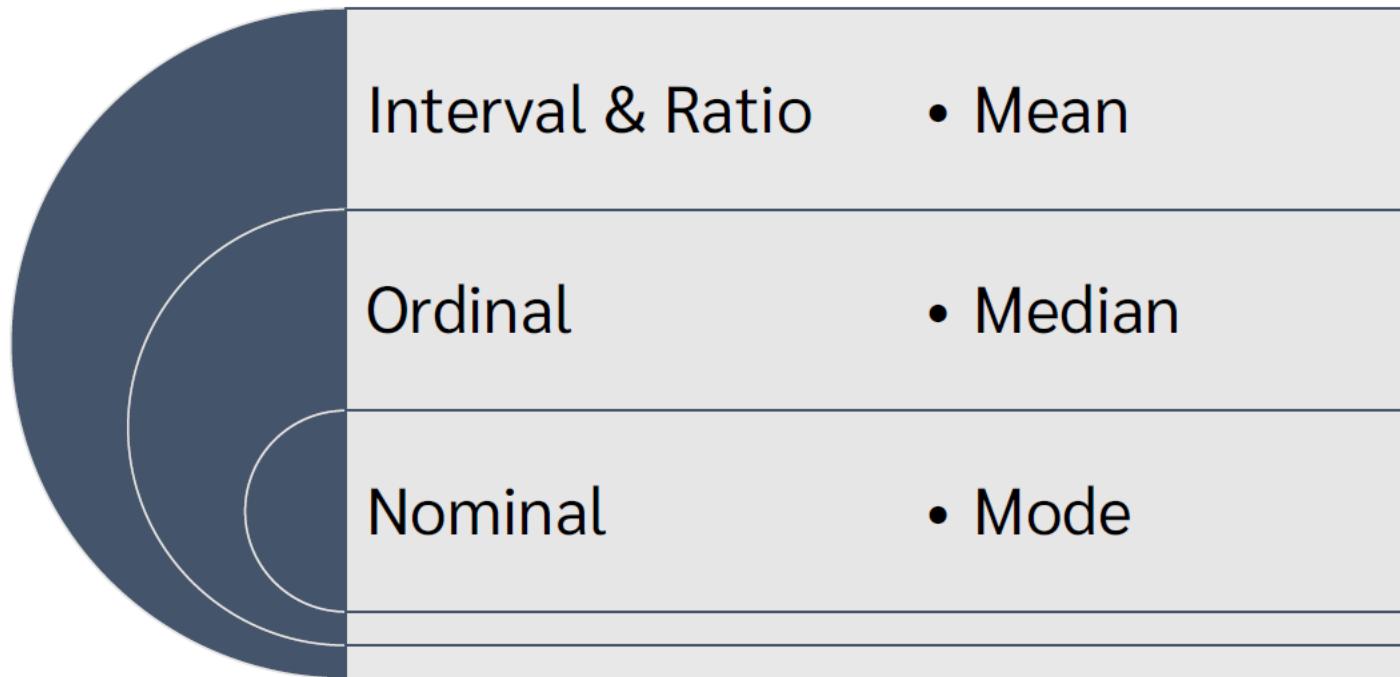
Arrange the data in ascending order (smallest value to largest value).

$$\text{Median} = \begin{cases} \frac{(N+1)^{\text{th}}}{2} \text{ term ; When } N \text{ is odd} \\ \frac{N^{\text{th}}}{2} \text{ term + } \left(\frac{N}{2}+1\right) \text{ term} \\ \hline \frac{1}{2} \end{cases} ; \text{ When } N \text{ is even}$$

Consider 46, 54, 42, 46, 32

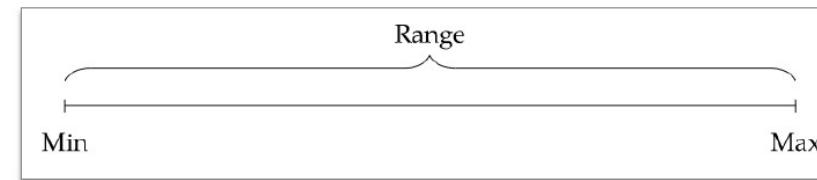
Sort >> 32, 42, 46, 54

# CENTRAL TENDENCY



# DISPERSION – RANGE

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



# DISPERSION - VARIANCE

The variance is a measure of variability that utilizes all the data.

The variance is based on the difference between the value of each observation ( $x_i$ ) and the mean.

## POPULATION VARIANCE

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$$

## SAMPLE VARIANCE

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

# DISPERSION - VARIANCE

Number of Students in Class ( $x_i$ )	Mean Class Size ( $\bar{x}$ )	Deviation About the Mean ( $x_i - \bar{x}$ )	Squared Deviation About the Mean ( $(x_i - \bar{x})^2$ )
46	44	2	4
54	44	10	100
42	44	-2	4
46	44	2	4
32	44	-12	144
		0	256
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{256}{4} = 64$$

# DISPERSION - STANDARD DEVIATION

The standard deviation is defined to be the positive square root of the variance.

## STANDARD DEVIATION

$$\text{Sample standard deviation} = s = \sqrt{s^2}$$

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2}$$

# POSITION - PERCENTILE

The  $p^{\text{th}}$  percentile is a value such that at least  $p$  percent of the observations are less than or equal to this value and at least  $(100 - p)$  percent of the observations are greater than or equal to this value.

**Step 1.** Arrange the data in ascending order (smallest value to largest value).

**Step 2.** Compute an index  $i$

$$i = \left( \frac{p}{100} \right) n$$

where  $p$  is the percentile of interest and  $n$  is the number of observations.

- Step 3.** (a) If  $i$  is not an integer, round up. The next integer greater than  $i$  denotes the position of the  $p^{\text{th}}$  percentile.  
(b) If  $i$  is an integer, the  $p^{\text{th}}$  percentile is the average of the values in positions  $i$  and  $i + 1$ .

# POSITION - PERCENTILE

How to find 85<sup>th</sup> percentile

**Step 1.** Arrange the data in ascending order.

1	2	3	4	5	6	7	8	9	10	11	12
3310	3355	3450	3480	3480	3490	3520	3540	3550	3650	3730	3925

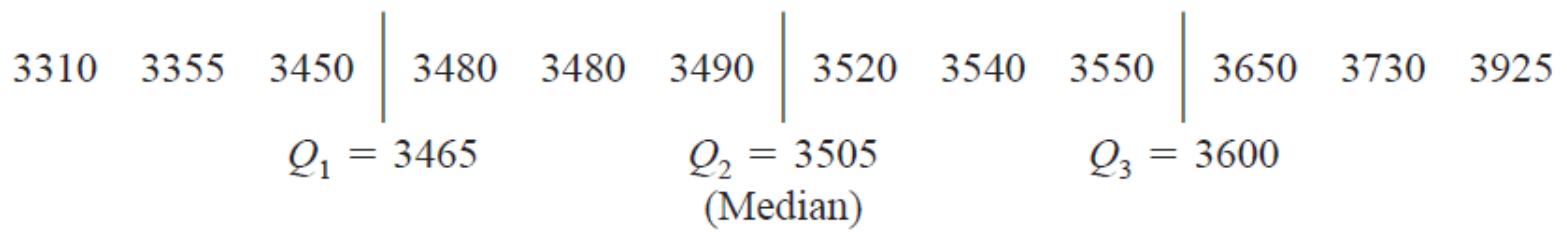
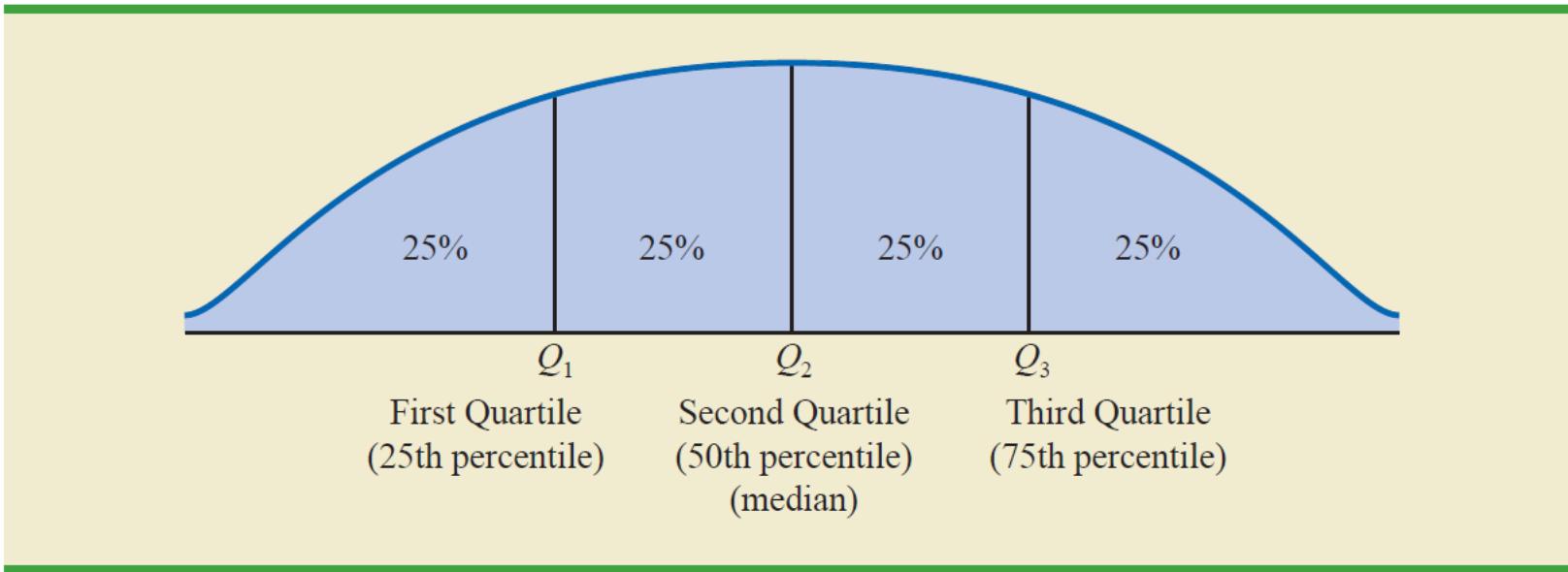
**Step 2.**

$$i = \left( \frac{p}{100} \right) n = \left( \frac{85}{100} \right) 12 = 10.2$$

**Step 3.** Because  $i$  is not an integer, *round up*. The position of the 85th percentile is the next integer greater than 10.2, the 11th position.

Returning to the data, we see that the 85th percentile is the data value in the 11th position, or 3730.

# POSITION - QUARTILE



# POSITION

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

The computations of quartiles  $Q_1$  and  $Q_3$  require the use of the rule for finding the 25th and 75th percentiles. These calculations follow.

For  $Q_1$ ,

$$i = \left( \frac{p}{100} \right) n = \left( \frac{25}{100} \right) 12 = 3$$

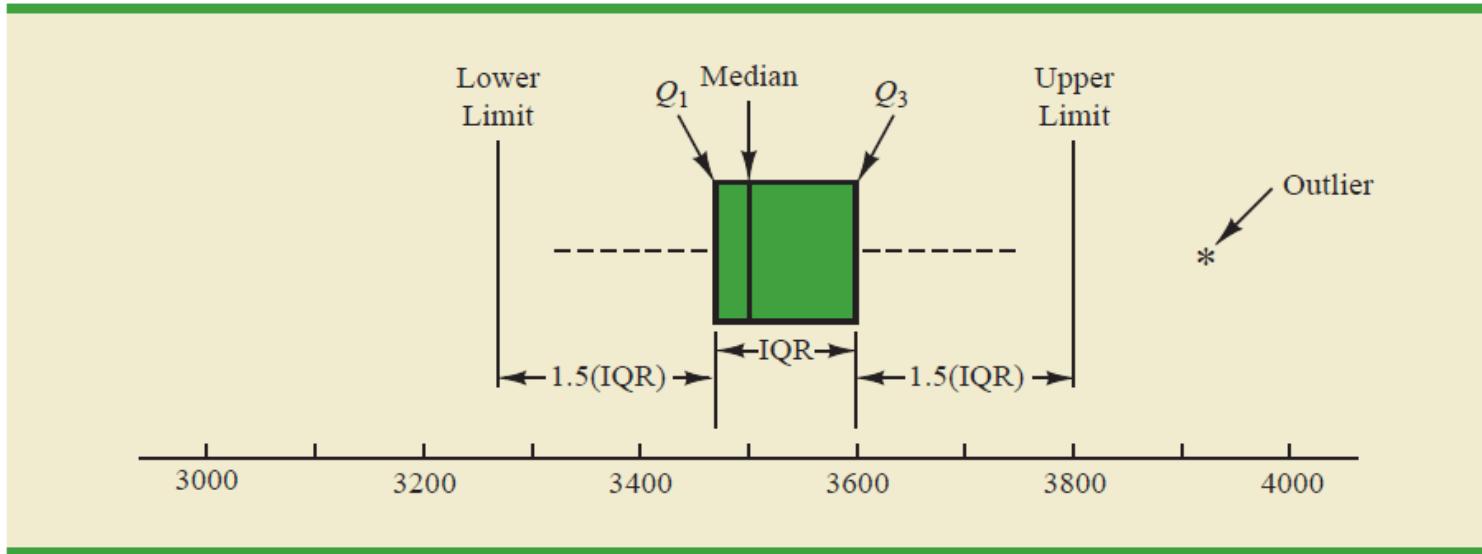
Because  $i$  is an integer, step 3(b) indicates that the first quartile, or 25th percentile, is the average of the third and fourth data values; thus,  $Q_1 = (3450 + 3480)/2 = 3465$ .

For  $Q_3$ ,

$$i = \left( \frac{p}{100} \right) n = \left( \frac{75}{100} \right) 12 = 9$$

Again, because  $i$  is an integer, step 3(b) indicates that the third quartile, or 75th percentile, is the average of the ninth and tenth data values; thus,  $Q_3 = (3550 + 3650)/2 = 3600$ .

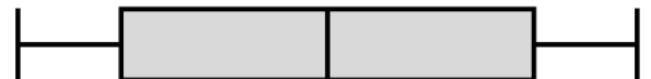
# IQR



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

## Normal Distribution

$(\text{Quartile 3} - \text{Quartile 2}) = (\text{Quartile 2} - \text{Quartile 1})$



## Positive Skew

$(\text{Quartile 3} - \text{Quartile 2}) > (\text{Quartile 2} - \text{Quartile 1})$



## Negative Skew

$(\text{Quartile 3} - \text{Quartile 2}) < (\text{Quartile 2} - \text{Quartile 1})$



# VARIATION OR DISPERSION

Nominal & Ordinal

Distribution tables

Bar Chart

Interval & Ratio

Range

Standard deviation

Variance

# RELATIVE STANDING

z-SCORE

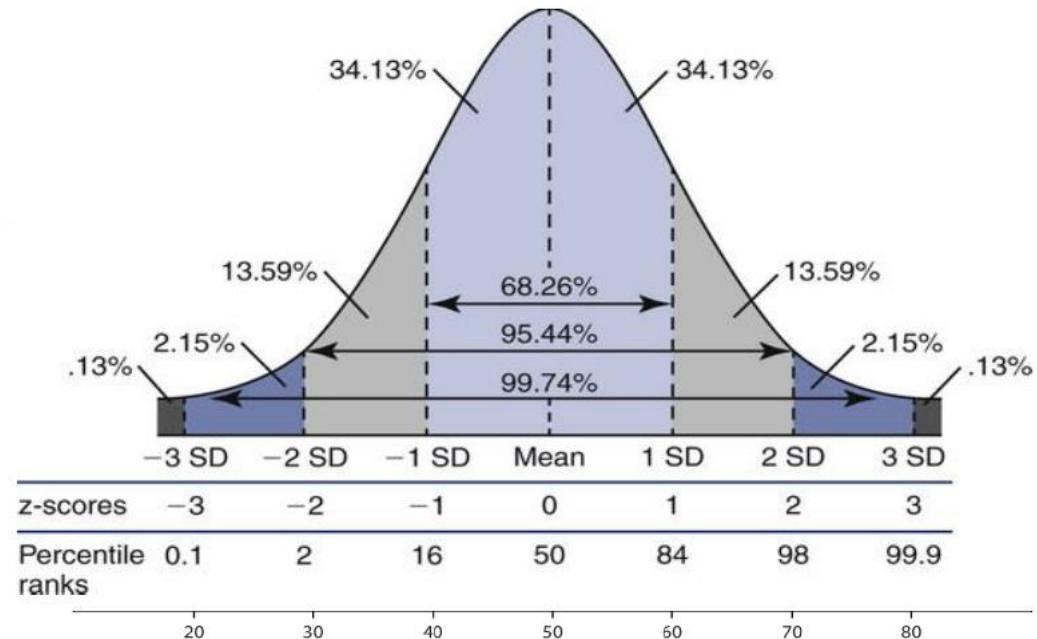
$$z_i = \frac{x_i - \bar{x}}{s}$$

where

$z_i$  = the z-score for  $x_i$

$\bar{x}$  = the sample mean

$s$  = the sample standard deviation



# RELATIVE STANDING

- **Example:** Comparing student scores on the SAT and ACT tests.
- Suppose that student A scored 1800 on the SAT, and student B scored 24 on the ACT.
- **Which student performed better relative to other test-takers?**

	SAT	ACT
Mean	1500	21
Standard deviation	300	5

$$\text{The z-score for student A is } z = \frac{x - \mu}{\sigma} = \frac{1800 - 1500}{300} = 1$$

$$\text{The z-score for student B is } z = \frac{x - \mu}{\sigma} = \frac{24 - 21}{5} = 0.6$$

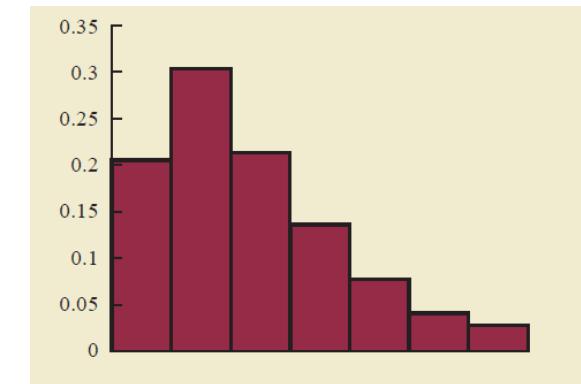
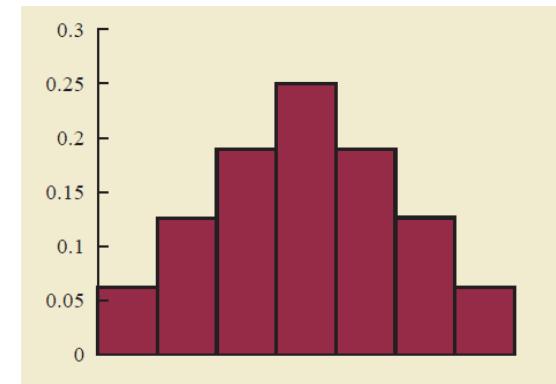
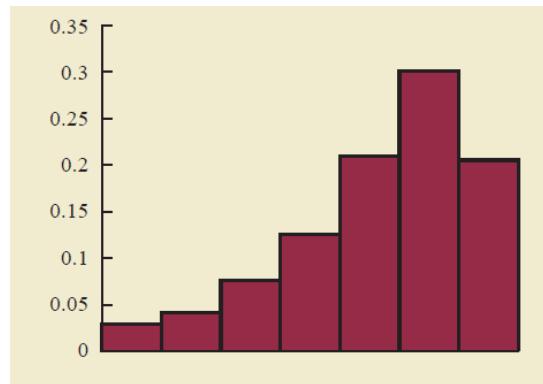
- Because student A has a higher z-score than student B, student A performed better compared to other test-takers than did student B.

# SHAPE - SKEWNESS

Extent to which the data values are not symmetrical around the mean.

Negative/Left-skewed : Mean < Median

Positive/Right-skewed : Mean > Median



# SHAPE - SKEWNWESS

Sample

$$\frac{n \sum_{i=1}^n (x_i - \bar{x})^3}{(n - 1)(n - 2)s^3}$$

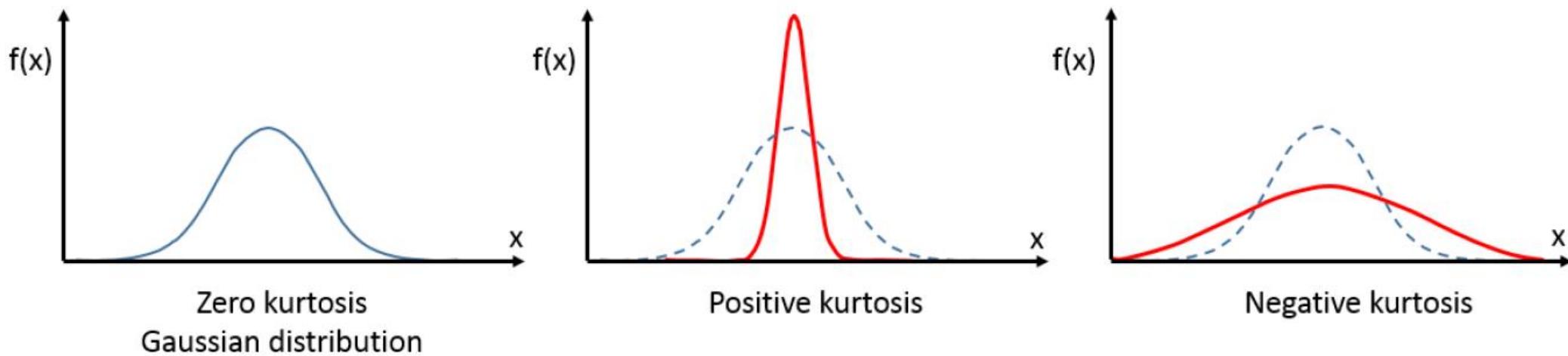
Population

$$\frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{\sigma} \right)^3$$

# SHAPE - KURTOSIS

Kurtosis ( $k$ ) is a unitless parameter or statistic that quantifies the distribution shape of a signal relative to a Gaussian distribution.

The distribution could be “sharper”, “flatter”, or equal to the Gaussian distribution



# SHAPE - KURTOSIS

Sample

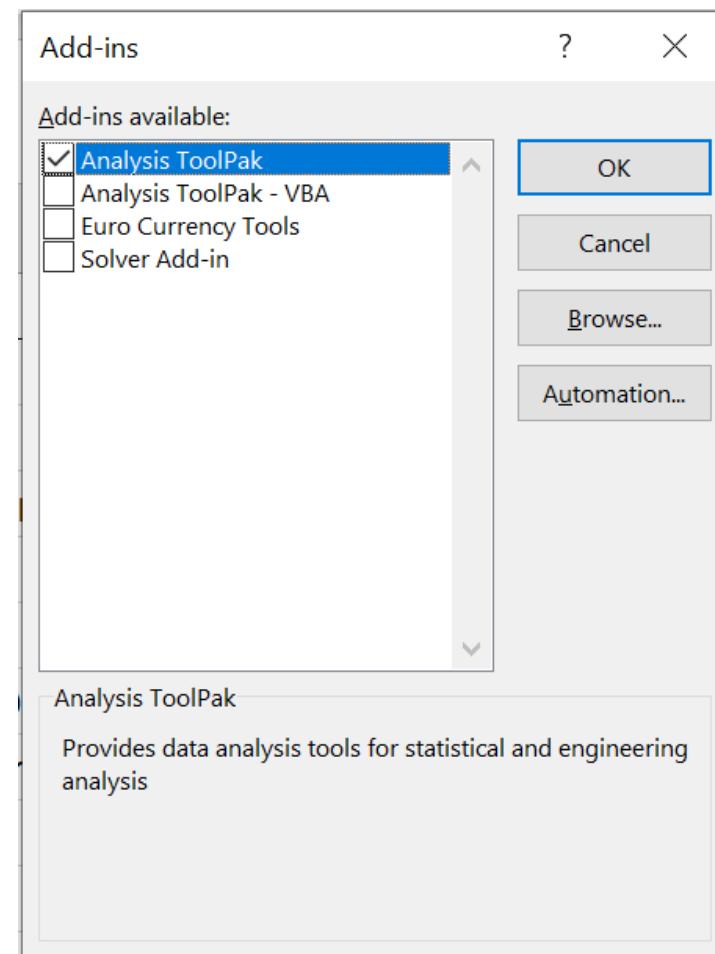
$$\frac{n(n + 1) \sum_{i=1}^n (x_i - \bar{x})^4}{(n - 1)(n - 2)(n - 3)s^4} - \frac{3(n - 1)^2}{(n - 2)(n - 3)}$$

Population

$$\frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \mu}{\sigma} \right)^4$$

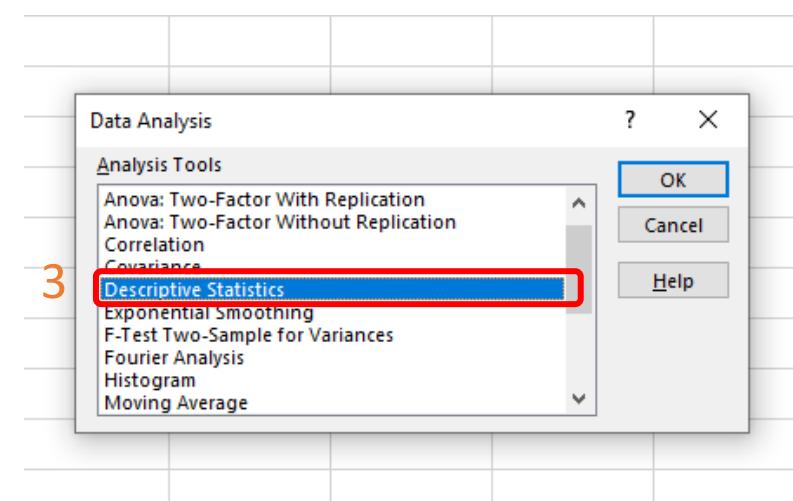
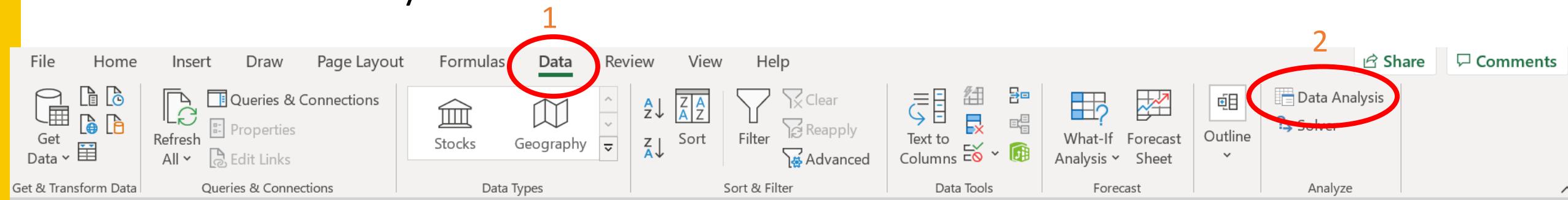
# EXCEL for DESCRIPTIVE STATISTICS

File >> More >> Options >> Add – ins >> Manage EXCEL Add – ins >> Analysis Toolpak



# EXCEL for DESCRIPTIVE STATISTICS

Data >> Data Analysis



# EXCEL for DESCRIPTIVE STATISTICS

Use “[Data01.xls](#)”

Age	
Mean	50.55
Standard Error	3.081694896
Median	51
Mode	51
Standard Deviation	30.81694896
Sample Variance	949.6843434
Kurtosis	-1.249871158
Skewness	-0.03066699
Range	98
Minimum	1
Maximum	99
Sum	5055
Count	100

3.64

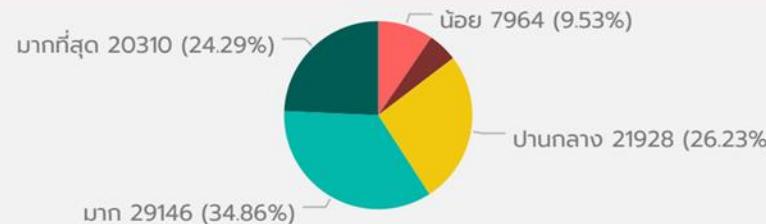
สามารถทำได้

83,611

ไม่สามารถทำได้

2,574

ก้าวใช้ภาษาไทยในการพัฒนาอ่านและเขียนได้อย่างดี



2.41

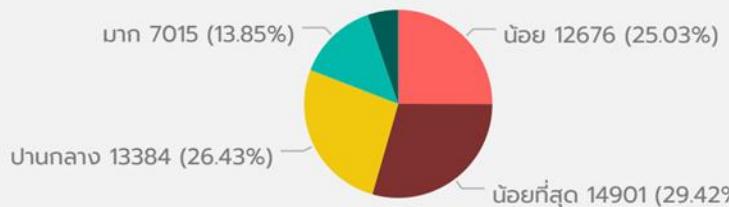
สามารถทำได้

50,647

ไม่สามารถทำได้

34,050

ก้าวใช้อินเทอร์เน็ตในการสืบค้นข้อมูลได้อย่างดี



2.21

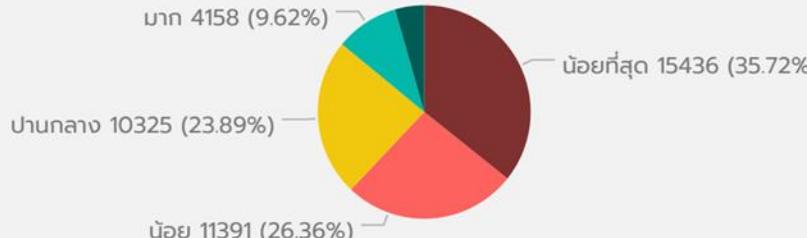
สามารถทำได้

43,219

ไม่สามารถทำได้

40,903

ก้าวใช้โปรแกรมแอพพลิเคชันเพื่อเป็นช่องทางในการสร้างรายได้อย่างดี

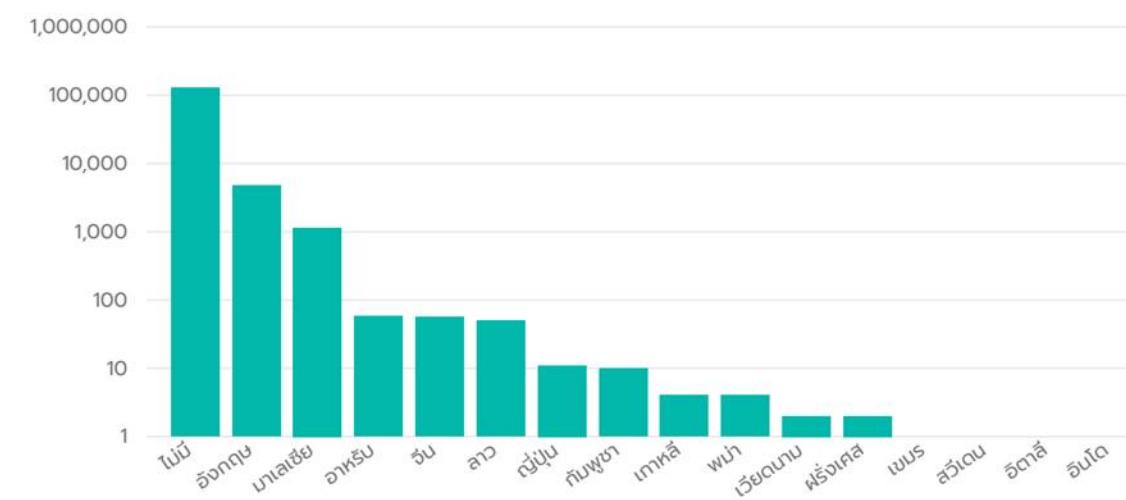
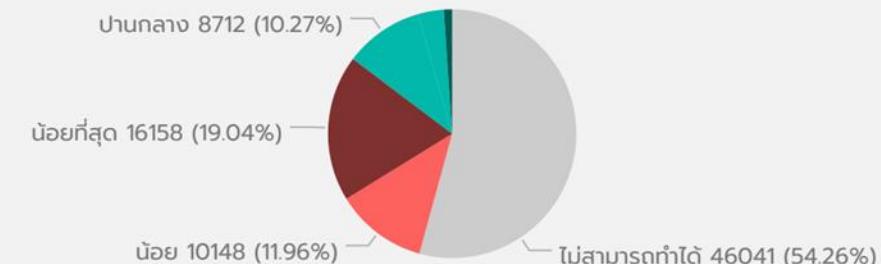


## ความสามารถในการทำงานของงานนอกระบบ

ก้าวสามารถใช้ภาษาต่างประเทศได้หรือไม่



ก้าวใช้ภาษาต่างประเทศเพื่อเป็นภาษาที่สองได้อย่างดี



# Workshop I

# ESSENTIAL COMMANDS

โหลด Package ที่จำเป็นและข้อมูลจากไฟล์ที่เตรียมไว้

```
import pandas as pd  
import numpy as np  
  
df = pd.read_csv('HR-Employee-Attrition.csv')
```

```
df.shape
```

```
(1470, 41)
```



```
df.info()
```

```
↳ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Age              1470 non-null    int64  
 1   Attrition        1470 non-null    object  
 2   BusinessTravel   1470 non-null    object  
 3   DailyRate         1470 non-null    int64  
 4   Department        1470 non-null    object  
 5   DistanceFromHome 1470 non-null    int64  
 6   Education         1470 non-null    int64  
 7   EducationField    1470 non-null    object  
 8   EmployeeCount     1470 non-null    int64  
 9   EmployeeNumber    1470 non-null    int64  
 10  EnvironmentSatisfaction 1470 non-null  int64
```

# PYTHON BASICS TEST

ถ้าเรออยากรู้ว่าพนักงานผู้หญิง (Gender = “Female”) มี อายุเฉลี่ยเท่ากับกี่ปีจะต้องใช้คำสั่งอะไร?

```
female_selector = df["Gender"]=="Female"  
df_female = df[female_selector]  
print("The average age of female employees is "+str(df_female['Age'].mean()))
```

The average age of female employees is 37.32993197278912

เรอาจจะใช้ groupby() method เพื่อแบ่งตามหลายๆ Category ได้ โดยหลักการทำงานคร่าวๆดังนี้

- แบ่งกลุ่มข้อมูลตามค่าของคอลัมน์ที่เราต้องการ (เช่น แบ่งตามเพศ)
- เลือกข้อมูลคอลัมน์ที่เราสนใจ (เช่น สนใจอายุ)
- คำนวณสถิติสรุป (เช่น สรุปด้วยค่าเฉลี่ย)

# DESCRIPTIVE STATISTICS BY CATEGORY

df.groupby("JobRole")["DailyRate", "TotalWorkingYears"].mean()

/usr/local/lib/python3.7/dist-packages/ipykernel\_launcher.py:1: FutureWarning: The 'Entry point for launching an IPython kernel.'

JobRole	DailyRate	TotalWorkingYears
Healthcare Representative	854.251908	14.068702
Human Resources	757.923077	8.173077
Laboratory Technician	796.617761	7.656371
Manager	782.950980	24.549020
Manufacturing Director	796.020690	12.786207
Research Director	802.450000	21.400000
Research Scientist	800.359589	7.715753
Sales Executive	802.098160	11.101227
Sales Representative	811.349398	4.674699

# GROUPBY MULTIPLE ASPECTS

```
df.groupby(["JobLevel","Gender"])["MonthlyIncome"].agg(['mean','std'])
```

		mean	std	edit
JobLevel	Gender			
1	Female	2780.487437	709.605053	
	Male	2790.633721	771.300650	
2	Female	5435.327273	1266.690800	
	Male	5549.184713	1502.541186	
3	Female	9962.702128	1892.555119	
	Male	9706.991935	1737.145905	
4	Female	15431.372549	1701.573119	
	Male	15570.927273	1929.704985	
5	Female	19129.916667	587.444052	
	Male	19224.844444	471.321930	

# PRACTICE

หากเราสนใจว่าแต่ละ JobRole มีความพึงพอใจกับงานโดยเฉลี่ยเท่าใด จะใช้คำสั่งอะไร

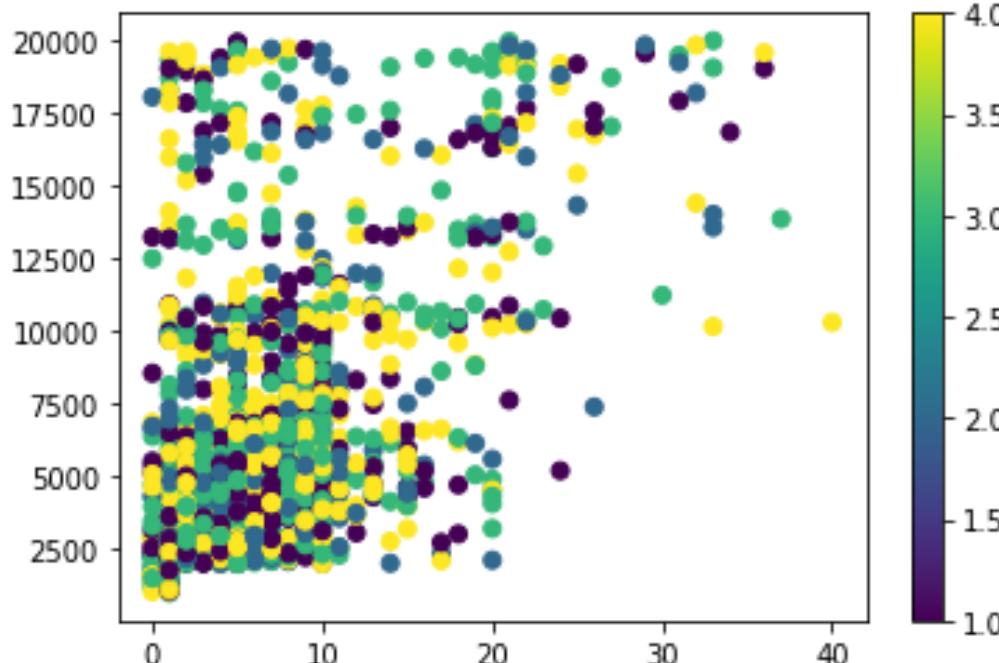
JobRole	mean	std
<b>Healthcare Representative</b>	2.786260	1.109453
<b>Human Resources</b>	2.557692	1.073997
<b>Laboratory Technician</b>	2.691120	1.126306
<b>Manager</b>	2.705882	1.130918
<b>Manufacturing Director</b>	2.682759	1.052137
<b>Research Director</b>	2.700000	1.072085
<b>Research Scientist</b>	2.773973	1.095260
<b>Sales Executive</b>	2.754601	1.139997
<b>Sales Representative</b>	2.734940	1.025104

# SCATTER PLOT

ใช้สำหรับมองหาความสัมพันธ์ของข้อมูลสองตัวแปร

```
import matplotlib.pyplot as plt  
plt.scatter(df['YearsAtCompany'], df['MonthlyIncome'], c= df['JobSatisfaction'])  
plt.colorbar()
```

<matplotlib.colorbar.Colorbar at 0x7f606a4a1b90>



มีจุดเยื่อะ ทำให้เห็น pattern ไม่ชัด

# SAMPLING

หากข้อมูลมีจำนวนใหญ่ไปเราอาจจะเลือก sample มาจำนวนน้อยๆ ก่อนได้

```
df.sample(5)
```



	Age	Attrition	BusinessTravel	DailyRate
696	45	No	Non-Travel	805
415	34	Yes	Travel_Frequently	296
399	31	No	Travel_Rarely	329
1102	36	No	Travel_Rarely	1157
746	41	No	Non-Travel	247

5 rows x 41 columns

# SAMPLING

ใช้ random\_state เพื่อทำให้การ sample 2 ครั้งได้ผลลัพธ์เดียวกัน

```
df.sample(5,random_state=26)
```

	Age	Attrition	BusinessTravel	DailyRate
1402	31	No	Travel_Rarely	1276
1223	47	Yes	Travel_Frequently	1093
1415	33	No	Non-Travel	1313
997	27	Yes	Travel_Rarely	135
724	24	No	Travel_Rarely	1206

# EACH SAMPLING GIVES DIFFERENT RESULTS

การสุ่มตัวอย่าง 2 ครั้งอาจจะได้ผลลัพธ์ที่ไม่เหมือนกัน

```
sample_1 = df.sample(100,random_state = 21)
sample_2 = df.sample(100, random_state = 43)
print("Mean of Group 1 is "+str(sample_1['MonthlyIncome'].mean()))
print("Mean of Group 2 is "+str(sample_2['MonthlyIncome'].mean()))
```

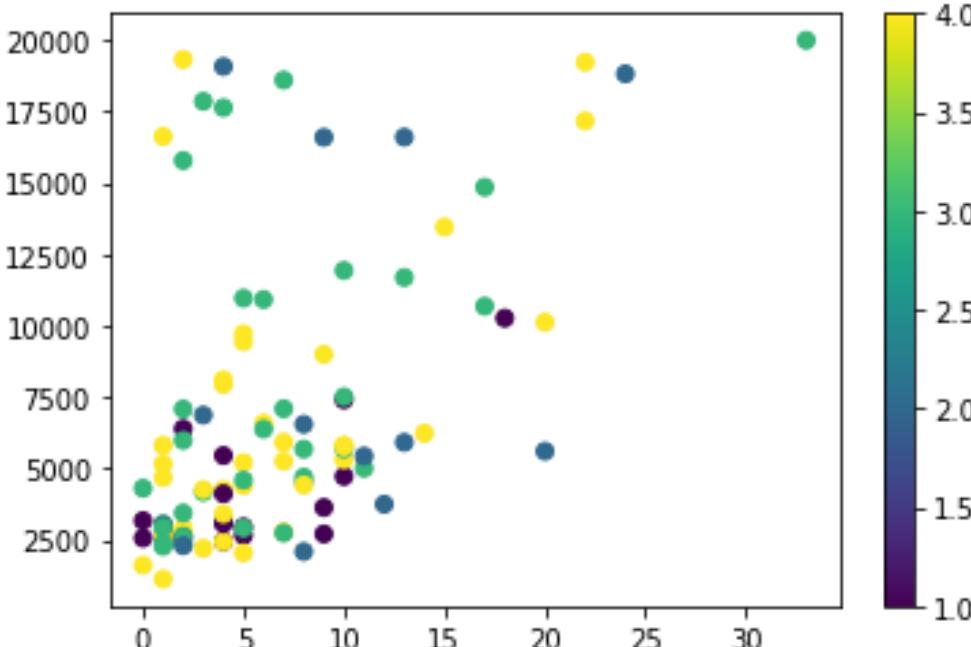
```
Mean of Group 1 is 6051.69
Mean of Group 2 is 7137.58
```

# SCATTER PLOT - RERUN

หากทำ scatter plot กับ sample 100 จะเห็นแนวโน้มขัดขึ้น

```
import matplotlib.pyplot as plt
df_sample = df.sample(100)
plt.scatter(df_sample['YearsAtCompany'], df_sample['MonthlyIncome'], c= df_sample['JobSatisfaction'])
plt.colorbar()
```

<matplotlib.colorbar.Colorbar at 0x7f60665f1250>

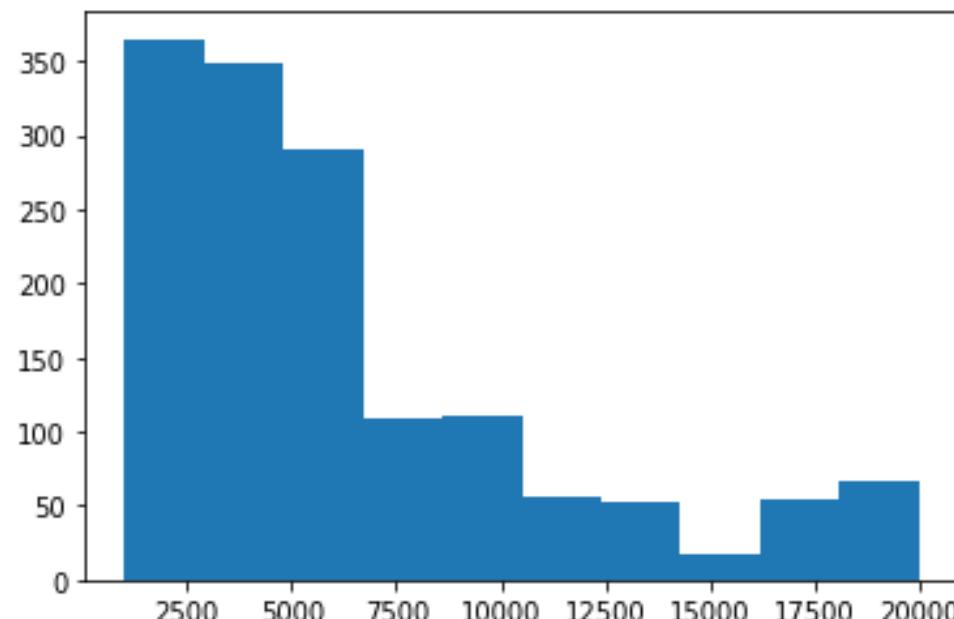


# HISTOGRAM – NOT EVERYTHING IS NORMAL

ข้อมูลที่มีอาจจะไม่ได้แจกแจงเป็น Normal Distribution เช่นอุป

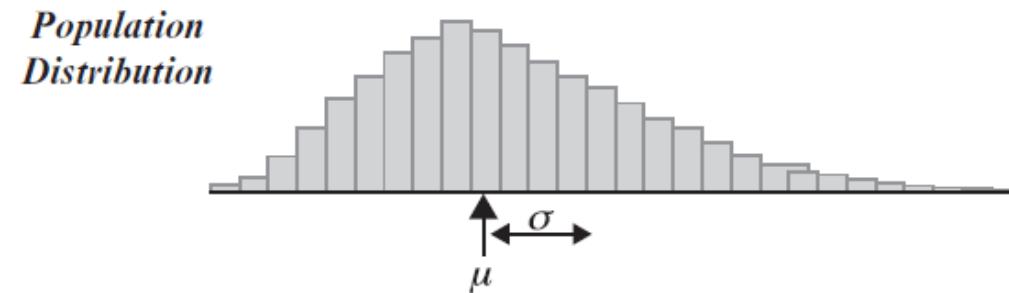
```
#not everything is normal distribution!!
plt.hist(df['MonthlyIncome'])
```

```
(array([365., 349., 290., 109., 110., 56., 52., 18., 54., 67.]),
 array([ 1009., 2908., 4807., 6706., 8605., 10504., 12403., 14302.,
        16201., 18100., 19999.]),
 <a list of 10 Patch objects>)
```

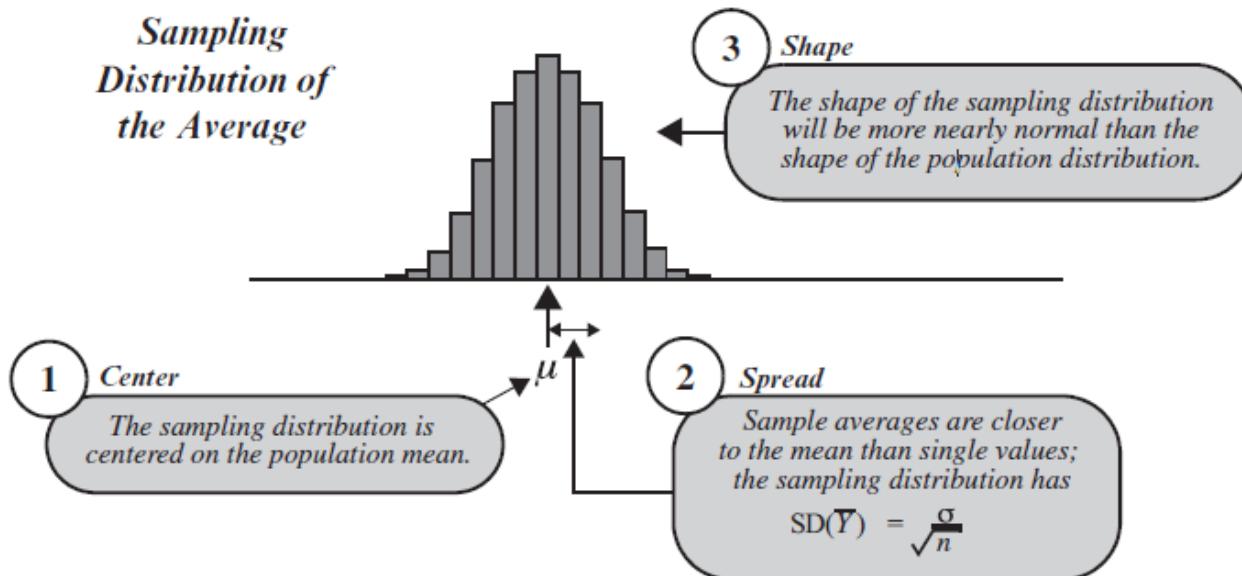


# CENTRAL LIMITING THEOREM

ถ้าเราเลือกจำนวน Sample มากพอ ค่าเฉลี่ยของ Sample จะมีการกระจายตัว



*Sampling Distribution of the Average*



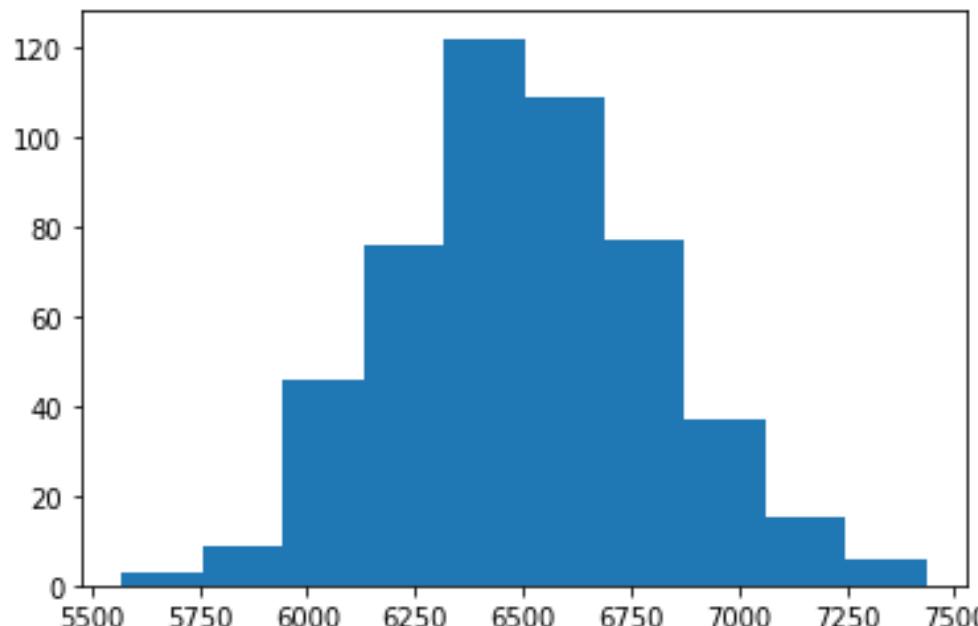
# CENTRAL LIMITING THEOREM

- ทำการ sample จาก df ทั้งหมด 500 ครั้ง โดยแต่ละครั้งให้เลือก 200 samples
- ให้หาค่าเฉลี่ยของ ‘MonthlyIncome’
- นำค่าเฉลี่ยเหล่านี้มา plot histogram

# CENTRAL LIMITING THEOREM

```
mean_list=[]
for i in range(500):
    average_income = df.sample(200)['MonthlyIncome'].mean()
    mean_list.append(average_income)
plt.hist(mean_list)
```

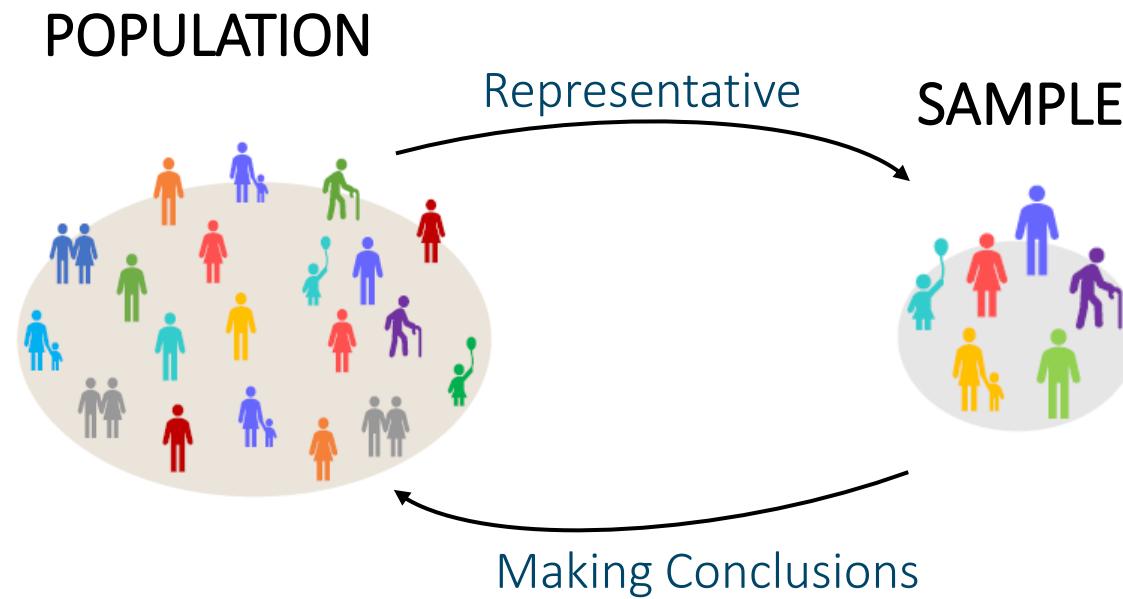
```
(array([ 3.,  9.,  46.,  76., 122., 109.,  77.,  37.,  15.,  6.]),
 array([5571.615 , 5757.9595, 5944.304 , 6130.6485, 6316.993 , 6503.3375,
        6689.682 , 6876.0265, 7062.371 , 7248.7155, 7435.06  ]),  
<a list of 10 Patch objects>)
```



# INFERENTIAL STATISTICS

# INFERENTIAL STATISTICS

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



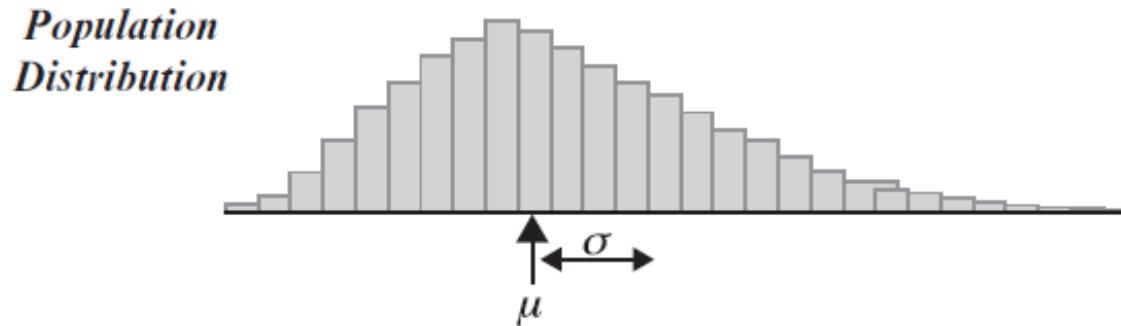
Description	Sample statistic
Mean	$\bar{x}$
Variance	$s^2$
Standard Deviation	$s$
Size	$n$
Correlation	$r$
Proportion	$\hat{p}$

# INFERENTIAL STATISTICS

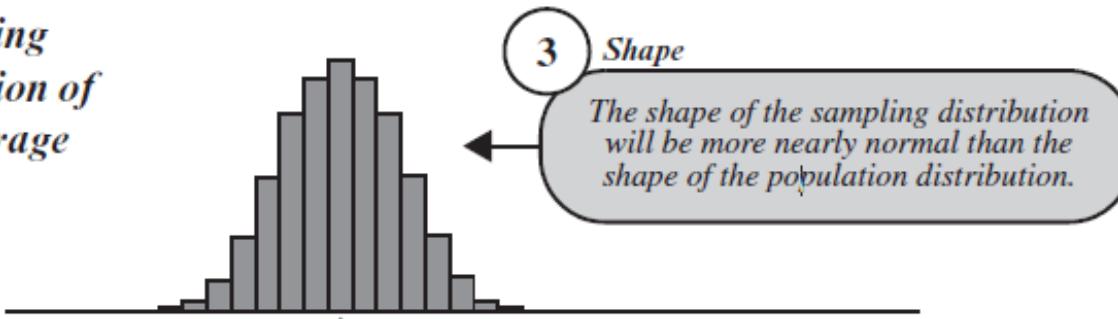


Different sampling results in different estimation – The smaller the sample size, the less reliable it is

# CENTRAL LIMIT THEOREM (CLT)



*Sampling Distribution of the Average*



3 Shape

The shape of the sampling distribution will be more nearly normal than the shape of the population distribution.

1 Center

The sampling distribution is centered on the population mean.

2 Spread

Sample averages are closer to the mean than single values; the sampling distribution has

$$SD(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$$

# CHECK LIST FOR CLT

*To use Central Limiting Theorem, we need to make sure that:*

**Samples are taken independently**

*The result of the previous samples don't influence the probabilities of the other samples*

**Samples come from the same distribution**

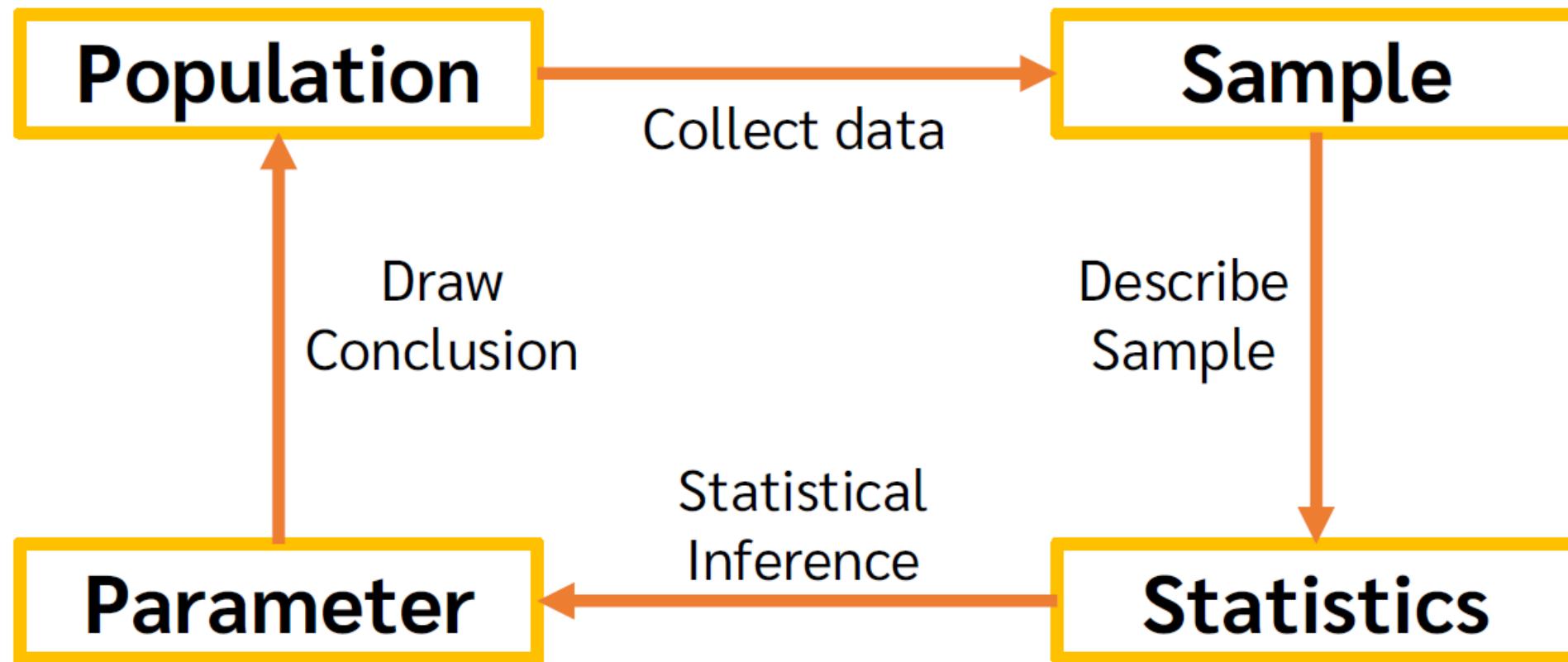
*The samples all have same probability rules*

**Flat tail behavior**

*The population distribution has finite second moment ( $E([X^2]) < \infty$ )*

*This is often relevant when it comes to risk management area where the distribution of the loss is heavy tailed*

# Using sample statistics to estimate population parameters.



# INFERENTIAL STATISTICS

What was the prevalence of smoking at Penn State University before the 'no smoking' policy?

---

The main campus at Penn State University has a population of approximately 42,000 students. A research question is "what proportion of these students smoke regularly?" A survey was administered to a sample of 987 Penn State students. Forty-three percent (43%) of the sampled students reported that they smoked regularly. How confident can we be that 43% is close to the actual proportion of all Penn State students who smoke?

- The population is all 42,000 students at Penn State University.
- The parameter of interest is  $p$ , the proportion of students at Penn State University who smoke regularly.
- The sample is a random selection of 987 students at Penn State University.
- The statistic is the proportion,  $\hat{p}$ , of the sample of 987 students who smoke regularly. The value of the sample proportion is 0.43.

# Inferential Statistical



## 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

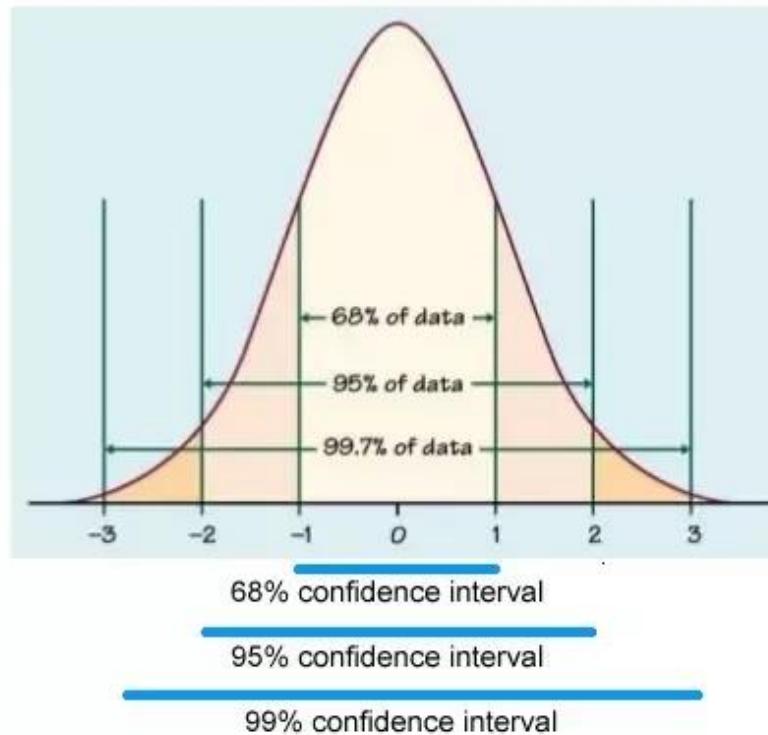
## Point Estimation

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 : $\hat{p}$

# PARAMETER ESTIMATION

## Interval Estimation

Confidence interval (CI)



# 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 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: Confidence Interval

95% Confidence Interval:

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

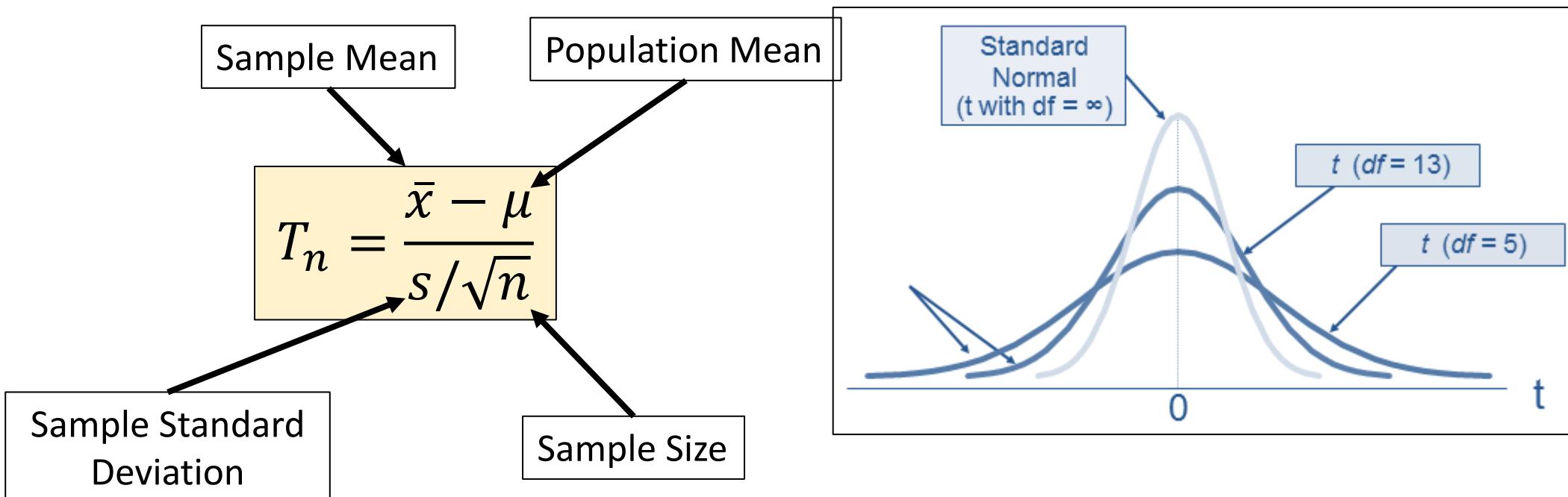
Conclusion:

*With 95% probability, we may estimate*

$1,000 \times (0.3 \pm 0.13) = 300 \pm 130$  students to have a laptop

# T-DISTRIBUTION

*If the sample size is not big enough, we cannot use Central Limiting Theorem. Instead, we will use t-distribution*



# 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: Sampling for Mean

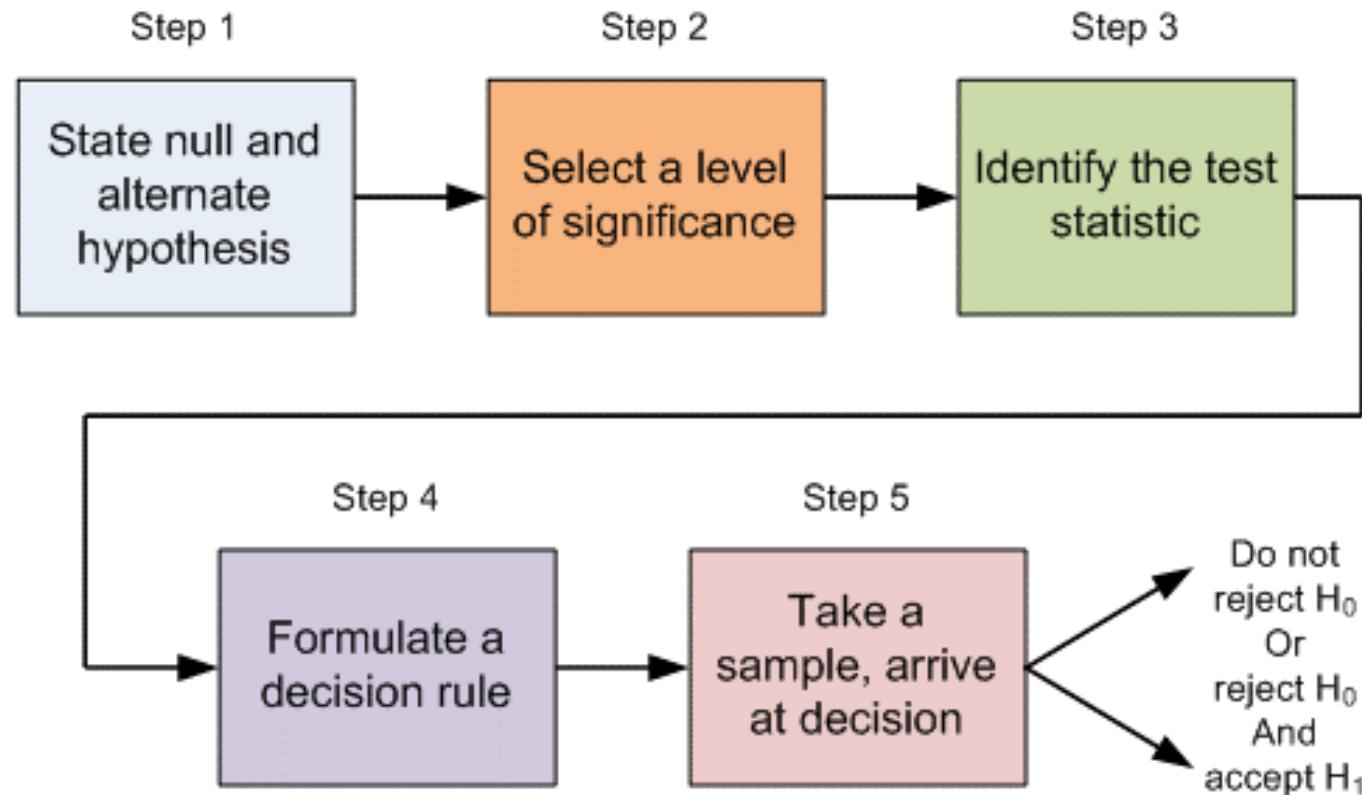
Using t-distribution,

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

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

*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.*

# HYPOTHESIS TESTING



# HYPOTHESIS TESTING

In hypothesis testing we begin by making a tentative assumption about a population parameter.

This tentative assumption is called the **null hypothesis** and is denoted by  $H_0$ .

We then define another hypothesis, called the **alternative hypothesis**,

$$H_0: \mu \geq \mu_0$$

$$H_a: \mu < \mu_0$$

$$H_0: \mu \leq \mu_0$$

$$H_a: \mu > \mu_0$$

$$H_0: \mu = \mu_0$$

$$H_a: \mu \neq \mu_0$$

# HYPOTHESIS TESTING

ตัวอย่าง 1 บริษัทผู้ผลิตหลอดไฟอ้างว่า หลอดไฟของเขามีอายุการใช้งานเฉลี่ยนานกว่า 1,000 ชั่วโมง

$$H_0 : \mu \leq 1,000$$

$$H_0 : \mu = 1,000$$

$$H_1 : \mu > 1,000$$

$$H_1 : \mu > 1,000$$

ตัวอย่าง 2 สัดส่วนของครอบครัวไทยที่เห็นด้วยว่าครอบครัวมีบุตรไม่เกิน 2 คน มีค่าอย่างน้อย 0.7

$$H_0 : p \geq 0.7$$

$$H_0 : p = 0.7$$

$$H_1 : p < 0.7$$

$$H_1 : p < 0.7$$

ตัวอย่าง 3 ความแปรปรวนของราคายางพารา ในปี 2562 มาากกว่า 12.25 บาท

$$H_0 : \sigma^2 \leq 12.25$$

$$H_0 : \sigma^2 = 12.25$$

$$H_1 : \sigma^2 > 12.25$$

$$H_1 : \sigma^2 > 12.25$$

# CHOOSE YOUR HYPOTHESIS

There are three types of hypothesis:

1. Less than:

$$\begin{aligned}H_0: \mu &= \mu_0 \\H_A: \mu &< \mu_0\end{aligned}$$

2. Greater than:

$$\begin{aligned}H_0: \mu &= \mu_0 \\H_A: \mu &> \mu_0\end{aligned}$$

3. Not Equal:

$$\begin{aligned}H_0: \mu &= \mu_0 \\H_A: \mu &\neq \mu_0\end{aligned}$$

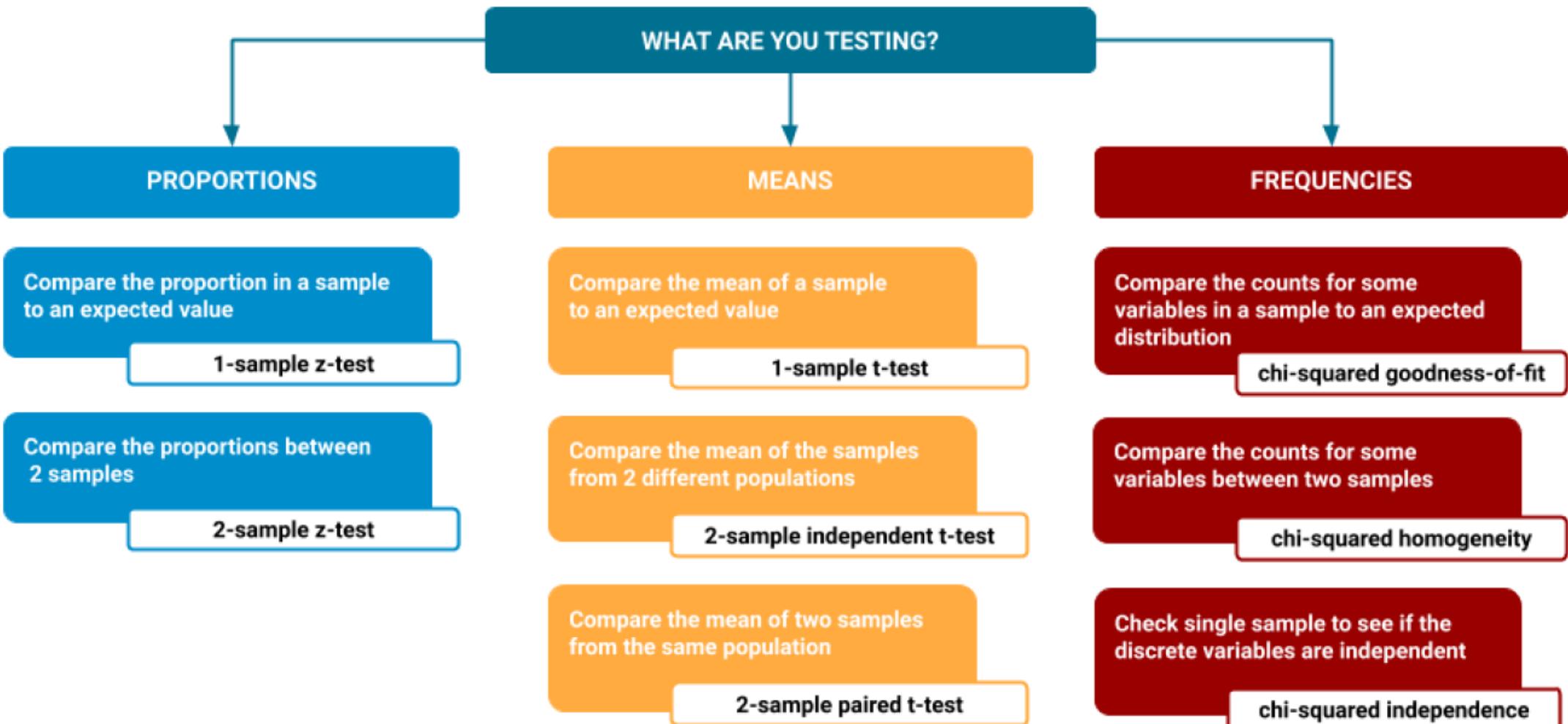


One-tailed



Two-tailed

# HYPOTHESIS TESTING



# INFERRENTIAL STATISTICS

จากการศึกษาการแก้ไขคุณภาพอาหารในพื้นที่แห่งหนึ่งโดยใช้มาตรการหนึ่ง พบว่ามีปริมาณก้าชคาร์บอนไดออกไซด์เฉลี่ย 9.4 ppm เพื่อที่จะตรวจสอบว่ามาตรการการแก้ไขดังกล่าวช่วยลดก้าชคาร์บอนไดออกไซด์ได้จริง จึงได้ตรวจวัดปริมาณก้าชคาร์บอนไดออกไซด์ที่จุดต่างๆ ทั้งหมด 18 จุดดังตารางต่อไปนี้

8.6	6.4	7.2	10.5	8.7	10.7	5.4	5.7	3.9
4.5	3.6	7.6	6.8	10.9	10.2	7.9	9.4	7.9

สมมุติว่าปริมาณก้าชคาร์บอนไดออกไซด์มีการแจกแจงปกติ อยากรู้ว่ามาตรการดังกล่าวแก้ไขได้ผลหรือไม่ ที่ระดับนัยสำคัญ 0.05

กำหนดให้  $\mu$  แทน ปริมาณก้าชคาร์บอนไดออกไซด์เฉลี่ยในพื้นที่ที่ทำการศึกษา

สมมติฐาน

$$H_0 : \mu = 9.4$$

จะได้ว่า  $\bar{x} = 7.75$        $\mu_0 = 9.4$

$$H_1 : \mu < 9.4$$

$$s^2 = 5.37 \quad df = n - 1 = 17$$

$$n = 18$$

# INFERRENTIAL STATISTICS

ค่าสถิติทดสอบ

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{7.55 - 9.4}{\sqrt{5.37 / 18}} = -3.38$$

นัยสำคัญ  $\alpha = 0.05$

ค่าวิกฤต  $-t_{\alpha, df} = -t_{0.05, 17} = -1.740$

บริเวณวิกฤต  $t < -1.740$

สรุปผลการทดสอบ  
ค่าสถิติทดสอบ -3.38 มีค่าน้อยกว่า -1.740  
นั่นคือ ค่าที่คำนวณได้ตกอยู่ในบริเวณวิกฤต  
สรุปผลได้ว่า ปฏิเสธ  $H_0$   
นั่นคือมาตราการดังกล่าวได้ผลจริง

# HYPOTHESIS TESTING

Errors and correct conclusions in hypothesis testing

		Population Condition	
		$H_0$ True	$H_a$ True
Conclusion	Accept $H_0$	Correct Conclusion	Type II Error
	Reject $H_0$	Type I Error	Correct Conclusion

# HYPOTHESIS TESTING

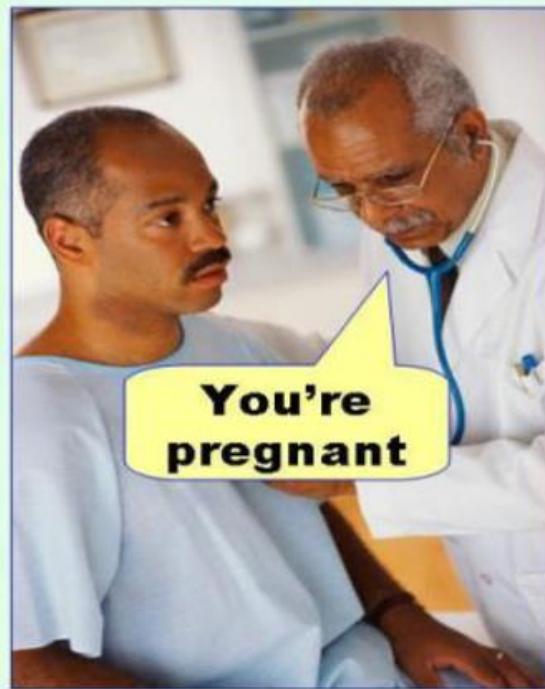
$H_0$  การดีมสุรา ไม่ทำให้เกิดโรคมะเร็งตับ

$H_1$  การดีมสุรา ทำให้เกิดโรคมะเร็งตับ

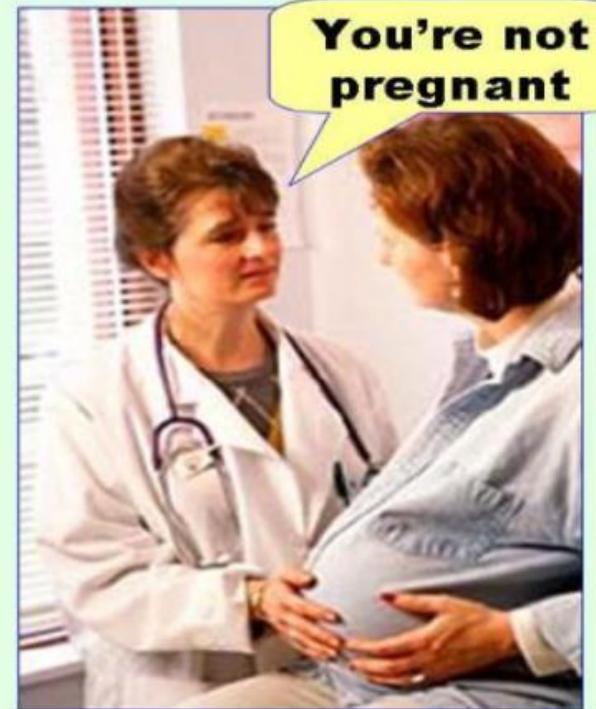
การตัดสินใจ	สมมติฐานหลักเป็นจริง	สมมติฐานหลักไม่เป็นจริง
ยอมรับสมมติฐานหลัก	<b>การตัดสินใจถูกต้อง</b> การดีมสุราไม่ทำให้เกิดโรคมะเร็งตับ	<b>ความผิดพลาดประเภทที่ 2</b> นั่นคือ การดีมสุราไม่ทำให้เกิดโรคมะเร็งตับ ทั้งที่ความเป็นจริง การดีมสุราส่งผลให้เกิดโรคมะเร็งตับ
ปฏิเสธสมมติฐานหลัก	<b>ความผิดพลาดประเภทที่ 1</b> นั่นคือ การดีมสุราทำให้เกิดโรคมะเร็งตับ ทั้งที่ความเป็นจริง การดีมสุรา ไม่ทำให้เกิดโรคมะเร็งตับ	<b>การตัดสินใจถูกต้อง</b> การดีมสุราทำให้เกิดโรคมะเร็งตับ

# HYPOTHESIS TESTING

**Type I error**  
(false positive)



**Type II error**  
(false negative)



# INFERENTIAL STATISTICS

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
<b>Hypotheses</b>	$H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$
<b>Test Statistic</b>	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
<b>Rejection Rule: <i>p</i>-Value Approach</b>	Reject $H_0$ if $p\text{-value} \leq \alpha$	Reject $H_0$ if $p\text{-value} \leq \alpha$	Reject $H_0$ if $p\text{-value} \leq \alpha$
<b>Rejection Rule: Critical Value Approach</b>	Reject $H_0$ if $z \leq -z_\alpha$	Reject $H_0$ if $z \geq z_\alpha$	Reject $H_0$ if $z \leq -z_{\alpha/2}$ or if $z \geq z_{\alpha/2}$

# INFERRENTIAL STATISTICS

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
<b>Hypotheses</b>	$H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$
<b>Test Statistic</b>	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
<b>Rejection Rule: <i>p</i>-Value Approach</b>	Reject $H_0$ if <i>p</i> -value $\leq \alpha$	Reject $H_0$ if <i>p</i> -value $\leq \alpha$	Reject $H_0$ if <i>p</i> -value $\leq \alpha$
<b>Rejection Rule: Critical Value Approach</b>	Reject $H_0$ if $t \leq -t_\alpha$	Reject $H_0$ if $t \geq t_\alpha$	Reject $H_0$ if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

# INFERRENTIAL STATISTICS

**Test Statistic for Hypothesis Tests About a Population Mean:  $\sigma$  Known**

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

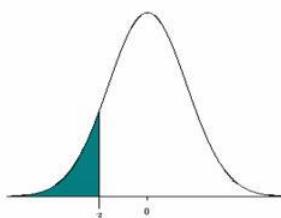
**Test Statistic for Hypothesis Tests About a Population Mean:  $\sigma$  Unknown**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

**Test Statistic for Hypothesis Tests About a Population Proportion**

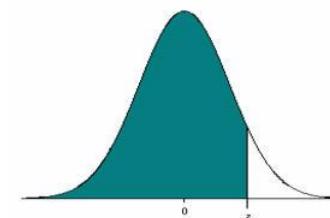
$$z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

**Table of Standard Normal Probabilities for Negative Z-scores**



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003	
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005	
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0007	0.0007	
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0010	0.0010	
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

**Table of Standard Normal Probabilities for Positive Z-scores**



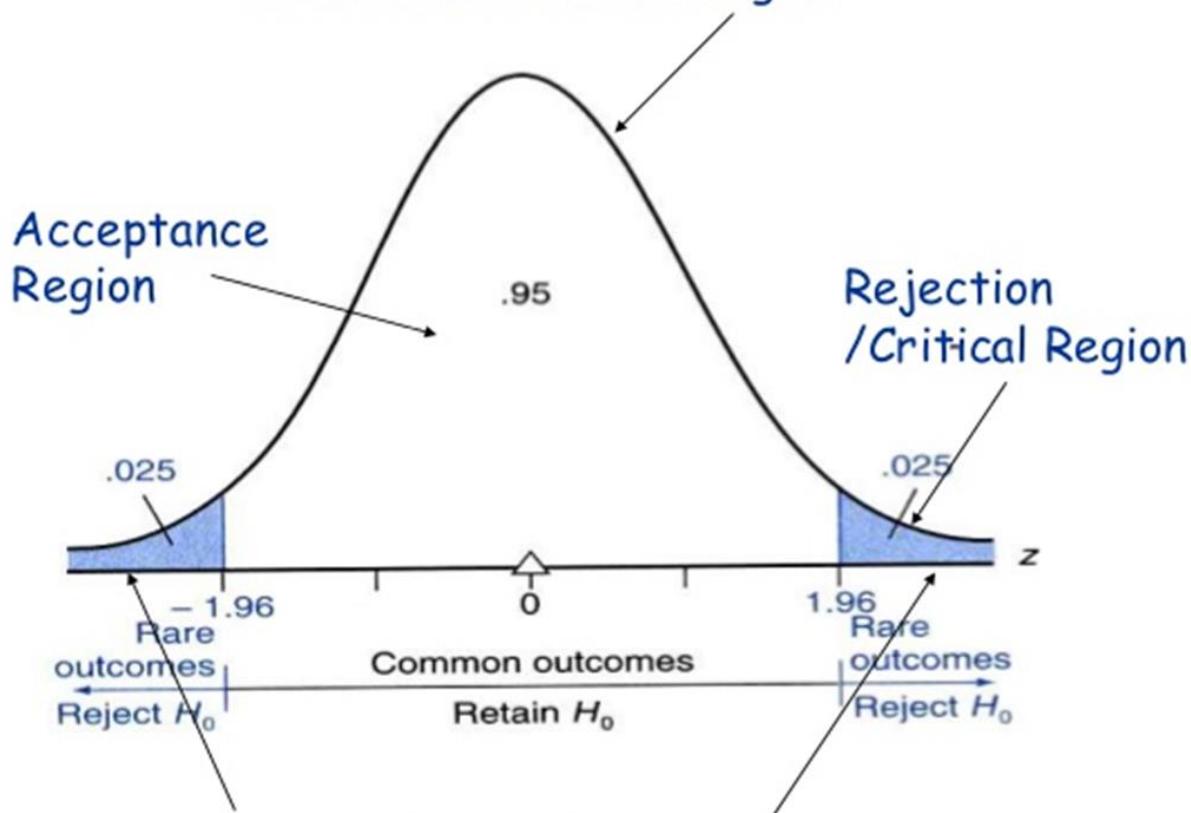
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9986	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Note that the probabilities given in this table represent the area to the LEFT of the z-score.

The area to the RIGHT of a z-score = 1 – the area to the LEFT of the z-score

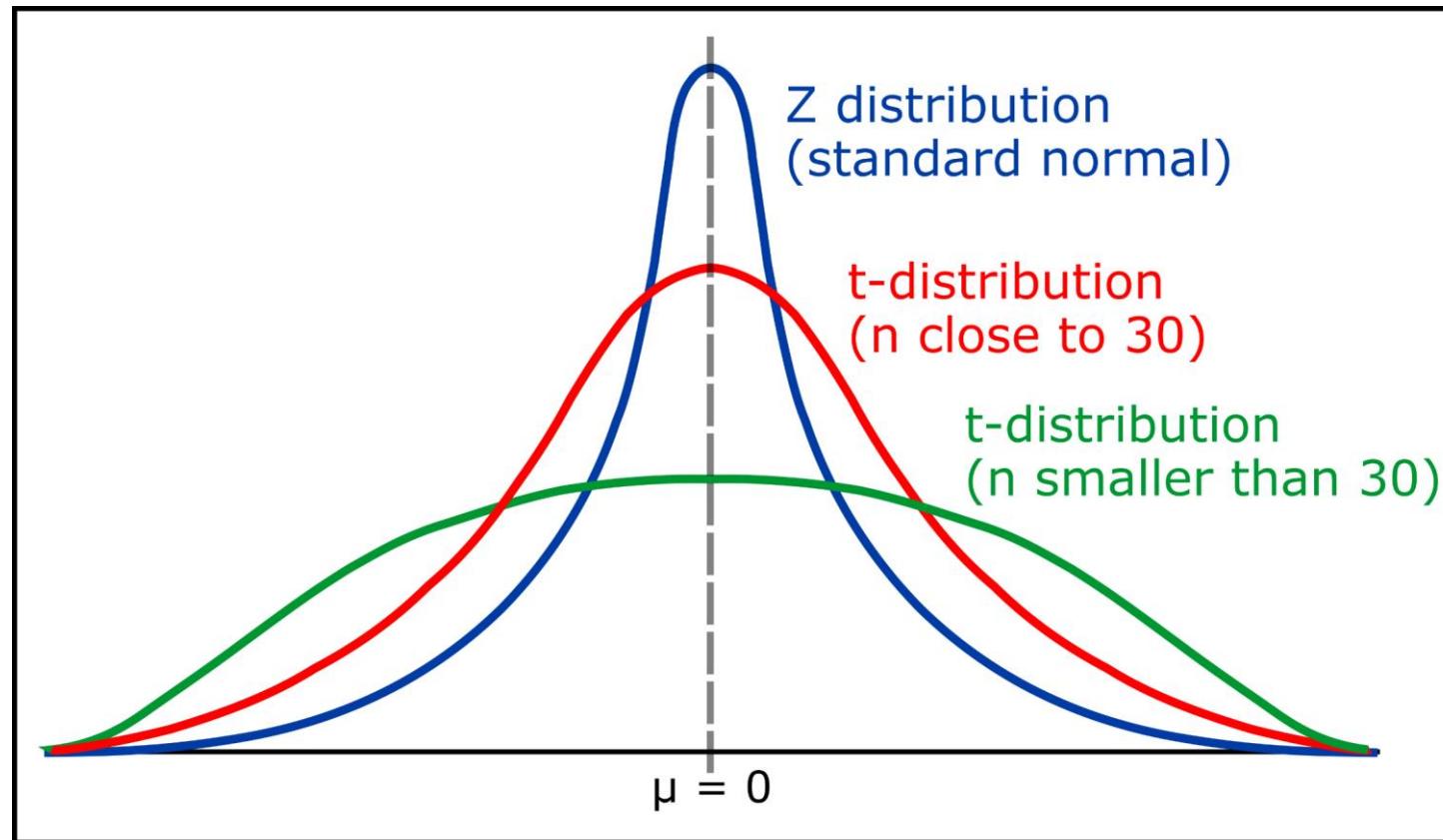
# INFERENTIAL STATISTICS

Accept the null hypothesis if the sample statistic falls in this region

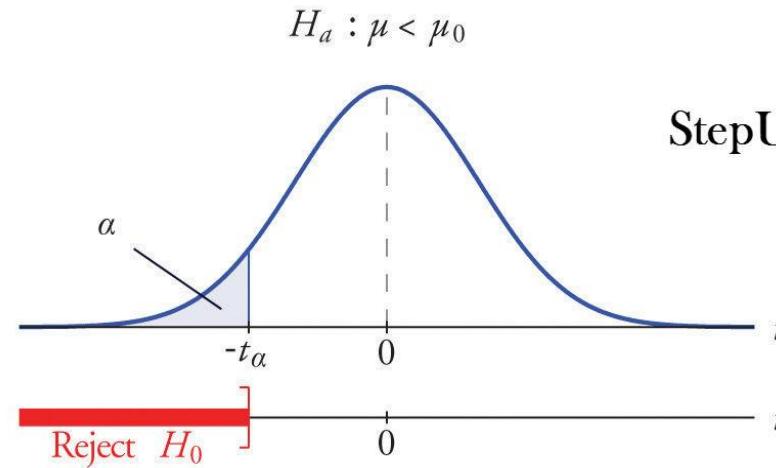


Reject the null hypothesis if the sample statistic falls in these two regions.

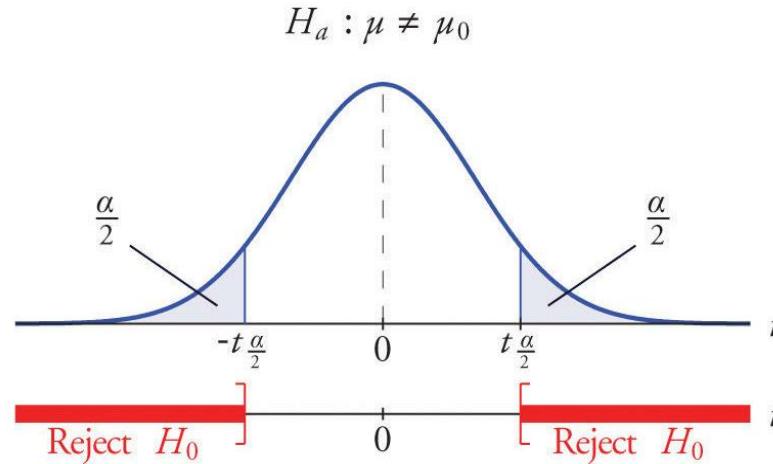
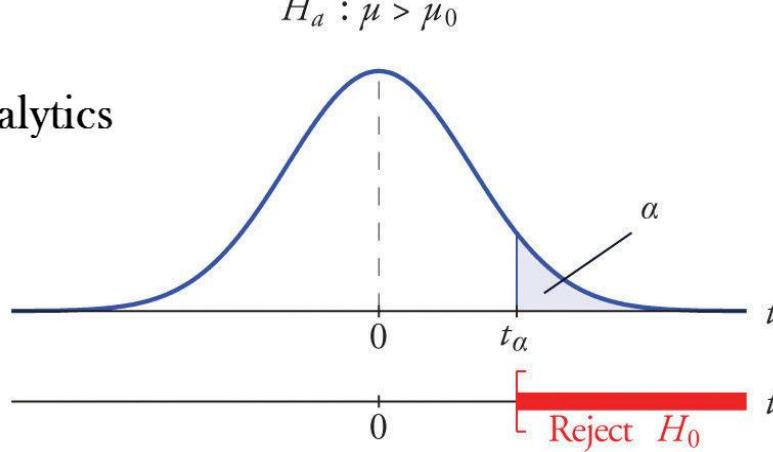
# INFERENTIAL STATISTICS



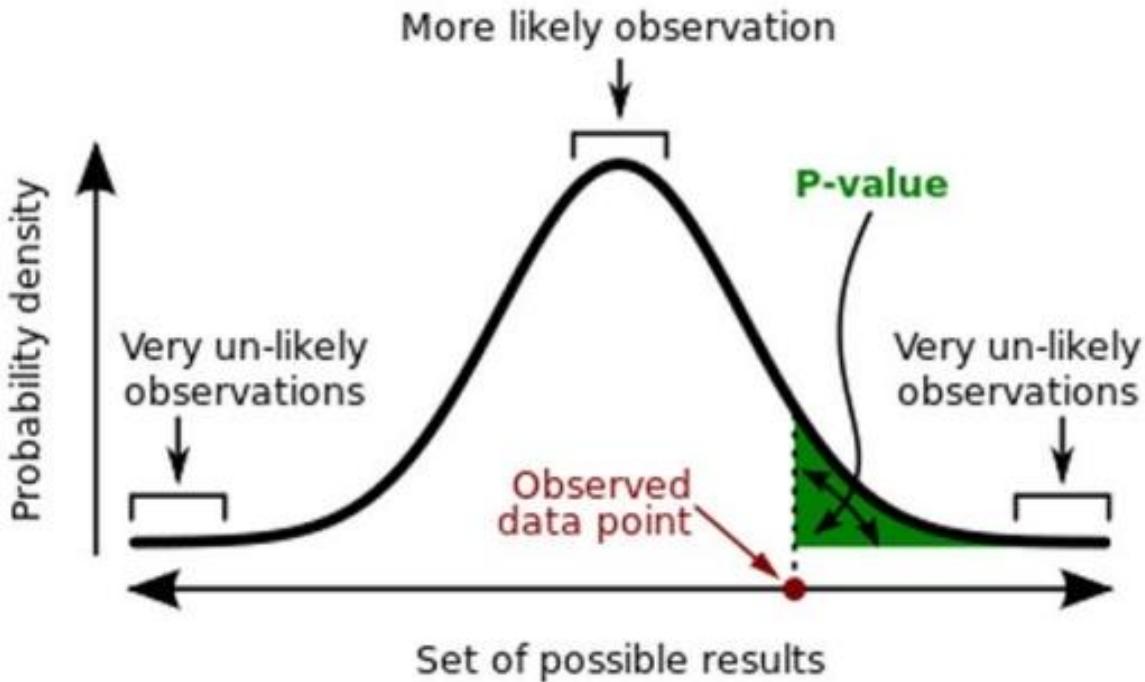
# INFERRENTIAL STATISTICS



StepUp Analytics



# INFERENTIAL STATISTICS



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

If  $p\text{-value} < \alpha$ , we **reject**  $H_0$

จำนวนและร้อยละปัจจัยส่วนบุคคลของนักศึกษา จำแนกตามเพศ

เพศ	จำนวน(คน)	ร้อยละ
ชาย	๕๖	๓๖.๐
หญิง	๑๔๑	๖๔.๐
รวม	๒๙๗	๑๐๐.๐

จำนวนและร้อยละปัจจัยส่วนบุคคลของนักศึกษา จำแนกตามอายุ

อายุ	จำนวน (คน)	ร้อยละ
น้อยกว่า ๒๐ ปี	๖	๒.๗
๒๑-๒๕ ปี	๖๖	๒๘.๗
๒๖-๓๐ ปี	๖๘	๒๔.๕
๓๑-๓๕ ปี	๙๐	๓๓.๗
๓๖ ปีขึ้นไป	๓๗	๑๓.๙
รวม	๒๙๗	๑๐๐.๐

ค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐานและระดับความพึงพอใจในการใช้บริการเทคโนโลยีสารสนเทศของนักศึกษาโดยรวม และด้านที่ส่งผลต่อความพึงพอใจ

ปัจจัยที่เกี่ยวข้อง	$\bar{X}$	SD	ระดับความพึงพอใจ
๑. ด้านระบบการให้บริการ	๓.๗๕	.๔๐๕	มาก
๒. ด้านขั้นตอนการให้บริการ	๓.๘๙	.๖๕๐	มาก
๓. ด้านบุคลากรผู้ให้บริการ	๔.๓๐	.๔๙๙	มาก
รวม	๓.๙๙	๐.๓๕	มาก

การเปรียบเทียบความแตกต่างระหว่างปัจจัยส่วนบุคคลกับความพึงพอใจในการใช้บริการ  
เทคโนโลยีสารสนเทศด้านบุคลากรผู้ให้บริการ จำแนกตามเพศ

เพศ	ด้านบุคลากรผู้ให้บริการ		ค่า t	ค่า Sig.
	X	SD		
ชาย	๔.๓๐	.๕๔๗	.๘๗๔	.๐๓๙
หญิง	๔.๒๙	.๖๒๗		

\* นัยสำคัญทางสถิติ 0.05

การเปรียบเทียบความแตกต่างระหว่างปัจจัยส่วนบุคคลกับความพึงพอใจในการใช้บริการ  
เทคโนโลยีสารสนเทศด้านระบบการให้บริการ จำแนกตามอายุ

อายุ	ด้านระบบการให้บริการ		ค่า	ค่า
	$\bar{X}$	$SD$		
น้อยกว่า ๒๐ ปี	๓.๙๑	.๕๖๑		
๒๑-๒๔ ปี	๓.๔๙	.๕๔๘		
๒๖-๓๐ ปี	๓.๙๘	.๔๕๙	๙.๔๘๔	.๐๐๐
๓๑-๓๕ ปี	๓.๙๐	.๔๙๙		
๓๖ ปี ขึ้นไป	๓.๖๙	.๒๗๓		
รวม	๓.๙๔	.๕๐๔		

\* นัยสำคัญทางสถิติ ๐.๐๕

# ONE GROUP TESTING

# 1-Sample z-test

ตัวอย่าง 1 นักเรียนคนหนึ่งทำข้อสอบแบบการถูกผิดทั้งหมด 20 ข้อ และตอบถูกทั้งหมด 14 ข้อ อยากรู้ว่าโอกาสที่นักเรียนคนนี้จะทำข้อสอบข้อหนึ่งถูกมีมากกว่า 50% ใช่หรือไม่

Step 1: State the Hypothesis

$$H_0: p = 0.5$$
$$H_A: p > 0.5$$

Step 2: Choose 95% confidence level

$$\alpha = 0.95$$

# WORKED EXAMPLE 1

Step 2: Calculate the test statistics

We will use the Z-statistics for the test:

Proportion test:

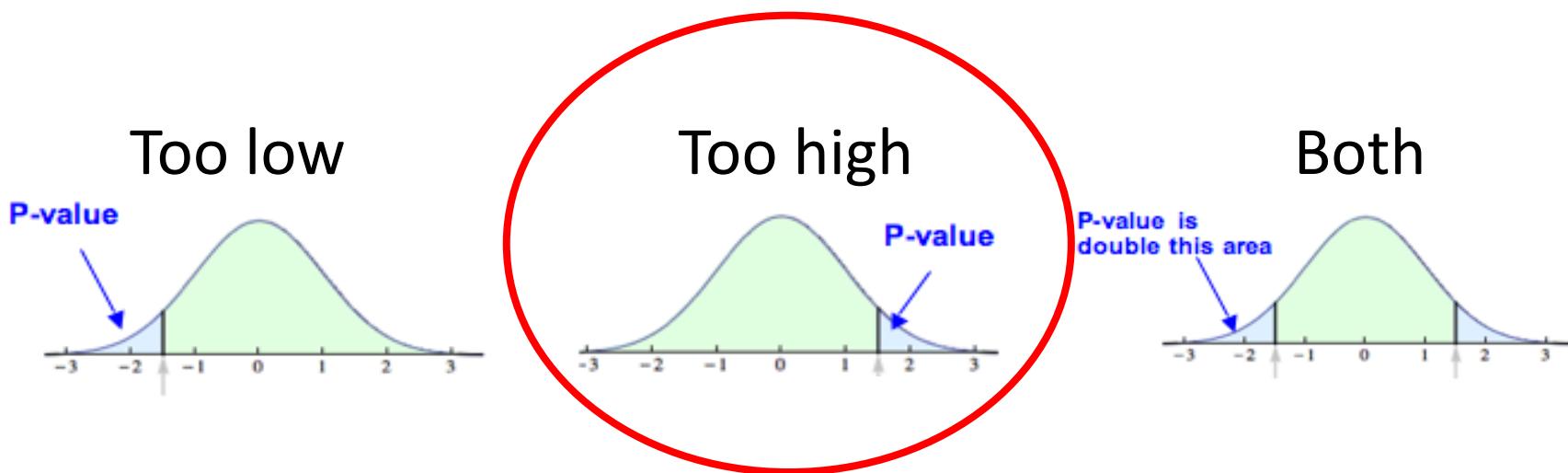
$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = 1.78$$

Notation	Meaning
$\hat{p}$	Sample proportion
$p_0$	Proportion in null hypothesis
$q_0$	$1 - p_0$
$n$	Sample size

# WORKED EXAMPLE 1

## Step 3: Calculate p-value

Referring to our hypothesis, we need the sample proportion to be too high to reject the null hypothesis



### Python Code

```
[1] from scipy.stats import norm  
norm.sf(1.78)  
0.03753798034851679
```

### P-value:

$$P(Z > 1.78) = \underline{0.038}$$



# WORKED EXAMPLE 2

ตัวอย่าง 2 คนไทยอ่านหนังสือน้อยกว่าวันละ 8 บรรทัดจริงหรือไม่?

เราทำการสำรวจตัวอย่างทั้งหมด 10 คน โดยในกลุ่มตัวอย่างนี้พบว่ามีค่าเฉลี่ยการอ่านหนังสืออยู่ที่วันละ 9 บรรทัด และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 4 บรรทัด อยากร้าบว่า ข้อมูลนี้เพียงพอต่อการโต้แย้งข้อความข้างต้นหรือไม่?

Step 1: State Hypothesis

$$H_0: \mu = 8$$

$$H_A: \mu > 8$$

Step 2: Use 95% Confidence Level

$$\alpha = 0.05$$

# WORKED EXAMPLE 2

ตัวอย่าง 2 คนไทยอ่านหนังสือน้อยกว่าวันละ 8 บรรทัดจริงหรือไม่?

Recap:  $n = 10$ ,  $\mu_0 = 8$ ,  $\bar{x} = 9$  และ  $s = 4$

Step 3: Calculate Test Statistics

$$t = \frac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{9 - 8}{\left(\frac{4}{\sqrt{10}}\right)} = 0.79$$

# WORKED EXAMPLE 2

ตัวอย่าง 2 คนไทยอ่านหนังสือน้อยกว่าวันละ 8 บรรทัดจริงหรือไม่?

Recap:  $n = 10$ ,  $\mu_0 = 8$ ,  $\bar{x} = 9$  และ  $s = 4$

Step 4: Compute the p-value

$$p\text{-value} = P(t_{10-1} > 0.79) = 0.22$$

```
from scipy.stats import t  
t.sf(0.79, 10-1, loc=0, scale=1)
```

```
0.22492022494259561
```

# WORKED EXAMPLE 2

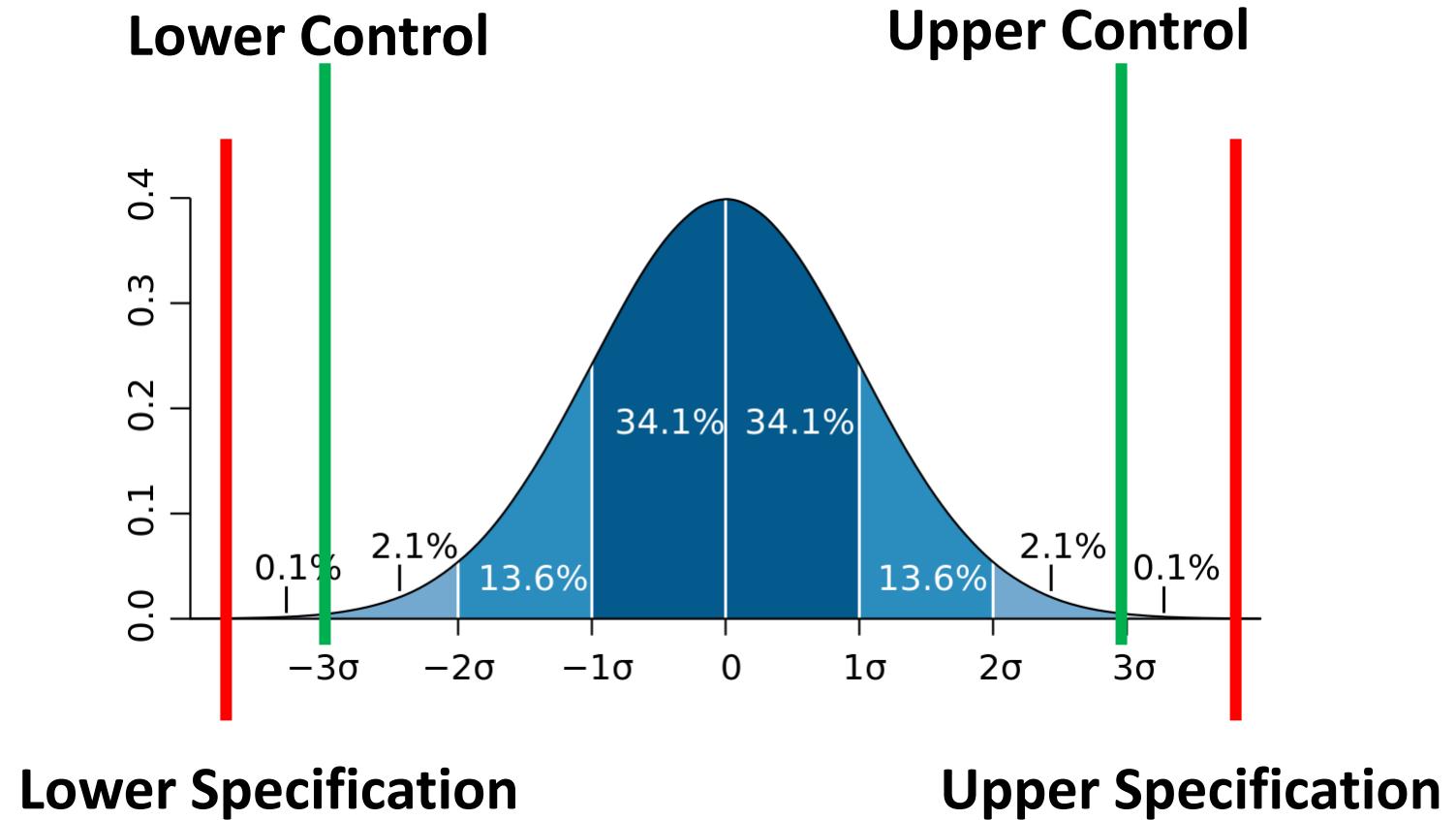
ตัวอย่าง 2 คนไทยอ่านหนังสือน้อยกว่าวันละ 8 บรรทัดจริงหรือไม่?

Step 5: Conclusion

$$p - value = 0.22 > 0.05$$

ข้อสรุป จากการสำรวจครั้งนี้พบว่าข้อมูลในการสำรวจของเรามีไม่เพียงพอที่จะโต้แย้ง null hypothesis ในกรณีนี้เราอาจต้องเพิ่ม sample size เพื่อทำการทดสอบเพิ่มเติม

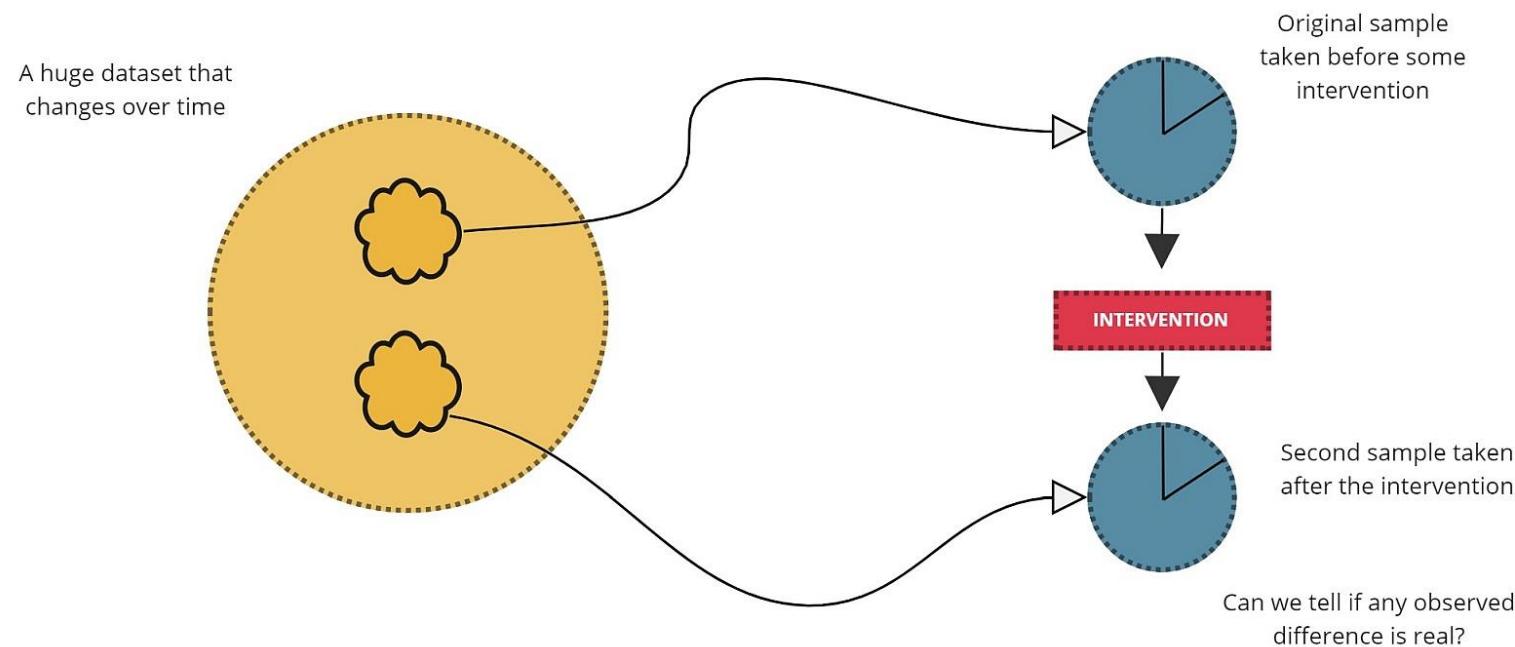
# Example: Quality Control



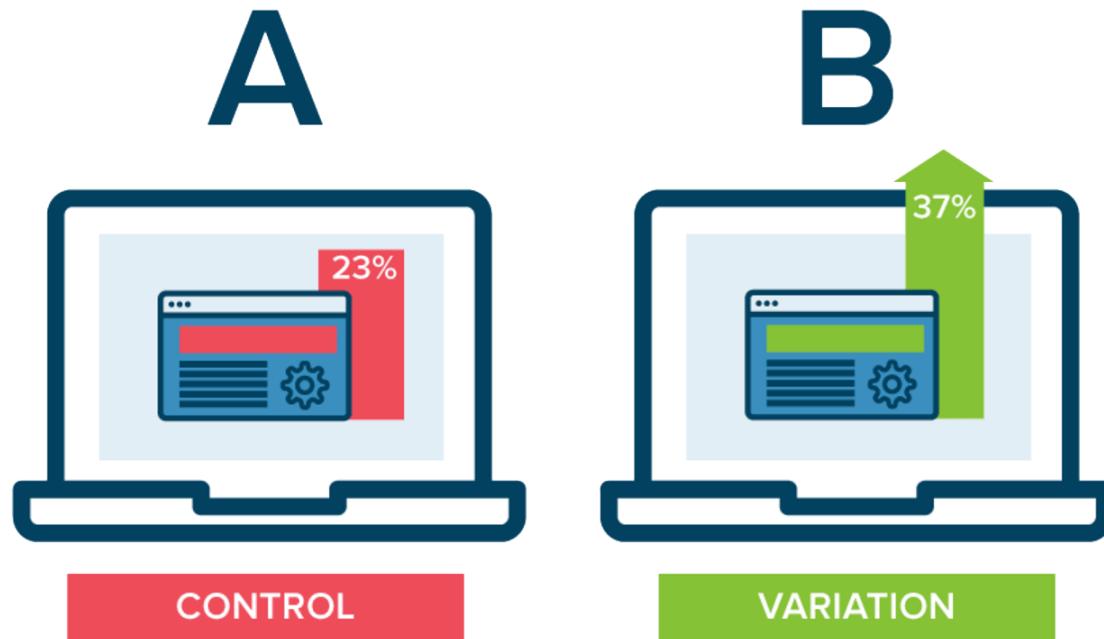
# TWO GROUPS TESTING

# Why 2 groups testing?

**Objective:** Compare 2 groups of samples that are given different “treatments” and see if the difference is statistically significant



# 2-Sample Z-test



## A/B Testing:

Compare 2 groups of sample with binary response

- Compare purchase rates between 2 web designs

$$H_0: p_A - p_B = 0$$
$$H_a: p_A - p_B \neq 0$$

# 2-Sample Z-test: Python Code



```
import pandas as pd
import numpy as np
from scipy import stats
from statsmodels.stats import weightstats as stests

# randomly generating 2 sample groups
# x1 is the data from design A
# x_2 is the data from design B
x1 = np.random.choice([0,1], 100,p=[0.3,0.7])
x2 = np.random.choice([0,1], 50, p=[0.5,0.5])

print("p-value is equal to "
+str(stests.ztest(x1, x2,value=0,alternative='two-sided')[1]))
```

```
⌚ p-value is equal to 5.3368864253269504e-05
```

# 2-Sample Mean Tests

## Paired Observation:

The individuals between the 2 groups **can be related** e.g.

- A person's height measured at different time
- Father's height vs Son's Height

## Unpaired Observation:

The individuals between two groups **are independent** e.g.

- Blood sugar levels of 2 different sample group following different diet plans

$$\begin{aligned} H_0: \mu_A - \mu_B &= 0 \\ H_a: \mu_A - \mu_B &\neq 0 \end{aligned}$$

# 2-Sample Mean Tests: Python Codes

Paired

```
from scipy.stats import ttest_rel
#Randomly generate fathers' and sons' heights
x1 = np.random.normal(170,5,size=10)
x2 = np.random.normal(175,10,size=10)
#Perform 2-sample paired tests
ttest_rel(x1,x2)

Ttest_relResult(statistic=-1.3077364017455502, pvalue=0.22337870651072805)
```

Unpaired

```
from scipy.stats import ttest_ind
#Randomly generate blood sugar of 2 groups
x1 = np.random.normal(90,5,size=100)
x2 = np.random.normal(95,10,size=50)
#Perform 2-sample independent tests
ttest_ind(x1,x2,equal_var=True)

Ttest_indResult(statistic=-2.302565741882167, pvalue=0.022697127461132692)
```

# FURTHER READINGS:

More details about hypothesis testing including sample Python codes can be found on the link below

[An Introduction to Hypothesis Testing](#)



# Inferential Statistics: Techniques and Types of Calculation:



## Linear Regression Analyses

Linear regression is a statistical method for studying relationships between one or more independent variables ( $X$ ) and one dependent variable ( $Y$ ).

## Logistic Regression Analysis

Logistic regression is conducted when the dependent variable is dichotomous (i.e. the dependent variable has only two possible values). Examples of dichotomous (binary) variables are: 0 and 1, Yes and No.

## Analysis of Variance (ANOVA)

A statistical method used to test and analyze differences between two or more means (averages of 2 or more groups). It searches significant differences between means.

## Analysis of Covariance (ANCOVA)

ANCOVA blends ANOVA and regression.

When a continuous covariate is included in an ANOVA we have ANCOVA (just to remind that a covariate is a continuous independent variable).

## Statistical Significance (T-Test)

The t-test compares two means (averages of 2 groups) and tells us if they are different from each other. The t-test also tells us how significant the differences are.

## Correlation Analysis

Correlation analysis studies the strength of a relationship between two variables. It shows that two or more variables have a strong correlation or a weak correlation.

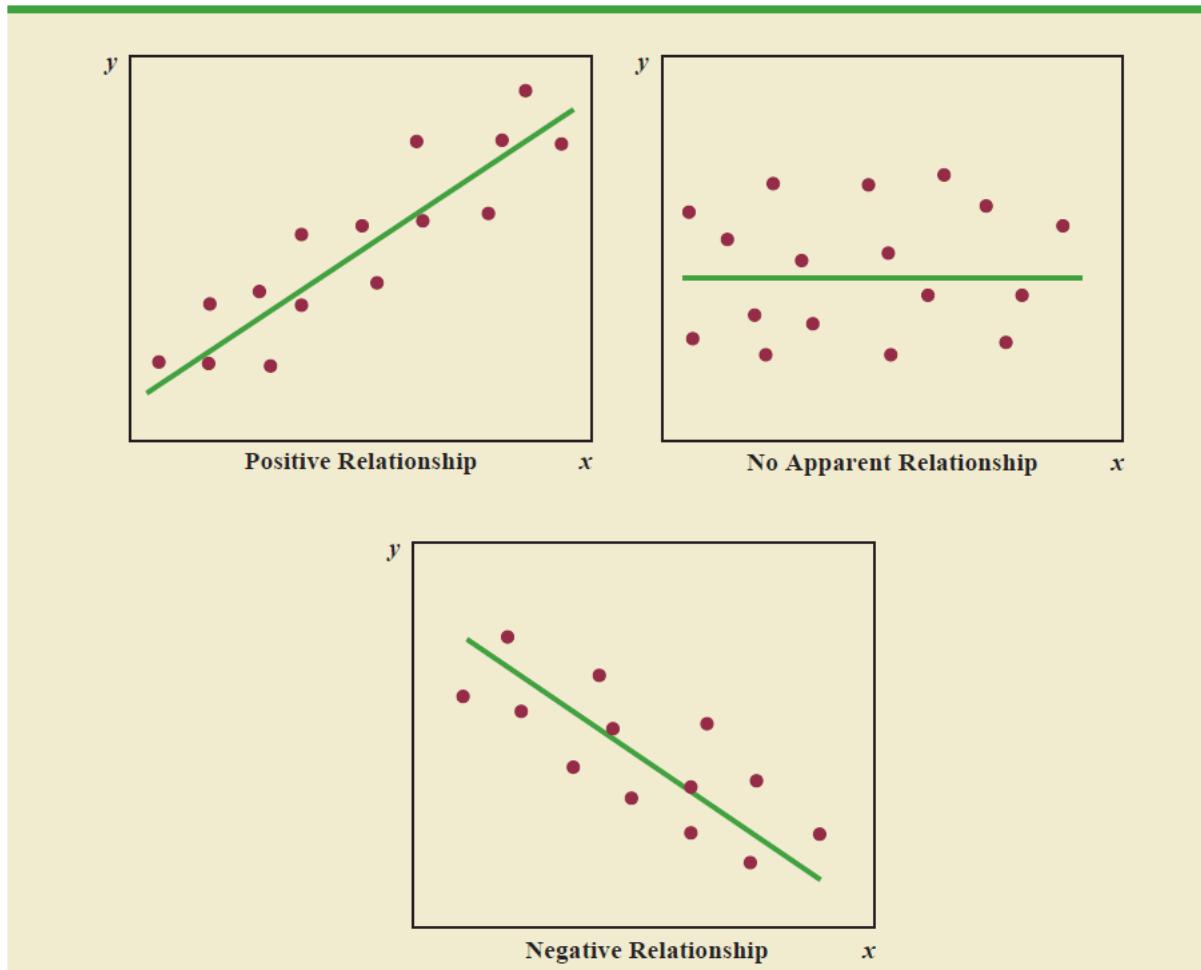
## Other common techniques:

Structural equation modeling, Survival analysis, Factor analysis, Multidimensional scaling, Cluster analysis, Discriminant function.



<http://intellspot.com>

# MEASURES BETWEEN TWO VARIABLES

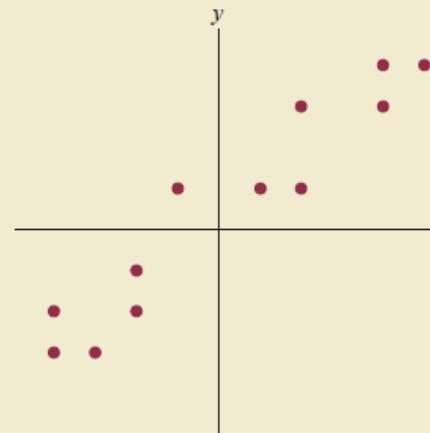


# MEASURES BETWEEN TWO VARIABLES

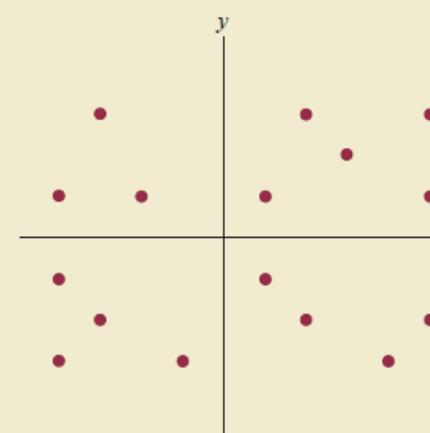
## SAMPLE COVARIANCE

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

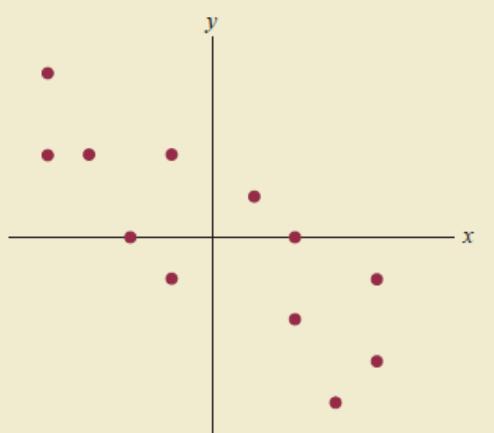
$s_{xy}$  Positive:  
( $x$  and  $y$  are positively linearly related)



$s_{xy}$  Approximately 0:  
( $x$  and  $y$  are not linearly related)



$s_{xy}$  Negative:  
( $x$  and  $y$  are negatively linearly related)



# CORRELATION ANALYSIS

## CORRELATION COEFFICIENT

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where

$r_{xy}$  = sample correlation coefficient

$s_{xy}$  = sample covariance

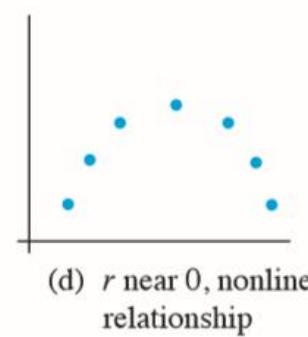
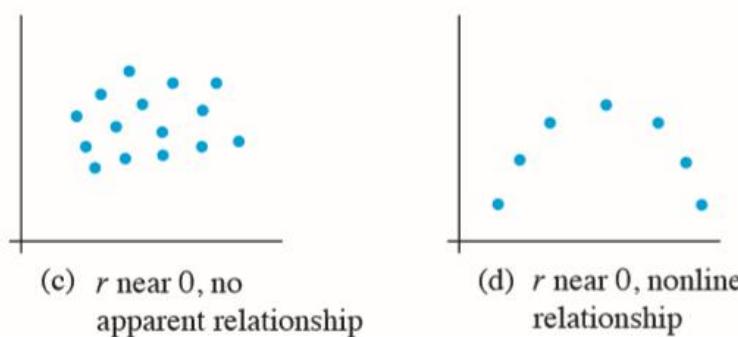
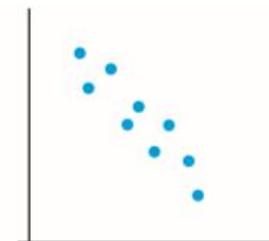
$s_x$  = sample standard deviation of  $x$

$s_y$  = sample standard deviation of  $y$

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

- ◆ Values in [-1, 1]
- ◆ Can be misleading for quadratic or non-linear relationship

# CORRELATION ANALYSIS

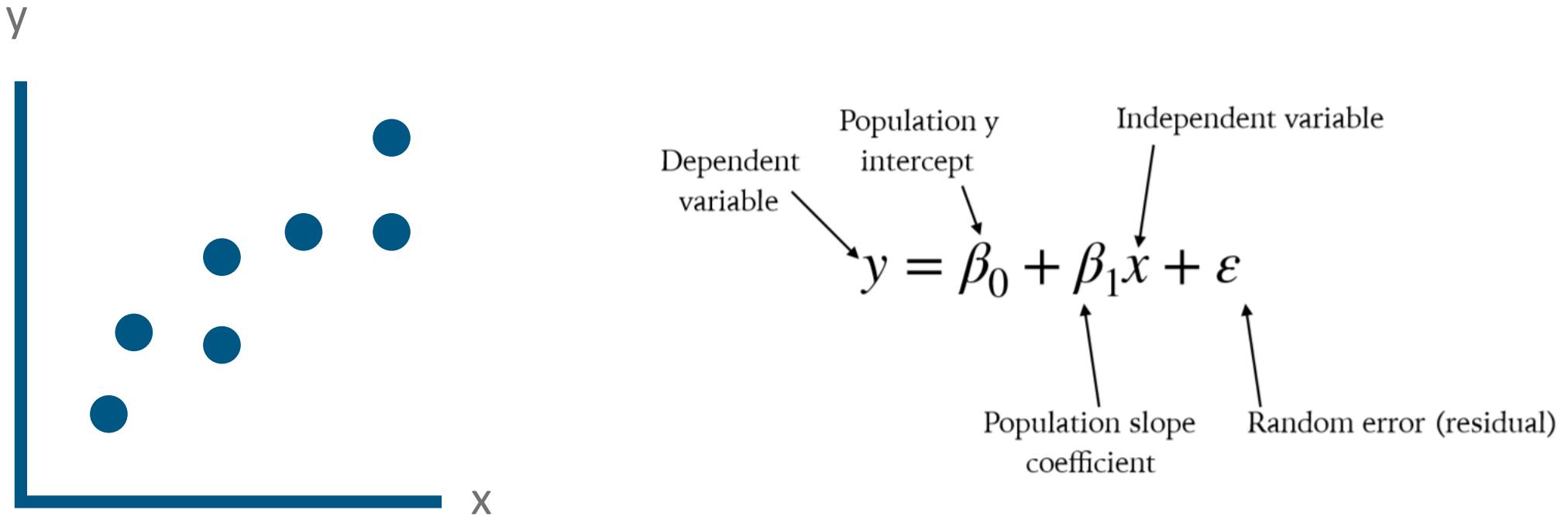


**Weak**  
 $-.5 \leq r \leq .5$

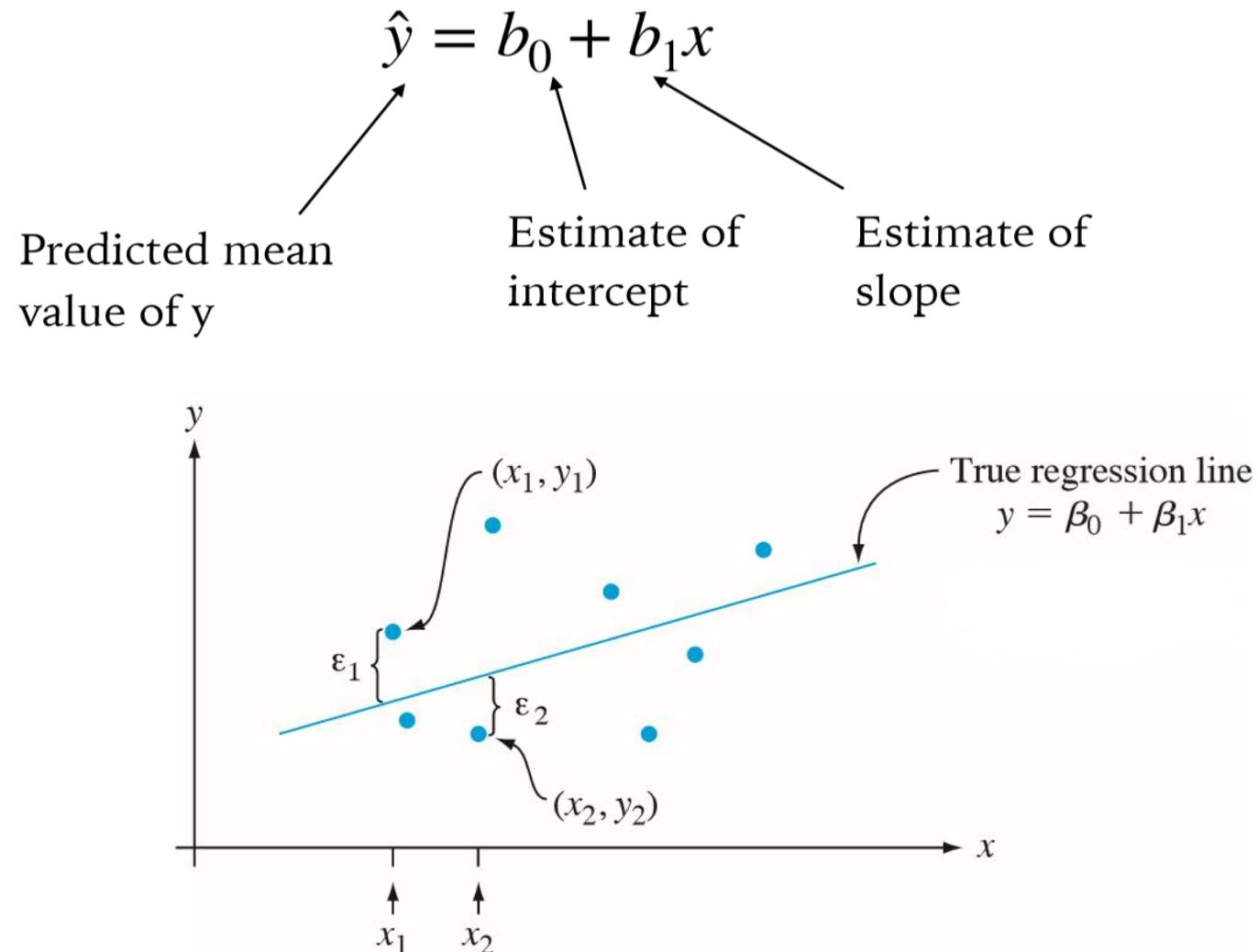
**Moderate**  
either  $-.8 < r < -.5$  or  $.5 < r < .8$

**Strong**  
either  $r \geq .8$  or  $r \leq -.8$

# SIMPLE LINEAR REGRESSION MODEL



# ESTIMATED REGRESSION MODEL



# ERROR

$$MSE = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}$$

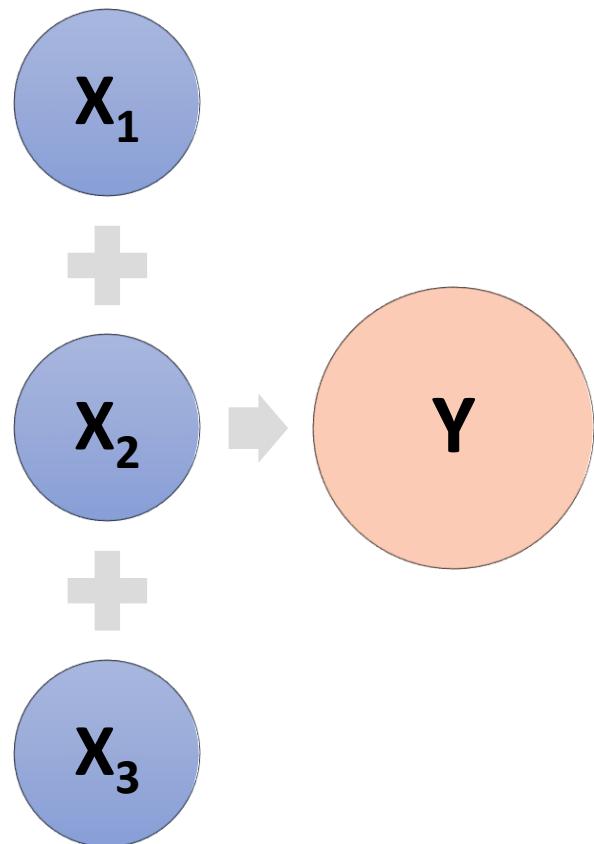
$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \times 100$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}}$$

# COEFFICIENT OF DETERMINATION

$$R^2 = \frac{\left( \sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y} \right)^2}{\left( \sum_{i=1}^n X_i^2 - n \bar{X}^2 \right) \left( \sum_{i=1}^n Y_i^2 - n \bar{Y}^2 \right)}$$

# MULTIPLE LINEAR REGRESSION MODEL



Dependent Variable

Intercept Value

First Independent Variable

Second Independent Variable

K-th Independent Variable

Coefficients/Weights

Error Term

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

# MULTIPLE LINEAR REGRESSION MODEL

Our Model  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

1. Generate Data Points

```
import numpy as np
import pandas as pd
X1 = np.random.uniform(0,1,1000)
X2 = np.random.uniform(0,1,1000)
X3 = np.random.uniform(0,1,1000)
X = pd.DataFrame({'X1':X1, 'X2':X2, 'X3':X3})
noise = (X1+0.2)*(X1+0.2)*np.random.normal(0,1,1000)
Y = 10*X1 - 2*X2 +0.2*noise
```

2. Fit Model

```
import statsmodels.api as sm
#add beta 0
X= sm.add_constant(X)
model = sm.OLS(Y,X).fit()
```

# MULTIPLE LINEAR REGRESSION MODEL

## 3. Statistical Significance

```
| model.summary()
```

OLS Regression Results

Dep. Variable:	y	R-squared:	0.998	
Model:	OLS	Adj. R-squared:	0.998	
Method:	Least Squares	F-statistic:	1.411e+05	
Date:	Tue, 24 May 2022	Prob (F-statistic):	0.00	
Time:	07:47:46	Log-Likelihood:	-550.12	
No. Observations:	1000	AIC:	-1092.	
Df Residuals:	996	BIC:	-1073.	
Df Model:	3			
Covariance Type:	nonrobust			
coef	std err	t	P> t	[0.025 0.975]
const	-0.0016	0.014	-0.112	0.010 0.000 0.005
X1	9.9714	0.016	637.690	0.000 9.941 10.002
X2	-1.9831	0.016	-122.264	0.000 -2.072 1.964
X3	-0.0006	0.015	-0.037	0.970 -0.031 0.029
Omnibus:	81.029	Durbin-Watson:	2.109	
Prob(Omnibus):	0.000	Jarque-Bera (JB):	411.463	
Skew:	0.115	Prob(JB):	4.49e-90	
Kurtosis:	6.134	Cond. No.	6.08	

F-test:

$H_0: Y = \text{Constant}$

$H_1: Y = \text{Model Prediction}$

t-test:

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

# MULTIPLE LINEAR REGRESSION MODEL

## 4. Make prediction

1D list input

```
model.predict([1,1,2,20])
```

```
array([5.99232563])
```

```
model.predict(X)
```

0	0.413832
1	2.240266
2	6.018805
3	6.298800
4	3.536102

...

995	2.712025
996	2.873105
997	0.806432
998	4.479468
999	3.167633

Length: 1000, dtype: float64

Data Frame input

# MULTIPLE LINEAR REGRESSION MODEL

5. For comparison with other models, we may use the followings

## Root Mean Square Error

```
prediction=model.predict(X)
np.sqrt(sum((Y-prediction)*(Y-prediction))/1000)
```

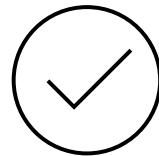
0.13958805804993368

## Mean Absolute Percentage Error

```
prediction=model.predict(X)
mean_percentage_error = 100*np.mean(np.abs(Y-prediction)/Y)
print("Mean Percentage Error is "+str(mean_percentage_error)+"%)
```

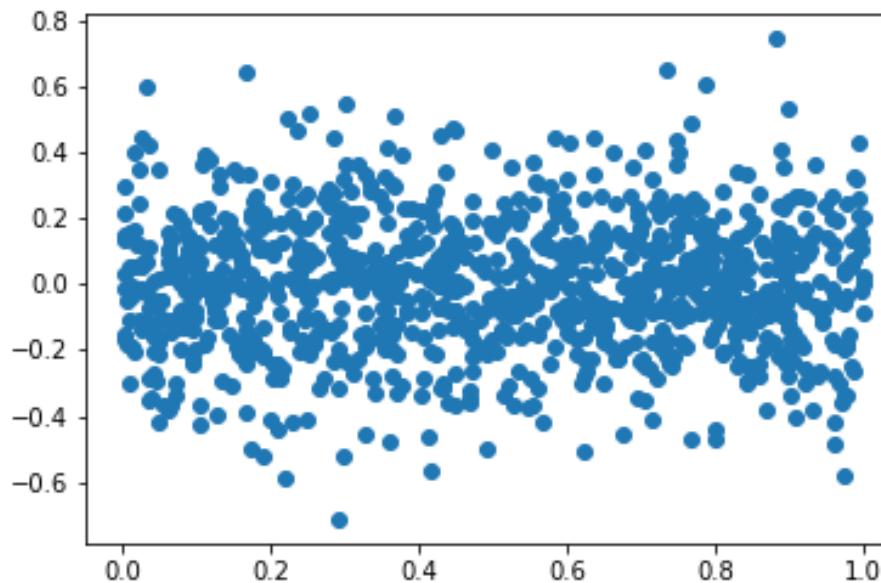
Mean Percentage Error is 8.592069555511923%

# Residuals Plots

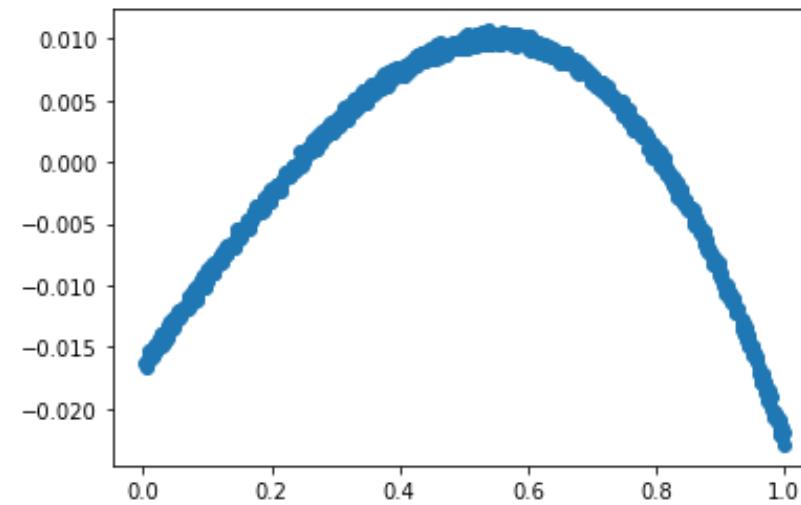
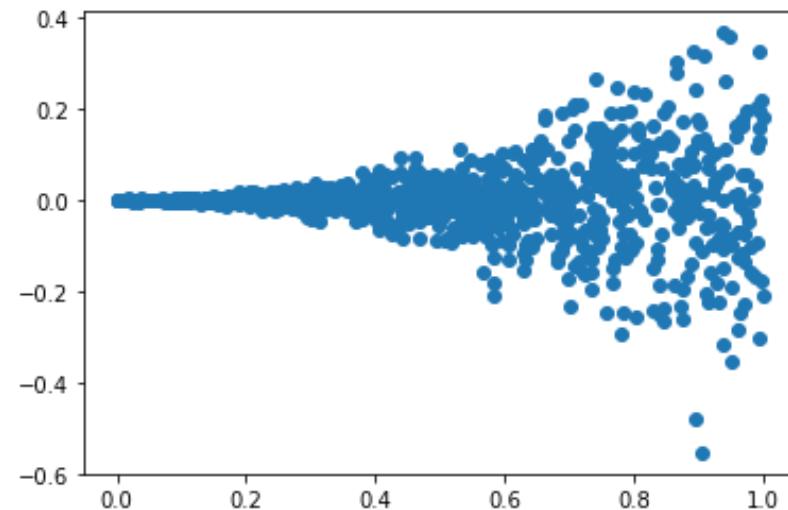
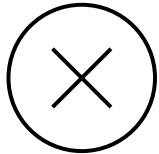


```
import matplotlib.pyplot as plt  
residuals = Y - model.predict(X)  
plt.scatter(X1, residuals)
```

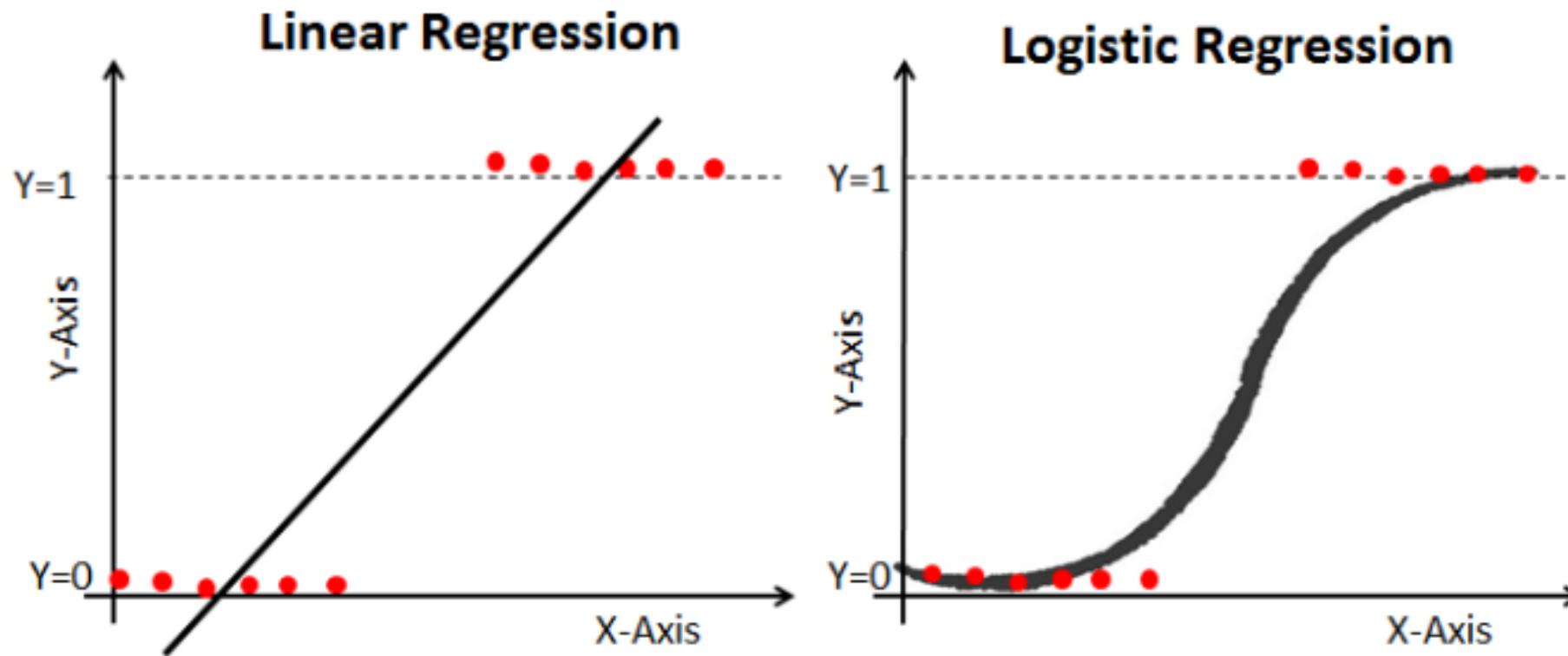
```
<matplotlib.collections.PathCollection at 0x7fec84607b10>
```



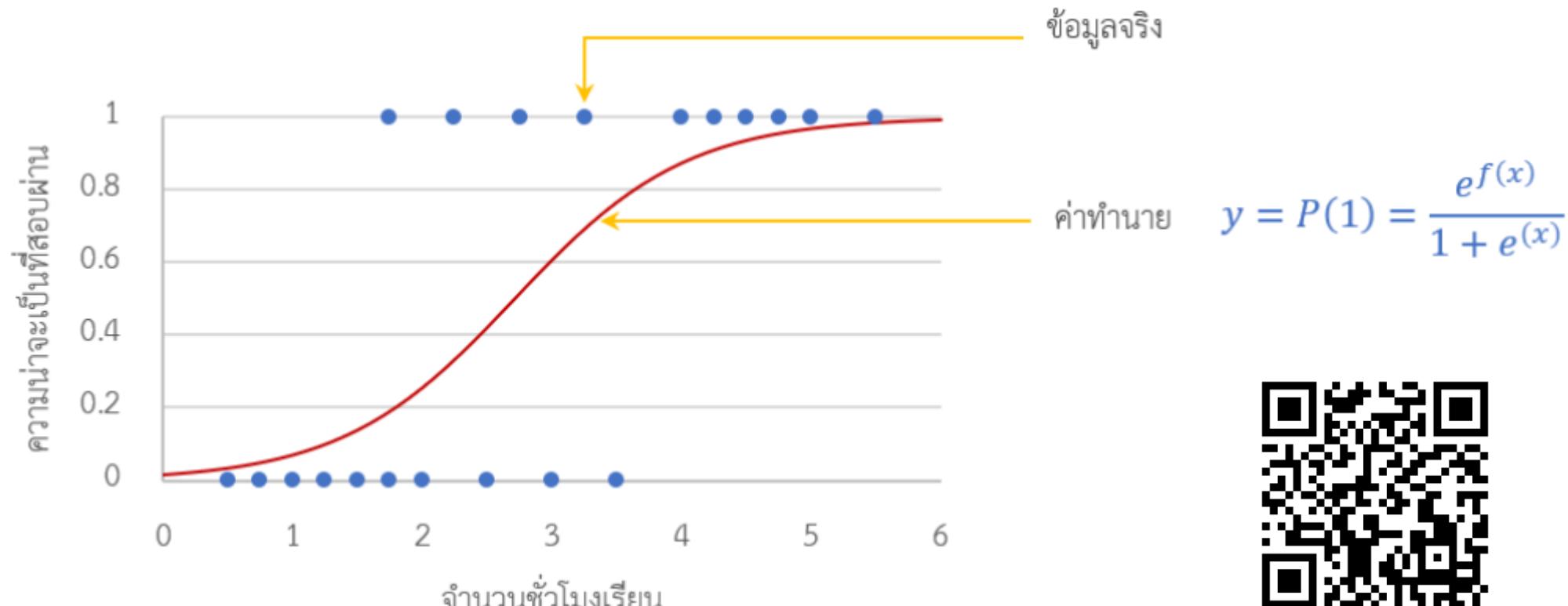
# Residuals Plots



# LOGISTIC REGRESSION MODEL

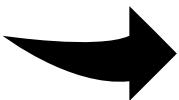
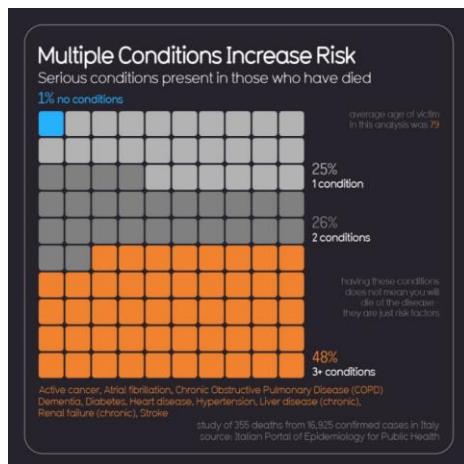
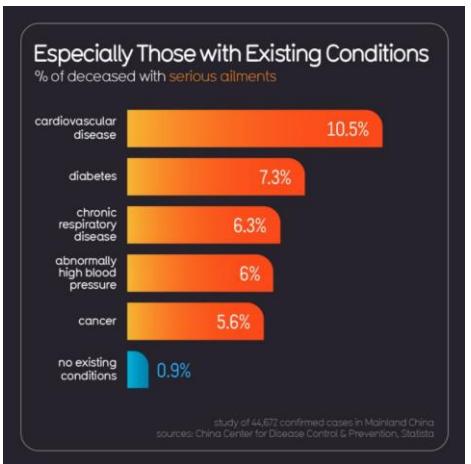
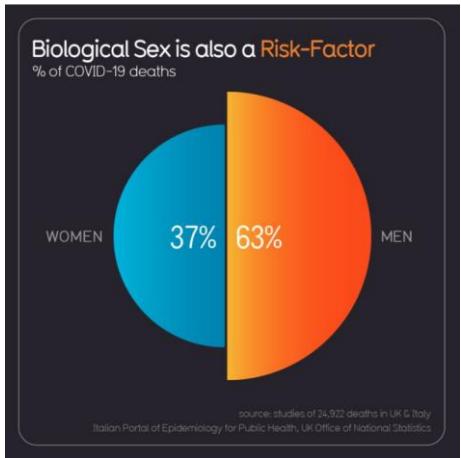
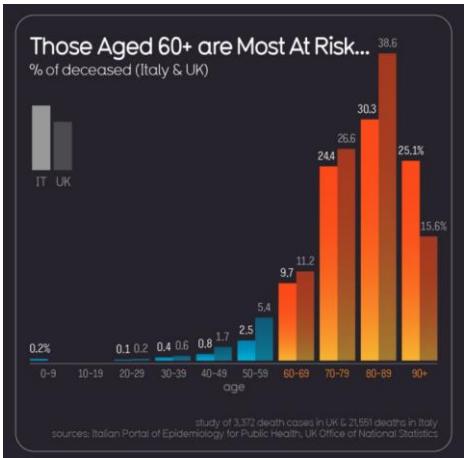


# LOGISTIC REGRESSION MODEL



[Logistic Regression ด้วย Microsoft Excel](#)

# APPLICATION



**You are at risk**

You have an estimated **61.50%** chance of **dying** from covid-19 if infected

Please note this is just an estimation, and not an absolute assessment of the effects covid-19 might have on you.

**Age group**

- 1 - 19
- 20 - 39
- 40 - 59
- 60 - 79**
- 80 - 100+

**Sex**

- Male**
- Female**

**Do you have any of these conditions?**

- No Cardiovascular disease
- Yes Diabetes
- No Chronic respiratory disease
- Yes Hypertension
- No Cancer

**Check**

**Link**



3.64

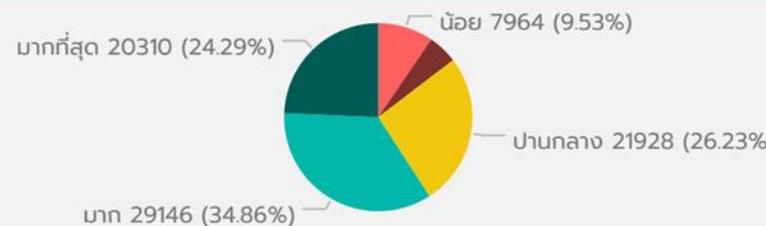
สามารถทำได้

83,611

ไม่สามารถทำได้

2,574

ก้าวใช้ภาษาไทยในการพัฒนาอ่านและเขียนได้อย่างดี



2.41

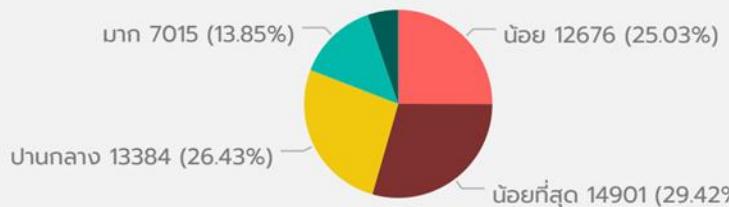
สามารถทำได้

50,647

ไม่สามารถทำได้

34,050

ก้าวใช้อินเทอร์เน็ตในการสืบค้นข้อมูลได้อย่างดี



2.21

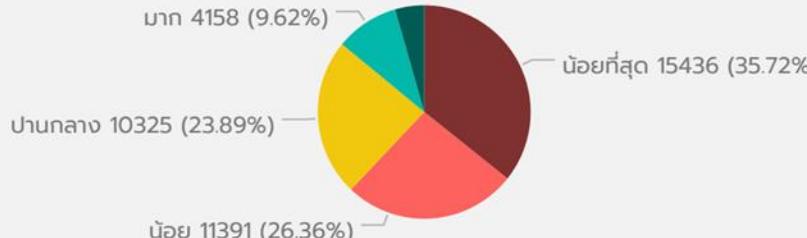
สามารถทำได้

43,219

ไม่สามารถทำได้

40,903

ก้าวใช้โปรแกรมแอพพลิเคชันเพื่อเป็นช่องทางในการสร้างรายได้อย่างดี

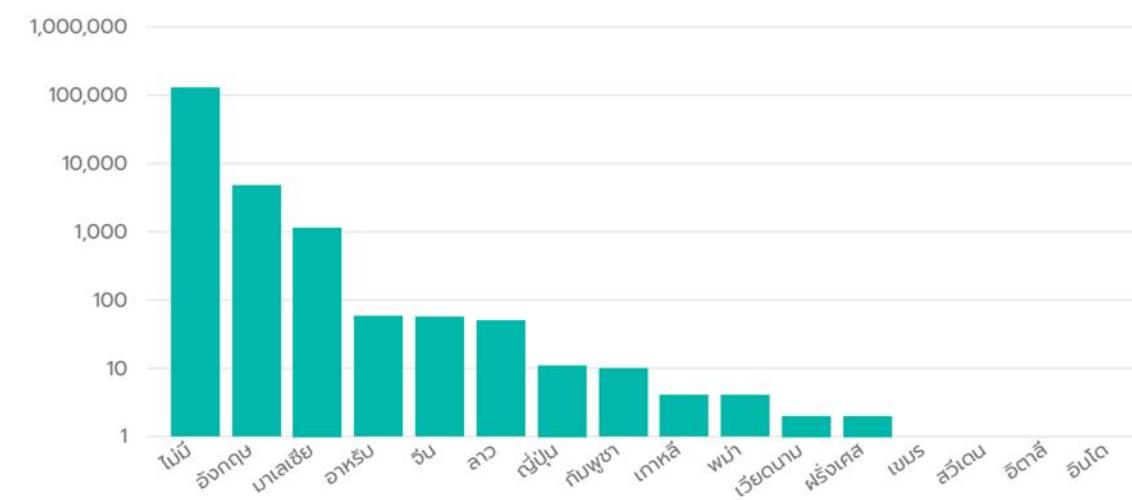
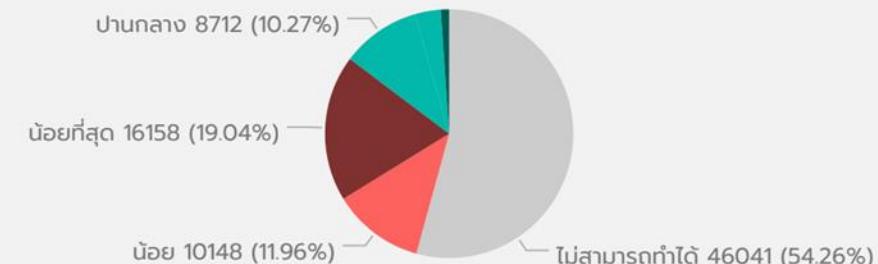


## ความสามารถในการทำงานของงานนอกระบบ

ก้าวสามารถใช้ภาษาต่างประเทศได้หรือไม่

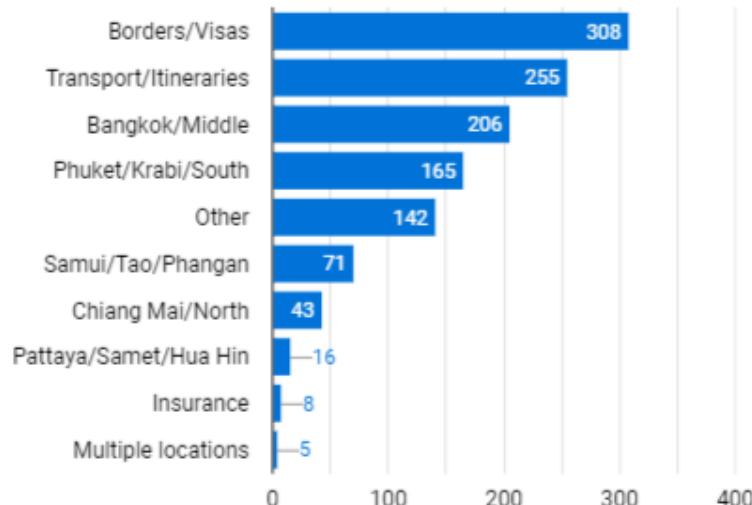


ก้าวใช้ภาษาต่างประเทศเพื่อเป็นภาษาที่สองได้อย่างดี





Tag ยอดนิยม

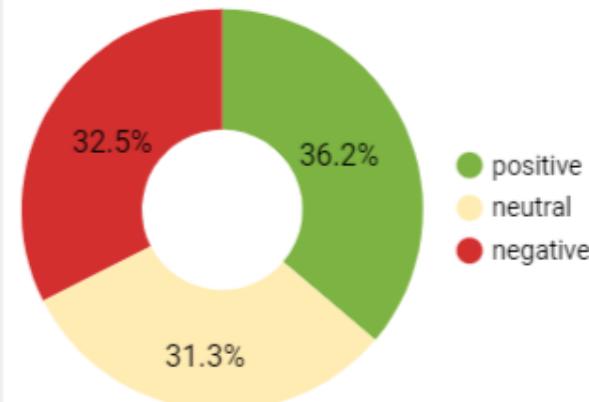


คำที่ถูกพูดถึง

	คำ	ค่าความสำคัญ	% Δ	Sentiment
1.	thailand	121.65	3.1% 	-0.28 
2.	bangkok	86.11	16.9% 	-0.11 
3.	thai	63.86	15.8% 	-0.43 
4.	pass	60.6	25.3% 	-0.62 
5.	phuket	60.21	2.5% 	0.05 

**ค่าความสำคัญ:** ค่าผลรวมคะแนน Term Frequency-Inverse Document Frequency (TF-IDF) ของแต่ละคำในกระทุ้นหน้า

## ความรู้สึก (Sentiment)



## จำนวนหัวข้อกระทุกสนทนา



## รายละเอียดกระทุ้นหนา

วันที่	กระทู้	Tag	ความเห็น	Sentiment
25 พ.ค. 2022	<a href="#">Is it ok if medical insurance just state "unlimited" coverage instead of \$10K for the Th...</a>	Other	0	
25 พ.ค. 2022	<a href="#">Thailand itinerary check</a>	Transport/Itineraries	0	
25 พ.ค. 2022	<a href="#">How crazy is driving.(car/scooter) in Thailand really? In countryside? Nature around C...</a>	Transport/Itineraries	0	
25 พ.ค. 2022	<a href="#">Visa/travel suggestions</a>	Phuket/Krabi/South	0	
25 พ.ค. 2022	<a href="#">New Thai rules check in</a>	Borders/Visas	0	
25 พ.ค. 2022	<a href="#">What does the script on 100 baht bill say?</a>	Other	0	
25 พ.ค. 2022	<a href="#">Thailand itinerary help</a>	Transport/Itineraries	0	
25 พ.ค. 2022	<a href="#">Hi I was wondering how I'd go about getting a refund from Thai airways ? They've canc...</a>	Transport/Itineraries	0	



# จัดกลุ่มทางสกิดิของนักท่องเที่ยว: ภาพรวม

จัดกลุ่มโดยใช้ข้อมูลจำนวนนักท่องเที่ยวและรายได้ตั้งแต่ ม.ค. 2019 ถึง ม.ค. 2022

เดือนที่สนใจ

1 ม.ค. 2019 - 31 ม.ค. 2022

ข้อมูลล่าสุด

ปีที่สนใจ

กุมภาพันธ์ 2022

31 มกราคม 2022

กลุ่มทางสกิดิของนักท่องเที่ยว [รายละเอียดการจัดกลุ่มล่าสุด](#)

1. กลุ่มที่มีจำนวนนักท่องเที่ยวและรายได้สูง

2. กลุ่มที่มีรายได้ต่อหัวสูง

3. กลุ่มที่ได้รับความสนับสนุนจากชาวไทยและชาวต่างชาติ

4. กลุ่มที่ได้รับความสนับสนุนจากชาวไทย

5. กลุ่มที่เริ่มเป็นที่นิยมของชาวต่างชาติ

6. กลุ่มที่มีจำนวนนักท่องเที่ยวและรายได้น้อย

## กลุ่มทางสกิดิแบ่งตามจังหวัด

จังหวัด ②	กุมภาพันธ์	กุญแจ	กุญสกิดิ ①
กรุงเทพมหานคร	กรุงเทพมหานคร		1
ภูเก็ต	ภาคใต้		2
ชลบุรี	ภาคตะวันออก		3
เชียงใหม่	ภาคเหนือ		3
กาญจนบุรี	ภาคตะวันตก		4
นครราชสีมา	ภาคตะวันออกเฉียงเหนือ		4
ประจวบคีรีขันธ์	ภาคตะวันตก		4
เพชรบุรี	ภาคตะวันตก		4
กระบี่	ภาคใต้		5
ขอนแก่น	ภาคตะวันออกเฉียงเหนือ		5

## จำนวนนักท่องเที่ยว

จำนวนนักท่องเที่ยวชาวไทยเฉลี่ย จำนวนนักท่องเที่ยวชาวต่างชาติเฉลี่ย

154.5K 32.6K

จำนวนนักท่องเที่ยวเฉลี่ยรวม

93.5K

เทียบระหว่างกุญสกิดิ



## รายได้

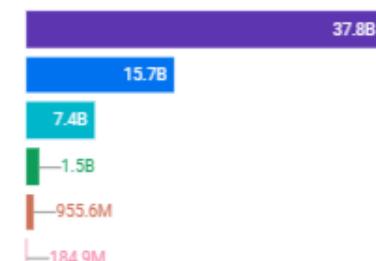
รายได้จากชาวไทยเฉลี่ย รายได้จากชาวต่างชาติเฉลี่ย

641.8M 698.5M

รายได้เฉลี่ยรวม

670.2M

เทียบระหว่างกุญสกิดิ



## รายได้ต่อหัว

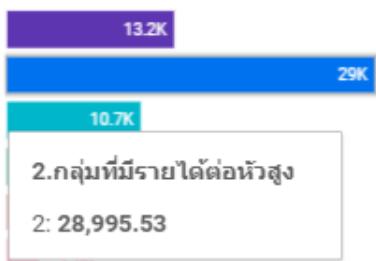
รายได้ต่อหัวชาวไทยเฉลี่ย รายได้ต่อหัวชาวต่างชาติเฉลี่ย

2.7K 6.2K

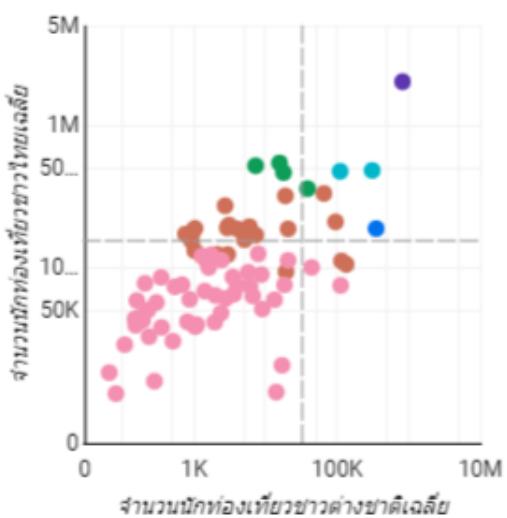
รายได้ต่อหัวเฉลี่ยรวม

7.2K

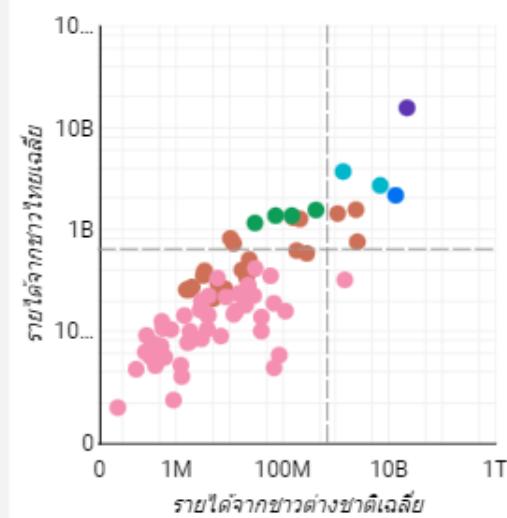
เทียบระหว่างกุญสกิดิ



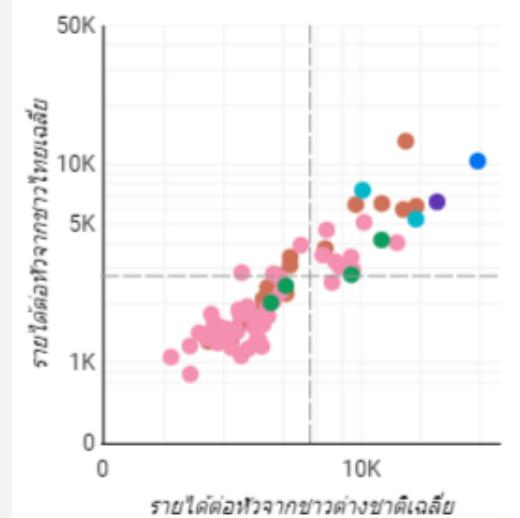
## เทียบระหว่างชาวไทยกับชาวต่างชาติ



## เทียบระหว่างชาวไทยกับชาวต่างชาติ



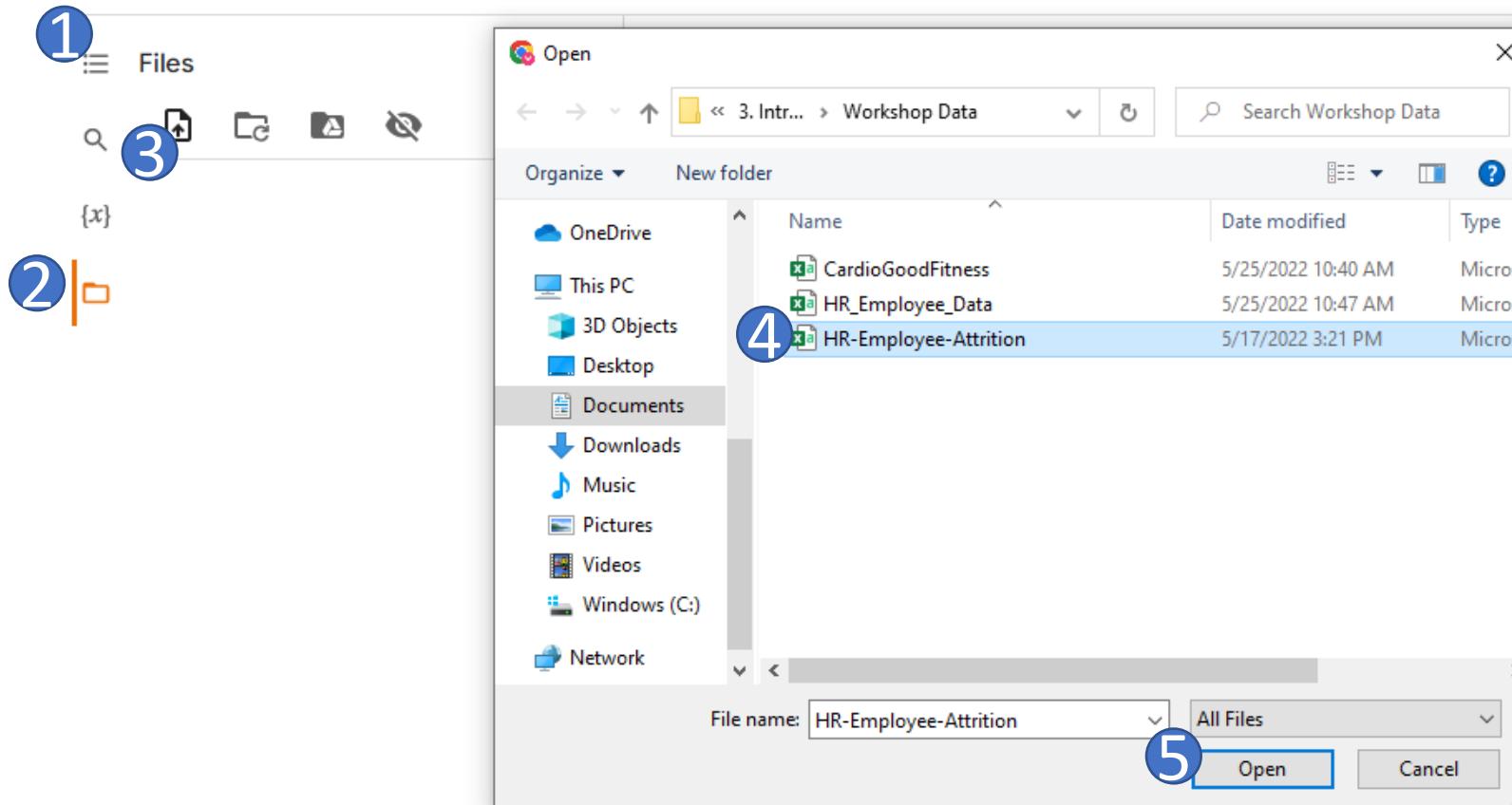
## เทียบระหว่างชาวไทยกับชาวต่างชาติ



# WORKSHOP II

# HR Attrition – Pay Equality

ให้เปิด Colab และอัปโหลดข้อมูลชุด HR-Employee-Attrition.csv



# HR Attrition – Pay Equality

ใช้คำสั่ง pd.read\_csv เพื่ออ่านไฟล์

```
df = pd.read_csv('HR-Employee-Attrition.csv')
```

ลองใช้ groupby

```
df.groupby(["Gender"])["MonthlyIncome"].agg(['mean', 'std'])
```

	mean	std	%
Gender			
Female	6686.566327	4695.608507	
Male	6380.507937	4714.856577	

# HR Attrition – Pay Equality

```
df.groupby(["JobLevel","Gender"])["MonthlyIncome"].agg(['count','mean','std'])
```

		count	mean	std	+
JobLevel	Gender				
1	Female	199	2780.487437	709.605053	
	Male	344	2790.633721	771.300650	
2	Female	220	5435.327273	1266.690800	
	Male	314	5549.184713	1502.541186	
3	Female	94	9962.702128	1892.555119	
	Male	124	9706.991935	1737.145905	
4	Female	51	15431.372549	1701.573119	
	Male	55	15570.927273	1929.704985	
5	Female	24	19129.916667	587.444052	
	Male	45	19224.844444	471.321930	

# HR Attrition – Two Group Testing

ให้เปิด Colab และอัพโหลดข้อมูลชุด HR-Employee-Attrition.csv

```
df = pd.read_csv('HR-Employee-Attrition.csv')
```

Hypothesis Testing:

$$H_0: \text{MonthlyIncome}_{\text{male}} = \text{MonthlyIncome}_{\text{female}}$$

$$H_1: \text{MonthlyIncome}_{\text{male}} \neq \text{MonthlyIncome}_{\text{female}}$$

```
from scipy.stats import ttest_ind
female_rates = df[(df['Gender']=='Female') ]["MonthlyIncome"]
male_rates = df[(df['Gender']=='Male') ]["MonthlyIncome"]
ttest_ind(female_rates,male_rates)
```

```
Ttest_indResult(statistic=1.2212617308870655, pvalue=0.22218303455087898)
```

Remarks: ตาม default จะเป็นการ test แบบสองทาง “two-sided” สามารถตั้งค่า alternative = “Greater” ได้  
แต่ต้องอัพเดท package scipy แต่ไม่สามารถทำได้ใน colab >> [pip install --upgrade scipy](#)

# HR Attrition – Two Group Testing

ทำการแยก Dataframe เป็นส่วนของ feature และ output จากนั้นตรวจสอบ dimension ด้วย .shape

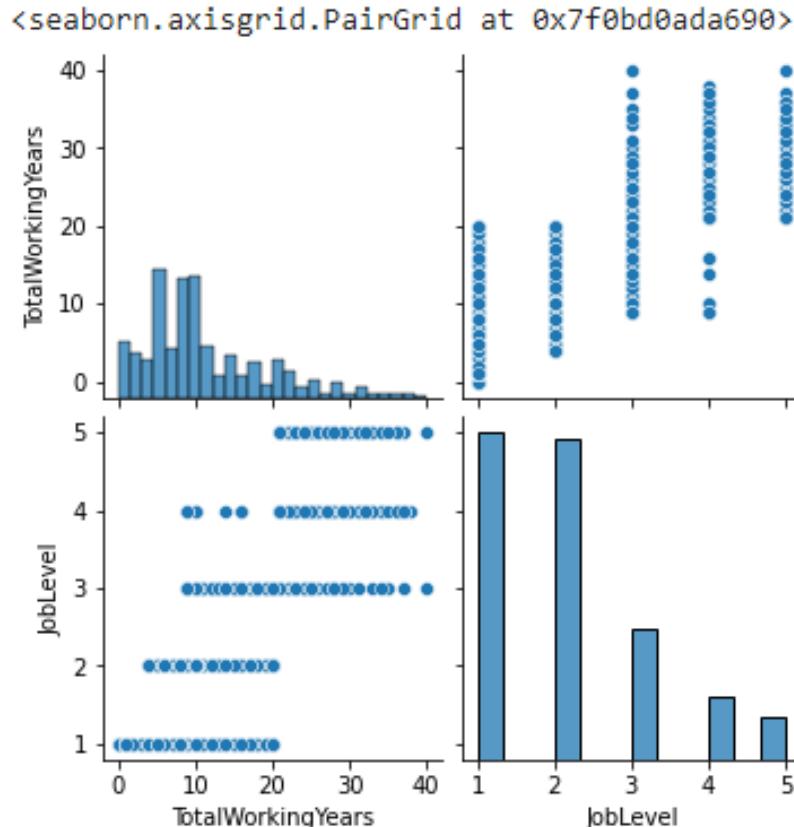
```
df.columns  
features = ['TotalWorkingYears', 'JobLevel']  
X = df[features]  
Y = df['MonthlyIncome']  
print(X.shape)  
print(Y.shape)
```

```
(1470, 2)  
(1470, )
```

# HR Attrition – Two Group Testing

ใช้ pair plot เพื่อหาความสัมพันธ์ระหว่างสองตัวแปร (หากคล้ายกันมากควรเลือกใช้แค่ตัวเดียว)

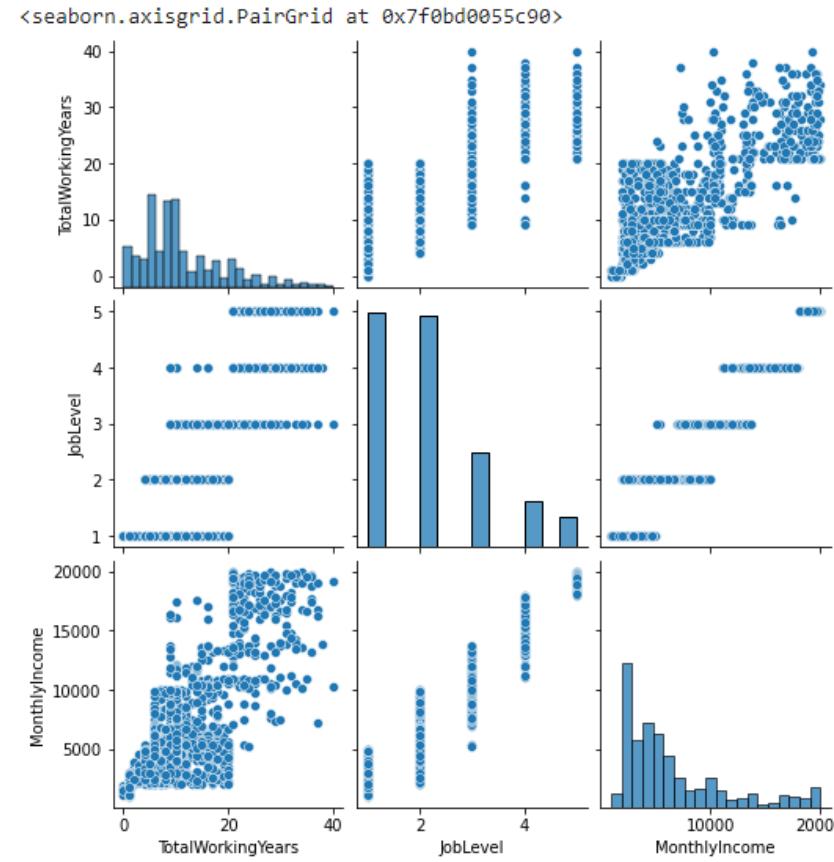
```
import seaborn as sns  
sns.pairplot(X)
```



# HR Attrition – Two Group Testing

ใช้ pair plot เพื่อหาความสัมพันธ์ระหว่างสองตัวแปร (หากคล้ายกันมากควรเลือกใช้แค่ตัวเดียว)

```
import seaborn as sns  
sns.pairplot(pd.concat([X,Y],axis=1))
```



# HR Attrition – Two Group Testing

- เพิ่มคอลัมน์เลข 1

```
import statsmodels.api as sm
X = sm.add_constant(X)
X.head()

/usr/local/lib/python3.7/dist-packages/statsmoc
x = pd.concat(x[::-order], 1)
```

	const	TotalWorkingYears	JobLevel	edit
0	1.0	8	2	
1	1.0	10	2	
2	1.0	7	1	
3	1.0	8	1	
4	1.0	6	1	

# HR Attrition – Two Group Testing

ใช้ OLS method เพื่อหา parameter

```
model = sm.OLS(Y,X).fit()
print(model.summary())
prediction = model.predict(X)
print("MAPE= " +str(100*np.mean(np.abs(((Y-prediction)/Y))))+"%)
```

```
OLS Regression Results
=====
Dep. Variable: MonthlyIncome R-squared:          0.905
Model:                 OLS  Adj. R-squared:        0.905
Method:                Least Squares F-statistic:     7014.
Date:      Wed, 25 May 2022 Prob (F-statistic):   0.00
Time:          15:08:10 Log-Likelihood:    -12785.
No. Observations:      1470 AIC:            2.558e+04
Df Residuals:         1467 BIC:            2.559e+04
Df Model:                   2
Covariance Type:    nonrobust
=====
            coef    std err          t      P>|t|      [0.025      0.975]
-----
const     -1835.8617    80.019     -22.943      0.000    -1992.825    -1678.898
TotalWorkingYears    46.0820     7.802       5.906      0.000      30.777      61.387
JobLevel      3788.3785    54.843      69.077      0.000    3680.800    3895.957
=====
Omnibus:             7.910 Durbin-Watson:        2.050
Prob(Omnibus):      0.019 Jarque-Bera (JB):    10.099
Skew:                  -0.047 Prob(JB):        0.00641
Kurtosis:                 3.395 Cond. No.:           32.2
=====
```

Warnings:

```
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.  
MAPE= 21.78938650334339%
```

# HR Attrition – One hot encoding

บางตัวแปรเป็นประเภท Categorical จำเป็นต้องใช้ One hot encoding

```
one_hot_level = pd.get_dummies(df['JobLevel'],prefix='JobLevel_').iloc[:, :-1]
df[one_hot_level.columns]= one_hot_level
one_hot_level.head()
```

	JobLevel_1	JobLevel_2	JobLevel_3	JobLevel_4	⋮
0	0	1	0	0	⋮
1	0	1	0	0	⋮
2	1	0	0	0	⋮
3	1	0	0	0	⋮
4	1	0	0	0	⋮

# HR Attrition – One hot encoding

ทำการ fit model

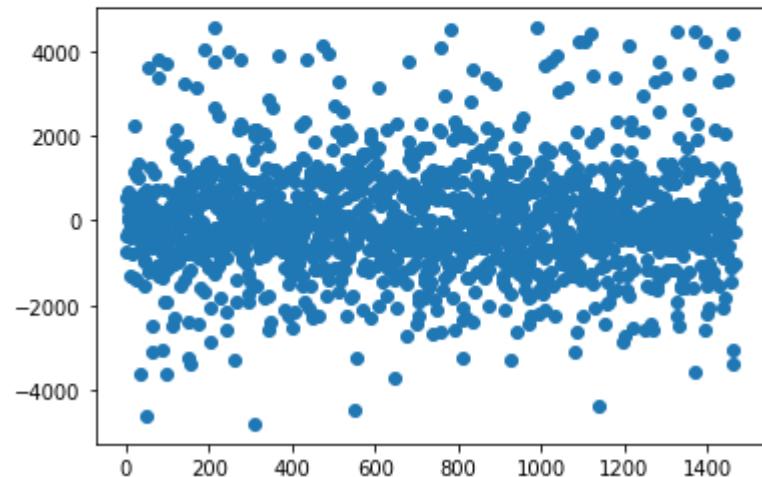
```
import statsmodels.api as sm
df.columns
features = ['TotalWorkingYears']+list(one_hot_level)
X = df[features]
X = sm.add_constant(X)
Y = df['MonthlyIncome']
model = sm.OLS(Y,X).fit()
print(model.summary())
prediction = model.predict(X)
print("MAPE= " +str(100*np.mean(np.abs(((Y-prediction)/Y))))+"%)
```

# HR Attrition – One hot encoding

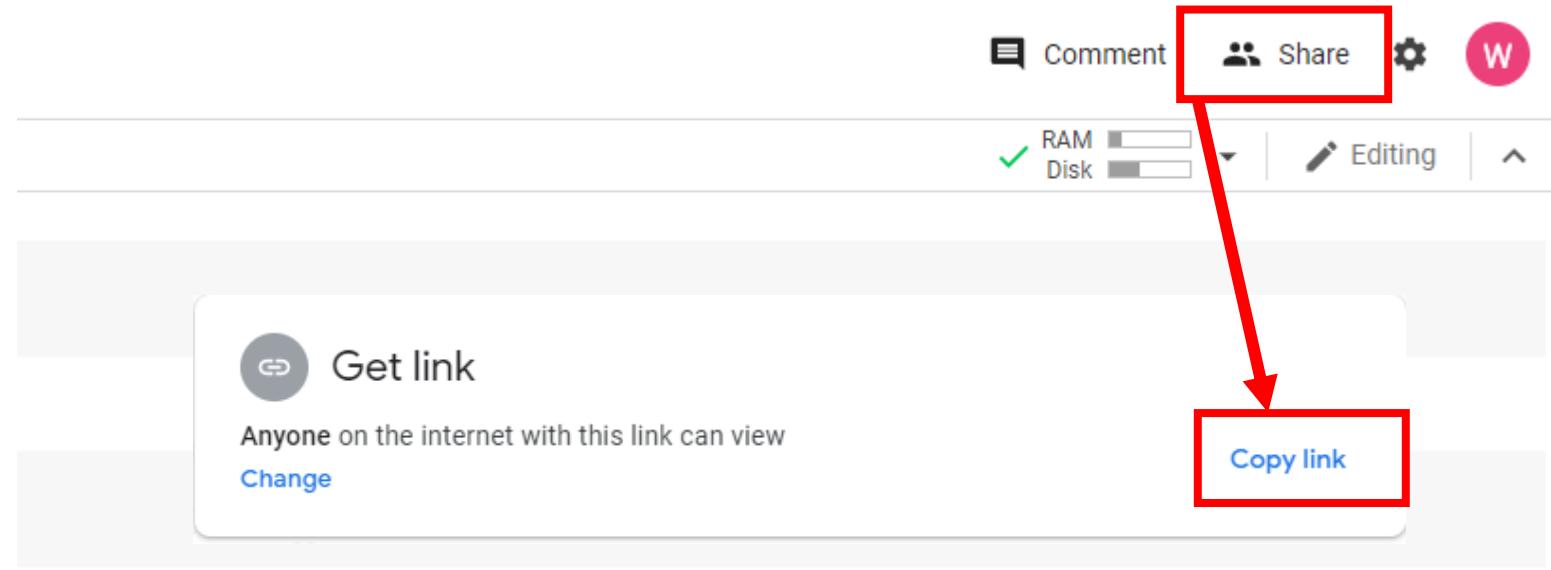
ลองทำ residual plot

```
import matplotlib.pyplot as plt
res = Y-prediction
plt.scatter(range(len(res)),res)
```

<matplotlib.collections.PathCollection at 0x7f0bcda84e50>



# How to get the link



# QUIZ

Follow us on



Facebook



Twitter



Blockdit



govbigdata

YouTube

Government Big Data Institute  
(GBDI)



Line Official  
@gbdi