

Essential Statistics Live Class

| | | |
|---------------|-------------------------------------|------------|
| Tags | Live Class | Statistics |
| Class | | |
| Finished Yet? | <input checked="" type="checkbox"/> | |
| Knowledge | Live Classes | |

Part 1

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/d9fa0cce-f1e3-4a19-8a14-780982ff3ded/Stats_Part_1.pdf

^Sketchwow จ่ายเงินครั้งเดียว ใช้ได้ตลอดไป

Stats Workbook:

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/0ddbfef06-a338-4823-8d3b-4ea9fd2b66c5/stats_workbook.xlsx

เนื้อหาหลัก: Descriptive Statistics *ต้องเข้าใจ 3 เรื่องนี้

1. Central Tendency (การวัดค่ากลางของข้อมูล)
2. Measure of Spread (การวัดการกระจายของข้อมูล)
3. Measure Position (การวัดตำแหน่งของข้อมูล)

Sampling: การสุ่มตัวอย่าง

การวัดค่าในเชิงสถิติ: Point (จุด) VS Interval (ช่วง)

-การเรียนสถิติให้รู้เรื่อง ให้ใช้คำศัพท์ภาษาอังกฤษไปเลย (เช่น Regression = สมการทดแทน ใช้คำแปลภาษาไทยแล้วจะงกกว่าศัพท์ภาษาอังกฤษตรง ๆ)

What is Statistics?

-สมมติว่าเรามีคนในประเทศไทย 70 ล้านคน (เป็นจำนวนประชากร [Population] ก็คงหมด) เราไม่สามารถสำรวจความคิดเห็นของคนทั้ง 70 ล้านคนได้

-ดังนั้น เราจึงต้องสุ่มตัวอย่าง [Sample] ขึ้นมา เช่นกลุ่มตัวอย่าง 1,000 คน เป็นตัวแทนของประชากรก็คงหมด

Simple Random Sampling (Population ทุกคนมีโอกาสในการถูก Sampling เท่ากัน)

=RAND() ใช้ในการสร้างเลขแบบสุ่มใน Excel หรือ Google Sheets ถ้าเรากด Double Click ค่าก็จะเปลี่ยนไปเรื่อย ๆ

-ขั้นตอนแรกของงานด้านสถิติคือการวางแผนการทำ Research และวางแผนการเก็บข้อมูล

-Framework ใน การเก็บข้อมูลและนำไปใช้สำคัญมาก สำคัญกว่าการสนใจแต่ Data Analysis เพียงอย่างเดียว

ตัวอย่างการแบ่งคนไทยเป็น 5 ภาค

| Stratified Random Sampling | | | |
|----------------------------|-----|------|-----------------------|
| Thaibev Chang n=1000 | | | |
| GBKK | 20% | 200 | |
| Central | 30% | 300 | จันทบุรี |
| South | 10% | 100 | |
| North | 10% | 100 | ลูกค้าเลือก |
| NE | 30% | 300 | KK, Ubon, Udon, Korat |
| | | 1000 | |

*แบ่งประชากรที่เราสนใจเป็นกลุ่ม ๆ (Stratify) → เช็ต Quota แล้วค่อยไปสุ่มคนในกลุ่มนั้น ๆ

Probability Sampling:

1. Simple
2. Systematic
3. Stratified

-ยิ่งเรามี Sample Size มากขึ้น ค่าที่ได้จะเข้าใกล้ความเป็นจริงมากขึ้น Margin of Error จะน้อยลง

-ในความเป็นจริงแล้ว ลูกค้าและผู้บริหารจะกังวลอยู่ 2 เรื่อง (Constraints):

1. Budget (งบประมาณ กำหนด Sample Size)
2. Timeline (เวลา)

CPI = Cost Per Interview

สูตรคำนวณอ. Taro Yamane: <https://datarockie.com/blog/yamane-sample-size-calculation/>

SurveyMonkey Sample Size Calc: <https://www.surveymonkey.com/mp/sample-size-calculator/>

The screenshot shows a sample size calculator interface. At the top, there are three input fields: 'Population Size' (7,000,000), 'Confidence Level (%)' (95), and 'Margin of Error (%)' (5). Below these, the calculated 'Sample size' is displayed prominently as '385'. A descriptive text below the result states: 'Doing market research? SurveyMonkey Audience gets you the right survey respondents fast and easy and helps you target them by demographics, consumer behavior, geography, or even designated marketing area.' A green button labeled 'Choose your audience' is visible at the bottom.

-ในชีวิตจริงจะมี Diminishing Return ยิ่งเรารอยากลด Margin of Error ให้น้อยลงเท่าไหร่ ก็จะต้องใช้เงินเยอะขึ้นมากเก่า�ัน (เช่น MoE 3% เสียเงิน 500k บาท แต่ 2% เสียเงินขึ้นหลักล้าน สูกค่าที่มีงบจำกัดคงไม่อยากที่จะเสียเงินเพิ่มขึ้นมากกว่าเก่าตัวเพื่อลด MoE แค่ 1%)

-สกิติคือกระบวนการคิด คือ Mindset คือ Framework

-Sample Size / Incidence (%) = Real Sample Size (คบก์ต่อ Survey เราชริง ๆ)

-Central Limit Theorem (CLT): the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough.

Sampling วิ่งจาก Population ไป Sample

Generalization วิ่งจาก Sample ไป Population

Descriptive Statistics: การหาค่าทางสถิติเพื่อธิบายข้อมูลชุดนั้น

3 broad categories:

1. Measures of Central Tendency (mean, median, mode)
2. Measures of Spread (Variability)

-Standard Deviation (SD)

-Variance

-Range

-IQR

3. Measures of Position

-Min, Max

-Percentile (50th, 75th)

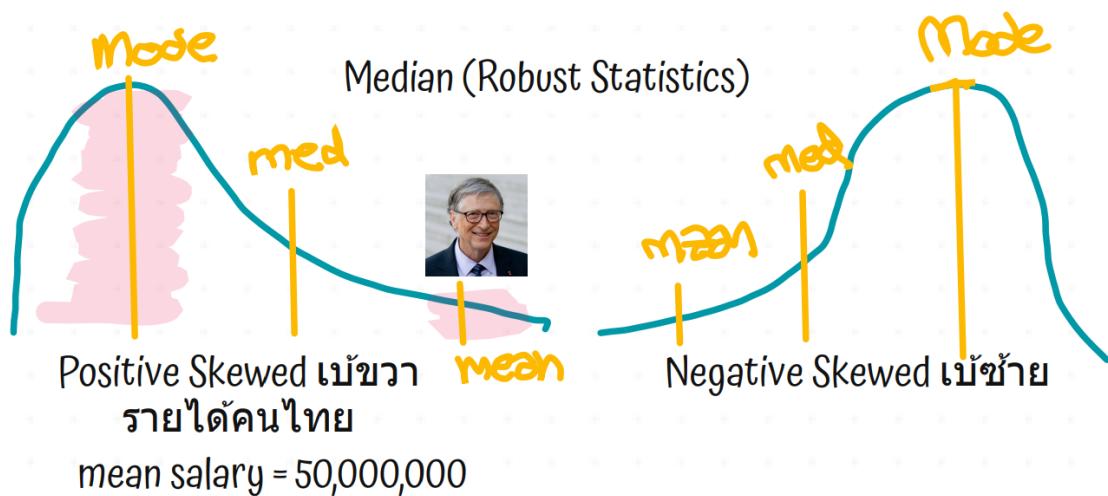
-Quartile (Q1, Q2, Q3, Q4)

Normal Distribution: กระจายแบบกราฟตระหุกตรงกลาง ไม่เบี้ยวหรือเบี้ยว (Mean = Median = Mode หรือค่าทั้งสามใกล้เคียงกันมาก)

| A | B | C | D | E | F | G | H | I |
|--------|------------|-------------|---------|------------------------------------|---|---|---|---|
| height | d^2 | | | | | | | |
| 168 | 9.87755 | n | | 7 | | | | |
| 170 | 1.30612 | mean | | 171.143 | | | | |
| 195 | 569.163 | median (Q2) | | 168 | | | | |
| 140 | 969.878 | mode | | 195 | | | | |
| 165 | 37.7347 | range | | 55 | | | | |
| 165 | 37.7347 | variance | | 365.81 | | | | |
| | | sd | 19.1261 | ส่วนเบี่ยงเบนมาตรฐาน(จากค่าเฉลี่ย) | | | | |
| | | IQR | | 17.5 | | | | |
| | | Min | | 140 | | | | |
| | | Max | | 195 | | | | |
| | Quartile | Q1 | | 165 | | | | |
| | | Q2 | | 168 | | | | |
| | | Q3 | | 182.5 | | | | |
| | | Q4 | | 195 | | | | |
| | Percentile | 50th | | 168 | | | | |
| | | 77th | | 185.5 | | | | |
| | | 24th | | 165 | | | | |

-การทำ Visualization เป็นต้นจะทำให้เราสามารถเห็นภาพการกระจายของข้อมูลได้ (เช่น เป็น Unimodal, Bimodal, หรือ Multimodal เป็นต้น)

การเบื้องต้นของกราฟ



Range (พิสัย) = Max - Min

Variance (ความแปรปรวน)

=VAR.P ใช้กับ Population

=VAR.S ใช้กับ Sample

*Sample Variance จะสูงกว่า Population Variance เช่นเดียวกัน

Sample Variance Formula:

Formula :

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

S^2 = sample variance

x_i = the value of one observation

\bar{x} = the mean value of all observations

n = the number of observations

S.D. (Standard Deviation) = $\text{SQRT}(\text{Variance})$ * S.D. ใช้บ่อยมาก เพราะเป็นส่วนเบี่ยงเบนมาตรฐานเมื่อเทียบกับค่าเฉลี่ย ดังนั้น เราจะ Report Mean และ S.D. คู่กันเสมอ

Empirical Rule

1. $+/-1\text{SD}$: ข้อมูล 68% จะตกลอยู่ในช่วงนี้
2. $+/-2\text{SD}$: ข้อมูล 95% จะตกลอยู่ในช่วงนี้
3. $+/-3\text{SD}$: ข้อมูล 99.7% จะตกลอยู่ในช่วงนี้

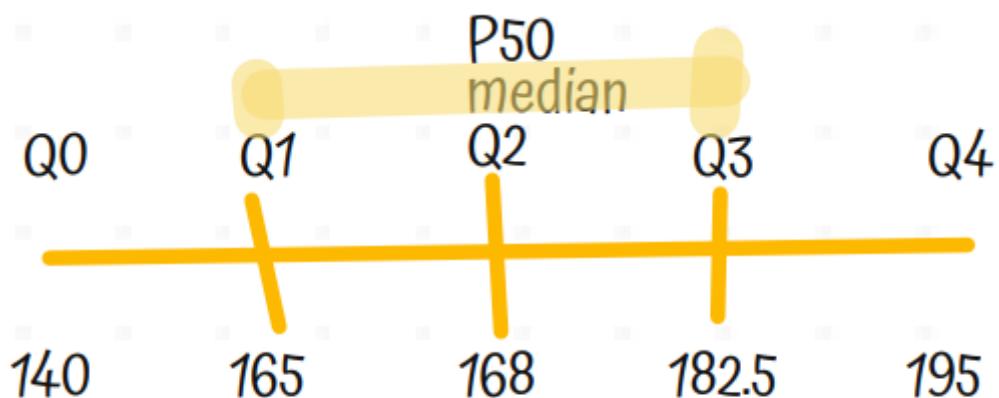
Measure of Position:

| | | |
|------------|------|-------|
| | Min | 140 |
| | Max | 195 |
| Quartile | Q1 | 165 |
| | Q2 | 168 |
| | Q3 | 182.5 |
| | Q4 | 195 |
| Percentile | 50th | 168 |
| | 77th | 185.5 |
| | 24th | 165 |

Min: ค่าต่ำสุด

Max: ค่าสูงสุด

Quartile (แบ่งข้อมูลเป็น 4 ส่วน)



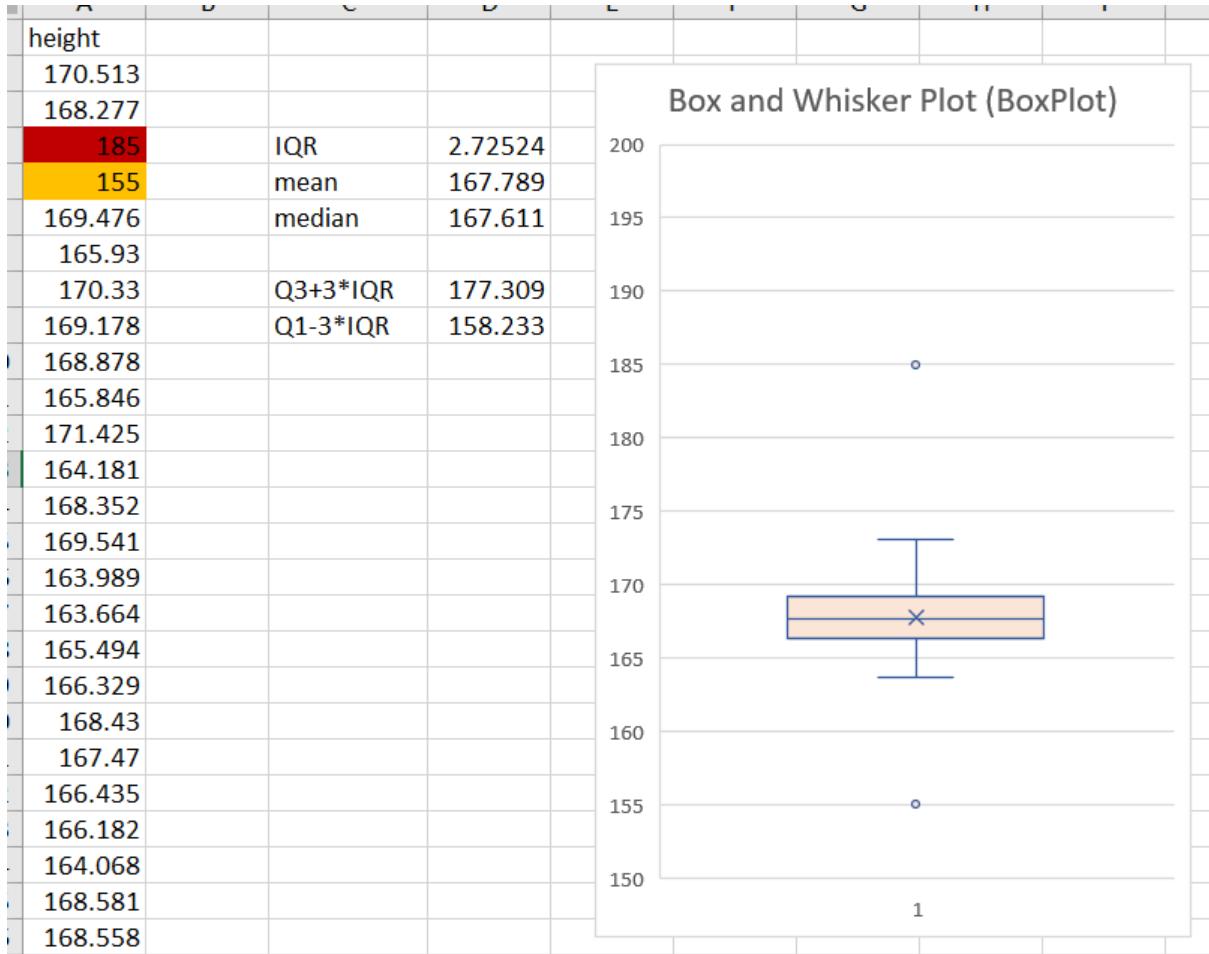
$$IQR = Q3 - Q1 = 182.5 - 165 = 17.5$$

1. Q0 [Min]
2. Q1
3. Q2 [Median]
4. Q3
5. Q4 [Max]

Percentile (ແບ່ງຂ້ອມລົບເປັນ 100 ສ່ວນ) ເຊັ່ນ P50 = Q2 = Median

IQR (Interquartile Range) = Q3 - Q1

Box and Whisker Plot (Boxplot)



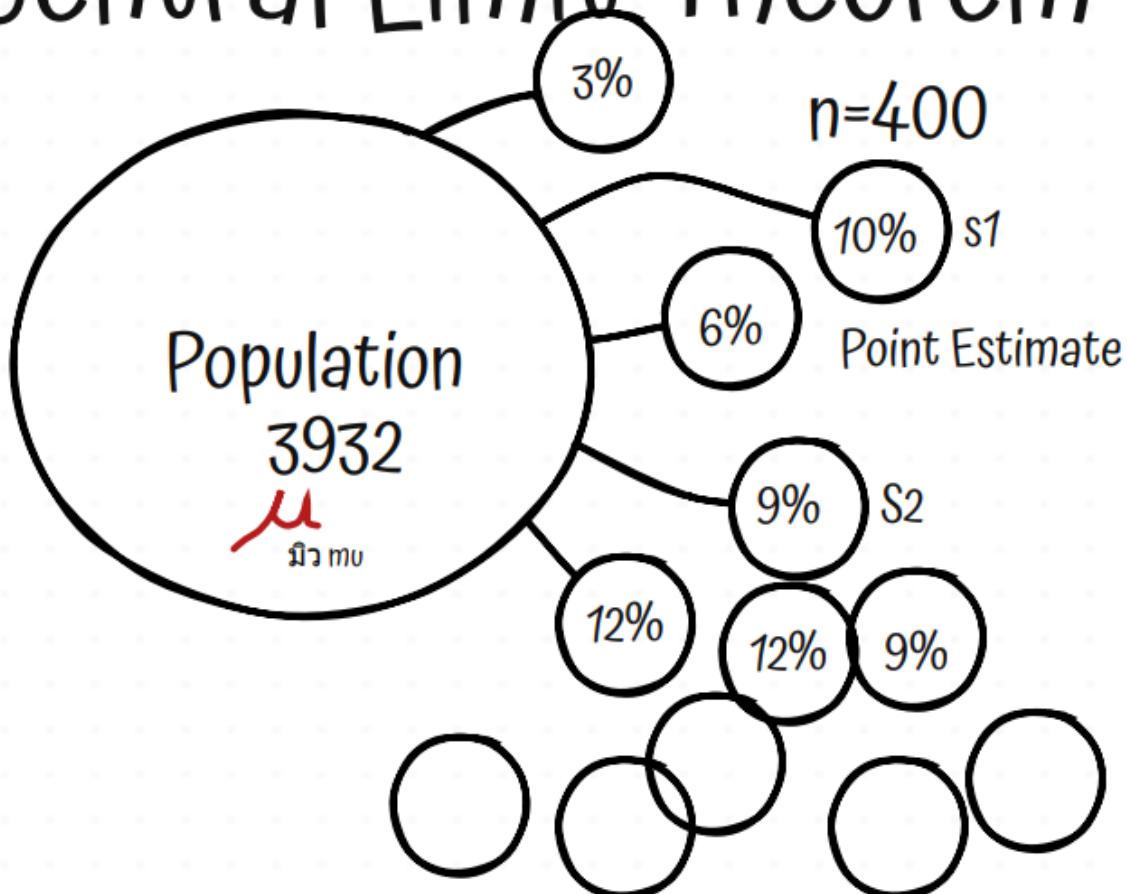
How to detect outliers by using boxplot?

How to detect outliers?

1. Boxplot
2. Formula: $Q3 + 3 \times IQR$, $Q1 - 3 \times IQR$

Central Limit Theorem

Central Limit Theorem



CTL holds true when

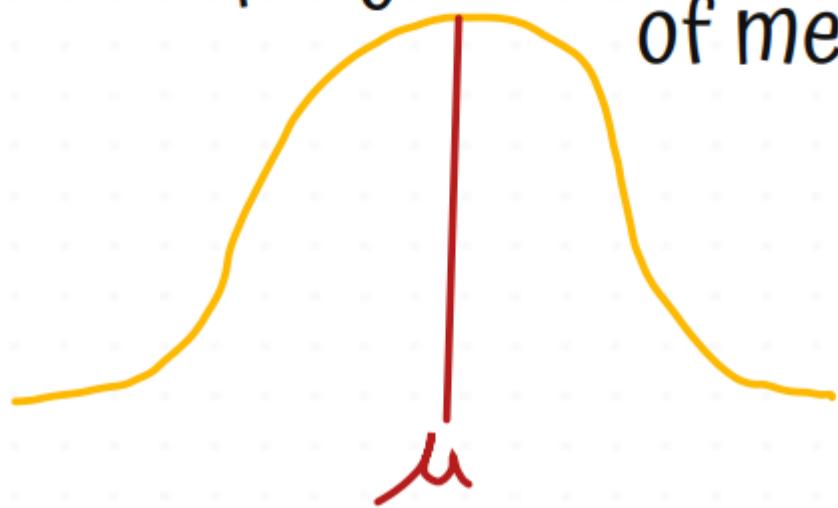
1. sample size ≥ 30
2. Random sampling

-ในการสุ่มแต่ละครั้ง เป็นไปไม่ได้ที่ค่าที่ได้ออกมาจะมีค่าเท่ากันทุกครั้ง เราจะเรียกจุดที่เกิดจาก การสุ่มแต่ละครั้งว่า Point Estimate

-ถ้าเรานำ Point Estimate แต่ละครั้งมา Plot เป็นกราฟ จะได้ Normal Distribution ที่นิยมเรียก พิเศษว่า Sampling Distribution

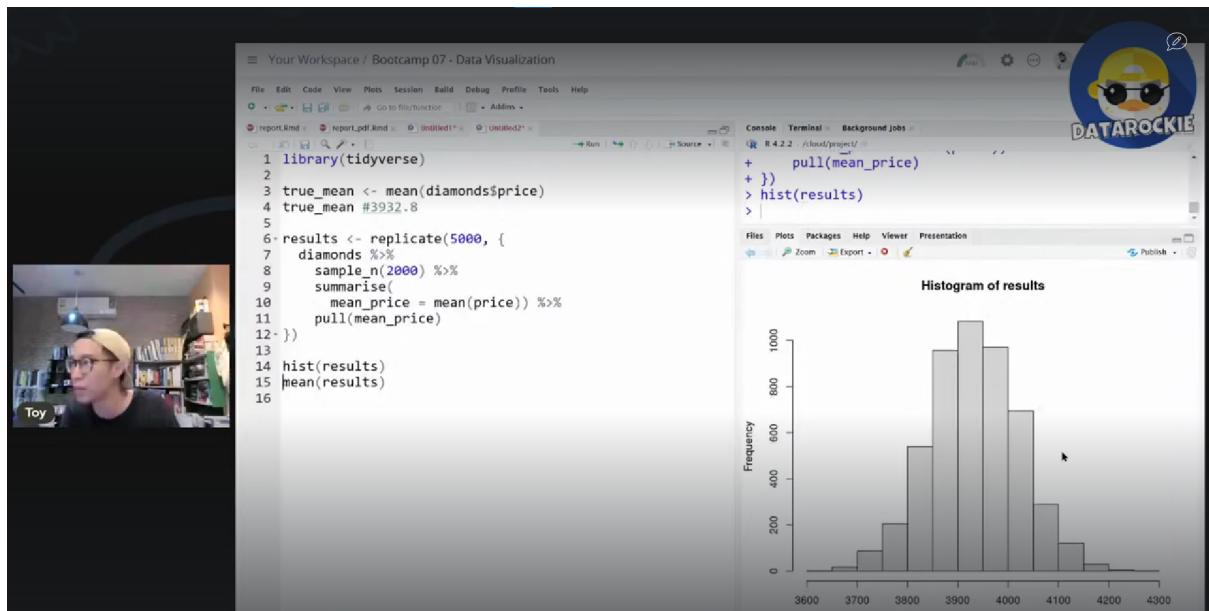
[Sampling Distribution of Mean = การกระจายของค่าเฉลี่ยที่ถูกสุ่มออกมาน]

Sampling Distribution of mean



*ในการทำงานจริง เราจะสุ่มตัวอย่างแค่ครึ่งเดียว

พิสูจน์ CLT ด้วย R:



Part 2

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/73e4c49d-01d9-4cb6-902f-83cc1fe12b12/Essential_Stats_Part_2.pdf

Confidence Interval:

Confidence Interval

สำหรับคลาสเนี้ยแอดคิตวารการค่าทาง confidence interval สำคัญที่สุดเลย เป็นเรื่องที่ data analyst เราต้องทำความเข้าใจอย่างต่องแท้ core of statistics

- Point vs. Interval Estimate ข้อดีของการใช้ interval คือมันมีเรื่องความน่าจะเป็นมากกว่าข้อด้วย เช่น 95% CI [3.2, 4.6]
- $1 - \alpha$ = confidence level โดยทั่วไป $\alpha = 0.05$ เราจะมีในผลลัพธ์ได้ $1 - 0.05 = 0.95$ หรือ 95%
- ถ้าอยากรู้ผลลัพธ์ที่แม่นยำมากขึ้น เก็บ sample size ให้เยอะขึ้น (more n, closer to the truth)

Confidence Interval จะมีสองแบบคือ

1. Confidence Interval for **Mean** ที่แอดสอนใน Live
2. Confidence Interval for **Proportion (%)** อันนี้แอดลองค่าทางเพิ่มให้ดูในไฟล์ Excel นักเรียนสามารถดาวน์โหลดไปแกะสูตรได้นะครับ

ขั้นตอนการค่าทาง Confidence Interval ของทั้ง Mean และ Proportion ไม่ต่างกันเลย

- ค่าทาง SE (standard error มาจาก central limit theorem)
- ค่าทาง Margin Error (ME) ค่าความคลาดเคลื่อน
- สร้าง Confidence Interval จาก mean-/-+ME หรือ proportion-/-+ME

Excel:

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/f3ad0012-6db9-4cae-b9cf-3a6773e16453/confidence_interval_calculation.xlsx

```
##Excel: Generate data from normal distribution with known Mean and Variance  
=NORM.INV(RAND(), mean, sd)
```

Non Propability Sampling: ไม่ค่อยได้ใช้ เพราะมี Bias in data collection

Convenient Sampling: ใช้งานระดับ Professional ไม่ได้ (เป็น Non Propability Sampling) แต่เป็นวิธีที่ง่ายและสะดวก

Snowball Sampling: สัมภาษณ์คนหนึ่ง แล้วขอ Contact ให้เราสามารถสัมภาษณ์คนรู้จักของ เขายได้ (เป็น Non Propability Sampling)

Mean of Population: mu (μ)

S.D. of Population: sigma (Σ)

Mean of Sample: mean

S.D. of Sample: S.D

*ถ้าเราเข้าถึง Population ได้ไม่หมด เราจะเรียก Arn กับ sigma ว่า Unknown Parameter แต่ เราสามารถสุ่มตัวอย่างเพื่อเข้าถึงค่า mean และ S.D. เพื่อประมาณค่า (Estimate) เพื่อให้เข้าใจ Population มากขึ้น

Statistics VS Analytics

-สุ่มข้อมูล = สถิติ

-ใช้ข้อมูลทั้งหมด = Big Data Analytics (เช่น ChatGPT [GPT-3] ใช้ Parameters มากถึง 175B Parameters)

Z-Score (Standardization)

$$Z = (x - \text{mean}) / \text{S.D.}$$

*ค่าโดยส่วนใหญ่จะอยู่ระหว่าง -3 ถึง +3

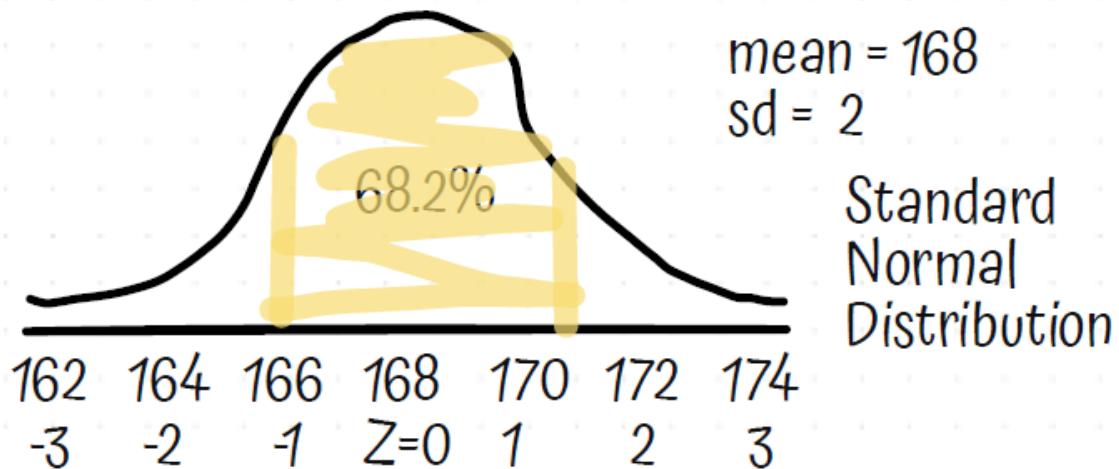
ความหมายของค่า Z

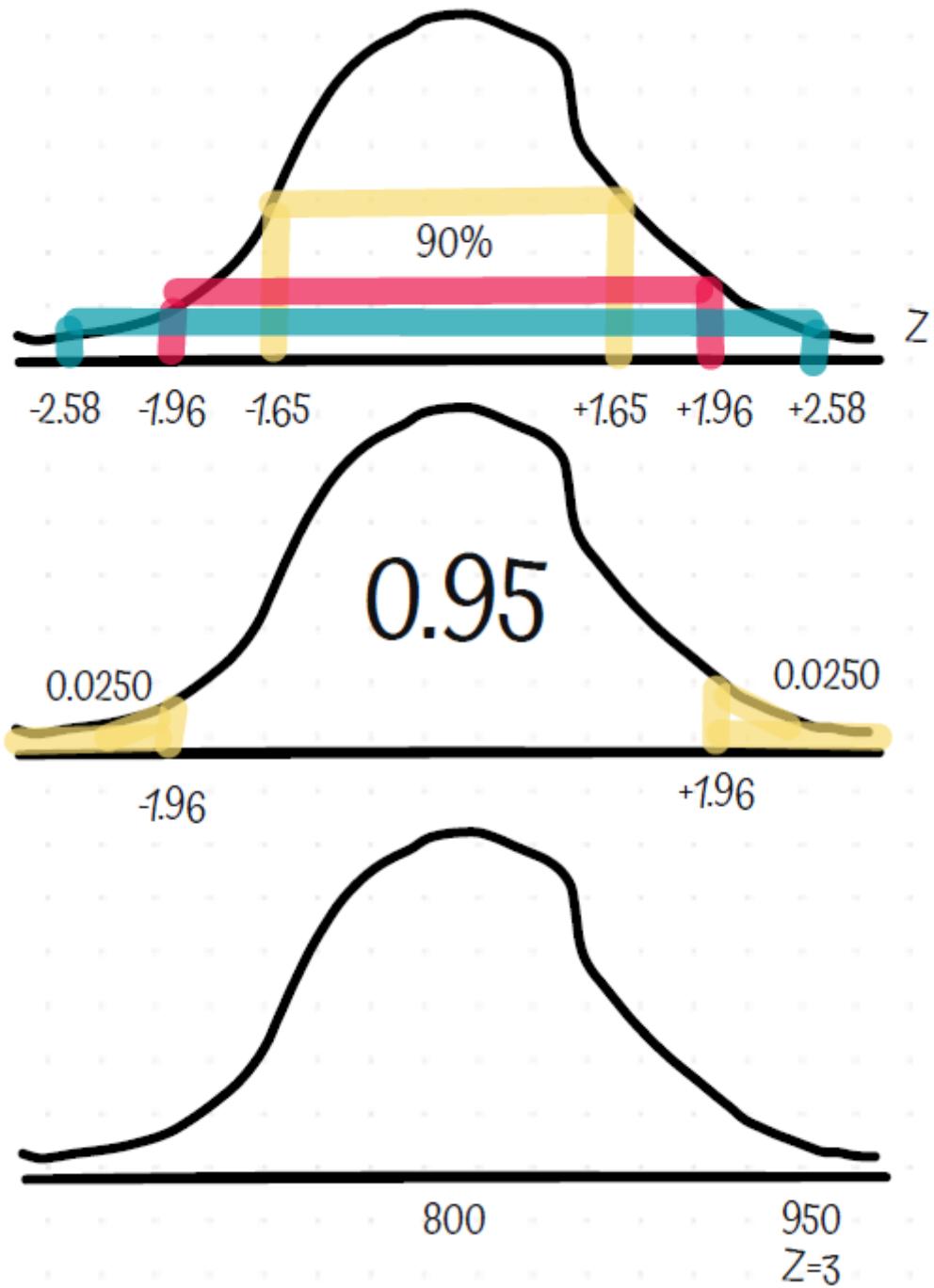
1. $Z = 0$: ค่าเท่ากับค่าเฉลี่ย
2. Z เป็นบวก: ค่ามากกว่าค่าเฉลี่ยอยู่ Z เท่าของ S.D. (เช่น $Z = +1.5$ คือ ค่า Z มากกว่าค่าเฉลี่ยอยู่ 1.5 S.D.)
3. Z เป็นลบ: ค่าน้อยกว่าค่าเฉลี่ยอยู่ Z เท่าของ S.D. (เช่น $Z = -2$ คือ ค่า Z น้อยกว่าค่าเฉลี่ยอยู่ 2 S.D.)

Standard Normal Distribution

Standardization (Z score)

$$Z = (X - \text{mean}) / \text{sd} \quad -3 \dots +3$$

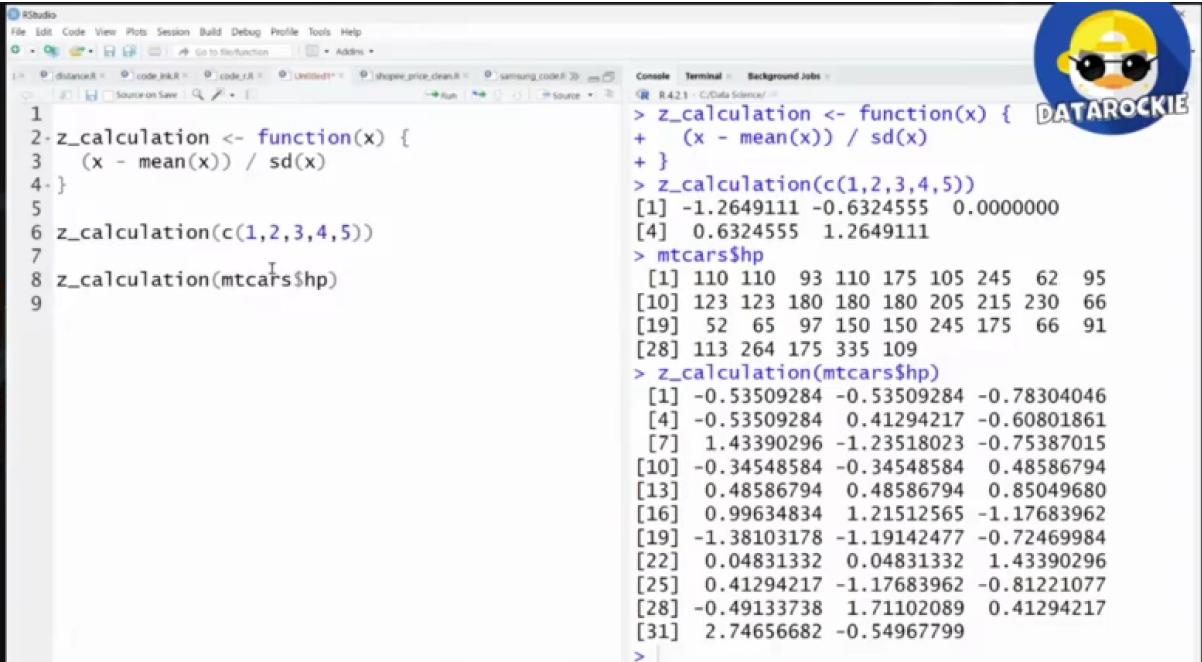




1. ข้อมูลระหว่าง $Z = -1.65$ ถึง $Z = +1.65$ เป็น 90% ของข้อมูลทั้งหมด
2. ข้อมูลระหว่าง $Z = -1.96$ ถึง $Z = +1.96$ เป็น 95% ของข้อมูลทั้งหมด
3. ข้อมูลระหว่าง $Z = -2.58$ ถึง $Z = +2.58$ เป็น 99% ของข้อมูลทั้งหมด

-นักสถิติจะชอบค่า Z มากกว่าค่าปกติ เพราะสามารถบอกรู้ว่าค่า Z มากกว่าหรือน้อยกว่าค่าเฉลี่ยอยู่เท่าไร

การคำนวณค่า Z ใน R:



```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
File Source on Save Run Source ...
Console Terminal Background Jobs
> z_calculation <- function(x) {
+   (x - mean(x)) / sd(x)
+ }
> z_calculation(c(1,2,3,4,5))
[1] -1.2649111 -0.6324555  0.0000000
[4]  0.6324555  1.2649111
> mtcars$hp
[1] 110 110 93 110 175 105 245 62 95
[10] 123 123 180 180 180 205 215 230 66
[19] 52 65 97 150 150 245 175 66 91
[28] 113 264 175 335 109
> z_calculation(mtcars$hp)
[1] -0.53509284 -0.53509284 -0.78304046
[4] -0.53509284  0.41294217 -0.60801861
[7]  1.43390296 -1.23518023 -0.75387015
[10] -0.34548584 -0.34548584  0.48586794
[13]  0.48586794  0.48586794  0.85049680
[16]  0.99634834  1.21512565 -1.17683962
[19] -1.38103178 -1.19142477 -0.72469984
[22]  0.04831332  0.04831332  1.43390296
[25]  0.41294217 -1.17683962 -0.81221077
[28] -0.49133738  1.71102089  0.41294217
[31]  2.74656682 -0.54967799
>

```

```

#Standardization (Z-Score)
z_calculation <- function(x) {
  (x - mean(x)) / sd(x)
}

z_calculation(c(1, 2, 3, 4, 5))

z_calculation(mtcars$hp)

#Min-Max Normalization (Feature Scaling)
#0-1

norm_calculation <- function(x) {
  #x - min / range
  (x - min(x)) / (max(x) - min(x))
}

norm_calculation(1:5)
#Output: 0.00 0.25 0.50 0.75 1.00

norm_calculation(mtcars$hp)

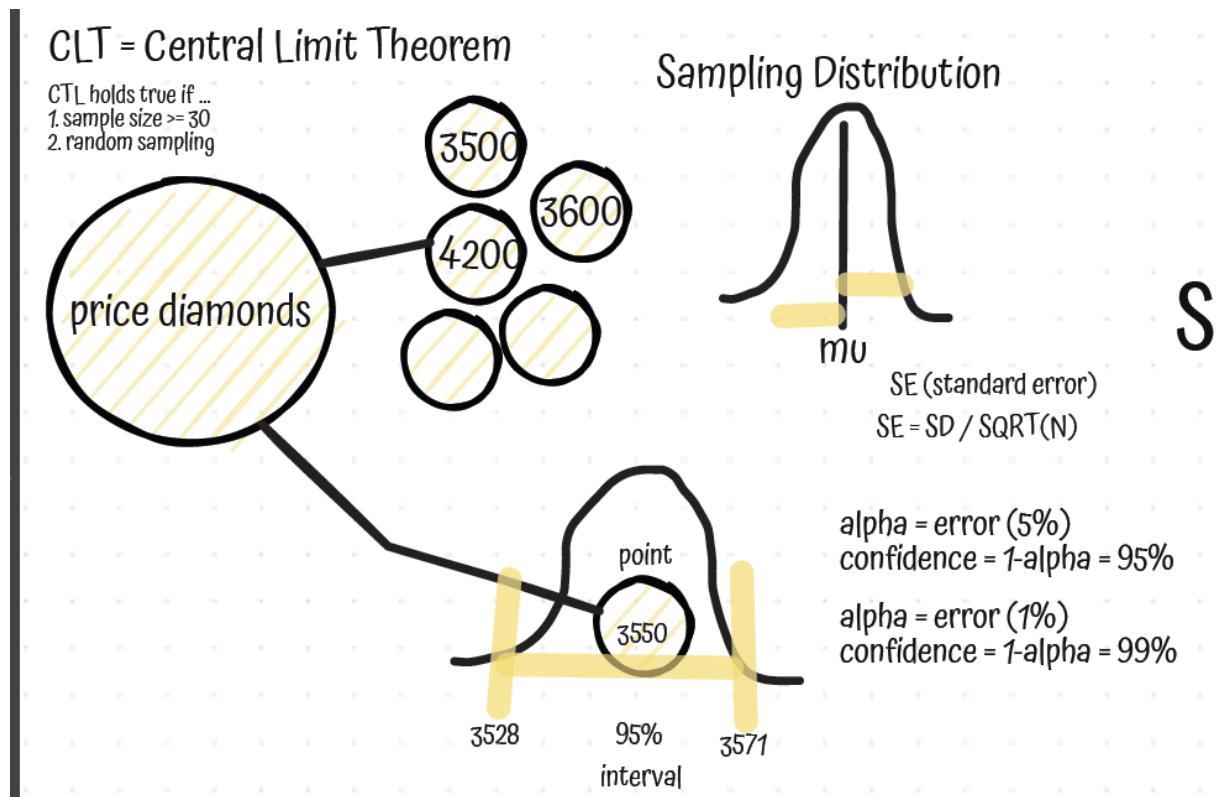
```

*ก่อนเราจะ Train ตัว Regression Model เราควรจะ Standardize Model ก่อน Fit เสมอ เพื่อให้ตัวแปรอยู่ในสเกลเดียวกันหมด ทำให้ Model ของเราระบุ Fair ขึ้น (เช่น ตัวแปรหนึ่งมีค่าอยู่ระหว่าง 0-100 แต่อีกตัวแปรมีค่าอยู่ระหว่าง 1-4 ก็ควรจะปรับให้ตัวแปรอยู่ในสเกลเดียวกันก่อน เป็นต้น)

-CLT (Central Limit Theorem) เป็น Backbone ของการทำงานทางสถิติ [Apply ได้กับแบบทุกอย่างในโลกของสถิติ เพราะในโลกความเป็นจริง เราสุ่มตัวอย่างมาแค่ครั้งเดียว แล้วคำนวณ Confidence Interval หรือช่วงความเชื่อมั่นขึ้นมาบวกกว่าค่า mean จะตกลอยู่ในช่วงระหว่าง ๆ เท่าไร และมี % ความเชื่อมั่นมากแค่ไหน]

-CLT holds true when:

1. Sample Size ≥ 30
2. Random Sampling



SE (Standard Error)

$$SE = SD / \sqrt{N}$$

Steps

1. Calculate SE (Standard Error): $=SD/\sqrt{N}$
2. Calculate ME (Margin of Error): $=T.INV.2T(0.05, 999) * SE$
3. Construct Confidence Interval: Upper Bound (mean + ME), Lower Bound (mean - ME)

* เราสามารถบีบให้ Confidence Interval แคบลงได้ด้วยการเพิ่ม Sample Size (เบื้องต้นเดียวที่ทำให้เข้าใกล้ความเป็นจริงได้มากขึ้น) แต่ต้องไม่ลืมที่จะสุ่มแบบ Random Sampling เพื่อลด Bias ให้น้อยที่สุดเท่าที่จะทำได้

สูตรหา ME ใน Excel: =CONFIDENCE.T(alpha, SD, Sample Size) * ปกติเราจะใช้ alpha = 0.05

-alpha = 0.05 → Confidence Interval = 95%

-alpha = 0.01 → Confidence Interval = 99%

วิธีที่ง่ายกว่า: ใช้ Analysis Toolpak (เลือก Descriptive Statistics และติ๊ก Summary Statistics และ Confidence Level for Mean = 95%)

| Height | | | |
|--|--------------------|---------|--------------------------|
| 168 | mean | 168.80 | =AVERAGE(A2:A6) |
| 170 | sd | 3.96 | =STDEV.S(A2:A6) |
| 175 | se | 1.77 | =D3/SQRT(5) |
| 165 | me | 4.92 | =T.INV.2T(0.05,4)*D4 |
| 166 | me (confidence.t) | 4.92 | =CONFIDENCE.T(0.05,D3,5) |
| | lower | 163.88 | =D2-D5 |
| | upper | 173.72 | =D2+D5 |
| <hr/> | | | |
| <i>Column1</i> | | | |
| <hr/> | | | |
| | Mean | 168.8 | |
| | Standard Error | 1.772 | |
| | Median | 168 | |
| | Mode | #N/A | |
| | Standard Deviation | 3.96232 | |
| | Sample Variance | 15.7 | |
| | Kurtosis | 0.87509 | |
| | Skewness | 1.08988 | |
| | Range | 10 | |
| | Minimum | 165 | |
| | Maximum | 175 | |
| | Sum | 844 | |
| | Count | 5 | |
| <hr/> | | | |
| Confidence Level(95.0%) 4.91987 Margin Error | | | |

-เราไม่ควรใช้ Point Estimate เพราะว่าในการสุ่มตัวอย่างแต่ละครั้ง ค่า Point Estimate ก็จะเปลี่ยนแปลงไปเรื่อยๆ ไม่แน่นอน การสร้างช่วงความเชื่อมั่นขึ้นมาจะเป็นวิธีที่ดีกว่าในการรับรู้ค่า Mean เพราะโอกาสที่ค่า Mean ใน การสุ่มตัวอย่างจะผิดไปจากช่วงที่เราสร้างไว้จะมีน้อยมาก (เช่น Confidence Level 95% ก็จะมีโอกาสทางค่า Mean ผิดแค่ 5% เป็นต้น)

```
#Confidence interval for HP
mean(mtcars$hp); sd(mtcars$hp)

sd_sample <- function(x) {
  sqrt(sum(x - mean(x)) ** 2) / (length(x) - 1)
}

sd_population <- function(x) {
  sqrt(sum(x - mean(x)) ** 2) / (length(x))
}

sd_sample(mtcars$hp)
sd_population(mtcars$hp)

#t.test
t.test(mtcars$hp)
```

-ตอบให้กับว่า ควรจะใช้ Confidence Interval = แบบจะทุก Moment

-ความหมายของ Confidence Interval: สมมติว่า Confidence = 95% แล้ว ถ้าเราทำ Survey 20 ครั้ง จะมี 1 ครั้งที่ค่าที่เก็บได้จะไม่ตรงกับช่วง Confidence Interval ที่สร้างไว้ (Error)

-Highest Precision = CI 90% (เพราะว่า ยิ่งความมั่นใจยะ = เก็บข้อมูลยะจะบีบช่วงกว้าง)

*ความแม่นยำ ≠ ความถูกต้อง การที่ช่วงมีความแม่นยำมากก็ยังมีโอกาสที่ข้อมูลจะผิดมาก

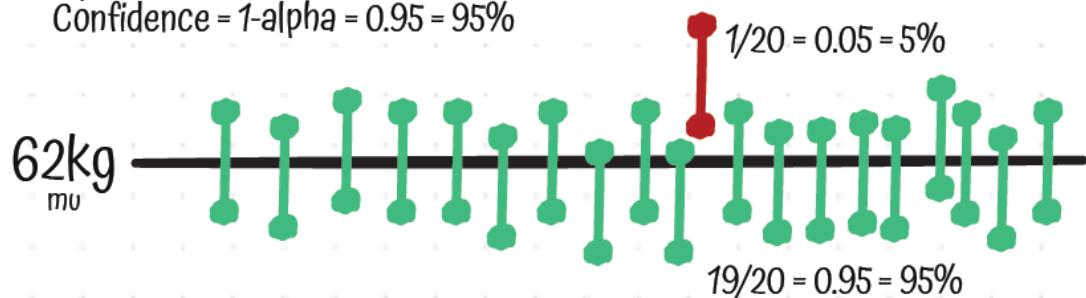
Confidence Interval

$$\alpha = 0.05 = 5\%$$

$$\text{Confidence} = 1 - \alpha = 0.95 = 95\%$$

$$1/20 = 0.05 = 5\%$$

$$19/20 = 0.95 = 95\%$$



CI 90%, 95%, 99%

Highest Precision



| | A | B | C | D | E | F |
|---|-----------|--------|----------|----------|-------|-------|
| 1 | SAT_SCORE | | ME | | Lower | Upper |
| 2 | 800 | CI 90% | 71.85304 | | 734 | 878 |
| 3 | 850 | CI 95% | 93.57897 | | 712 | 900 |
| 4 | 700 | CI 99% | 155.1792 | | 651 | 961 |
| 5 | 780 | | | | | |
| 6 | 900 | mean | | 806 | | |
| 7 | | sd | | 75.36577 | | |

Inferential Statistics

1. Comparison (การเปรียบเทียบ)
2. Association (การหาความสัมพันธ์)

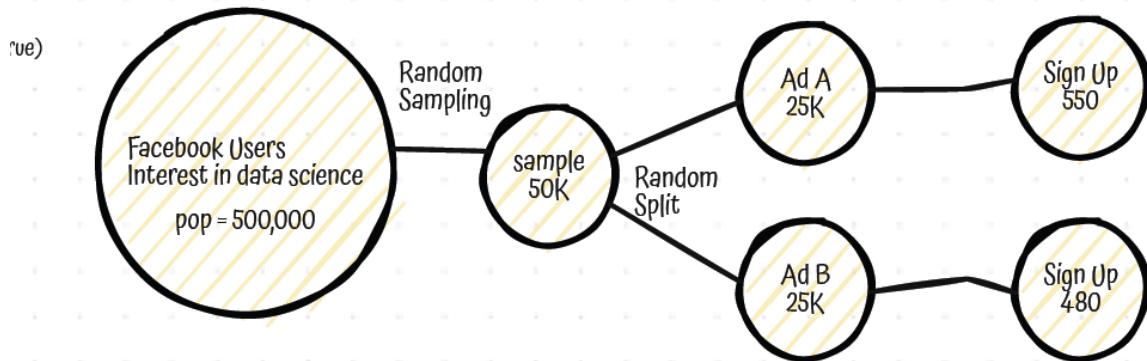
3. Prediction (การทำนายผล)

What is hypothesis testing and p-value?

AB Testing

AB Testing

นัยสำคัญทางสถิติ แปลว่า ข้อมูลที่เรารีบัตงหน้าไม่ใช่เรื่องฟลักๆ



Hypothesis

$$H_0 \text{ (null)}: \text{average sign up } A = B$$

$$H_a \text{ (alternative)}: \text{average sign up } A \neq B$$

Hypothesis Two-Tailed Test

$$H_0 \text{ (null)}: \text{average sign up } A - B = 0$$

$$H_a \text{ (alternative)}: \text{average sign up } A - B \neq 0$$

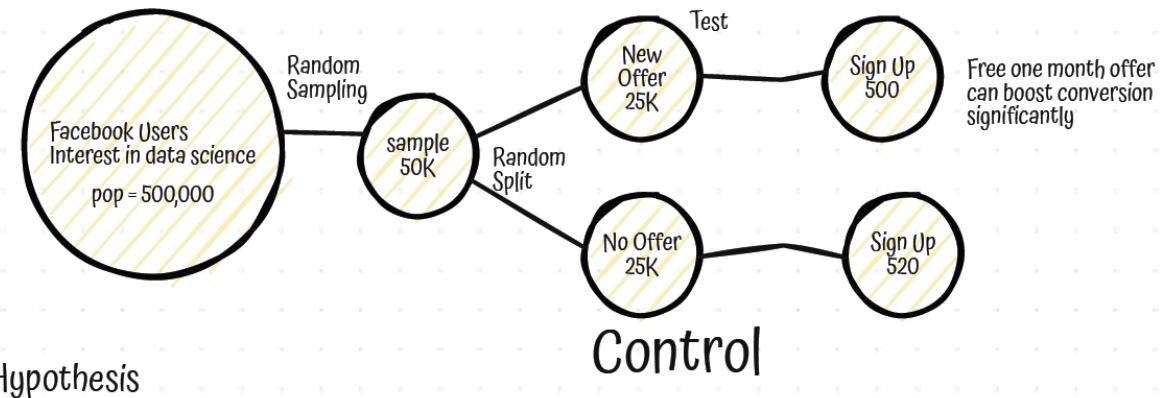
ถึงแม้ว่าเราทำ Experiment ไปแล้วครึ่งหนึ่ง แต่รูปที่เห็นข้างบนคือ Point Estimate ดังนั้นเราจะต้องหาวิธีก่อให้เกิดข้อมูลที่เราได้มาบันทึก ไม่ได้มาแบบฟลุค ๆ เลยต้องทำ Hypothesis testing เพื่อพิสูจน์หาความจริง

-null = 0, alternative $\neq 0$

RCT (Randomized Control Trial)

Randomized Control Trial - RCT

นัยสำคัญทางสถิติ แปลว่า ข้อมูลที่เรารีบัตต์ต้องหน้าไม่ใช่เรื่องฟลีค่า

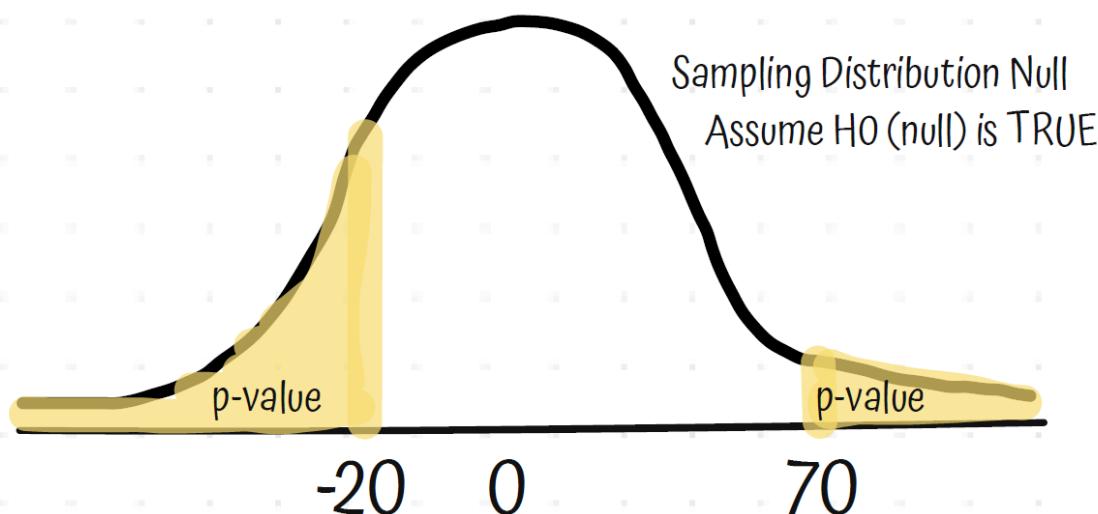


$$H_0 \text{ (null): } \text{average sign up test} - \text{control} = 0$$
$$H_a \text{ (alternative): } \text{average sign up test} - \text{control} \neq 0$$

- How to test significance?
- Reject H_0 if $p\text{-value} \leq \alpha$ (5%)

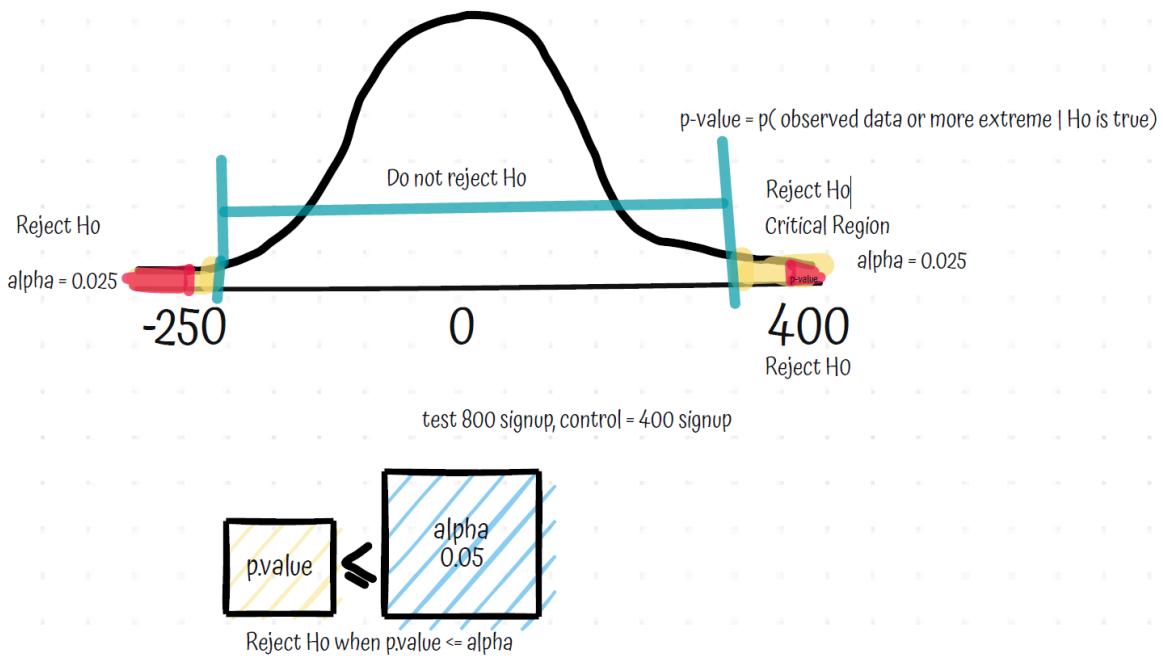
ความแตกต่างระหว่าง AB Testing และ RCT: ในการทำโน้มน้าว AB Testing จะเป็นการเปรียบเทียบประสิทธิภาพของ Ads 2 ตัวว่าตัวไหนได้ผลลัพธ์ดีกว่ากัน (Ads ไหนดีที่สุด ?) แต่ RCT จะเป็นการทดสอบว่า มี Ads หรือไม่มี Ads อย่างไหนได้ผลลัพธ์ดีกว่ากัน (Ads มีประสิทธิภาพหรือไม่เกียบกับการไม่มี Ads?)

How to test significance? [Reject H_0 if $p\text{-value} \leq \alpha$ (5%)]



$p\text{-value} = \text{พื้นที่ใต้กราฟ สบู่ติ } \text{test} - \text{control} = 70 \rightarrow p\text{-value} = \text{พื้นที่ใต้กราฟตั้งแต่ } 70 \text{ ขึ้นไป}$
 หรือ $\text{test} - \text{control} = -20 \rightarrow p\text{-value} = \text{พื้นที่ใต้กราฟตั้งแต่ } -20 \text{ ลงมา}$

*ปกติเราจะ Set alpha ไว้ที่ 5% เป็น default



*เราจะ Set โซน Critical Region ขึ้นมา หาก difference ของ test และ control ($\text{test} - \text{control}$) มาตกลอยู่ในช่วงสีเหลืองนี้ เราจะ Reject H_0 กันที่ และสรุปว่า Test มีนัยยะสำคัญจริง ๆ = [การทำ Offer ใน Test Group ได้ผลลัพธ์ที่ดีกว่า Control Group]

ยกตัวอย่างเช่น การที่คน sign up ใน Test Group มากกว่า Control Group ถึง 400 คน ไม่น่าจะเป็นเรื่องฟลุค แล้ว ตกอยู่ในช่วง Critical Region และ $p\text{-value}$ น้อยกว่า α เราจะ reject H_0 กันที่ หรือถ้าคน sign up ใน Test Group น้อยกว่า Control Group ถึง 250 คน เราอาจจะ reject H_0 กันที่เช่นกัน เป็นต้น ยิ่ง difference ตกไกลจากค่า 0 มากเท่าไรก็จะมีโอกาส Reject H_0 ได้ง่ายขึ้นเท่านั้น

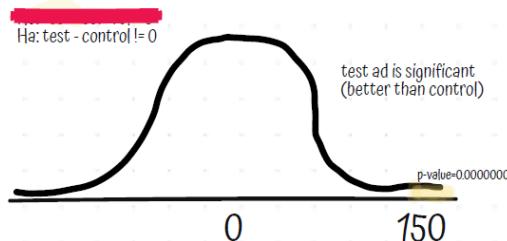
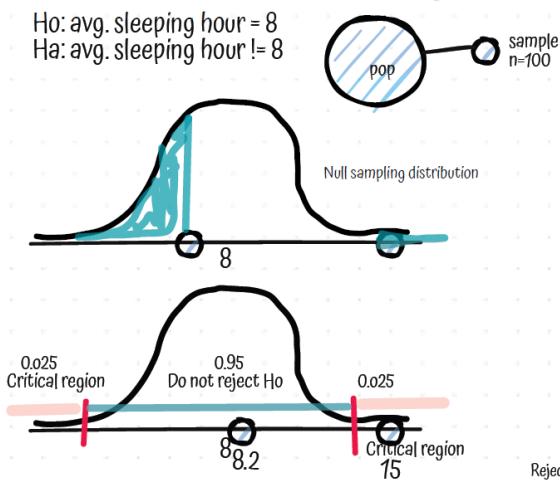
*****Key สำคัญ: ถ้าค่า $p\text{-value} \leq \alpha$ ให้ Reject H_0 กันที่**

ตัวอย่างการใช้งาน: คนไทยนอนวันละ 8 ชั่วโมงจริงหรือไม่?

H_0 = คนไทยนอนวันละ 8 ชั่วโมง

Hypothesis testing

$H_0: \text{avg. sleeping hour} = 8$
 $H_a: \text{avg. sleeping hour} \neq 8$



p-value = 0.051
 $\alpha = 0.05$

Do not reject H_0

สรุป: **ไม่ Reject H_0 เพราะ $p\text{-value} > \alpha$**

*การเข้าใจผลลัพธ์ทางสถิติเป็นเรื่องสำคัญมาก เพราะเราจะรู้ได้ว่าการทดลองของเราเมื่อผลอย่างมีนัยยะสำคัญทางสถิติหรือไม่

การวัดผลค่า p-value ใน R:

```
#Create Promo A VS Promo B
promo_a <- rnorm(100, mean = 550, sd = 10)
promo_b <- rnorm(100, mean = 400, sd = 8)

#H0: Promo A - Promo B = 0
#H1: Promo A - Promo B != 0
result <- t.test(promo_a, promo_b, alternative = "two.sided")

ifelse(result$p.value <= 0.05, "Significant", "Not Significant")
```

Linear Regression ใน Excel: แบ่งเป็น Score และ Group (เช่น 1 = test, 0 = control) และใช้ Linear Regression ใน Data Analysis (Analysis Toolpak) ก็จะได้ค่าตอบอย่างรวดเร็ว (ดู p-value ได้เหมือนกับ t.test แค่ต้อง Format Data บิดหน่อย)

*ในโลกความเป็นจริง เราสามารถทำงานกับข้อมูลได้หลากหลายแบบ ให้เลือกวิธีที่เหมาะสมที่สุดในแต่ละเคส

Summary:

Summary Key Learning

- Z Score (Standardization vs. Normalization)
- CLT
- Confidence Interval
- Point vs. Interval Estimate
- $\alpha + \text{confidence level} = 100\%$
- CI more precision 90% better than 99%
- CI more precision \rightarrow Increase sample size N
- Hypothesis testing p-value
- $p\text{-value} = p(\text{observed data or more extreme} \mid H_0 \text{ is true})$
- Reject H_0 if $p\text{-value} \leq \alpha$ ($\alpha 5\%$)
- AB test vs. RCT / Linear Regression

*AB test และ Linear Regression จะได้ใช้บ่อยมากในการทำงานจริง

- ถูเรื่อง p-value และ Linear Regression เพิ่มเติมได้ที่:

 [Statistics 101 \(Intro to Statistics\)](#)

Part 3

Statistical Conclusions

สรุปผลการทดลอง Hypothesis Testing ในทางสถิติ (Frequentist Approach) สามารถทำได้
สามวิธี ได้ผลลัพธ์เหมือนกัน

- Critical Region

- p-value

- Confidence Interval

ถ้าข้อมูลที่เราเก็บมา (Test Statistics) ไปตกอยู่ในพื้นที่ Critical Region

- p-value จะน้อยกว่า 0.05 ($p\text{-value} < 0.05$)

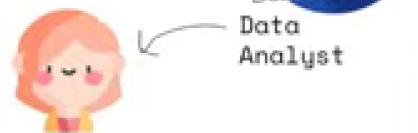
- Confidence Interval จะไม่เก็บค่า Null (H_0)



ตัวอย่างเช่น ถ้าค่า null (H_0) = **260g** เมื่อตัวอย่างขั้นมเลย์ที่แอดสอนในคลาส >> (observed data) เราเก็บข้อมูลได้ค่าเฉลี่ยน้ำหนักขันมที่ **330g** ไปตกใน critical region, $p\text{-value} \leq 0.05$, 95% CI [320, 340]

ความรู้ทางด้านสถิติที่ Data Analyst ต้องรู้

Essential Knowledge



- Descriptive
- Inferential
 - Hypothesis Test/ Basic Model
 - T-Test
 - F-Test
 - Correlation
 - Linear Regression
 - Logistic Regression

- $p\text{-value}$ เป็นหนึ่งใน 3 วิธีที่ใช้ในการทดสอบสมมติฐานทางด้านสถิติ (Hypothesis Test)

1. Critical Region
2. $p\text{-value}$
3. Confidence Interval

- $p\text{-value}$ มีจุดอ่อน คือ Sensitive to sample size

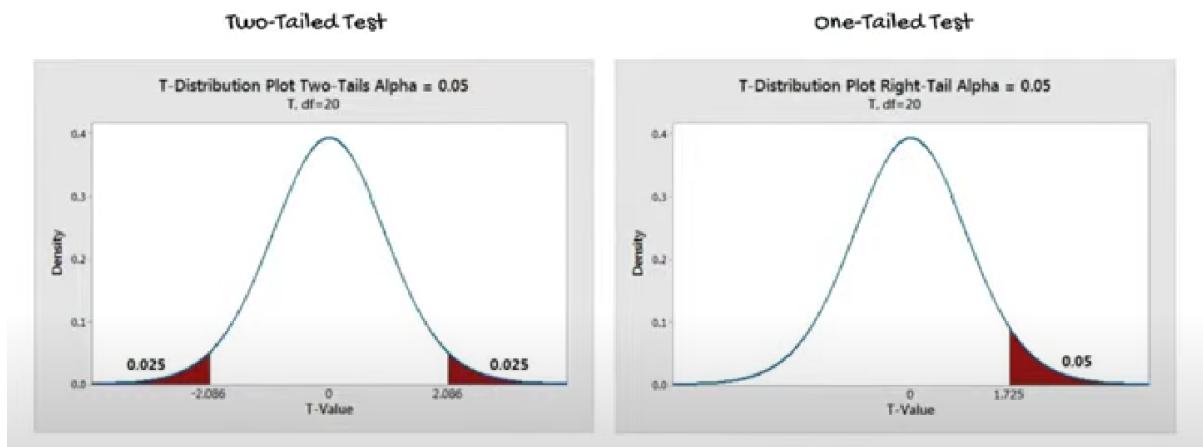
-เราไม่สามารถเข้าถึงประชากรก็ต้องหาตัวอย่างมาเป็นตัวแทนของประชากรที่เราสนใจ

Hypothesis Test

1. One-tailed Test
2. Two-tailed Test *Default

*การเลือกใช้ One-tailed กับ Two-tailed ส่งผลต่อผลลัพธ์ได้

Alpha (False Positive): ค่า Error ที่เรายอมรับได้ มีค่า Default ที่ 5%



-ใช้แบบ Two-tailed จะ Conservative มากกว่า

Confusion Matrix:

| Things can go wrong | Reject H0 | Do not reject H0 |
|---------------------|-------------------------------|------------------------------|
| H0 is TRUE | Type 1 [False Positive] Alpha | True Negative |
| H0 is FALSE | True Positive | Type 2 [False Negative] Beta |

Things can go wrong

| | Reject H ₀ | Do not reject H ₀ |
|-------------------------|----------------------------|------------------------------|
| H ₀ is TRUE | Type 1 <False Positive> | True Negative |
| H ₀ is FALSE | True Positive | Type 2 <False Negative> |

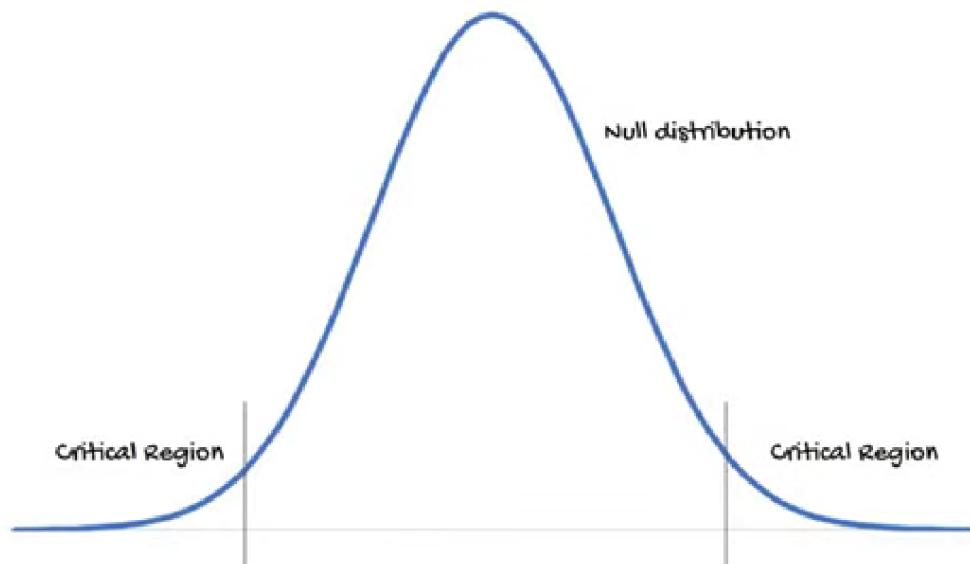
-Reject H₀ \Rightarrow Significant

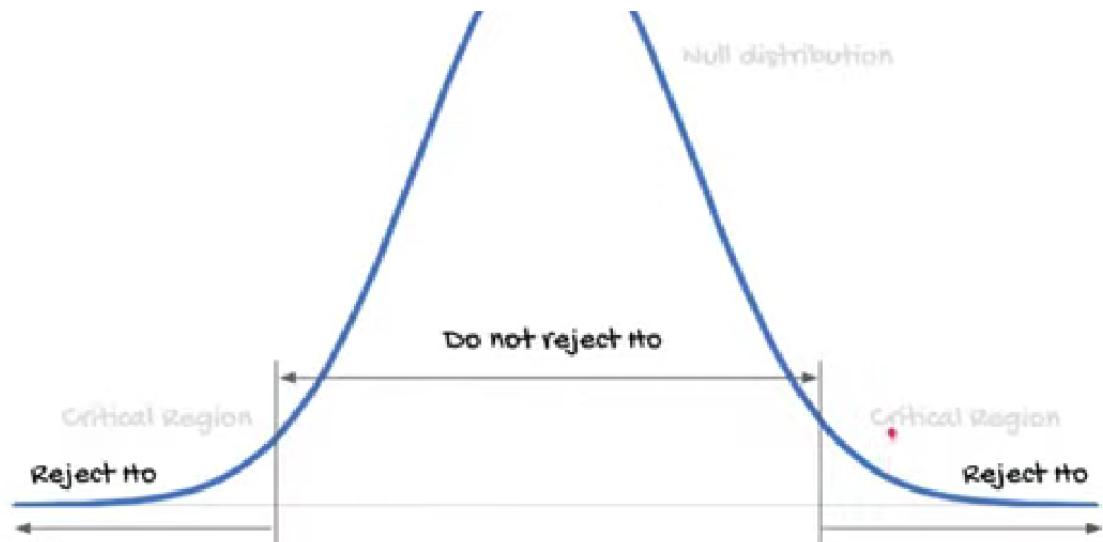
-อย่าเชื่อผลลัพธ์ทางสถิติมากเกินไป เพราะมีโอกาสที่ค่าทางสถิติจะผิด

* $1 - \alpha$ = Confidence Level

Critical Region

1. Critical Region 🔥





p-value

2. p-value 🔥



$p(\text{observed data or more extreme} \mid H_0 \text{ is TRUE})$

Assuming H_0 is true

Given

probability

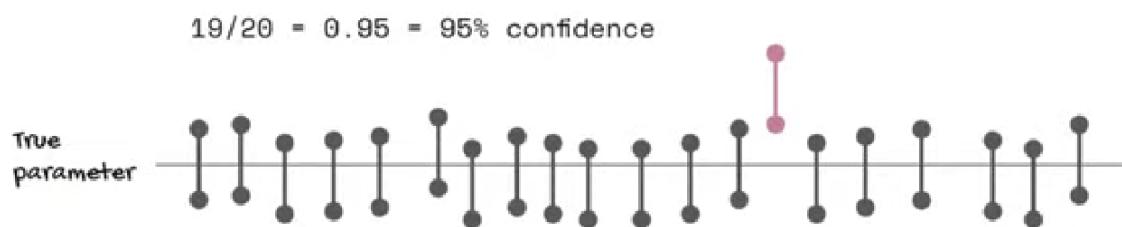
p = Propability (ความน่าจะเป็น)

-  DataRockie p-value ของแบบ two-tailed จะต้องคูณสองนะคับ ถึงจะได้ค่าถูกต้อง
-  DataRockie ปกติโปรแกรมสติติจะคูณสองให้เราเรียบง่ายไว้
-  DataRockie เพราะชาร์ทเรามันสมมาตร normal distribution (sampling distribution)

p-value: The likelihood of getting test findings that are at least as extreme as the result that was actually observed, assuming that the null hypothesis is true.

Confidence Interval *Recommended Method

3. Confidence Interval 🔥



-ถ้า Confidence Interval เก็บค่า Null เราจะ Fail to reject H₀ แต่ถ้า Confidence Interval ไม่เก็บค่า Null เราจะ Reject H₀

3. Confidence Interval 🔥



-Confidence Interval สามารถทำให้เราเห็น Uncertainty ได้ ทำให้เป็น Method ที่น่าใช้กว่าการใช้ p-value (ช่วงทำให้เราเห็น Insight 多得多)

Conclusion

The same conclusion

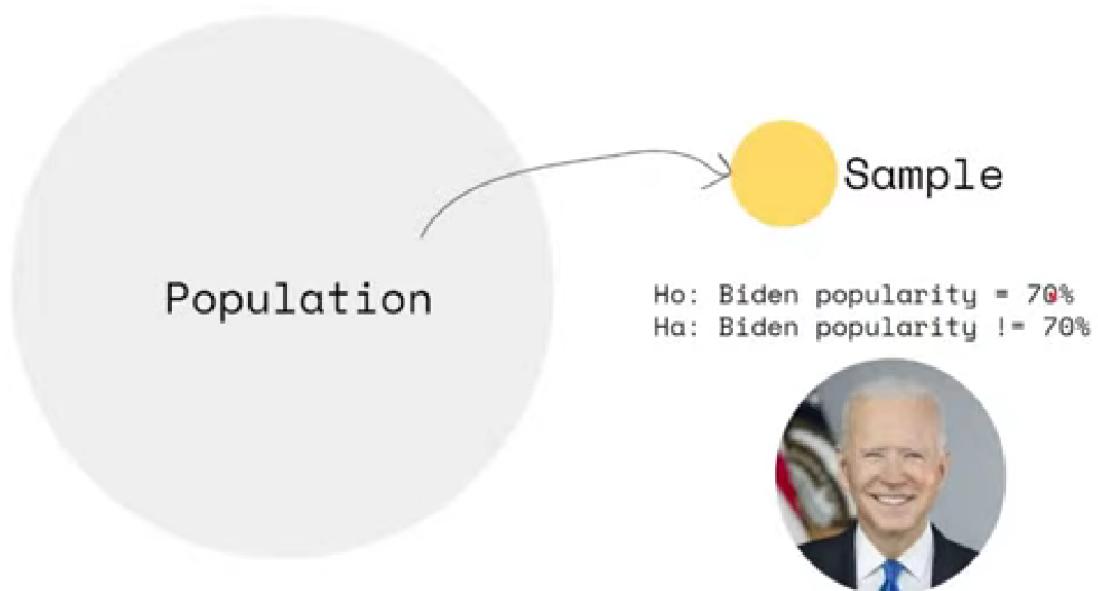
- If our data falls in critical region
 - p-value ≤ 0.05
 - 95% CI will not contain null value



Limitation of p-value

-p-value is sensitive to sample size (ยิ่ง sample size มาก แนวโน้มของ p-value ก็จะยิ่งลดลง)

-บางที Hypothesis ของเราราจอาจผิดตั้งแต่แรก ผิดตั้งแต่ตอนที่เราตั้งแล้ว เช่น:



^{^Popularity = 70% แน่ ๆ แต่ในความเป็นจริงแล้ว มันไม่ใช่}

ต่อให้ Hypothesis ของเราราจผิดตั้งแต่แรก เราจะ Test Significance และหาค่า p-value ไปเพื่ออะไร ?

- เรายากจะเข้าใจ Population มากขึ้นจากการสุ่ม Sample
- Hypothesis test tests whether the difference exist, but does not test the impact (effect size) of the difference.
- * คำความที่คุณถาม ไม่ใช่การถามว่า ต่างกันหรือไม่ ? แต่คุณถามว่า ต่างกันมากน้อยแค่ไหน ? (หา effect size ไม่ใช่ existence of difference)
- * เก็บข้อมูลให้เยอะที่สุดเท่าที่จะทำได้ ในงบและเวลาที่จำกัด

- ยิ่ง n มากขึ้น SE ก็จะยิ่งลดลง เพราะเราเน้นใจในผลลัพธ์ของเรามากขึ้น
- * Big Data ต้องระวังเรื่อง Spurious Correlation (ตัวแปรดูเหมือนเกี่ยวข้องกันแต่ที่จริงแล้วไม่เกี่ยว กับ)

The Earth is Round ($P < 0.05$)

- เป็นไปไม่ได้ที่โลกเราจะ 'กลม'
- ถ้าเราตั้งให้ $H_0 = \text{โลกกลม}$ และ $H_1 = \text{โลกไม่กลม}$ ก็จะต้อง Reject H_0
- * การที่เราไม่ Reject H_0 ในตอนนี้ ไม่ได้หมายความว่าเราจะไม่ Reject H_0 ในอนาคต

[H_0 ถูกตั้งไว้เพื่อรอวัน Reject (Falsifiable) เราจึงไม่ใช้คำว่า 'ยอมรับ' (Accept)]

- Karl Popper [Falsification]

Key Takeaways:

Key Takeaways

- All methods produce the same result
 - Critical region
 - p-value
 - Confidence interval
- Confidence interval is highly recommended in modern research
- p-value is sensitive to sample size

