



Essential Stats for Data Analyst

Part I

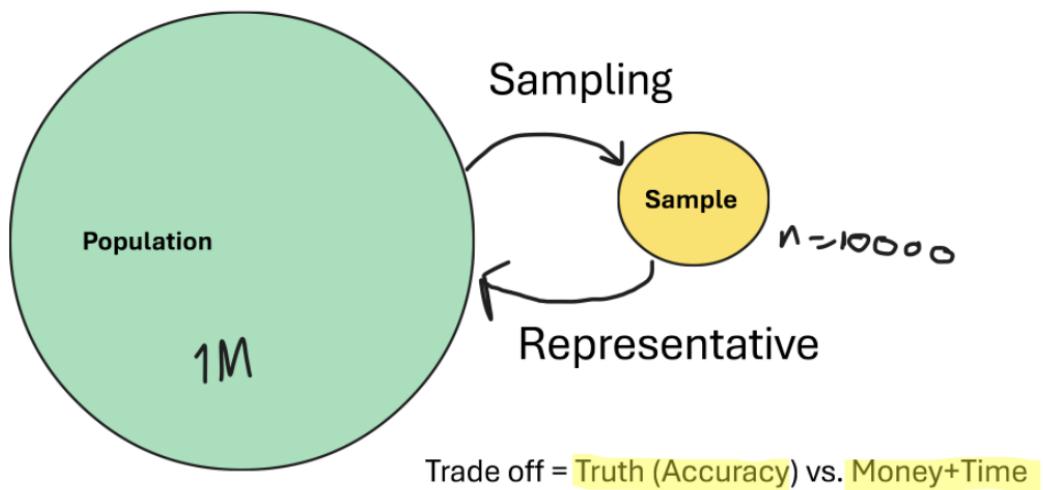
- What and why of statistic
- Population vs. Sample
- Sampling
- Descriptive statistics
- How to compute Z-score
- Normal Distribution

What and why of statistic



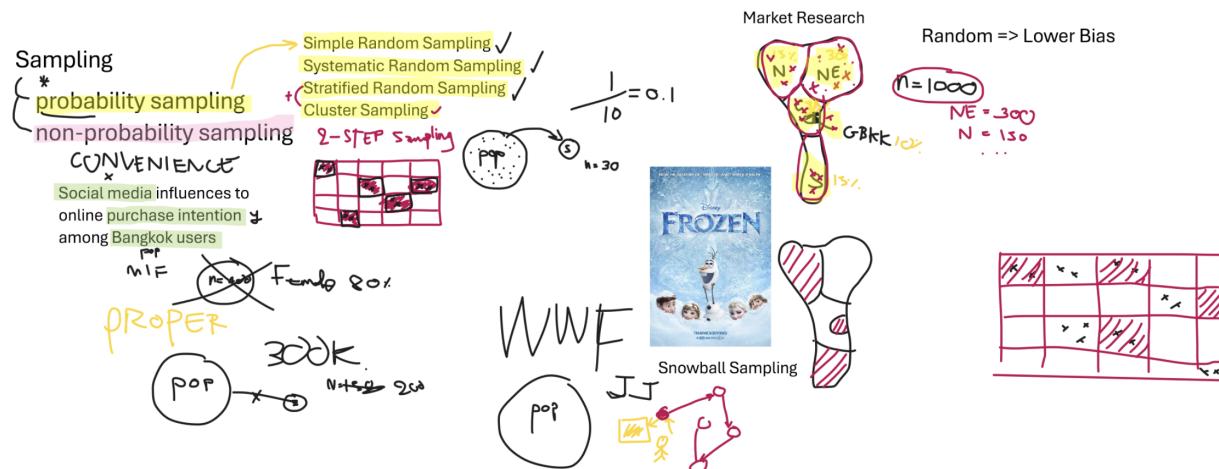
Stats = collect data, analyze data to make better decision. We want to know/understand the truth.

Population vs. Sample



Trade คือการคิดระหว่างความแม่นยำที่เราจะได้แลกกับเงิน ยิ่งกลุ่มตัวอย่างเยอะ มันแม่นขึ้นก็จริงแต่เงินที่ใช้ก็มากตาม

Sampling



1. Probability sampling

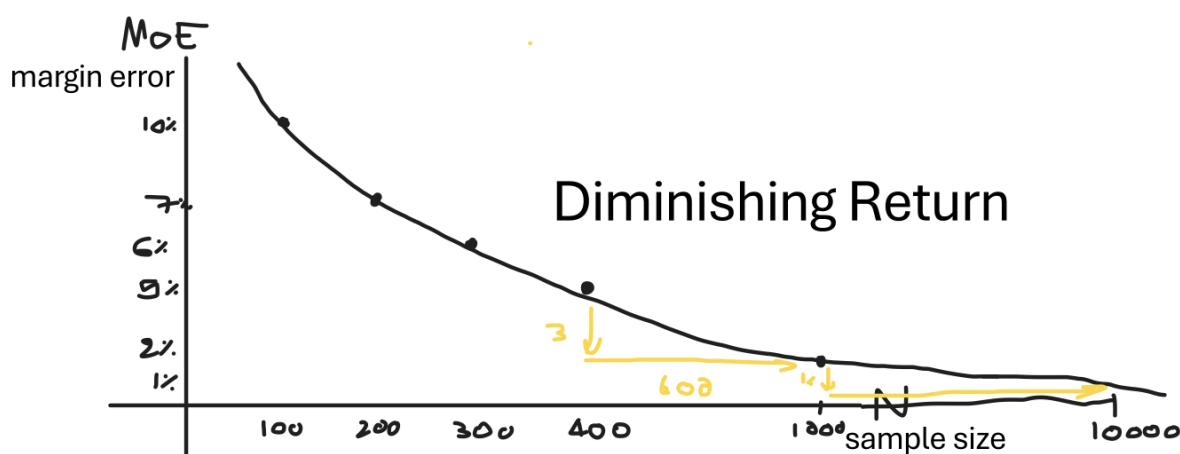
- sample random sampling คือประชากรที่เราจะสุ่มอุปกรณ์มา มีโอกาสที่จะถูกสุ่มอุปกรณ์มากทั้งหมด เก่ากัน ใน excel มีคำสั่ง `RAND()` ในการสุ่มอุปกรณ์
- systematic random sampling คือการสร้างกฎในการสุ่มอุปกรณ์ เช่น สุ่มทุกๆ 4 คน สุ่มคนแรกเสร็จไปสุ่มคนที่ห้าต่อและไปสุ่มคนที่เก้า **แต่ต้อง random ข้อมูลก่อน**
- stratified random sampling คือแบ่งประชากรเป็นกลุ่มๆ และสุ่มอย่างอุปกรณ์ในกลุ่มนั้นๆ และนำมารวบกัน
- cluster sampling คือการแบ่งประชากรเป็นกลุ่มๆ และเลือกมากลุ่มหนึ่งเพื่อกำหนดที่จะเข้าไปเก็บข้อมูลทั้งหมด

2. Non-probability sampling

- คือการสุ่มแบบ convenience การสุ่มแบบสะดวก เพราะฉะนั้นข้อมูลที่ได้มาไม่สามารถเป็นตัวแทนของประชากรจริงได้

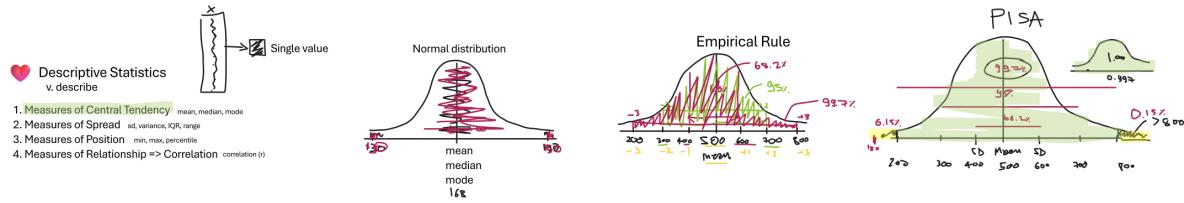
#snowball sampling คือการถามต่อไปเรื่อยๆ

#sample size



ทุกๆครั้งที่เราเพิ่ม sample size ขึ้นมา margin error จะลดลงเรื่อยๆ เรียกหลักการนี้ว่า Diminishing Return ต้องคิดเสมอว่า ทุกๆ % ที่เราได้น้ำหนักับการเพิ่ม sample size หรือป่าว สามารถลงได้ในเว็บ survey monkey

Descriptive statistics



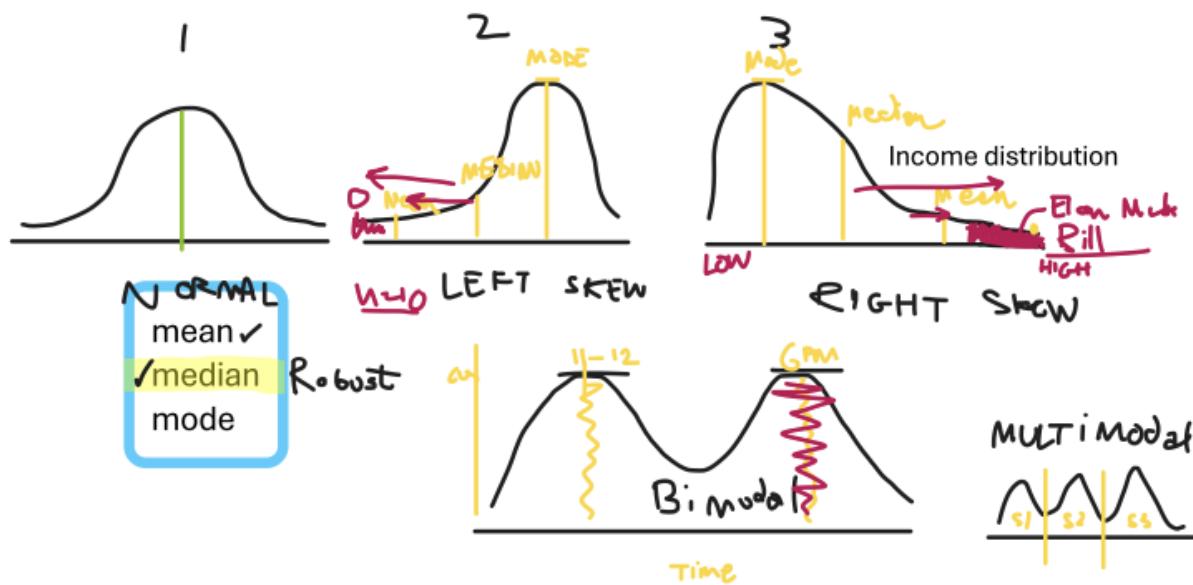
คือการอธิบายหน้าตาของข้อมูลที่เรามีว่าเป็นอย่างไร ให้เหลือแค่ 1 value หลักๆจะมี 3แบบ แต่เพิ่มมาอีกแบบหนึ่ง

- Measure of Central Tendency
- Measure of Spread
- Measure of Position
- Measure of Relationship \Rightarrow correlation

1. Measure of Central Tendency

คือการวัดค่ากลางข้อมูล (mean, median, mode)

Central Tendency



Left skew ⇒ มีค่าบางค่าที่ดึงค่า mean ให้ต่ำ ส่วน Right skew ⇒ มีค่าบางค่าที่ดึงค่า mean ให้สูง หรือกี่เราเรียกว่า outlier

Median คือค่ากลางสกัดที่กับต่อ outlier

Bimodal คือมีค่า mode ส่องตัว และถ้าเป็น multimodal คือมี mode หลายตัว

2. Measure of Spread

คือการวัดการกระจายตัวของข้อมูลเมื่อเกียบกับค่ากลาง (mean) ของข้อมูล (Standard variable, IQR, range)

Spread

* Range = Max - Min
 $= 60 - 20$
 $= 40$

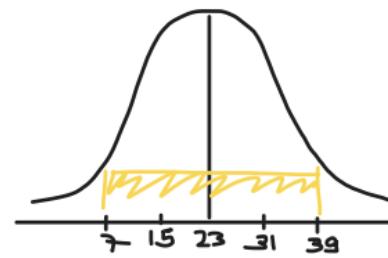
VAR | SD

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

ຮວມຢາຍ

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

ຢອງ



- Range = Max-Min គឺជាការគុណករណ៍តាមលក្ខណៈទូទៅនៃបញ្ហាបុគ្គលិក។ Range មិនមែនតាមលក្ខណៈទូទៅនៃបញ្ហាបុគ្គលិកដែលមានចំណោមខ្លះ។
- SD and Var គឺជាការគុណករណ៍តាមលក្ខណៈទូទៅនៃបញ្ហាបុគ្គលិកដែលមានចំណោមខ្លះ។



$$\text{Sample VAR} = \frac{\sum (x - \bar{x})^2}{n-1}$$

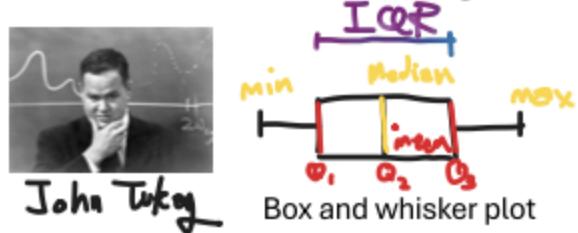
$$\text{SD} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

$x \rightarrow$

#ប៉កសតិແណះបានឱ្យប្រើប្រាស់ SD មាត្រក្នុង Var ហេតុផលគឺអនុវត្តន៍យកមុនគ្នា

- IQR (Inter Quartile Range) គឺជាការគុណករណ៍តាមលក្ខណៈទូទៅនៃបញ្ហាបុគ្គលិក។

IQR = Inter Quartile Range



3. Measure of Position

Position

min
max
percentile
quartile



- min คือค่าที่น้อยที่สุด
- max คือค่าที่มากที่สุด
- percentile คือการแบ่งเดต้าให้เป็นสองส่วน ส่วนที่มากกว่าเป็น 100 หรือ 100 % แต่อย่าลืม sort เดต้าก่อนนะ
- Quartile คือการแบ่งข้อมูลออกเป็น 4 ส่วน Q1(min), Q2, Q3, Q4(max)

#สังเกตว่า Q2 และ percentile 50th และค่า median จะมีค่าเท่ากัน

Part II

- Z-Score & Standard Normal Distribution
- Central Limit Theorem + Standard Error
- Confidence Interval
- Hypothesis Testing AB test
- What is Regression vs. Correlation

Z-Score & Standard Normal Distribution

การเปลี่ยน Normal Distribution ที่เป็นข้อมูลดิบ ให้เป็นการกระจายตัวของ Z-Score จะเรียกว่า Standard Normal Distribution หรือการทำ Normalize Data

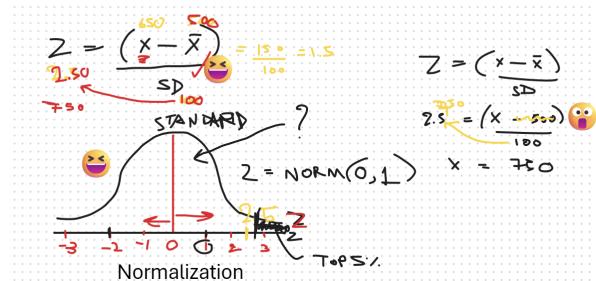
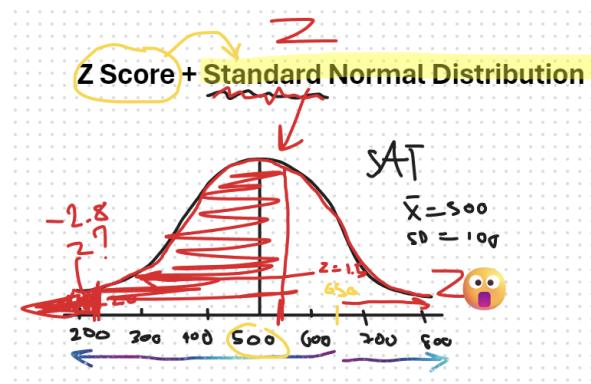


Empirical Rule คือกฎที่เป็นจริงเสมอเป็นการถูกความบ่าจะเป็นของข้อมูลว่ามีกี่ % ใน Normal Distribution

โดย ข้อมูลที่อยู่ระหว่าง -1 และ +1 SD จะมีข้อมูลอยู่ตຽรงนั้น 68.2 %

ข้อมูลที่อยู่ระหว่าง -2 และ +2 SD จะมีข้อมูลอยู่ตຽรงนั้น 95 %

ข้อมูลที่อยู่ระหว่าง -3 และ +3 SD จะมีข้อมูลอยู่ตຽรงนั้น 99.7 %



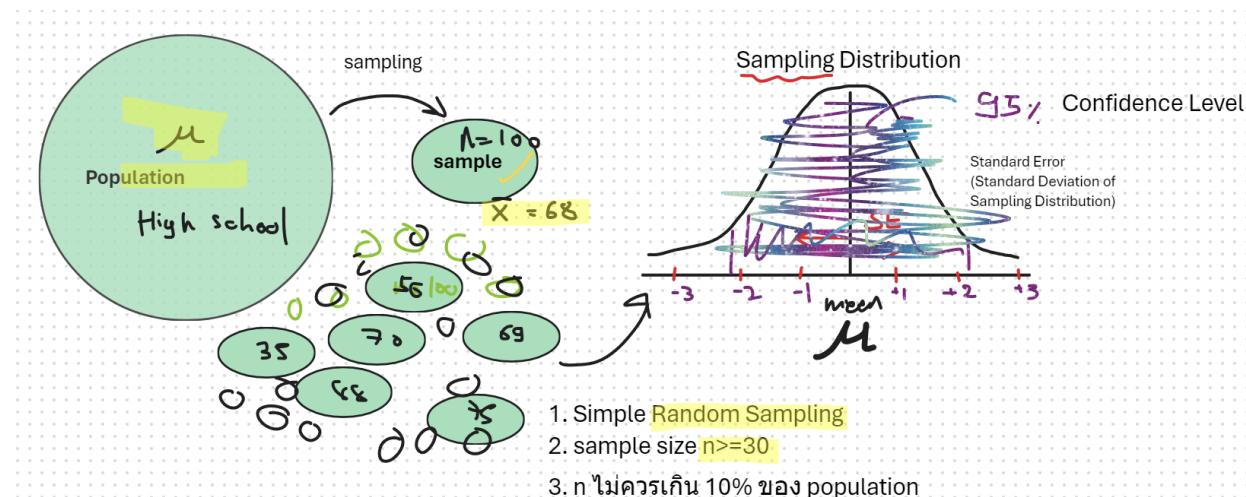
และสูตรการเปลี่ยนข้อมูลดิบ ให้เป็น Z-Score แสดงดังรูปด้านบน

แล้วเราเป็นให้เป็น Z-Score ทำไม... โดยนิยามของ Z-Score คือ Normal Dis ก็มีค่า mean เป็น 0 และมี SD เป็น 1

จนถ้าการที่เราเปลี่ยนเป็น Z-Score เพราะต้องการทราบว่า มีครึ่งหนึ่งมากหรือน้อยกว่าเราเท่าไหร่ โดยหาพื้นที่ใต้กราฟ ซึ่งการหาพื้นที่ใต้กราฟทางชัยจาก การเปิดตาราง หรือใช้สูตร

`NORM.S.DIST(ค่า Z-Score, TRUE)` หรือ `NORM.DIST(X, X_bar, SD, TRUE)` ใน excel และถ้าจะหาพื้นที่ใต้กราฟทางขวา ก็แค่เอาไป ลบ 1 เพราะพื้นที่ใต้กราฟ จะมีค่าเท่ากับ 1

Central Limit Theorem + Standard Error

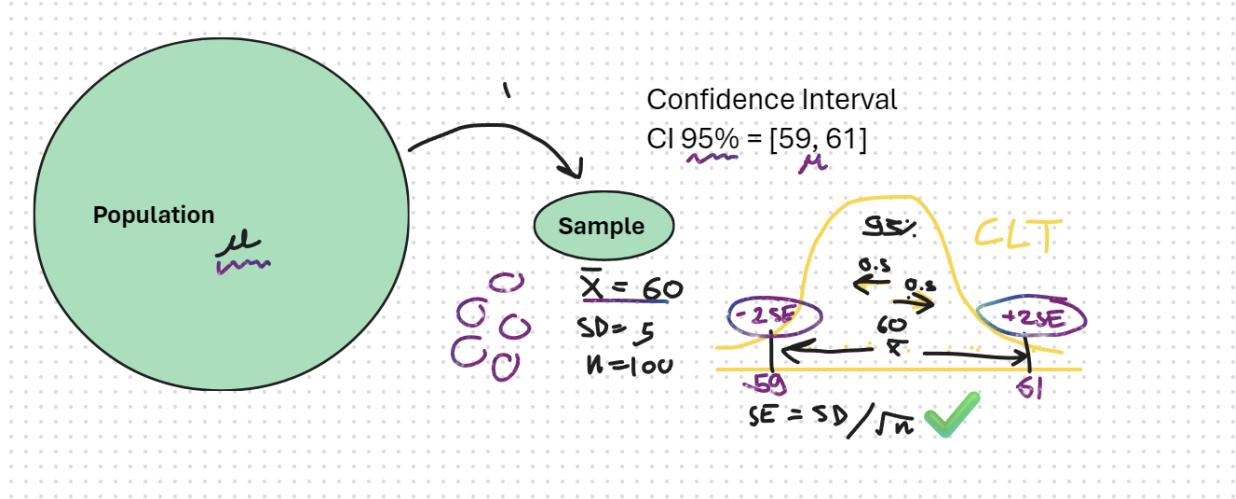


Central Limit Theorem คือภาพในหัวที่เราสุ่มข้อมูลออกมาเรื่อยๆ เป็น 1000 ครั้งหรือมากกว่านั้น และจะได้ค่าเฉลี่ยแต่ละครั้งของการสุ่มซ้ำที่ไม่เท่ากัน คำถามคือแล้วค่าเฉลี่ยตัวไหนที่นำไปใช้ได้จริง?

เราต้องนำค่าเฉลี่ยทั้งหมดจากการสุ่มซ้ำมาพล็อตกราฟ (จะเป็นตัวเลข หรือ % หรือค่าอื่นๆได้หมด) ผลที่ได้คือจะได้กราฟ Normal Distribution และซึ่งของมันจะเปลี่ยนเป็น Sampling Distribution และค่าเฉลี่ยของกราฟนี้ ก็คือค่ามัธย ส่วน SD จะถูกเรียกว่า Standard Error (SE) แทน และ ± 2 SE จะมีช่วงความมั่นใจอยู่ 95 %

แต่มีข้อแม้มสามข้อ ตามรูปด้านบน คือ 1. การสุ่มต้องเป็นแบบ Simple Random Sampling 2. จำนวนที่สุ่มต้องมากกว่าหรือเท่ากับ 30 3. จำนวนที่สุ่มไม่ควรเกิน 10 % ของ population

แต่ในความเป็นจริงเราสุ่มแค่ครั้งเดียว



เพราจะนั่นหลักการใช้ Central limit คือ เราจะนำค่าเฉลี่ยมา และ assume ว่า เราได้สุ่มมาเป็น 1000 ครั้ง หรือมากกว่านั้นแล้วนะ จากนั้นเราก็วาดกราฟ normal distribution ครอบ และคำนวณหาค่า SE ตามสูตรดังรูปด้านบน

วิธีการสรุปคือ เราบันใจ 95 % (โดยคิดจาก $+2SE$ ซึ่งอาจจายังไม่เป๊ะๆ) ว่า ค่ามิว หรือค่าเฉลี่ยประชากรจะอยู่ในช่วง $[59, 61]$ ตามรูปด้านบน

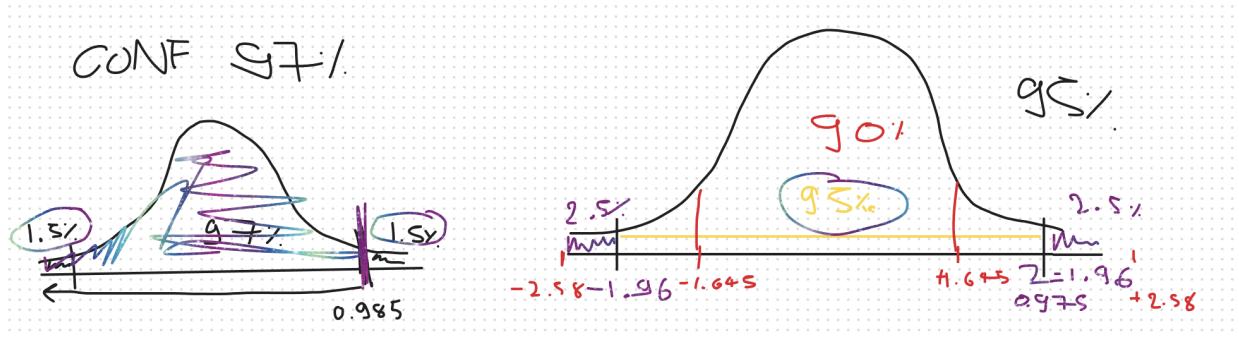
แต่ถ้าเราอยากได้ช่วงความเชื่อมั่น 95% (คือถ้าสุ่ม 100 ครั้ง จะได้ผลแบบเดิม ถึง 95 ครั้ง)
แบบเป๊ะๆ...

เราจะต้องคำนวณ Margin Error หรือ Confidence Level = $SE * T$ ใน google sheet หาจาก
`=CONFIDENCE.NORM(0.05, SD, N)`

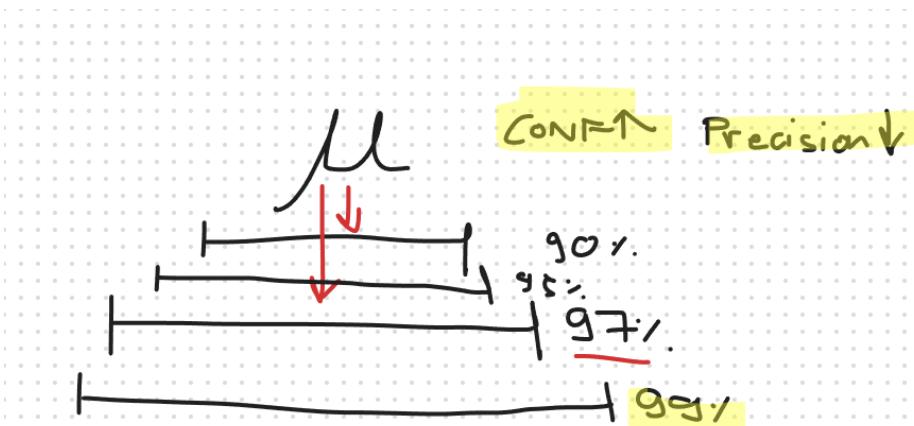
ซึ่ง T สามารถคำนวณใน excel ได้เลย จากสูตร `=T.INV.2T(0.05, n-1)`

จากนั้นถ้าอยากได้ช่วงให้สร้าง `Lower = mean - margin error, Upper = mean + margin error`

`X bar +- margin error = CI`



การหาค่า Z ที่ค่าความเชื่อถือ อย่างลึกลับค่าทางซ้ายของกราฟด้วย เช่น 95% จะคิดที่ $95\% + 2.5\%$



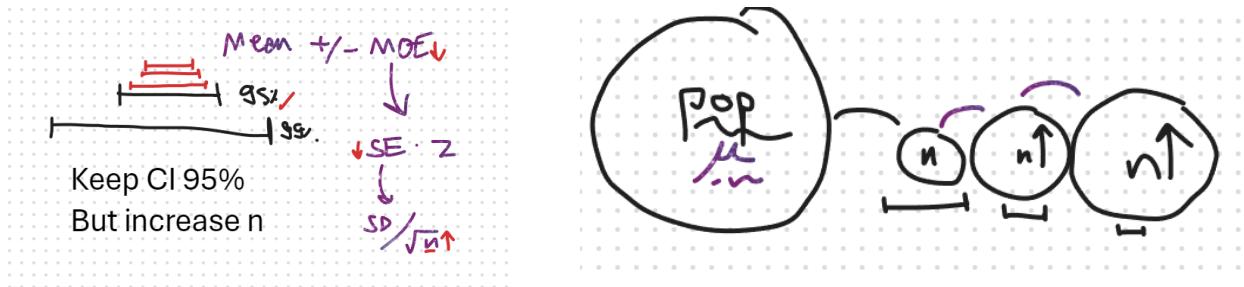
Forecast Sales.

- ① 95% [100, 120]
- ② 99% [80, 140]

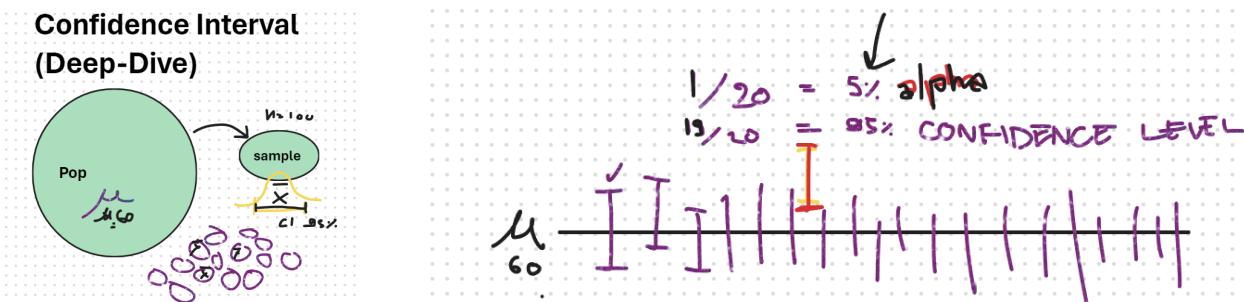
สังเกตว่ายิ่งค่าความมั่นใจเพิ่มค่า Precision จะลดลง (Keep default 95%)

แต่ถ้าเราอยากจะได้ช่วงที่แคบลงและใช้ CI อยู่ที่ 95% เนื้องเดิมเราควรทำอย่างไรดี....

คำตอบคือเพิ่ม n หรือ sample size



Confidence Interval



ถ้าเราสุ่มไปเรื่อยๆ ช่วงความเชื่อมั่นจะเปลี่ยนไปเรื่อยๆ สมมุติเราสุ่ม 20 ครั้ง ยอมพลาดได้หนึ่งครั้ง(หมายความว่าค่าจริง ๆ ของมันจะไม่ได้อยู่ในช่วง) และมันใจในผลลัพธ์ได้ 19 ช่วง คือ confidence level กี่ 95% -Fisher concept

$$\text{Confidence level} + \text{Alpha} = 1$$

Hypothesis Testing (AB test)

AB test คือ เห็นโฆษณาทั้งคู่และใช้โฆษณาที่ต่างกัน

คือการทดสอบค่าๆหนึ่ง ด้วยการตั้งสมมุติฐาน เพื่อกำหนดว่าจะยอมรับหรือปฏิเสธมัน

Case Study ขนม Cheetos



Hypothesis Testing

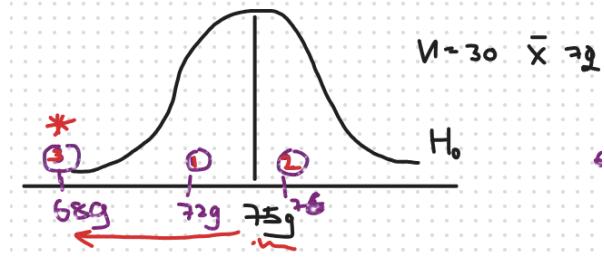
$N = 30$



H_0 : Cheetos avg. weight = 75g

H_a : Cheetos avg. weight \neq 75g

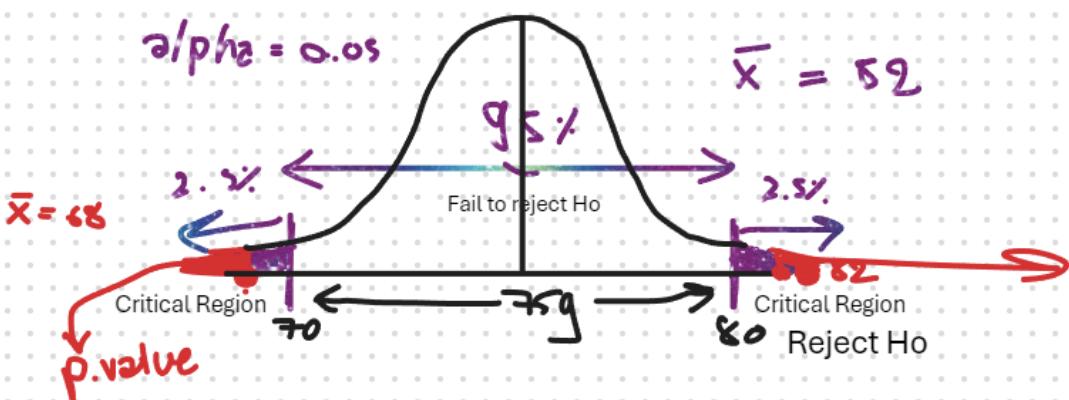
<>



ทาง Cheetos เขา claim ว่า ขนาดของเขานี่ค่าเฉลี่ยน้ำหนักอยู่ที่ 75 g

เราจึงตั้งสมมุติฐาน H_0 กับ H_a ซึ่งก็สองจะต้องขัดแย้งกัน

จากนั้นเราสร้างกราฟ Normal ขึ้นมาและ set ค่า 75 g อยู่ตรงกลาง สมมุติว่าเราไปสุ่มตัวอย่าง ออกมา 30 ช่อง พบร่วมค่าเฉลี่ยอยู่ที่ 72 g และสุ่มมาอีก ค่าเฉลี่ยอยู่ที่ 68 g และ 76 g



จากนั้นเราก็คำนวณพื้นที่ critical region ออกมายโดยใช้ $\alpha = 0.05$ ตาม concept ของ Fisher ถ้าค่าที่เราสุ่มมาตกอยู่ใน critical region เราจะปฏิเสธ H_0 ทันที

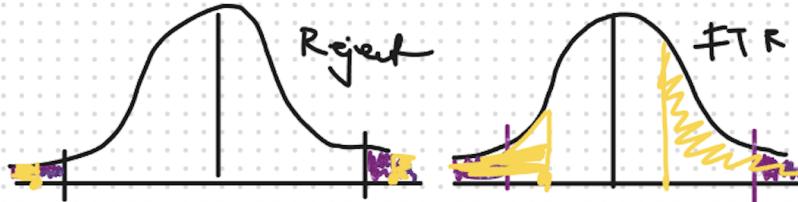


$p.value = p(\text{observed data or more extreme} \mid H_0 \text{ is TRUE})$



Reject H_0 if $p.value \leq \alpha$

Fail to Reject H_0 if $p.value > \alpha$



P-value คือความน่าจะเป็นของข้อมูลที่เราเก็บมาหรือมีค่าไปทางนี้อย หรือทางมาก สังเกตอุகศร สีเดงธูปด้านบน

$$\begin{aligned} n &= 30 \\ \bar{x} &= 78.2 \\ SD &= 2 \end{aligned}$$

$H_0: \mu = 75$
 $H_a: \mu \neq 75$

$$H_0: \mu = 75$$

เรา焉งสามารถใช้ช่วงความเชื่อมันเพื่อกดสอบสมมุติฐานได้

ถ้าค่าที่อยู่ในสมมุติฐานอยู่ในช่วงความเชื่อมัน = Fail to reject H_0 แต่ถ้าค่าไม่ได้อยู่ในช่วง = Reject H_0

- AB Testing

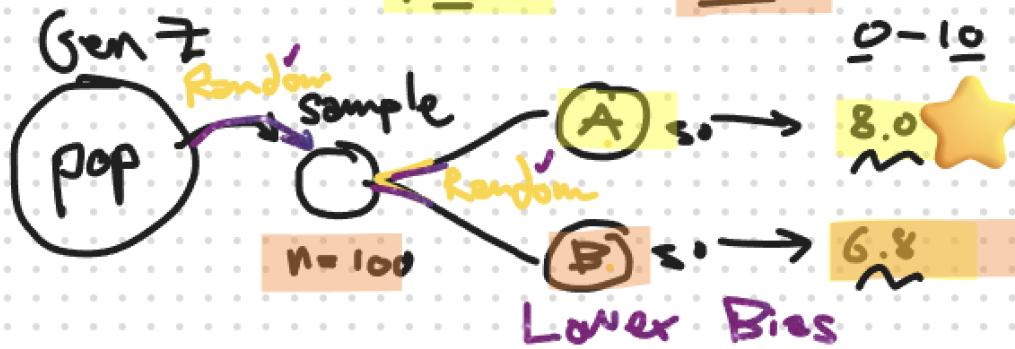
มันจะต่างกับ RCT ตรงที่ RCT จะควบคุมให้ออกฝั่งไม่เห็นโฆษณาแต่อีกฝั่งเห็น ส่วน AB test คือการที่ก้มสองฝั่งเห็นโฆษณาแต่โฆษณาที่ให้เห็นจะไม่เหมือนกัน จึงเรียกว่าเป็น A/B ระหว่างเดียว



AB Testing

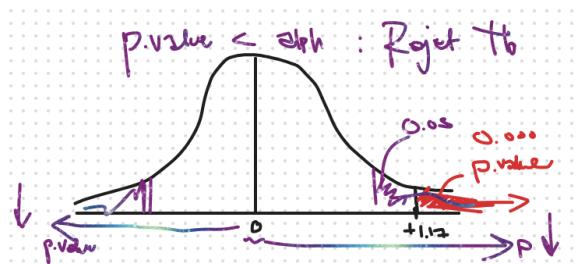
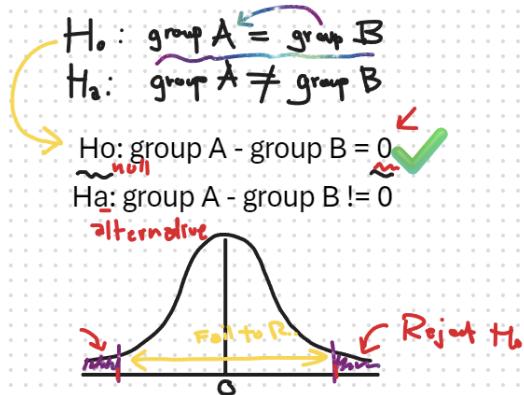
Cheetos → Gen-Z

ลูกปัด vs. ลูกฟูก



ในกรณีที่เราต้องการเกี้ยบผลิตภัณฑ์ในกลุ่มตัวอย่างเดียวกัน Gen-Z การสุ่ม sample และ สุ่มแต่ละชุดจะต้องใช้การ random เพื่อลด bias

ประเด็นก็คือเมื่อเราสุ่มอักเสบๆรอบ ค่าคะแนนที่เราจะพึงพอใจนิดกว่ากันจะเกิดการเปลี่ยนค่าไปเรื่อยๆ เราจึงต้องทำการตั้งสมมุติฐาน Ho และ Ha



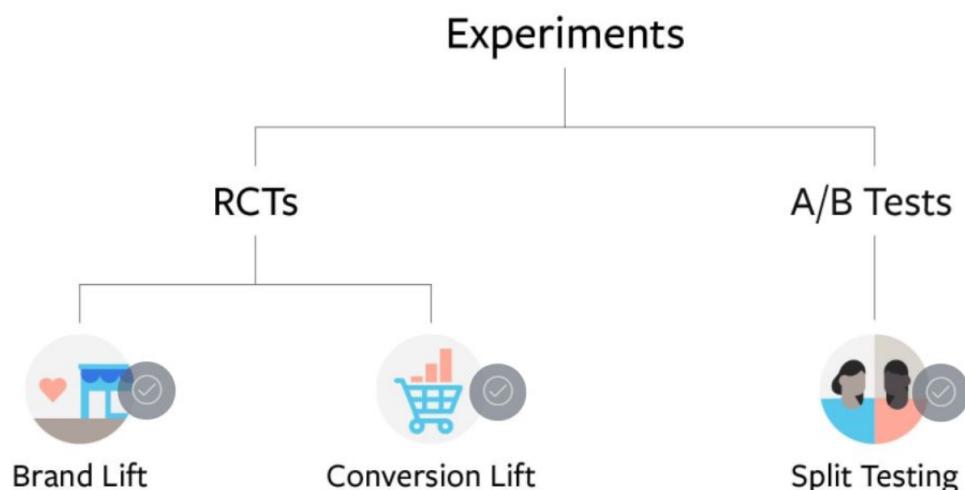
เราสมบูติว่า Ho คือ รสลางและรสตันยำดึงดูดลูกค้า gen-z เท่าๆกัน ส่วน Ha คือไม่เท่ากันแสดงว่า มีอย่างใดอย่างหนึ่งที่เดียวกับ จากนั้นเราต้องย้ายสมการ

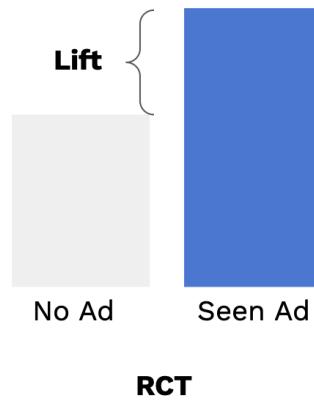
แนะนำให้ใช้ data analys tool ใน excel ในฟังก์ชันที่ชื่อว่า T-test two sample

จากนั้นก็ทดสอบโดย critical region หรือ p-value

- **Type of Study**

1. Observational > Correlation การทำแบบสอบถาม
2. Experimental > Causation เช่นการทดลองยิงไข่ชน คือตัวแปร x มีผลกับ y ด้วย ตรรษ
 - a. RCT (random control trials) คือการทดลองโดยการแยกเป็นสองกลุ่มและจะมี 1 กลุ่มที่ถูก control หรือปิดกั้นไว้ โดยใส่ action หรือ treatment อีก 1 กลุ่มก็ปล่อยตามปกติ หรือ test group และดูค่าความน่าจะเป็นของกั้งสองกลุ่ม เอาจมาเปรียบ เกี้ยบดู





RCT

Correlation and Linear Regression

- ก้าวสองใช้ Data Analyze in excel สร้างอุปกรณ์แบบง่าย ๆ
 - Regression คือ ถ้า x เปลี่ยน 1 หน่วย y จะเปลี่ยนไปเท่าไหร่ ตามข้อมูลที่เรามี
 - R square คือ correlation ยกกำลังสอง จะวัดว่า ค่า x อธิบายการเกิดของ y เท่าไหร่ เช่น 0.6 หมายความว่า x อธิบายการเกิดของ y ได้ 60%
 - Multiple R = Correlation เกิดจาก correlation ของ actual y กับ predicted y ก้าวสองบวกกันจะได้ error และต้องยกกำลังสองเพื่อให้ค่าที่เป็นบวก กล้ายเป็นบวก เพื่อที่จะรวมผลแล้วแม่นยำ ผลกระทบที่ได้คือค่า error ที่น้อยที่สุดแล้ว แม้ว่าจะเปลี่ยนตัวแปรเพิ่มหรือลดก็ตาม
 - Significance F ถ้า < 0.05 แสดงว่า x ใช้กำหนด ค่า y ได้ อย่างมีนัยสำคัญทางสถิติ หรือถ้าเป็น multiple linear regression หมายความว่ามีตัวแปรอย่างน้อย 1 ตัวที่ ใช้กำหนด ค่า y ได้ อย่างมีนัยสำคัญทางสถิติ
 - ค่า Coefficient คือ ตัวแปรตัวนี้ที่ใช้สร้างสมการ Linear Regression
 - P-value คือใช้ดูตัวไหนมีนัยสำคัญกับ y ต้องมากกว่า 0.05 แต่ถ้าคิดแล้วมันสมเหตุสมผลแล้วค่า P-value มันน้อย ต้องเก็บข้อมูลเพิ่ม
 - RMSE คือเอาค่า เฉลี่ยของ errors กำลังสอง และคิดรูราก หรือดูจากสูตร

Error หรืออีกชื่อคือ Residual

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

- In R

- Correlation

```
cor(mtcars$mpg, ,mtcars$hp) # -0.77

#scatter
plot(mtcars$hp, mtcars$mpg, pch=16) # ตบ.การวางแผนสำหรับ ด้านหน้าเป็น

#correlation matrix
cor(mtcars[ , c("mpg", "wt", "hp")])

library(dplyr)
mtcars %>%
  select(mpg, wt, hp, am) %>%
  cor()

# เช็คเม็ดนัยสำคัญทางสถิติด้วยดู P-value
cor.test(mtcars$mpg, ,mtcars$hp)
```

- Regression

```
lmfit <- lm(mpg ~ hp, data = mtcars) #แคลคเดียวจบๆ

summary(lmfit) # แสดงค่าหมดเลย อ่านที่ p-value และ *** คือ < , ** คือ

# predict hp = 200
lmfit$coefficients[[1]] + lmfit$coefficients[[2]] * 200 # มันจะคำนวณให้เรา
```

```

# predict hp หลายๆค่า
new_cars <- data.frame(
  hp = c(250, 320, 400, 410, 450)
)

new_cars$mpg_pred <- predict(lmfit, newdata = new_cars)
# ถ้าได้ติดลบแสดงว่า ข้อมูลที่เรานำมาคำนายมัน beyond ข้อมูลที่เรานำมาเทรน วิธีแก้คือ

# Predict follow by excel
lmfit_v2 <- lm(mpg ~ hp + wt + am, data = mtcars)

coefs <- coef(lmfit_v2)

coefs[[1]] + coefs[[2]]*200 + coefs[[3]]*3.5 + coefs[[4]]*1

# Built full model
lmfit_full <- lm(mpg ~ . , data mtcars)
mtcars$predicted <- predict(lmfit_Full)

squared_error <- (mtcars$mpg - mtcars$predicted)**2
rmse <- sqrt(mean(squared_error)) # train rmse

```

Logistic Regression

- ใช้กับปัญหา **Classification** โดยใช้สมการ **Sigmoid Function** ในการบีบค่าให้อยู่ 0-1 เพราะว่าถ้าเราใช้ **linear regression** เนี่ย เวลาคำนวณจะได้ค่าที่ติดลบและค่าที่เกิน 1 ขึ้นไป ถูกต้องหรือหมายความก็จะใช้งานมัน
 - Google Sheet

SUMMARY OUTPUT						
Regression Statistics						
Chi Square	20.88799261					
Residual Dev.	6.837894614					
# of iterations	9					
Observations	20					
	Coefficients	Standard Error	P-value	Odds Ratio	Lower 95%	Upper 95%
Intercept	9.921966078	5.684653887	0.08091614743	20373.0062	0.2953037462	1405533750
happiness	-1.816681696	1.004640145	0.07056122758	0.1625642947	0.02269205995	1.164598982

- ซึ่งตรง Coefficient เรากนำมาสร้างสมการให้เหมือนกับ linear regression เลย $y = \text{intercept} + \text{slope}(x)$

			(0-1)
happiness	divorce	LoReg(z)	Sigmoid
10	0	-8.244851	0.000263
8	0	-4.611487	0.009839
9	0	-6.428169	0.001613
7	0	-2.794806	0.057606
8	0	-4.611487	0.009839

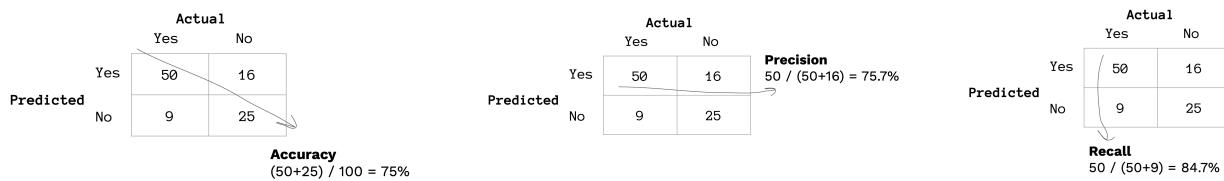
- จากนั้นก็เอาเข้าสูตร sigmoid เพื่อบีบค่าเป็น 0-1 มีสองสูตร คือ $=1 / (1+\text{EXP}(-\text{C4}))$ และ $=\text{EXP}(\text{C4}) / (1+\text{EXP}(\text{C4}))$
- จากนั้นลองปรับให้เป็น % และใช้ conditional formatting และปรับเทียบค่า x ดู จะเห็นว่าความสุขน้อยๆ มีโอกาสในการหย่าสูงมาก สามารถใช้ if มาช่วยอ่านค่าก็ได้ เช่นถ้า $\geq 50\%$ คือหย่าແບ່ງ $=\text{IF}(\text{D4}>=\$G\$3, 1, 0)$ 1 คือหย่าແບ່ງ

happiness	Sigmoid
10	0.03%
8	0.98%
9	0.16%
7	5.76%
8	0.98%
5	69.82%
9	0.16%
6	27.33%
8	0.98%
7	5.76%
1	99.97%
1	99.97%
3	98.87%
1	99.97%
4	93.43%
5	69.82%
6	27.33%
3	98.87%
2	99.81%
0	100.00%

- สร้างคอลัมน์ predicted divorce โดยใช้ if มาช่วยอ่านค่าก็ได้ เช่นถ้า Threshold $\geq 50\%$ คือ หยาบๆ `=IF(D4>=G3, 1, 0)` 1 คือ หยาบๆ จากนั้นหา accuracy เทียบกับ คอลัมน์ divorce `โดยจับมา =` และหาค่าเฉลี่ยของ accuracy true กับ false โดย `=ArrayFormula(AVERAGE(F4:F23*1))` คือ *1 และ ทำเป็น Array หรือจะใช้ IF มาช่วยแสดง เป็นเป็น 0,1 โดยถ้าเท่ากับก็แสดง 1 และว้าหาผลรวมหารด้วย ท คือได้ accuracy เมื่อลงกับ

happiness	divorce	Sigmoid	Predicted_divorce	Accuracy
10	0	0.03%	0	TRUE
8	0	0.98%	0	TRUE
9	0	0.16%	0	TRUE
7	0	5.76%	0	TRUE
8	0	0.98%	0	TRUE
5	0	69.82%	1	FALSE
9	0	0.16%	0	TRUE
6	0	27.33%	0	TRUE
8	0	0.98%	0	TRUE
7	0	5.76%	0	TRUE
1	1	99.97%	1	TRUE
1	1	99.97%	1	TRUE
3	1	98.87%	1	TRUE
1	1	99.97%	1	TRUE
4	1	93.43%	1	TRUE
5	1	69.82%	1	TRUE
6	1	27.33%	0	FALSE
3	1	98.87%	1	TRUE
2	1	99.81%	1	TRUE
0	1	100.00%	1	TRUE
Accuracy				90.00%

- **Precision** คือ ทุกครั้งที่เรา预言ว่า Yes ถูกถูกกี่ครั้งดูที่ Predicted, **Recall** คือ model สามารถ预言 Yes ที่เป็น actual ได้กี่คน, F1 score หาค่าเฉลี่ยระหว่าง recall และ precision



- การคำนวณใน Spread Sheet ใช้ **COUNTIFS** มาช่วย

	A	B	C	D	E	F	G
1	Actual Y	Predicted Y					
2	0	0					
3	0	0					
4	1	1					
5	0	1					
6	0	0					
7	1	1					
8	0	0					
9	1	0					
10	1	0					
11	1	1					
12	0	0					
13	0	0					
14	0	0					

Actual	Predicted	
	No	Yes
No	? =COUNTIFS(Actual,0,Predicted,0)	
Yes	3	7
Accuracy	0.75	=SUM(E4+F5) /SUM(E4:F5)
Precision	0.78	=F5/SUM(F4:F5)
Recall	0.70	=F5/SUM(E5:F5)
F1-Score	0.74	=2*((E9*E10)/(E9+E10))

- การ Split data สามารถใช้ `RAND()` ในการสุ่มตัวเลขอุปกรณ์จากนั้นก็ใช้ `FILTER` เรียงค่า บ้านมาก หรือบานไปบ้านอย เพื่อสลับແກວ แต่ก่อนหน้านั้นต้องໄລໂລກສະແບ່ງຂ້ອງມູນ Train Test ก່ອນ

ID	happiness	divorce	random
15	4	1	? =RAND()
18	3	1	0.9924
1	10	0	0.3159
6	5	0	0.5894
20	0	1	0.1969
3	9	0	0.4727
5	8	0	0.1710
9	8	0	0.2051
11	1	1	0.3152
8	6	0	0.8061
10	7	0	0.4276
19	2	1	0.1362
4	7	0	0.2200
16	5	1	0.6392
7	9	0	0.4701
13	3	1	0.1413
2	8	0	0.5081
14	1	1	0.7245
17	6	1	0.2440
12	1	1	0.2531

- R Programming

```

# Logistic Regression Example

happiness <- c(10, 8, 9, 7, 8, 5, 9, 6, 8, 7, 1, 1, 3, 1, 4, 5,
divorce <- c(0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1,
df <- data.frame(happiness, divorce)

# Fit model
model <- glm(divorce ~ happiness, data = df, family = "binomial"

summary(model) # ได้ตาราง P-value

# Predict and Evaluate model
df$prob_divorce <- predict(model, type = "response") # ได้ช่องที่เป็น
df$pred_divorce <- ifelse(df$prob_divorce >= 0.5, 1, 0)

# confusion matrix
conM <- table(df$pred_divorce, df$divorce, dnn = c("Predicted", "Actual"))

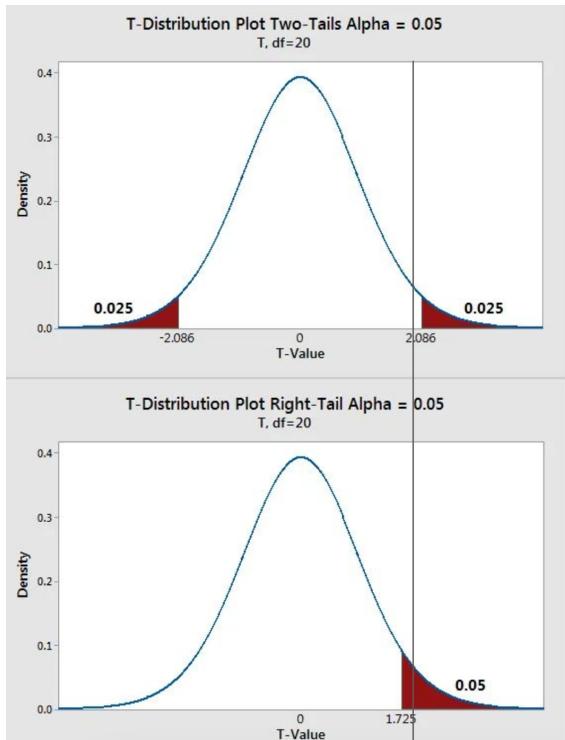
# Model Evaluation
cat("Accuracy:", (conM[1, 1] + conM[2, 2]) / sum(conM))
cat("Precision:", conM[2, 2] / (conM[2, 1] + conM[2, 2]))
cat("Recall:", conM[2, 2] / (conM[1, 2] + conM[2, 2]))

F1 <- 2 * (0.9 * 0.9) / (0.9 + 0.9)

```

Deep Dive into Significance Testing

- **Recap Hypothesis Tests**
 - One-Tailed Test
 - Two-Tailed Test นิยมใช้แบบนี้มากกว่า เพราะไม่อยากให้ Reject H₀ ง่ายเกินไป



Two tailed test

$H_0: \text{mean A} - \text{mean B} = 0$
 $H_a: \text{mean A} - \text{mean B} \neq 0$

One tailed test

$H_0: \text{mean A} - \text{mean B} \leq 0$
 $H_a: \text{mean A} - \text{mean B} > 0$

- **Two Types of Error**

- Type 1 (False Positive) คือ H_0 is True แต่เรา Reject H_0 (บักสกิติจึงนัยมใช้ $\alpha=0.05$)
- Type 2 (False Negative) คือ H_0 is False แต่เรา Do not reject H_0 (β)

- **Hypothesis Testing**

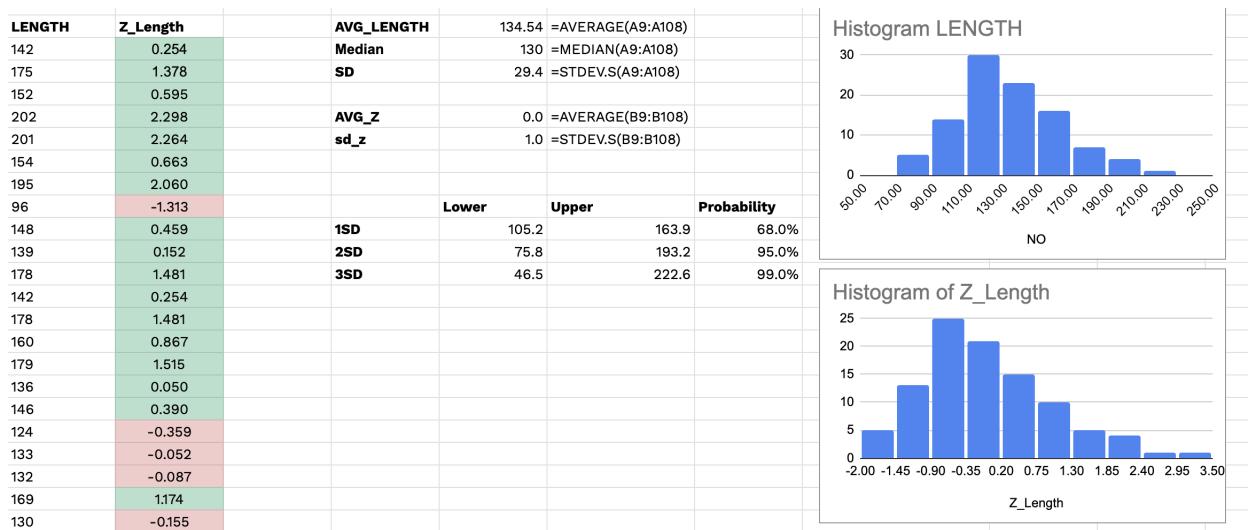
- Critical Region ข้อมูลที่เก็บมาได้ตกลงพื้นที่ Critical และเราจะ Reject มัน
- p-value คือความน่าจะเป็นของข้อมูลที่เราเก็บได้หรือมากกว่านั้น ถ้า H_0 is True และเรา จะคุณ 2 ในกรณี Two-tails เพราะเราคิดสองด้าน
- Confidence Interval แนะนำให้ใช้เนื่องจากค่าความเชื่อมั่นที่เป็นช่วงสามารถนำไปใช้ต่อได้หลากหลายมากกว่าเป็นค่า ๆ เดียว

- **Limitation of P-value**

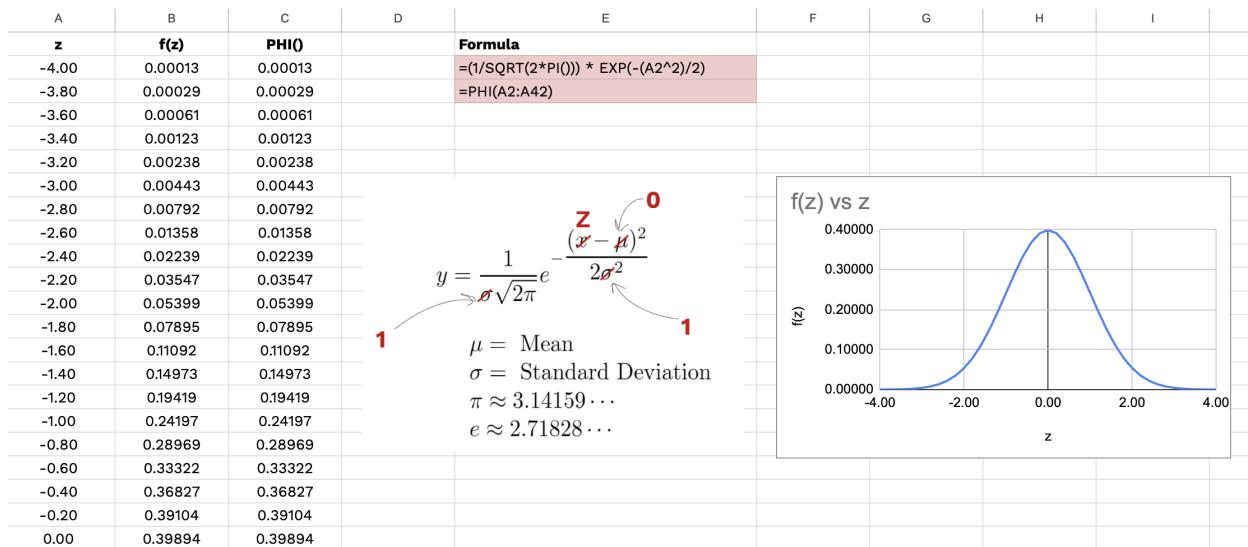
- ยิ่งเก็บ Sample size เพิ่มขึ้น P-value จะลดลง เนื่องจากเราเข้าใกล้ค่าความจริงมากขึ้น หมายความว่ายิ่งปัจจุบันข้อมูลมีมากขึ้น มันจะ Reject H_0 อย่างเดียว

Stat with Google sheets(stats101 - AB test)

- Normal Distribution** การคำนวณ Lower & Upper ใช้ array ง่ายๆ คือ `={mean-SD, mean+SD}` ถ้าเป็น 2SD ก็ใส่เลขสองหน้า SD มันก็จะสร้างมาสองคอลัมน์คือคอลัมน์ที่เป็น Lower และ Upper ส่วนการคำนวณ Probability ให้ใช้ `=COUNTIFS(ข้อมูลทั้งหมดล้อคเซลล์, ">="&Lower, ข้อมูลทั้งหมดล้อคเซลล์, "<="&Upper) / n` ปรับให้เป็น % จบ
 - ตาม Empirical rules → จากรูปด้านล่าง 68% หมายความว่ามีข้อมูลห่างจากค่าเฉลี่ยตรงกลางอยู่ $+/-1\text{SD}$ มีข้อมูลอยู่ในช่วง 105.2 ถึง 163.9



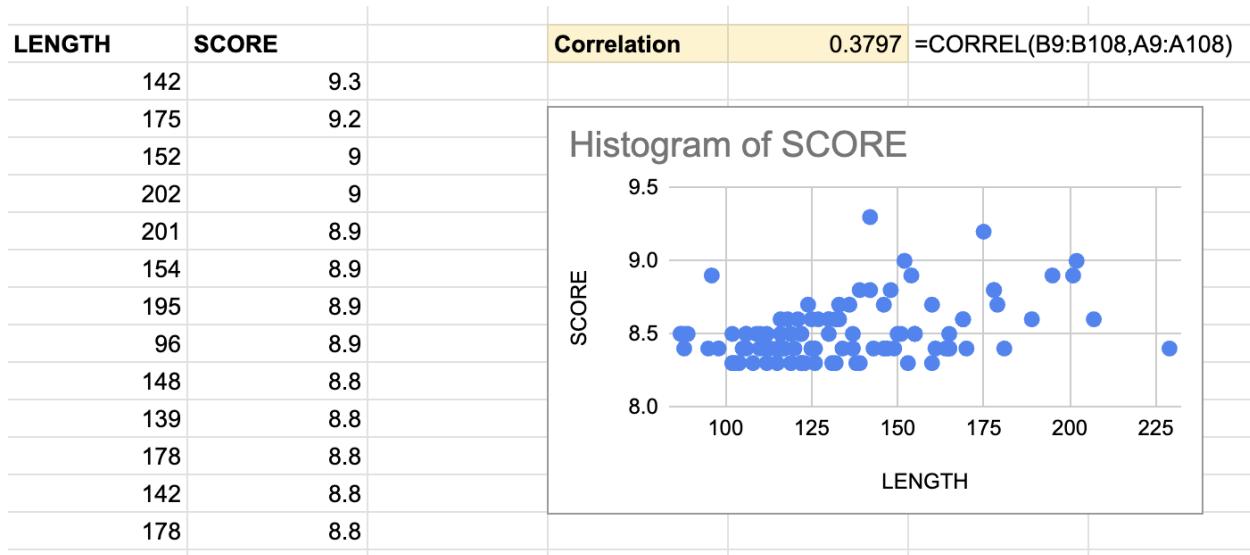
- Plotting Normal Distribution** ทำได้สองวิธี คือ 1. การคำนวณโดยใช้สูตร 2. ใช้ฟังก์ชันที่มีอยู่ คือ `PHI()`



- **Measure of Spread and Position**

MOVIE_NAME	LENGTH	Spread	
The Shawshank Redempti	142	SD	29.35
The Godfather	175	VAR	861.64
The Dark Knight	152	IQR	37.75 Q3-Q1
The Godfather: Part II	202		
The Lord of the Rings: The	201		
Pulp Fiction	154	Position	
Schindler's List	195	Min(Per 0th)	87 =PERCENTILE(\$B\$9:\$B\$108, 0)
12 Angry Men	96	Per 25th	114.5 =PERCENTILE(\$B\$9:\$B\$108, 0.25)
Inception	148	Per 50th	130 =PERCENTILE(\$B\$9:\$B\$108, 0.5)
Fight Club	139	Per 75th	152.25 =PERCENTILE(B9:B108,0.75)
The Lord of the Rings: The	178	Max(Per 99.99)	229 =PERCENTILE(\$B\$9:\$B\$108, 1)
Forrest Gump	142		
The Good, the Bad and the	178		
Hamilton	160		

- **Correlation** จะวัดได้แค่ข้อมูลที่เป็นเส้นตรง



- **Descriptive** คือการดูข้อมูลเบื้องต้น สามารถใช้เครื่องมือที่อยู่ใน google sheet ที่ชื่อว่า XLminer ต้อง install ก่อนน้าา

แล้วจะมีอีกฟังก์กันหนึ่งที่ใช้ในการวนด้อมข้อมูลที่เป็นแบบ Normal distribution คือ `=NORMINV(RAND(), mean, sd)` ถ้าเรา Mean และ SD ก็ลองวนด้อมดูข้อมูลเล่นๆได้แล้ว เมื่อนเราสร้างข้อมูลขึ้นมา Trick คือ สร้าง Tick-box ขึ้นมา กดแล้ว วนด้อม

- **Confidence Interval**

- ในความเป็นจริงเราอาจจะไม่ได้ เพราะต้องไปคำนวณทั้งโลก เราจึงสร้างช่วงความเชื่อ มั่นคงมา
- เราสามารถหา Margin Error ได้โดย `=CONFIDENCE.NORM(0.05, SD, N)` จากนั้นก็สร้างเป็นช่วง Lower กับ Upper

Mean	7.606	
SD	0.727964663	
N	50	
Confidence Interval		
95%	Lower Bound	Upper Bound
	7.404	7.808
	7.404	7.808
Margin Error	0.2017778021	

- **Independence T-Test** คือใช้ในการเปรียบเทียบค่าเฉลี่ยทั้ง RCT และ AB test ซึ่งเราจะทดสอบสมมุติฐาน โดยขั้นแรกคือ Test การกระจายตัวของ Variances ก่อน โดย F.test โดยสูตร `=IF(F.TEST(ช่วงข้อมูลที่หนึ่ง, ช่วงข้อมูลที่สอง) <= 0.05, "Reject Ho", "Do not reject")` จากนั้นเราจะทำ T-Test (ค่าเฉลี่ย) โดยสูตร `=IF(T.TEST(ช่วงข้อมูลที่หนึ่ง, ช่วงข้อมูลที่สอง, 2, 3) <= 0.05, "Reject Ho", "Do not reject")` เลข 2 ใบสูตรคือ Two-tails เลข 3 คือ Type โดยตัวอย่างนี้คือ Type 3 คือความแปรปรวนของทั้งสองไม่เท่ากัน จานวนเพิ่มเติมใน `Help` และก็ใช้ `mean_dif` ของสองข้อมูล สังเกตว่าถ้ากด Tick box ไปเรื่อยๆ มีโอกาสที่ เลข 31.8 จะเปลี่ยนไปเรื่อยๆ จนน้อยกว่า 0 มาก

Test equality of variances		if p-value <= 0.05, reject Ho
F.Test	Reject Ho	HO: equal vars assumed
H1: equal vars not assumed		
Test equality of means		
T.Test	Reject Ho	if p-value <= 0.05, reject Ho
		Ho = mean store 1 = mean store2
Mean_dif	31.8	H1 = mean store1 != mean store 2

- **Pair T-Test** คือ เราจะเกียบก่อนและหลัง(ของสิ่งเดียวกัน) ว่า จำนวนที่เพิ่มขึ้นมาเพิ่มขึ้นแบบมีนัยสำคัญหรือป่าว ทำเหมือนเดิมเลยคือ หาผลต่าง `mean_dif` และหา T-Test โดยสูตร

```
=IF(T.TEST(ช่วงข้อมูลก่อน, ช่วงข้อมูลหลัง, 2, 1) <= 0.05, "Reject Ho", "Do not reject Ho") โดยที่
```

1 คือ Type ของ Pair T-Test และเรา Reject Ho คือแสดงว่าจำบวนที่เพิ่มขึ้นมาเปลี่ยนแปลงอย่างมีนัยสำคัญ

Test equality of means	
T.Test	Reject Ho
Average Diff	7.52
<input checked="" type="checkbox"/>	
$H_0 = \text{mean before} = \text{mean after}$	
H1 = mean before != mean after	
if p-value <= 0.05, reject Ho	