STASTICS FOR PHYSICAL SCIENCE (2603284/2565)

- Lecture notes
- VIVATSATHORN THITASIRIVIT

Statistics

Lecture

① การเก็บข้อมูล : เก็บมาได้ → วิเคราะห์ได้ง่าย (สำคัญที่สุด)

- PRIMARY, ปฐมภูมิ
  - Observational
  - Experimental
- SECONDARY, ทุติยภูมิ : not raw data
  (ข้อมูลที่ถูกรวบรวมไว้แล้ว หรือถูกวิเคราะห์แล้ว) (processed)

* สำมะโน (census)
สำรวจ (sample survey) + sampling design
- simple
- classified
- etc...

ทดลอง (exp. design)

เกิด "sampling error"

Avg.(N) = x_0

Avg.(n) = x

S.E. = Δx ; $N > n$

n : sampling quantity

N : total quantity

Goal : Minimize $\Delta x \to 0$
$(n \to N)$

* CONTROLLABLE
~ systematic error

"non-sampling error" เช่น misleading questions,
inaccurate (untrue) answers,
mistyping, wrong record, ...

→ unable to detect

* UNCONTROLLABLE ~ random (stat.) error

+ศัพท์ : incentive → motivation

Quality of Data

① Consistency : หา Outliers → ตัดทิ้ง / แยกวิเคราะห์ต่างหาก
→ มักจะไม่เจอ Outlier จนกว่า α สูงเกินไป

② Accuracy

③ Completeness อาจทำให้เกิด Bias → worst case : ผลเปลี่ยน (กลับทิศ) / ผลเคลื่อน

④ Auditability

⑤ Validity : ชัดเจน, ไม่ misleading ไม่คลุมเครือ

⑥ Uniqueness : Repeating = bad (per person)

⑦ Timeliness

② Data Analysis

Level of Statistics
- Descriptive (พรรณนา)
- Inferential (อนุมาน)

Population (ประชากร) = กลุ่มทั้งหมด/ข้อมูล (Group of) / หน่วยศึกษา "Units of Study"

Population Size (ขนาดประชากร) :
- N = ... (Finite) ← จำนวนสมาชิกในประชากร
- Infinite

Sample (ตัวอย่าง) : กลุ่มของสมาชิกที่เป็นสมาชิกในประชากร * ~~กลุ่ม~~ตัวอย่าง

Sample Size (ขนาดตัวอย่าง) : - n = ... | แบบสอบถามที่เก็บ เช่น เก็บตัวอย่างขนาด 100 คน

* เก็บตัวอย่างขนาด 100 ~~ตัวอย่าง~~

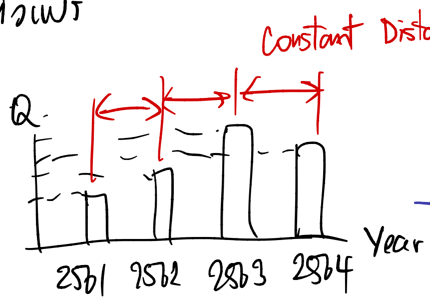
Variable (ตัวแปร) : Characteristics of unit study of more than 1 value.

- Quantitative Variable
- Qualitative (Categorical) Variable

Data (ข้อมูล) : ค่าของตัวแปร

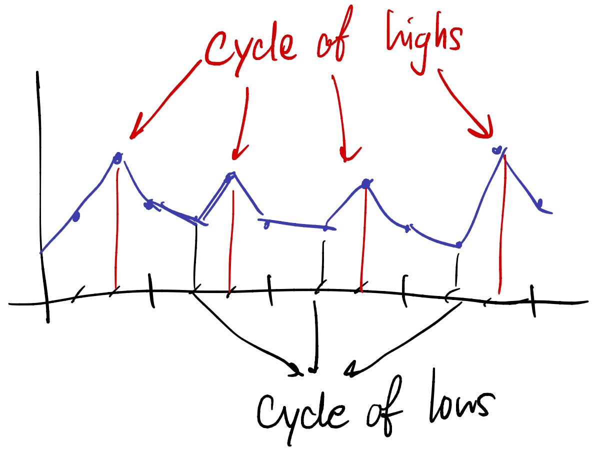
(Monthly) (Annual)
เช่น ทุกๆ เดือน, ทุกๆปี
ทุกๆ 3 ปี,...

Time series Data (ข้อมูลอนุกรมเวลา)



→ เรื่อง Trend

Season



Cross-sectional Data : ข้อมูลเฉพาะระยะเวลาหนึ่ง

↳ เก็บข้อมูลครั้งเดียว

มาตรวัดข้อมูล

- Qualitative — ไม่เป็นตัวเลขโดยธรรมชาติ (แทนด้วยเลขรหัสได้หรือเลขที่คำนวณไม่ได้)
  - Nominal scale * discrete/indiv.
    เช่น เพศ, โรงเรียน, สถานะสมรส
  - Ordinal scale * comparable/ranked
    เช่น ระดับการศึกษา, ความพึงพอใจ 1-5

- Quantitative — ข้อมูลเชิงปริมาณ เป็นตัวเลข
  - Interval scale $(A_1 - A_2)$
    คำนวณผลต่างได้
    เช่น อุณหภูมิ $(\Delta T = 5°C)$

  เพิ่มความยาก scale เป็นการเพิ่มความเป็นปริมาณ

  - Ratio scale (A_1/A_2)
    คำนวณทั้งผลต่างและอัตราส่วนได้
    เช่นความยาว $(\Delta L = 10\ m)$
    Ratio L = 11 เท่า
    จำนวนเงิน

  0 จริง, 0 เทียม   * เชิงสถิติ *

  เช่น เงิน 0 บาท = ไม่มี
  เช่น อุณหภูมิ 0°C ยังมี

Central Tendency (First Degree)

1. Arithmetic Mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$
→ *ค่าที่สรุปจากประชากร = parameters เป็นค่าคงที่

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
→ * ค่าที่สรุปจากตัวอย่าง = statistics มีความผันผวน เป็นตัวแปร

$$\mu_0 = \frac{\sum^{N} x_i}{N} \rightarrow N\mu_0 = \sum^{N} x_i$$

$$\mu = \frac{\sum^{N+1} x_i}{N+1}$$

$$(N+1)\mu = \sum^{N} x_i + \Delta x$$

$$(N+1)\mu = N\mu_0 + \Delta x$$

$$\mu = \frac{N}{N+1}\mu_0 + \frac{\Delta x}{N+1}$$

- นำไปคิด SE (Standard Error, ความคลาดเคลื่อนมาตรฐาน)

  $$SE = \frac{\bar{x}}{\sqrt{n}}$$

  *ไม่ใช่ sampling error
  ไม่ใช่ stdev.

  ↳ การวัดความผันผวนของค่าสถิติ
- ใช้คู่กับ STDEV ในการรายงาน.
  (σ/s)

② Median

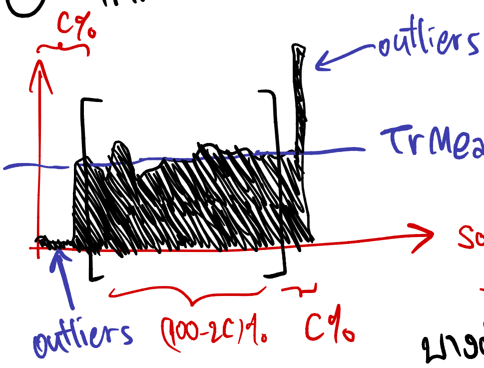
$M = x_{กลาง}$; กลาง $= \left(\frac{N+1}{2}\right)$

$m = x_{กลาง}$; กลาง $= \left(\frac{n+1}{2}\right)$

*อย่าลืมวงเล็บใน subscript เช่น $x_{(1)}$, $x_{(25)}$

- เมื่อเรียงจากน้อยไปมาก
- ใช้คู่กับ QD

③ Trimmed Mean (ค่าเฉลี่ยตัดปลาย)



TrMean (C%) → เฉลี่ยหลังตัดซ้ายออก C% a ตัว
ขวาออก C% a ตัว

จะเหลือข้อมูลใช้หา TrMean (100-2C)%

~~บางตำรา C% → ซ้ายออก C/2 % , ขวาออก C/2 % } เหลือ (100-C)%~~ ไม่ใช้

C ส่วนใหญ่ : 10%, 25%, ...

④ Mode (ฐานนิยม)

เช่น การนับคะแนนเลือกตั้ง

⑤ Quantiles Q_0, D_0, P_0 = min ~ Q_4, D_{10}, P_{100} = max — ซึ่งไม่ใช้

- Quartiles Q_1, Q_2, Q_3
- Deciles $D_1, \ldots, D_9$
- Percentiles $P_1, \ldots, P_{99}$
- n-tiles

"Inclusive"
แต่เราใช้ (เรียน) Exclusive.
"Exclude ตัวขอบทั้ง"

e.g. P_{30} :

ได้น้อยกว่า 40k ← 30% | 70% → ได้มากกว่า 40k

P_{30} = 40,000 บาท P_{80} = 75,000 บาท

50%

ในช่วง 40k < x < 75k

ขั้นตอน หา Position : $\frac{a}{a_0}(N+1)$

กรณี pos. $= x$ (ไม่เป็น $\mathbb{Z}$)

คือ $x - \lfloor x \rfloor = \{x\}$

โดย x point ไปที่ $A_{(\lfloor x \rfloor)}$, $A_{(\lfloor x \rfloor + 1)}$

จะได้ $P_x = A_{(\lfloor x \rfloor)} + \{x\}\left(A_{(\lfloor x \rfloor + 1)} - A_{(\lfloor x \rfloor)}\right)$

Second Degree Central Tendency (Dispersion / Variation)

① พิสัย (Range, R)

$R = x_{max} - x_{min}$: มีปัญหาเรื่องมี outliers

↳ เอาไปประมาณค่า $\sigma \approx \frac{R}{6}$ (มาจาก $R = 6\sigma$ ใน Normal Distribution)

* ช่วงของข้อมูลเป็น $\mu \pm 3\sigma$

② Average Deviation (A.D.) : ค่าเบี่ยงเบนเฉลี่ย

$$A.D. = \frac{\sum |x_i - \mu|}{N} \quad \text{(ประชากร)}$$

$$A.D. = \frac{\sum |x_i - \bar{x}|}{n} \quad \text{(ตัวอย่าง)}$$

[อย่างง่าย]

$$\% A.D. = \frac{A.D.}{\mu} \times 100\%$$

$< 50\%$: ความผันผวนต่ำ (ที่นิยมใช้กัน)

$< 100\%$: ปานกลาง (ยอมรับได้, ไม่แม่นยำ)

$\geq 100\%$: สูง (ใช้ไม่ได้!)

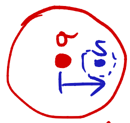
ยิ่ง $|S.D. - A.D.|$ มากเท่าไร แปลว่า signal / ข้อมูลมีปัญหา / outlier อยู่

③ Standard Deviation (S.D.) : ค่าเบี่ยงเบนมาตรฐาน ← เป็น RMS ของ $|x_i - \mu|$

$$S.D. = \sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} = \sqrt{\frac{\sum x_i^2}{N} - \mu^2}$$

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

- Biased Estimator of σ: $\sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$ เพี้ยนจาก σ — ไม่ใช้



- Unbiased Estimator of σ: $\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$ ใกล้ๆ σ มากกว่า

⊕ Recm.

$n > 30$: ตัวอย่างใหญ่ s, s_{n-1} ใกล้เคียง

$n < 30$: ตัวอย่างเล็ก

$$\% S.D. = \frac{S.D.}{\mu} \times 100\%$$

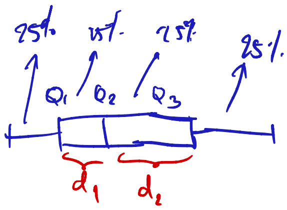
↳ coefficient of variation

④ Quartile Deviation (Q.D.): ส่วนเบี่ยงเบนควอไทล์

ใช้แก้ปัญหา outlier โดยใช้ค่ากลางเป็น Median (M/m)

Deriv. $M = Q_2$

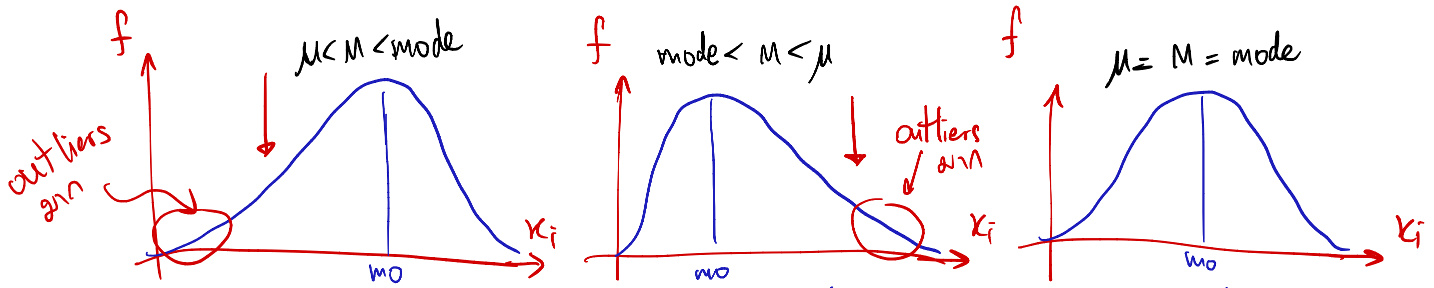
$$Q.D. = \frac{(Q_3 - Q_2) + (Q_2 - Q_1)}{2}$$ ← Average of



$$\therefore \; Q.D. = \frac{Q_3 - Q_1}{2}$$ บ่งบอกการกระจายของข้อมูล 50% ว่ากระจายออกจาก M เท่าใด

Interquartile Range (IQR) $= Q_3 - Q_1 = 2QD.$

Third Order : Skewness & Kurtosis



Left-Skew'd (เบ้ซ้าย) Right-Skew'd (เบ้ขวา) Symmetric

* S_k : สัมประสิทธิ์ความเบ้ / Coefficient of skewness
* k : สัมประสิทธิ์ความโด่ง / Coefficient of kurtosis

When : $S_k \in (-1, 1)$, $k \in (-1, 1)$ } ชุดข้อมูลมาจากประชากรที่มี Normal Dist. (มีความเป็นไปได้) ← possible ไม่ใช่ guarantee.

* เพราะสิ่งที่เห็นเป็นแค่ตัวอย่าง (Inference ของ scope ที่ใหญ่)

- $n(\bar{x} \pm \sigma) \approx 70\%$
- $n(\bar{x} \pm 2\sigma) \approx 95\%$
- $n(\bar{x} \pm 3\sigma) \approx 98\%$

} Normal Dist.

① Skewness (S_k): ความเบ้

(ประชากร)

- $S_k = 0$: Symmetric
- $S_k > 0$: Right
- $S_k < 0$: Left

(ตัวอย่าง)

- $S_k \in [-1, 1]$: Symmetric
- $S_k > 1$: Right
- $S_k < -1$: Left

② Kurtosis (k) : ความโด่ง

		(ประชากร)	(ตัวอย่าง)
– Leptokurtic	: กระจุกมาก	$(k > 0)$	$(k > 1)$
– Mesokurtic	: แบบ Normal Distribution	$(k = 0)$	$(k \in [-1, 1])$
– Platykurtic	: กระจายมาก	$(k < 0)$	$(k < -1)$

Coefficient of Variation

$C.V. = \frac{\sigma}{\mu}$ (ประชากร)

$C.V. = \frac{s}{\bar{x}}$ (ตัวอย่าง)

$C.V. \to 0$: กระจายน้อย (0 — 0.5 — 1) ต่ำ ปานกลาง

$C.V. \to \infty$: (>1) กระจายมาก

ผลลัพธ์ไม่มีหน่วย สามารถนำ C.V. ของข้อมูลต่างชนิด (หน่วย) มาเปรียบเทียบกันได้ (σ เทียบกันไม่ได้, ไม่ make sense)

*Empirical : นับข้อมูลว่าอยู่ในช่วง $\bar{x} \pm n\sigma$ กี่ตัว ← เป็น...

การจัดการ Outliers

① • Z-Score (คะแนนมาตรฐาน) : $z = \frac{x_i - \mu}{\sigma}$ ซึ่งใช้ได้ดีใน Normal dist. $z = \frac{x_1 - \bar{x}}{s}$

หาค่า Outliers.

μ : ประมาณด้วย $\bar{x}$ ได้

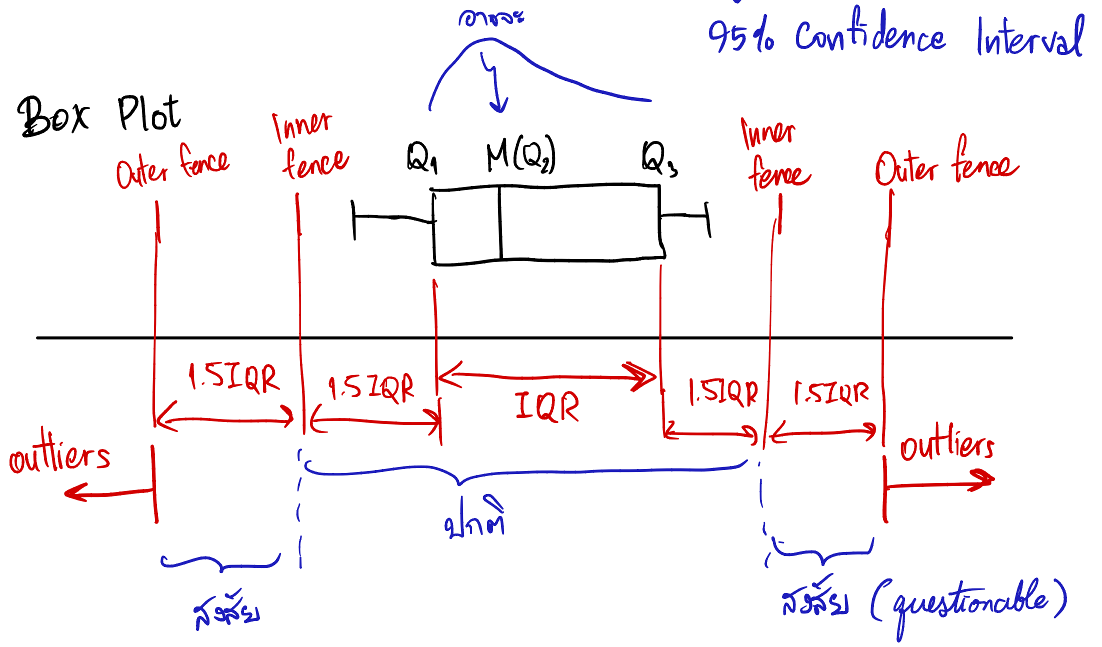
σ : ประมาณด้วย s ได้

outlier จะออกนอกช่วง $(-3, 3)$: ทั่วไป ** สอบใช้ทั่วไป (98%)

$(-2.5, 2.5)$: เข้มงวด

$(-2, 2)$: เข้มงวดมาก → 95% Confidence Interval

② Box Plot

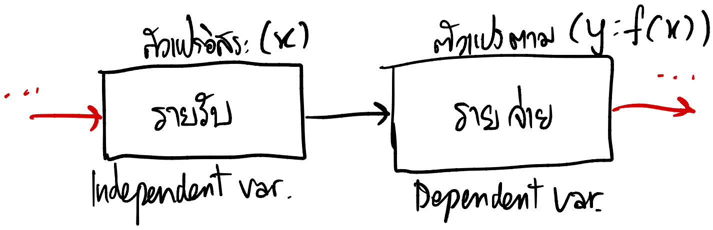


การจัดระดับความสัมพันธ์ระหว่างตัวแปรเชิงปริมาณ 2 ตัวแปร → "Correlation", etc.

กรณี "เชิงคุณภาพ" : ความเป็นอิสระกัน

"ความเกี่ยวข้องกัน"

eg. (ความสุข, น้ำหนัก)



"Structure Equation Model"

① สัมประสิทธิ์สหสัมพันธ์ (Coefficient of correlation, ρ) ← ประชากร

r : ค่าสถิติ (ตัวอย่าง)

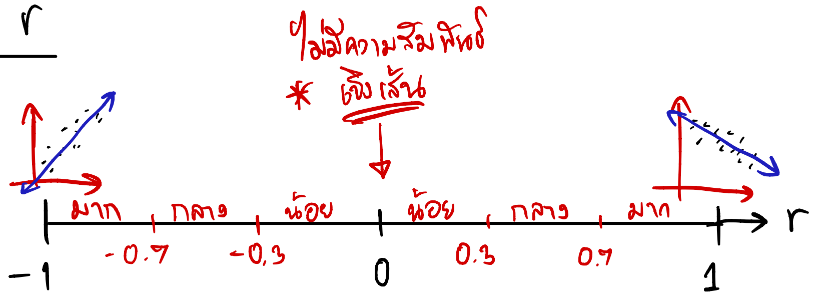
$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}\sqrt{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}}$$

$$r = \frac{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}}{\sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}\sqrt{\frac{\sum (y_i - \bar{y})^2}{n}}}$$

: linear relation

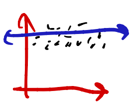
การพิจารณาค่า r



สัมพันธ์แบบผกผัน

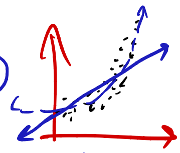
สัมพันธ์กันโดยตรง

①



f เป็นค่าคงที่

②



f ไม่ใช่เส้นตรง

(r ต่ำ ๆ)

ความน่าจะเป็น (Probability)

↳ ฟังก์ชันของระดับความเป็นไปได้

$$P(A) = P \quad ; \quad P \in [0, 1]$$

1. หา ___
2. หา จากการทดลอง : $f_A = \dfrac{n_A}{n_{\text{ทดลอง}}}$: converges approach P
3. หาวิธี classical

$$P(E) = \frac{n(E)}{n(S)} \quad : \quad E \subseteq S \begin{cases} 0 : E = \{\ \} \\ 1 : E = S \end{cases}$$

$$P(A \cap B) = P(A|B) \cdot P(B)$$

$$\frac{2}{36} \stackrel{?}{=} \frac{2}{6} \cdot \frac{6}{36} \quad \checkmark$$

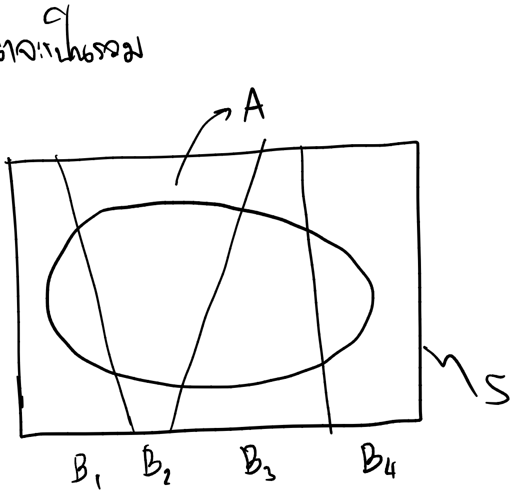
→ BAYES :

$$P(A \cap B) = P(B \cap A)$$

$$P(A|B)P(B) = P(B|A)P(A)$$

$$\therefore P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

ความน่าจะเป็นรวม



$$B_i \cap B_{j\neq i} = \emptyset \quad , \quad \bigcup_{i=1}^{N} B_i = S$$

$$P(A \cap (B_1, B_2)) \rightarrow P((A\cap B_1)\cup(A\cap B_2))$$

$$= P(A\cap B_1) + P(A\cap B_2)$$

$$= \underbrace{P(A|B_1)}_{\text{Prob.}}\underbrace{P(B_1)}_{\text{weight ของ Prob.}} + \underbrace{P(A|B_2)}_{\text{Prob.}}\underbrace{P(B_2)}_{\text{weight ของ Prob.}}$$

$\sim$ คล้ายคลึงกับ $\dfrac{\sum W_i X_i}{\sum W_i}$ (ค่าถ่วงน้ำหนัก)

<u>e.g.</u>

I	II
ดำ : 5	ดำ : 8
ขาว : 4	ขาว : 5

หยิบบอล 1 ลูก หา Prob. ที่จะเป็นลูกบอลสีดำ.

การเลือกกล่องหยิบจะสุ่มโดยโยนเหรียญ 2 เหรียญ. หากขึ้น H อย่างน้อย 1 เหรียญ จะหยิบจากกล่อง I. อื่นๆ กล่อง II.

$$P(\text{ดำ}|I) = \frac{C(5,1)}{C(9,1)} = \frac{5}{9}$$

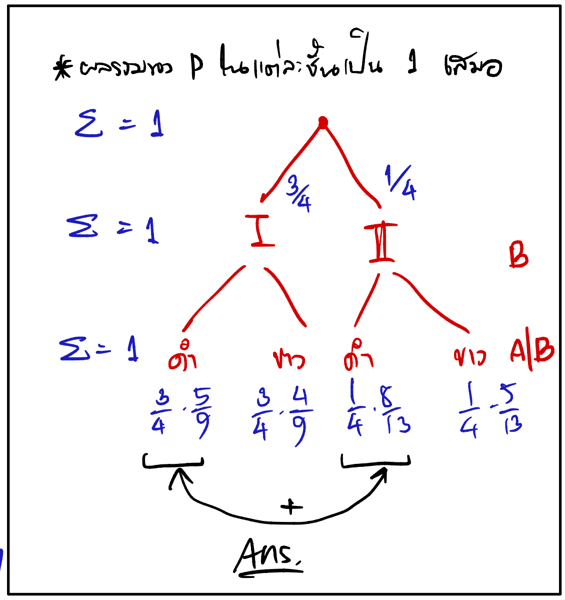
$$P(\text{ดำ}|II) = \frac{C(8,1)}{C(13,1)} = \frac{8}{13}$$

$P(I)$: H อย่างน้อย 1 : HH, HT, TH, TT $\left.\right\}$ $\frac{3}{4}$

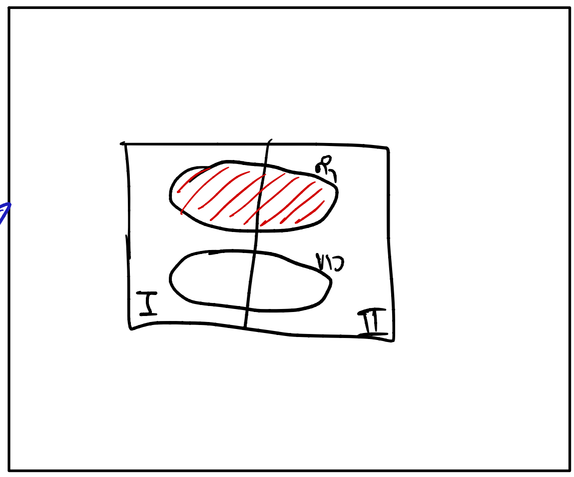
$$\therefore P(I) = \frac{3}{4}$$

$$P(II) = \frac{1}{4}$$

$$P(\text{ดำ}) = \frac{5}{9}\left(\frac{3}{4}\right) + \frac{8}{13}\left(\frac{1}{4}\right)$$



สังเกต แต่ละ Root (Parent)
จะไม่ overlap กัน $\left(\bigcup_{j\neq i}(B_i\cap B_j) = \emptyset\right)$
(เป็นอิสระกัน)



กรณีพิจารณา 2 เหตุการณ์ B , B' จะได้

$$P(A) = P(A|B)P(B) + P(A|B')P(B')$$

$$P(A) = P(A|B)P(B) + P(A|B')(1-P(B))$$

Bayes' Theorem

ชุด A ต่อ 60% พลาด 1/40

ชุด B ต่อ 40% พลาด 1/25

$$\text{โอกาสพลาดรวม} = \underbrace{0.6 \times \frac{1}{40}}_{P(\text{พลาด}|A)} + \underbrace{0.4 \times \frac{1}{25}}_{P(\text{พลาด}|B)} = \underbrace{0.031}_{P(\text{พลาด})}$$

หากนาย ก. โวยว่ามีของเล่นพลาด นาย ก. จะคาดเดาว่าของชิ้นนั้นต่อโดย A = ?
โดย B = ?

$$P(A|\text{พลาด}) = \frac{P(A \cap \text{พลาด})}{P(\text{พลาด})} = \frac{P(\text{พลาด} \cap A)}{P(\text{พลาด})}$$

$$P(A|\text{พลาด}) = \frac{P(\text{พลาด}|A)P(A)}{P(\text{พลาด})}$$

$$= \frac{P(\text{พลาด}|A)P(A)}{P(\text{พลาด}|A)P(A) + P(\text{พลาด}|B)P(B)}$$

$$P(A|\text{พลาด}) = \frac{\frac{0.6}{40}}{0.031}$$

General Form

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum P(A|B_k)P(B_k)} \qquad \sum P(A|B_k)P(B_k) = P(A)$$

เจอ ไม่เจอ

e.g. ขุดน้ำมัน 50:50 (Blind ขุด), ขุดทดสอบ เจอน้ำมัน (Y) มีโอกาสเจอดิน A = 0.2

$P(Y)=0.5$, $P(Y')=0.5$ ไม่เจอน้ำมัน (Y') มีโอกาสเจอดิน A = 0.7

จากการขุดครั้งนี้พบดิน A จะเจอน้ำมัน = กี่% → หา $P(Y|A)$

$$P(A|Y)=0.2 \quad , \quad P(A|Y')=0.7$$

$$P(Y|A) = \frac{P(A|Y)P(Y)}{\cancel{P(A)}\ P(A|Y)P(Y) + P(A|Y')P(Y')}$$

$$P(Y|A) = \frac{(0.2)(\cancel{0.5})}{(0.2)(\cancel{0.5})+(0.7)(\cancel{0.5})}$$

$$P(Y|A) = \frac{2}{9}$$

$\therefore P(Y|A) = 22.2\%$ (ความน่าจะเป็นที่จะเจอน้ำมัน หากเจอดิน A)

$\therefore$ ไม่ควรขุดต่อ

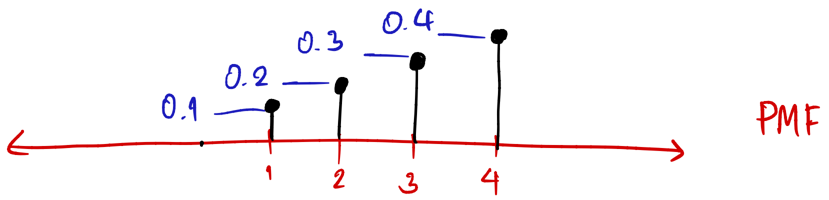
Probability Density Function (Probability Distribution)

- ตัวแปรสุ่ม (Random Variable) → ต่อเนื่อง (Continuous)
  ↘ ไม่ต่อเนื่อง (Discrete)

- Probability Mass Function (PMF) : Discrete X ; $P(X=c)$ = ค่าหนึ่ง

  ↳ $f(x) = P(X=x) \geq 0$; $\sum_{\text{all } x} f(x) = 1$

  e.g. $f(x) = \begin{cases} \frac{x}{10} & 1,2,3,4 \\ 0 & \text{else.} \end{cases}$ ⤳ เป็น PMF $\begin{cases} \sum = \frac{1}{10}+\frac{2}{10}+\frac{3}{10}+\frac{4}{10} = 1 & ✓ \\ f(x) \geq 0 & ✓ \end{cases}$

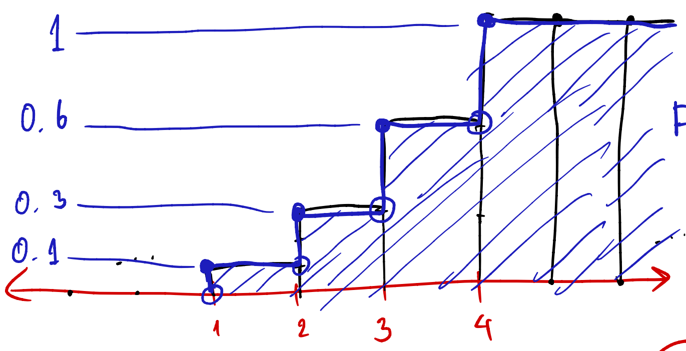


↳ หา $P(X=x)$, $P(a \leq x \leq b)$, $E[X] = \sum_{\text{all } x} x \cdot f(x)$, $Var(X) = \sum_{\text{all } x} x^2 f(x) - (E(x))^2$

$\quad = f(x)$, $\quad = \sum_{x=a}^{b} f(x)$ ← $F(b) - F(a-1)$, $\quad S = \sqrt{Var(X)}$

- Cumulative Distribution Function (CDF) → $F(x) = f(x \in [a,b])$;

  ↓ Discrete X.



$F(x) = f(a) + f(a+1) + \dots + f(b-1) + f(b)$

เช่น $P(x \leq 4) = F(4)$

กรณีนี้ : $F(x) = \sum_{i=0}^{x} f(i)$ → $f(x) = F(x) - F(x-1)$

- Probability Density Function (PDF) : Continuous X ; $P(X \in [a,b])$ = ค่าหนึ่ง

  ↳ $P(X=c) = 0$

Uniform Distribution Function : "f(x) เท่าๆ กัน" เช่น ทอยลูกเต๋า

(แบบไม่ต่อเนื่อง) :

$$f(x) = \frac{1}{k} \quad ; \quad \mu = \frac{1}{k}\sum_{\text{all } x} x_i \quad , \quad \sigma^2 = \frac{1}{k}\sum_{\text{all } x}(x_i - \mu)^2$$

แผ่นป้าย $X = \{1, 2, 3, 5, \underline{8}, \underline{10}\}$

e.g. นาย A และ นาย B เปิดแผ่นป้ายคนละ 1 ครั้ง ครั้งละ 1 แผ่น จาก 6 แผ่น (ส่วนแต่ละขาเปิด)

ความน่าจะเป็นที่ A หรือ B จะได้เงินมากกว่า 5000 บาท $P(A \cup B)$

$$E[>5000] = \frac{1}{3} = P(A) = P(B)$$

$P(A \cap B)$

$$E[>5000 \cap >5000] = P(A) \cdot P(B) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$$

$$\therefore P(A \cup B) = \frac{2}{3} - \frac{1}{9} = \frac{5}{9} \text{ ※}$$

Bernoulli Distribution

↳ สำหรับตัวแปรที่มี 2 ลักษณะ : Success & Fail → Spectrum

0: Fail : $1-p$

1: Success : p

ต้องรู้ค่าใดค่าหนึ่ง เช่น

โอกาสสำเร็จ = 0.98 → p

$\therefore 1-p = 0.02$

$$\boxed{f(x) = p^x (1-p)^{1-x}} \quad ; \quad x = 0, 1$$

Binomial Distribution

↳ จาก Bernoulli ทำต่อกันหลายครั้ง (n ครั้ง) และสำเร็จ x ครั้ง

$$\boxed{f(x) = C(n, x)\, p^x (1-p)^{n-x}} \quad ; \quad x = 0, 1, 2, \ldots, n$$

↳ ทั่วคมแท้ : p คงที่ทุกครั้ง เช่น ทอยลูกเต๋า 10 ครั้ง : $p = \frac{1}{6}$

↳ ไม่แท้ : p ใช้ความสัมพันธ์ เพื่อให้ p มีค่าเดียว คือค่าเฉลี่ยของ p ทุกครั้ง.

$$\rightarrow \mu = np \quad , \quad \sigma^2 = np(1-p)$$

Use : $P(X) = 1 - P(X')$

Negative Binomial Distribution

→ มองหา " ต้องทำกี่ครั้ง ถึงจะได้ความสำเร็จ r ครั้ง "

$$\boxed{f(x) = C(x-1,\, r-1)\, p^{r}(1-p)^{x-r}} \;;\quad x = r, r+1, \ldots$$

x : จำนวนครั้งที่ทำ จนเกิดความสำเร็จครั้งที่ $r=3$ (e.g.)

$x = 3, 4, 5, \ldots$ → or " ยิงปืนนัดที่ $\overset{x}{7}$ เข้าเป้าเป็นนัดที่ $\overset{r}{4}$ "

$f(x)$ เป็นโอกาสที่ เมื่อกระทำ x ครั้ง แล้วสำเร็จ r ครั้ง (เป๊ะๆ)

$$\rightarrow \mu = \frac{r}{p}\,, \qquad \sigma^2 = \frac{r(1-p)}{p^2}$$

Geometric : $r = 1$ (" โอกาสที่ทำ x ครั้ง แล้วสำเร็จเป็นครั้งแรก ")

$$\rightarrow \mu = \frac{1}{p}\,, \qquad \sigma^2 = \frac{1-p}{p^2}$$

Poisson Distribution

X : จำนวนครั้ง ของความสำเร็จ ภายใต้เงื่อนไขจำกัด เช่น ช่วงเวลา, พื้นที่, จำนวนครั้ง, ...

$$\boxed{f(x) = \frac{e^{-\mu}\mu^{x}}{x!}} \;;\quad x = 0, 1, 2, 3, \ldots$$

$$\frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$\rightarrow \text{ต้องรู้ } \mu, \qquad \sigma^2 = \mu$$

$$\sum_{x\geq 0} f(x) = e^{-\mu}\sum_{x\geq 0}\frac{\mu^{x}}{x!} = e^{-\mu}\cdot e^{\mu} = 1 \quad \square$$

FINALS

Hypergeometric Distribution

$$f(x) = \frac{C(K,x)\,C(N-K,\,n-x)}{C(N,n)}$$

: N เป็นสิ่งของทั้งหมด
หยิบตัวอย่างมา n สิ่ง

: K เป็นอัตรา success ใน N

: x เป็นตัวแปรสุ่ม

$$\mu = np \,, \quad \sigma^2 = np(1-p)\left(\frac{N-n}{N-1}\right)$$

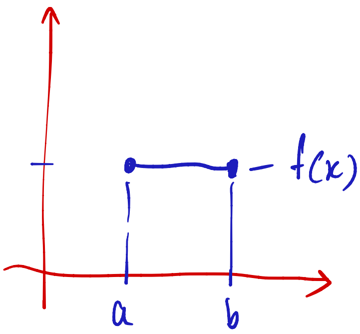
Continuous Distribution

- Uniform Distribution :

$$f(x) = \frac{1}{b-a}$$

$$\mu = E[X] = \frac{a+b}{2}$$

$$\sigma^2 = \frac{(b-a)^2}{12}$$



- Normal Distribution :

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mu = \mu, \quad \sigma^2 = \sigma^2$$

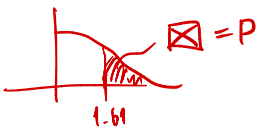
Standardization : $\mu = 0, \quad \sigma^2 = 1 \quad : \quad Z = \frac{X-\mu}{\sigma}$, $\boxed{z = \frac{x-\mu}{\sigma}}$

$T : X \to Z \; : \; z = T(x) \,, \; T(x) = \frac{x-\mu}{\sigma}$

การเปรียบเทียบ $P(X \le x) = P(Z \le z)$

0.01

$*\,P(Z > 1.61)$ อ่านที่ 1.6 ☒ = P



Exponential Distribution (Inverse ของ Poisson)

เช่น Poisson : จำนวน.../ชม.

Exp : กี่ชม./ชิ้น

$$f(x) = \lambda e^{-\lambda x}$$ (λ ใน Poisson) ; x = ระยะเวลา/หน่วยที่จะเกิดความสำเร็จในครั้งถัดไป

$$\mu = \frac{1}{\lambda}, \quad \sigma^2 = \frac{1}{\lambda^2}$$

$$F(x) = 1 - e^{-\lambda x}$$

Gamma Distribution → $P(X > x) = \sum_{k=0}^{r-1} \frac{e^{-\lambda x}(\lambda x)^k}{k!}$ (Poisson)

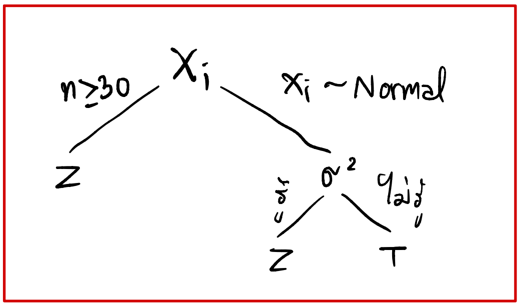
$$f(x) = \frac{1}{\Gamma(r)} \lambda^r x^{r-1} e^{-\lambda x}$$; $x > 0$, x = ระยะเวลา/หน่วยที่จะเกิดความสำเร็จเป็นครั้งที่ r

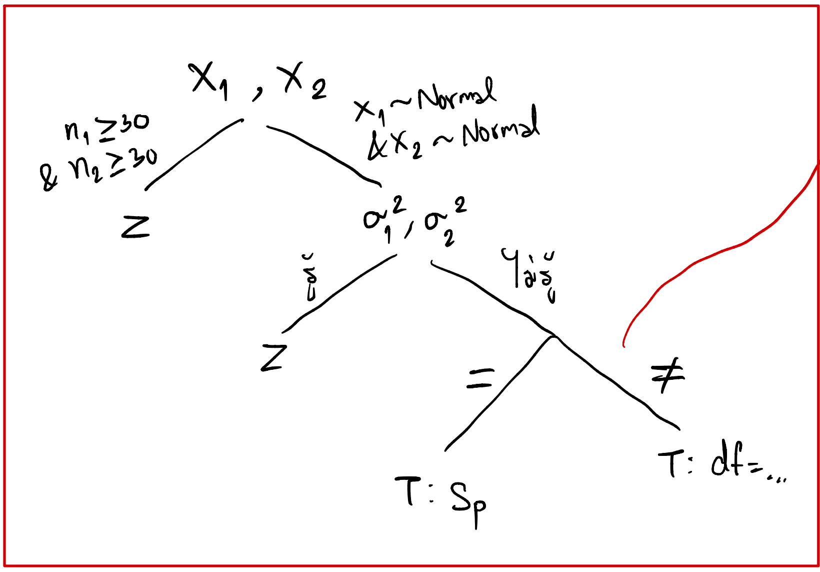
$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x}\,dx = (r-1)!, \quad \Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi}$$

$$\mu = \frac{r}{\lambda}, \quad \sigma^2 = \frac{r}{\lambda^2}$$

Chi-squared Distribution (Γ where $\lambda = \frac{1}{2}$, $r = \frac{n}{2}$), χ^2

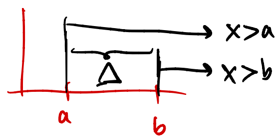
SAMPLING CHART





✱ $P(X<a)$, $P(X>b)$;

$$P(a<X<b) = P(X>a) - P(X>b)$$

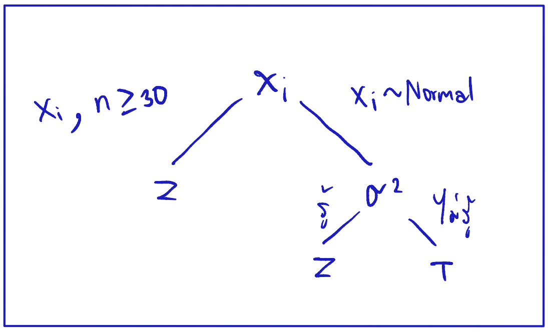


$$0.25 \leq \frac{S_1^2}{S_2^2} \leq 4 : \sigma_1^2 = \sigma_2^2$$

$$\text{else} : \sigma_1^2 \neq \sigma_2^2$$

Sampling

- $\bar{x}$, s^2, $\hat{p} = \frac{x}{n}$
- $\bar{x}_1 - \bar{x}_2$, $\frac{S_1^2}{S_2^2}$, $\hat{p}_1 - \hat{p}_2$



$\boxed{\bar{X}}$: ค่าเฉลี่ยตัวอย่าง

ให้ตัวอย่างสุ่ม $X_1, X_2, \ldots, X_n$: ขนาด n : เป็นอิสระกัน และเป็นการแจกแจงปกติ
ค่าเฉลี่ยเป็น μ, ความแปรปรวนเป็น σ^2
แล้วค่าเฉลี่ยตัวอย่าง $\bar{X}$ เป็น Normal Dist ที่ $\mu_{\bar{x}} = \mu$ และ $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$

$$X \sim N(\mu, \sigma^2) \rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

ด้วย $T: \bar{X} \rightarrow Z$

$$Z = \frac{\bar{x} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

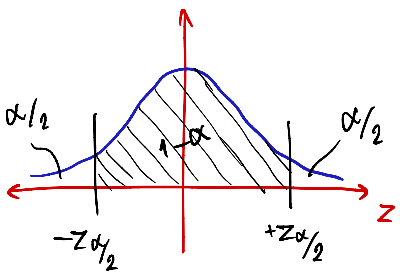
□ Confidence Interval (on mean $\bar{x}$) : $L < \mu < U$ with confidence level $1-\alpha$ (100%)

$$X \sim N(\mu, \sigma^2) \rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$, ใช้ $T: \bar{X} \rightarrow Z$

unknown (μ), known (σ^2)

Thm.

$$\boxed{\mu = \bar{x} \pm Z_{\frac{\alpha}{2}} (S.E.)}$$



* $Z_{\frac{\alpha}{2}}$ นับจาก Tail

99%	: $Z_{0.005}$	=	2.575
98%	: $Z_{0.01}$	=	2.33
95%	: $Z_{0.025}$	=	1.96
90%	: $Z_{0.05}$	=	1.645

ใช้ α เป็น error, e.g., 5%

$$P(-Z_{\frac{\alpha}{2}} < Z < +Z_{\alpha/2}) = 1 - \alpha$$

$$P\left(-Z_{\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}\right) = 1 - \alpha$$

$$\therefore P\Big(\underbrace{\bar{X} - \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}}_{L} < \mu < \underbrace{\bar{X} + \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}}_{U}\Big) = 1 - \alpha$$

$$\boxed{\mu = \bar{X} \pm \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}}$$ ก็ต่อเมื่อ $X \sim N(\mu, \sigma^2)$

Central Limit Theorem

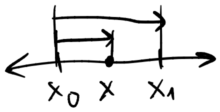
$T: \bar{X} \to Z$ เมื่อขนาดตัวอย่าง $n \to \infty$ (เหมือนด้านบน)

ใช้เมื่อไม่ทราบ σ แต่ใช้ $\hat{\sigma} = s$ แทน. (เอา s ไปประมาณเป็น σ ได้เลย)

t-Distribution : centered at 0. Like z but with kurtosis. k (degree of freedom)

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)} \cdot \frac{1}{\left(\frac{x^2}{k}+1\right)^{\frac{k+1}{2}}} \quad ; \; x \in \mathbb{R}$$

$$T = \frac{\bar{X}-\mu}{s/\sqrt{n}} \quad , \quad * \; P(T(df) > t_{\frac{\alpha}{2},df}) = \frac{\alpha}{2} \; *$$



$* \; df = n-1$

ใช้ linear approx (Interpolate) ได้

$$f(x) = f(x_0) + \frac{f(x_1)-f(x_0)}{x_1-x_0}(x-x_0)$$

$$x = x_0 + \frac{x_1-x_0}{f(x_1)-f(x_0)}\left(f(x)-f(x_0)\right)$$

– for Dist. table $(x_1 > x_0)$

$$\boxed{\mu = \bar{x} \pm \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}, n-1}}$$

$(df = n-1)$

Distribution ของ $\bar{X}_1 - \bar{X}_2$ (เอาไปประมาณค่าเฉลี่ยความแตกต่างของประชากร X_1 และ X_2)

Z สุ่มตัวอย่าง $X_{11}, \ldots, X_{1n_1}$ ขนาด n_1 จาก $N(\mu_1, \sigma_1^2)$; ทราบ σ_1^2

และ $X_{21}, \ldots, X_{2n_2}$ ขนาด n_2 จาก $N(\mu_2, \sigma_2^2)$; ทราบ σ_2^2

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_{\bar{X}_1-\bar{X}_2}, \sigma^2_{\bar{X}_1-\bar{X}_2})$$

โดย

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$$

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = (S.E.)^2$$

$$\therefore Z = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \xrightarrow{n>30} \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$

App. แตกต่างเกิน k : $P(|\bar{X}_1-\bar{X}_2| > k) = P(\bar{X}_1-\bar{X}_2 > k) + P(\bar{X}_1-\bar{X}_2 < -k)$

Confidence Interval : $\mu_1-\mu_2 = (\bar{X}_1-\bar{X}_2) \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_1}}$

(T) □ กรณี $\sigma_1^2 = \sigma_2^2 = \sigma^2$

มี $df = (n_1-1)+(n_2-1) = n_1+n_2-2$

หา Pooled variance : $S_p^2 = \dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ (เฉลี่ยแบบ weight = df)

$$\rightarrow \quad T = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Confidence Interval : $\mu_1-\mu_2 = (\bar{x}_1-\bar{x}_2) \pm t_{\frac{\alpha}{2},\, n_1+n_2-2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

□ กรณี $\sigma_1^2 \neq \sigma_2^2$

$$\text{มี } df = \frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}} = \frac{(n_1-1)(n_2-1)\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{(n_2-1)\left(\frac{S_1^2}{n_1}\right)^2+(n_1-1)\left(\frac{S_2^2}{n_2}\right)^2}$$

$$\rightarrow \quad T = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$

Confidence Interval : $\mu_1-\mu_2 = (\bar{x}_1-\bar{x}_2) \pm t_{\frac{\alpha}{2},\, df}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$

<u>จับ $\bar{X}_1, \bar{X}_2$ เป็นคู่ ๆ</u>

$D_j = \bar{X}_{1,j} - \bar{X}_{2,j}$, $D_j \sim$ Normal , $\mu_D = \mu_1-\mu_2$, σ_D^2 หรือ $\sigma_D^2 \approx S_D^2$

T : $df = n-1$ โดย $T = \dfrac{\bar{D}-\mu_D}{\frac{\sigma_D}{\sqrt{n}}}$; $\sigma_D = S_D = \sqrt{\dfrac{\sum_{j=1}^{n} D_j^2 - nD^2}{n-1}}$

Confidence Interval : $\mu_D = \bar{D} \pm t_{\frac{\alpha}{2},\, df}\dfrac{S_D}{\sqrt{n}}$

$\boxed{S^2}$: ความแปรปรวนของตัวอย่าง

ใช้ Chi-squared Distribution (χ^2)

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$ เป็น χ^2 ที่ $df = n-1$

$$f(x) = \frac{2^{-k/2}}{\Gamma(k/2)} x^{k/2-1} e^{-x/2} \; ; \; x>0 \, , \; k = df$$

$\mu = k, \sigma^2 = 2k$

- การประมาณค่า σ ในช่วง (Lower, Upper) : $P(L<\sigma^2<U) = 1-\alpha$

  ใช้ Transformation $\sigma^2 \to \chi^2$

  Confidence Interval : $\sigma^2 \in \left(\dfrac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}, n-1}} , \dfrac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}, n-1}} \right)$

Distribution ของ S_1^2/S_2^2

สุ่มตัวอย่าง $X_{11}, X_{12}, \dots, X_{1n_1}$ ขนาด n_1 จาก $N(\mu_1, \sigma_1^2)$

และ $X_{21}, X_{22}, \dots, X_{2n_1}$ ขนาด n_2 จาก $N(\mu_2, \sigma_1^2)$

→ ตัวแปรสุ่ม $F = \dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ โดย df_1 (numerator) $= n_1 - 1 = \nu_1$

df_2 (denomenator) $= n_2 - 2 = \nu_2$

Confidence Interval : $\dfrac{\sigma_1^2}{\sigma_2^2} \in \left(\dfrac{S_1^2}{S_2^2} \cdot \dfrac{1}{f_{\frac{\alpha}{2}, n_1-1, n_2-1}} , \dfrac{S_1^2}{S_2^2} \cdot f_{\frac{\alpha}{2}, n_2-1, n_1-1} \right)$

* $f_{k,a,b} = \dfrac{1}{f_{1-k,b,a}}$ (ตารางไม่มี $f_{0.975}$ แต่มี $f_{0.025}$)

$\boxed{\hat{p}}$: สัดส่วนตัวอย่าง (P : สัดส่วนประชากร) Population Proportion : $p = \dfrac{X}{N}$ (A) → Count of interests, → Pop. Size

- ตัวอย่างสุ่ม $X_1, X_2, \ldots, X_n$ เป็นอิสระกัน
- 0 = หน่วยตัวอย่าง <u>ไม่มี</u>คุณลักษณะที่กำหนด
- 1 = หน่วยตัวอย่าง <u>มี</u> คุณลักษณะที่กำหนด

} Bernoulli Dist. : $f(x) = p^x(1-p)^{1-x}$

$$\mu = p, \quad \sigma^2 = p(1-p)$$

โดย p = ความน่าจะเป็นที่หน่วยตัวอย่าง <u>มี</u> คุณลักษณะที่กำหนด

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{A}{n} = \hat{p}$$ และ $\hat{p} \sim \text{Normal}$ เมื่อ $n \geq 30$

$$\therefore \mu_{\hat{p}} = p, \quad \sigma^2_{\hat{p}} = \frac{p(1-p)}{n} \longrightarrow \text{Transform } \hat{p} \to N : \quad Z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}}$$

$$\therefore \quad Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Confidence Interval : $$p = \hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

<u>Distribution ของ $\hat{p}_1 - \hat{p}_2$</u>

มี $\hat{p}_1$ และ $\hat{p}_2$: $$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

Confidence Interval : $$p_1 - p_2 = (\hat{p}_1 - \hat{p}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Hypotheses Testing

- Statistical Hypotheses : สมมติฐานที่คาดการณ์เกี่ยวกับพารามิเตอร์
  - สมมติฐานหลัก (Null)
  - สมมติฐานรอง (Alt.)

- Null hypothesis (H_0) ใช้เป็นทางเลือกในการสรุปผล กรณีตัวอย่าง <u>ยืนยันไม่ได้</u> ว่าสมมติฐานรองเป็นจริง

  → $H_0 : \theta = \theta_0$, $H_0 : \theta \le \theta_0$, $H_0 : \theta \ge \theta_0$

- Alternative hypothesis (H_1) ใช้ตรวจสอบว่าเป็นจริง / ยืนยันได้

  → $H_1 : \theta \ne \theta_0$, $H_1 : \theta > \theta_0$, $H_1 : \theta < \theta_0$

} θ_0 เป็นตัวเลข

* H_1 เป็นพระเอก
H_0 กำหนดมาแย้ง H_1

<u>ขั้นตอน</u> กำหนด H_0, H_1 → กำหนด α → เก็บตัวอย่าง, วัดค่า, คำนวณ

→ กำหนดเกณฑ์ (θ_0) → สรุปผล
- ปฏิเสธ H_0 (H_1 จริง)
- ไม่ปฏิเสธ H_0 (H_1 ยืนยันไม่ได้ว่าจริง)

* นัยสำคัญ (Level of Significance, α) = ความน่าจะเป็นในการปฏิเสธ H_0 ทั้งที่ H_0 เป็นจริง

* หาก θ ห่าง θ_0 มาก ๆ จะปฏิเสธ H_0 ทันที (ต้องชัดเจน และไม่มีข้อสงสัย)

→ มีโอกาสผิดพลาด <u>Error</u>

- Type I : ปฏิเสธ H_0 ทั้งที่ H_0 จริง (θ ห่าง θ_0 เกินไป) → α
- Type II : ไม่ปฏิเสธ H_0 ทั้งที่ H_0 เท็จ (θ แถวรอบ θ_0 ชิดเกินไป) → β

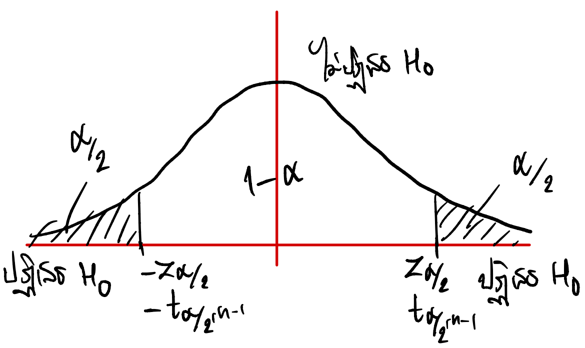
<u>Forms of hypotheses</u>

1. ค่าเฉลี่ย 1 ประชากร ($\bar{X}$)

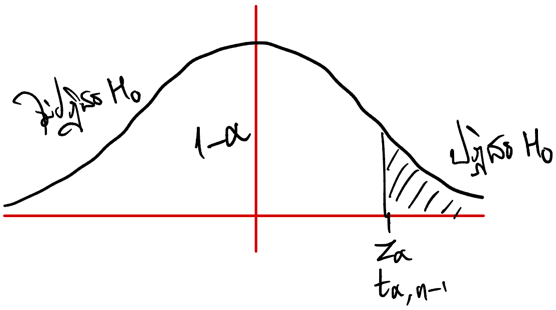
- $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \ne \mu_0 \end{cases}$ $\begin{cases} H_0 : \mu \le \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$ $\begin{cases} H_0 : \mu \ge \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$
- หา $\bar{X}$, S จากตัวอย่างขนาด n
- เลือก Z หรือ T มาใช้ตามเกณฑ์

→ ถ้า $H_1 : \mu \ne \mu_0$ แล้ว ปฏิเสธ H_0
เมื่อ $|Z| > Z_{\alpha/2}$ หรือ $|T| > t_{\alpha/2, n-1}$

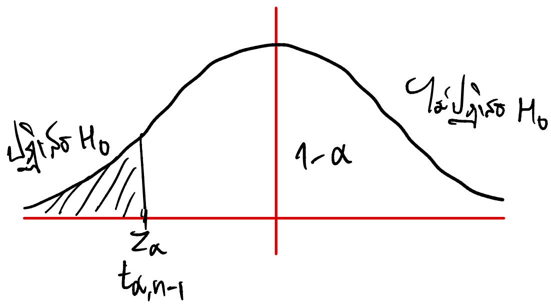
→ $\boxed{Z, T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}}$



→ ถ้า $H_1: \mu > \mu_0$ ปฏิเสธ H_0
เมื่อ $Z > Z_\alpha$ หรือ $T > t_{\alpha,n-1}$



→ ถ้า $H_1: \mu < \mu_0$ ปฏิเสธ H_0
เมื่อ $Z < Z_\alpha$ หรือ $T < t_{\alpha,n-1}$



* หากปฏิเสธ H_0 แปลว่า H_1 เป็นจริง
* หากไม่ปฏิเสธ H_0 แปลว่า ยืนยันไม่ได้ว่า H_1 จริง หรือ H_0 เป็นไปได้
* สามารถดู p-value (sig) ได้ : p-val $< \alpha$: ปฏิเสธ H_0
  p-val $\geq \alpha$: ไม่ปฏิเสธ H_0

2. ค่าเฉลี่ย 2 ประชากร $(\bar{X}_1 - \bar{X}_2)$ same as 1.

$$\begin{cases} H_0: \mu_1 - \mu_2 = d_0 \\ H_1: \mu_1 - \mu_2 \neq d_0 \end{cases} \quad \begin{cases} H_0: \mu_1 - \mu_2 \leq d_0 \\ H_1: \mu_1 - \mu_2 > d_0 \end{cases} \quad \begin{cases} H_0: \mu_1 - \mu_2 \geq d_0 \\ H_1: \mu_1 - \mu_2 < d_0 \end{cases}$$

3. ค่าเฉลี่ย 2 ประชากรที่เก็บมาเป็นคู่ (อิสระ) : D same as 1.

$$\begin{cases} H_0: \mu_d = d_0 \\ H_1: \mu_d \neq d_0 \end{cases} \quad \begin{cases} H_0: \mu_d \leq d_0 \\ H_1: \mu_d > d_0 \end{cases} \quad \begin{cases} H_0: \mu_d \geq d_0 \\ H_1: \mu_d < d_0 \end{cases}$$

4. ความแปรปรวน 1 ประชากร (s^2)

$$\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 \neq \sigma_0^2 \end{cases} \quad \begin{cases} H_0: \sigma^2 \leq \sigma_0^2 \\ H_1: \sigma^2 > \sigma_0^2 \end{cases} \quad \begin{cases} H_0: \sigma^2 \geq \sigma_0^2 \\ H_1: \sigma^2 < \sigma_0^2 \end{cases}$$

→ หา $\bar{X}, s$

→ ใช้ χ^2-Dist.

→ ถ้า $H_1: \sigma^2 \neq \sigma_0^2$ ปฏิเสธ H_0 เมื่อ $\chi^2 > \chi^2_{\frac{\alpha}{2},n-1}$ หรือ $\chi^2 < \chi^2_{1-\frac{\alpha}{2},n-1}$

→ ถ้า $H_1: \sigma^2 < \sigma_0^2$ ปฏิเสธ H_0 เมื่อ $\chi^2 < \chi^2_{1-\alpha,n-1}$

→ ถ้า $H_1: \sigma^2 > \sigma_0^2$ ปฏิเสธ H_0 เมื่อ $\chi^2 > \chi^2_{\alpha,n-1}$

5. ความแปรปรวน 2 ประชากร (s_1^2/s_2^2)

$\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 \\ H_1: \sigma_1^2 \neq \sigma_2^2 \end{cases}$ $\begin{cases} H_0: \sigma_1^2 \leq \sigma_2^2 \\ H_1: \sigma_1^2 > \sigma_2^2 \end{cases}$ $\begin{cases} H_0: \sigma_1^2 \geq \sigma_2^2 \\ H_1: \sigma_1^2 < \sigma_2^2 \end{cases}$

→ ใช้ F-Dist.

→ ถ้า $H_1: \sigma_1^2 \neq \sigma_2^2$ ปฏิเสธ H_0 เมื่อ $F > f_{\frac{\alpha}{2}, n_1-1, n_2-1}$ หรือ $F < f_{1-\frac{\alpha}{2}, n_1-1, n_2-1}$

→ ถ้า $H_1: \sigma_1^2 < \sigma_2^2$ ปฏิเสธ H_0 เมื่อ $F < f_{1-\alpha, n_1-1, n_2-1}$

→ ถ้า $H_1: \sigma_1^2 > \sigma_2^2$ ปฏิเสธ H_0 เมื่อ $F > f_{\alpha, n_1-1, n_2-1}$

6. สัดส่วน 1 ประชากร

$\begin{cases} H_0: p = p_0 \\ H_1: p \neq p_0 \end{cases}$ $\begin{cases} H_0: p \leq p_0 \\ H_1: p > p_0 \end{cases}$ $\begin{cases} H_0: p \geq p_0 \\ H_1: p < p_0 \end{cases}$

→ ใช้ Bernoulli Dist. $\xrightarrow{\text{norm}}$ Z-Dist.

→ ถ้า $H_1: p \neq p_0$ ปฏิเสธ H_0 เมื่อ $|Z| > z_{\alpha/2}$

→ ถ้า $H_1: p < p_0$ ปฏิเสธ H_0 เมื่อ $Z < z_\alpha$

→ ถ้า $H_1: p > p_0$ ปฏิเสธ H_0 เมื่อ $Z > z_\alpha$

7. สัดส่วน 2 ประชากร

$\begin{cases} H_0: p_1 - p_2 = d_0 \\ H_1: p_1 - p_2 \neq d_0 \end{cases}$ $\begin{cases} H_0: p_1 - p_2 \leq d_0 \\ H_1: p_1 - p_2 > d_0 \end{cases}$ $\begin{cases} H_0: p_1 - p_2 \geq d_0 \\ H_1: p_1 - p_2 < d_0 \end{cases}$

หาก $d_0 = 0$: $Z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$; $\bar{p} = \dfrac{a_1 + a_2}{n_1 + n_2}$

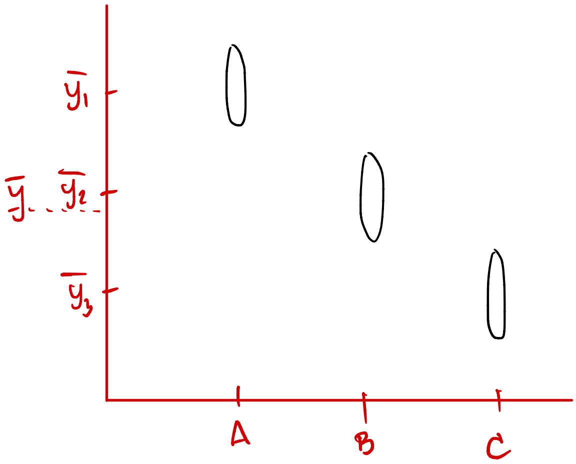
$d_0 \neq 0$: $Z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$ (ปกติ)

→ ถ้า $H_1: p_1 - p_2 \neq d$ ปฏิเสธ H_0 เมื่อ $|Z| > z_{\alpha/2}$

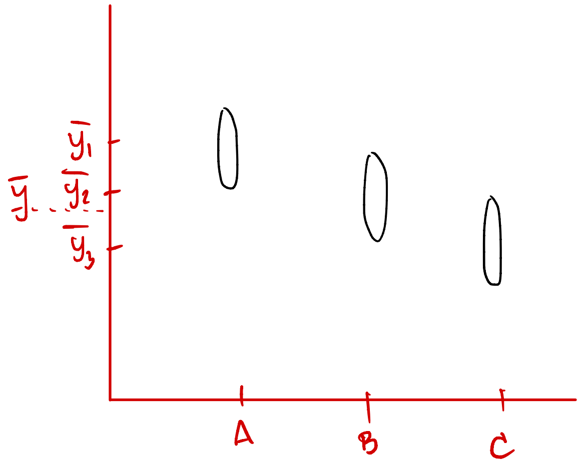
→ ถ้า $H_1: p_1 - p_2 < d$ ปฏิเสธ H_0 เมื่อ $Z < z_\alpha$

→ ถ้า $H_1: p_1 - p_2 > d$ ปฏิเสธ H_0 เมื่อ $Z > z_\alpha$

1 ANOVA : One-way Analysis of Variance



มีความแตกต่างของ mean



ไม่มีความแตกต่างของ mean

ความผันแปรรวม (Total sum of squared) :

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = SSB + SSW$$

$$SST = \sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_1}(y_{ij}-\bar{y}_i)$$

ความผันแปรระหว่างกลุ่ม (Between-groups SS) :

$$SSB = \sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2$$

ความผันแปรภายในกลุ่ม (Within-group SS) :

$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_1}(y_{ij}-\bar{y}_i)^2$$

- เปรียบเทียบ SSB, SSW → Normalize

→ $MSB = \dfrac{SSB}{df_B}$, $MSW = \dfrac{SSW}{df_W}$; $df_B = k-1$, $df_W = N-k$, $df_T = N-1 = df_B + df_W$

→ $F = \dfrac{MSB}{MSW}$; $F \sim f(df_B, df_W)$ ← f-Dist.

$N = n_1 + n_2 + \dots + n_k$ = ขนาดตัวอย่างทั้งหมด

k = จำนวนกลุ่มเปรียบเทียบ

n_i = ขนาดตัวอย่างกลุ่มที่ i $(i \le k)$

* F ใหญ่กว่า ค่า f ที่เป็นจุดวิกฤตมากๆ → MSB >> MSW → มีความแตกต่างระหว่างกลุ่ม อย่างมีนัยสำคัญทางสถิติ.

Hypothesis:
1. $X_k \sim$ Normal หรือคล้าย Normal
2. $S^2_{max}/S^2_{min} \in (0.25, 4)$
3. $X_i, X_{j \ne i}$ ต้องเป็นอิสระกัน

ANOVA : 1. $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$

$H_1 : \mu_i \neq \mu_j : \exists i,j : i \neq j \quad , i,j = 1, 2, \dots, k$

2. กำหนด α

3. สุ่มตัวอย่างขนาด n_i จากกลุ่มที่ i และวัด $y_{i,1}, y_{i,2}, y_{i,3}, \dots, y_{i,n_i}$

4. ทำเป็นตารางผลรวม

X_1	X_2	...	X_k	
y_{11} y_{12} y_{13} $\vdots$ y_{1n_1}	y_{21} y_{22} y_{23} $\vdots$ y_{2n_2}	...	y_{k1} y_{k2} y_{k3} $\vdots$ y_{kn_k}	
$T_1 = \sum_{j=1}^{n_1} y_{1j}$	$T_2 = \sum_{j=1}^{n_2} y_{2j}$	...	$T_k = \sum_{j=1}^{n_k} y_{kj}$	$T = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \sum_{i=1}^{k} T_i$
$\sum_{j=1}^{n_1} y_{1j}^2$	$\sum_{j=1}^{n_2} y_{2j}^2$	...	$\sum_{j=1}^{n_k} y_{kj}^2$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2$

* $SST = SSB + SSW$

* $SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{T^2}{N}$ * $SSB = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - \frac{T^2}{N}$

; $N = n_1 + \dots + n_k$, $SSW = SST - SSB$

5. ตาราง F-test

Source	df	SS	MS	F_{Test}
B	$k-1$	SSB	$MSB = SSB/(k-1)$	$F = \frac{MSB}{MSW}$
W	$N-k$	SSW	$MSW = SSW/(N-k)$	
T	$N-1$	SST		

6. ปฏิเสธ H_0 เมื่อ $F_{Test} > F_{\alpha, df_B, df_W}$

Simple Regression & Correlation

Regression : ตัวแปรตาม y ขึ้นกับตัวแปรอิสระ $x_1, x_2, \ldots, x_p$

- $y = f(\vec{x})$
- แบบเส้นตรง : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon_i$

Correlation coefficient (ρ, r)

→ $r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$;

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

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

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

→ $|r| \geq 0.7$: สัมพันธ์มาก

$|r| \geq 0.3$: สัมพันธ์ปานกลาง

$|r| < 0.3$: สัมพันธ์น้อย

$r > 0$: ตาม

$r < 0$: ตรงข้าม

Hypothesis :

1. $\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$
2. α
3. r
4. $T = r\sqrt{\dfrac{n-2}{1-r^2}}$
5. ปฏิเสธ H_0 เมื่อ $|T| > t_{\frac{\alpha}{2}, n-2}$

Regression : $y = E(y) + \varepsilon$; $E(y) = \beta_0 + \beta_1 x$ สปส. ถดถอย

ให้ y depends on independent x

มีข้อมูลจากตัวอย่างสุ่มขนาด n : $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

หา β_0, β_1 ด้วย Least-square method

ผลลัพธ์ :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Standard error of estimate (SEE)

$$SEE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} \quad ; \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$SEE = \sqrt{\frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}} \quad ✓$$

โดย $SEE = 0 \rightarrow$ Perfect linearity

$SEE \rightarrow 0 \rightarrow$ High linearity

$SEE \rightarrow \infty \rightarrow$ Low linearity

Coefficient of determination (R^2)

$$S_{yy} = SSE + SSR \longrightarrow \underbrace{\frac{SSR}{S_{yy}}}_{R^2} + \frac{SSE}{S_{yy}} = 1$$

$$\sum (y_i - \bar{y})^2 = \underbrace{\sum (y_i - \hat{y}_i)^2}_{SSE} + \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{SSR}$$

หรือ $\boxed{R^2 = r^2}$ ก็ได้

$R^2 = 1$	100 %	ความน่าเชื่อถือ
$R^2 \rightarrow 1$	มาก	
$R^2 \rightarrow 0$	น้อย	
$R^2 = 0$	0 %	

Chi-squared Independence test (For ordinal scale / nominal scale)

1. Hypothesis $\begin{cases} H_0: & X_1, X_2 \text{ อิสระกัน} \\ H_1: & X_1, X_2 \text{ ไม่อิสระกัน} \end{cases}$

   หรือ $\begin{cases} H_0: & X_1 \text{ affects } X_2 \\ H_1: & X_1 \text{ \underline{not} affect } X_2 \end{cases}$

2. α
3. Contingency table

$\#X_1$	$\#X_2$				tot.
	1	2	...	c	
1	O_{11}	O_{12}	...	O_{1c}	$TR1$
2	O_{21}	O_{22}	...	O_{2c}	$TR2$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	O_{r1}	O_{r2}	...	O_{rc}	TR_r
tot.	TC_1	TC_2	...	TC_c	n

* O_{ij} = จำนวนหน่วยศึกษา $=i$ ใน X_1, $=j$ ใน X_2 (Observed frequency)

4. หาค่า E_{ij} (ความคาดหวัง) ในช่องเดียวกับ O_{ij}

   $\hookrightarrow E_{ij} = \dfrac{T_{R_i} \times T_{C_j}}{n}$

5. หาค่า $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \dfrac{(O_{ij} - E_{ij})^2}{E_{ij}}$

6. ปฏิเสธ H_0 เมื่อ $\chi^2 > \chi^2_{\alpha, r-1, c-1}$