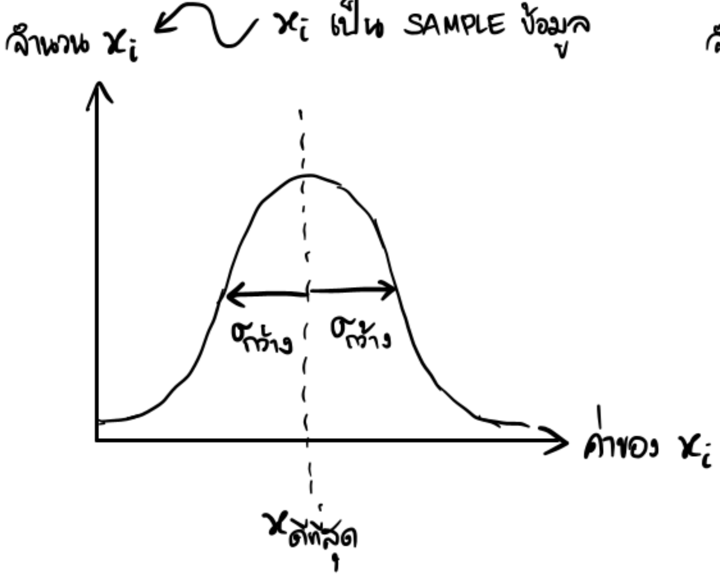
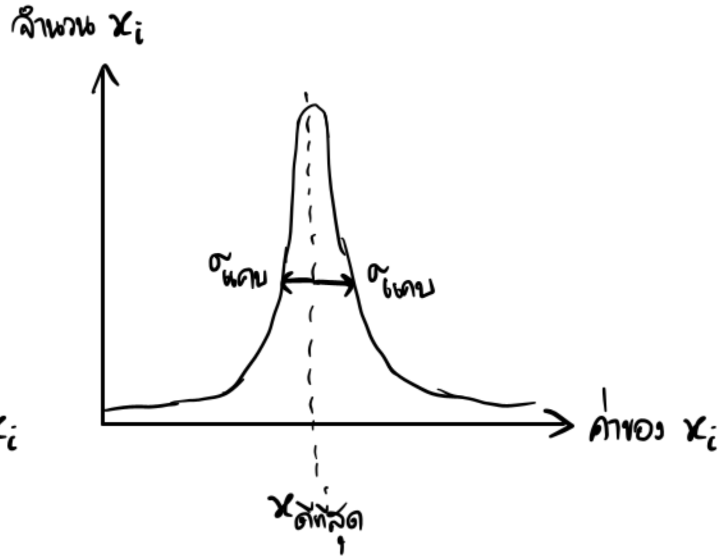
# *Data Analysis* (DA101) : ข้อมูลตัวแปรเดียว และหลายตัวแปร

*Measurement of Average*

- *First Order* (*central tendency*)
  - $\bar{x}$   *Median*   *Mode*
  - $\boxed{\text{พื้นฐาน}}$
- * *Second Order* (*dispersion*)
  - $\boxed{\text{ประยุกต์}}$
- *Third Order* (*skewness & kurtosis*)
  - $\boxed{\text{ขั้นสูงชำนาญ}}$

การกระจาย/ส่วนเบี่ยงเบนเป็นการวัดข้อมูลทางสถิติลำดับที่สอง ใช้บอกความเป็นกลุ่มของข้อมูล




สมมติเหตุการณ์การวัดค่า โดยทั่วไปนั้นเครื่องมือที่ใช้วัดค่า มักมีความคลาดเคลื่อนในตัวเสมอ ไม่มีเครื่องมือไหนบนโลกที่เที่ยงตรง (หรือแม่นยำ) (วัดค่าได้ตรงกับความจริงที่สุด เรียกว่า *Accuracy*) และวัดละเอียด (วัดได้ค่าเดิมซ้ำ ๆ เรียกว่า *Precision*) โดยทั้ง 2 ค่า มีผลกับส่วนเบี่ยงเบนโดยตรง กราฟยิ่งแคบ (ด้านขวา) ยิ่งดี ค่าเกาะกลุ่มกัน ทั่วไปมักใช้ส่วนเบี่ยงเบนมาตรฐาน (STANDARD DEVIATION) แต่ก็มีค่าอื่น ๆ ที่นำมาใช้แทนได้ แต่ไม่ค่อยนิยมมากนัก

<u>หมายเหตุ</u> ในบทนี้อาจมีการนำแคลคูลัสมาใช้ หากไม่ชินก็อ่านไปแล้วเชื่อไปก่อนว่าจริงก็ได้ แต่หากต้องการรู้เนื้อหา สามารถอ่านเองได้ ไม่ต้องพึ่งคนสอน

โดยก่อนเริ่มพูดถึงค่าวิลิศมาหราเหล่านั้น ต้องมาพูดถึงหัวใจของค่าต่าง ๆ ทางสถิติก่อน นักสถิติมองว่า การเก็บหรือวัดข้อมูลครั้งเดียว ไม่มีความน่าเชื่อถือ เพราะเรามิอาจทราบได้ว่าค่านั้นสามารถบ่งชี้ความเป็นจริงได้หรือไม่ นักสถิติจึงเลือกเก็บข้อมูลหลาย ๆ "ตัวอย่าง" จากกลุ่ม "ประชากร" และไม้บรรทัด แต่แล้วก็ยังไม่ทราบวิธีนำข้อมูลเหล่านั้นมาใช้ประโยชน์ได้อย่างเต็มที่ ก็เพราะมันเยอะไปหมด ต้องคิดหาวิธีหาตัวแทนข้อมูลเหล่านั้นออกมา เพื่อให้คนที่ขี้เกียจอ่านข้อมูลดิบมาก ๆ เข้าใจได้อย่างลึกซึ้งถึงสิ่งที่ข้อมูลต้องการแสดงออกมา หรือสิ่งที่คาดว่าจะได้เห็นจากข้อมูลนั้น ๆ จึงคิด "ค่าคาดหวัง" (EXPECTED VALUE) ออกมา โดยใช้แนวคิดของความน่าจะเป็นในการพบข้อมูลนั้น กระจายออกมาในทุก ๆ ข้อมูล (PROBABILITY DISTRIBUTION) นำมาคูณเป็น *factor* ของค่านั้น ๆ แล้วนำมารวมกัน จะได้ตัวแทนข้อมูลออกมา เขียนในรูปคณิตศาสตร์ว่า

ให้เข้าใจเสมอว่า 1 = 100%

$$E[X] = \sum_{i=1}^{N} p_i x_i$$

EXPECTED VALUE — $E[X]$

PROBABILITY OF $x_i$ — $p_i$

DISCRETE VARIABLE $i$ — $x_i$

โดย $\sum p_i = 1$ เสมอ

และ $E[X] = constant$

คือ $E[YE[X]] = E[X]E[Y]$

ซึ่งค่าค่านี้เป็นที่มาของค่าทางสถิติหลาย ๆ ค่ามาก ซึ่งในที่นี้จะยกมาบางค่าเท่าที่อยากเกริ่นนำให้รู้จัก หากต้องการเปิดโลกการวิเคราะห์ข้อมูลทางสถิติในแบบต่าง ๆ (แปลก ๆ) ขอให้ผู้อ่านทำการศึกษาค้นคว้าด้วยตนเอง เพราะมันจะทำให้เราอยากอ่าน และหาไปเรื่อย ๆ และขอให้ตั้งคำถามไว้เสมอว่า <u>ทำไม?</u> เพราะวิชาสถิติหลายคนมองว่าเป็นวิชาท่องสูตรแล้วทำโจทย์ แต่จริง ๆ แล้ว ทุกอย่างย่อมมีที่มา และนำมาใช้ประโยชน์ได้ทั้งนั้น

สูตร 1 สูตร 4
สูตร 2 ⋮
สูตร 3 สูตร N

ซั้ปังโต้

ที่มา + sense ทาง math

ฉลาด

1. STANDARD DEVIATION (S.D. หรือ $\sigma$)

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}} \longrightarrow \sigma^2 = \text{ความแปรปรวน}$$

(VARIANCE)
บางครั้งอาจเห็นในรูป $var(x)$

2. MEAN DEVIATION (M.D.)

$$M.D. = \frac{\sum |x_i - \bar{x}|}{N}$$

3. INTERQUARTILE RANGE (I.Q.R.)

$$I.Q.R. = Q_3 - Q_1$$

4. INTERQUARTILE DEVIATION (Q.D.)

$$Q.D. = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

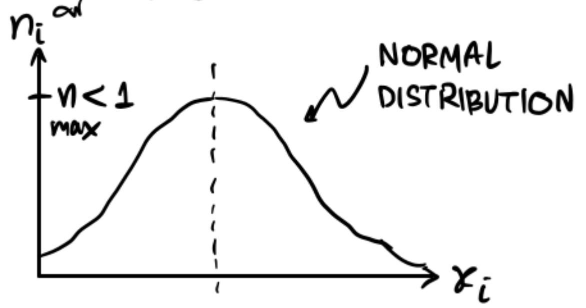
## <u>การประยุกต์ใช้ความแปรปรวนในข้อมูลทางสถิติ</u>

① NORMALIZATION เบื้องต้น : ค่า $z$

โดยทั่วไปข้อมูลจะกระจายเป็นหย่อมๆ เมื่อนำมาพล็อตใน *histogram* แล้วจะมีแอมพลิจูดไม่คงที่ (ผลรวมของพื้นที่ใต้กราฟเป็นผลรวมของข้อมูล ($\sum x_i$)) แต่เมื่อผ่านกระบวนการ ผลรวมของข้อมูลจะเป็นแบบเทียบกับ 100% หรือ 1.00 (ทั้งหมด) โดยเทียบอัตราส่วนความเบี่ยงเบนจากค่าเฉลี่ย (ERROR) กับส่วนเบี่ยงเบนมาตรฐาน (DEVIATION)

ผลต่าง

จะได้ค่า $z_i = \frac{x_i - \bar{x}}{\sigma}$ ซึ่งมีรูปร่างหน้าตา $\longrightarrow$



PDF (PROBABILITY DISTRIBUTION FUNCTION)

โดยเขียนสมการโค้งระฆังคว่ำ NORMAL ได้เป็น $P(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$ ⊛

“95% CONFIDENCE INTERVAL”

และในทางที่นิยมนำมาหากลุ่มข้อมูลส่วนใหญ่ (95% - 97.5% ของข้อมูล) ซึ่งอยู่บริเวณกึ่งกลาง (ในที่นี้ $\bar{x}$) จะรายงานว่า $A \pm B$ ซึ่งหมายถึงข้อมูลส่วนใหญ่มีค่าตั้งแต่ $A - B$ ไปถึง $A + B$ ซึ่งค่า A คือ $\bar{x}$ (กึ่งกลาง) และ B คือขอบเขต คำนวณจากการหาพื้นที่ใต้กราฟ $P(x)$ คือ

แคลคูลัสเล็กน้อย

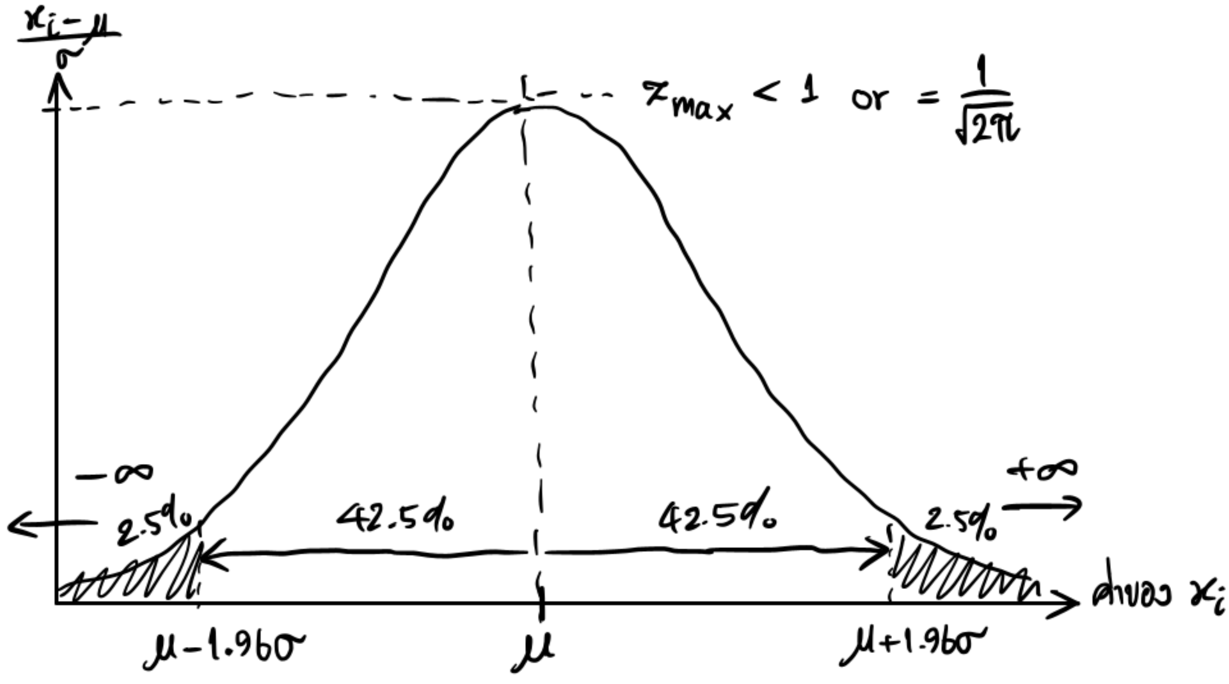
$$\mathcal{A} = \int_{-y}^{y} P(x)\,dx = 0.95$$

เมื่อแก้ด้วยวิธีทางคณิตศาสตร์ จะได้ $y = 1.96$ ซึ่งฟังก์ชัน $P(x)$ เป็นกรณีพิเศษสำหรับ $\mu$ (เขียนแทน $\bar{x}$) $= 0$ และ $\sigma = 1$ สำหรับกรณีทั่วไป จะใช้ฟังก์ชัน

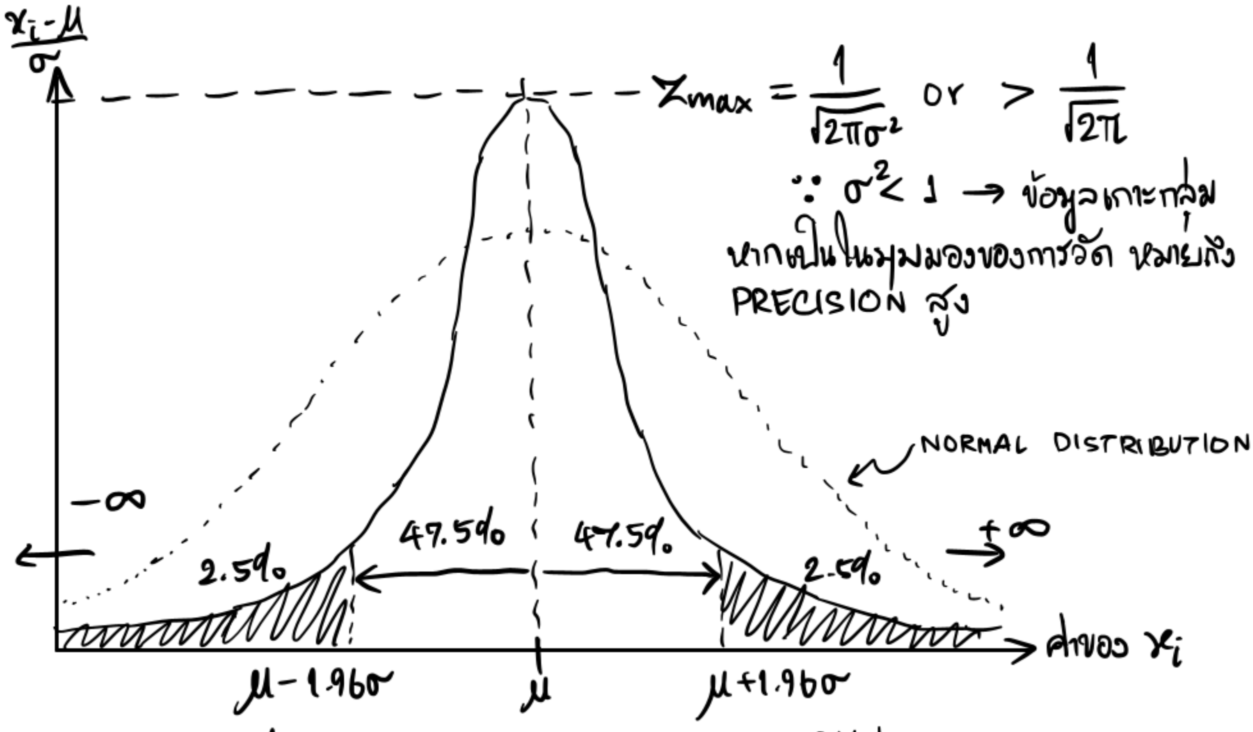
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$ ⊛ เพิ่มเติม

เรียกว่า SYMMETRICAL DISTRIBUTION (GENERAL FORM)

ดังนั้นเมื่อคิด *factor* $\sigma$ เข้ามาเกี่ยวข้อง จะได้ $A \pm B = \mu \pm 1.96\sigma$




(การกระจายแบบ NORMAL)




(กรณีไม่กระจายแบบ NORMAL / สมมาตรทั่วไป)

กลุ่มข้อมูลจำพวก 5% ในความเป็นจริงมักมองว่าเป็นประชากรส่วนน้อย, ค่า ERROR, ค่า NOISE, ความผิดพลาดในการวัด, ฯลฯ มักไม่นำมาคิด

## ② VARIANCE ($\sigma^2$, $Var(x)$, $Cov(x,x)$)

ทั่วไปเราอาจใช้สูตร STDEV$^2$ มาคิด แต่แท้จริงแล้ว สูตรพวกนี้มาจากไหน?

หาก X เป็นตัวแปรสุ่มและไม่ต่อเนื่อง (*random discrete variable*) ความแปรปรวน นิยามเป็นผลรวมของผลคูณของ กำลังสองของผลต่างกับโอกาสที่จะเจอตัวแปรที่มีค่านั้นในทางสถิติ หรือเขียนในรูปทั่วไปว่า $Var(x) = \sum_{i=1}^{N} p_i (x_i-\mu)^2$ ซึ่ง $p_i = \frac{n_i}{N}$ ← จำนวน $x_i$

หาก $x_i$ เป็นตัวแปรเดี่ยว จะได้ $p_i = \overset{\text{constant}}{\frac{1}{N}}$ ซึ่งจะได้รูปพิเศษ (SPECIAL CASE) ที่เราเรียนกันในวิชาสถิติ ม.ปลาย – มหาลัยฯ ว่า $Var(x) = \frac{\sum(x_i-\mu)^2}{N} \geq 0$ เสมอ

เราอาจมอง VARIANCE ในรูปแบบดั้งเดิมในเทอมของค่าคาดหวัง (EXPECTED VALUE) ว่า $Var(X) = E[(X_i-\mu)^2] = \sum p_i(x_i-\mu)^2$ ก็จะได้รูปเดิม

และหากเราใช้ค่าคาดหวังมาหาที่มาของค่าเฉลี่ย : $\mu = E[X] = \sum p_i x_i$ ; $p_i = \frac{1}{N}$

$\therefore \mu = \bar{x} = \frac{\sum x_i}{N}$ ได้เช่นกัน

ในอีกมุมมอง VARIANCE , จาก $\mu = E[X]$ และ $Var(X) = E[(X_i - \mu)^2]$

ได้ว่า $Var(X) = E[(X - E[X])^2]$

$$= E[X^2 - 2XE[X] + (E[X])^2] \quad ; \; E[c\xi] = cE(\xi)$$

$$= E[X^2] - E[X]^2$$

$$= \frac{\sum X^2}{N} - \mu^2 = \frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2$$

ซึ่งเป็นสูตรที่เราเรียนกันมาโดยไม่ทราบที่มาที่ไป

ค่าอื่น ๆ ที่มาจาก $E[X]$

- WEIGHTED MEAN : $\bar{x} = E[X]$

$$= \sum p_i x_i \quad ; \; p_i = \frac{w_i}{\sum w_i}$$

$$= \sum \frac{w_i}{\sum w_i} x_i$$

$$= \frac{\sum w_i x_i}{\sum w_i}$$

MATCHING $\square^2$ AND $\sqrt{\square}$

- ROOT MEAN SQUARED : $x_{RMS} = \sqrt{E[X^2]}$

$$= \sqrt{\sum p_i x_i^2}$$

$$= \sqrt{\frac{\sum x_i^2}{N}}$$

ใช้ PDF แทนเพราะนิยามเดียวกัน

- CONTINUOUS CASE : $E[X] = \lim_{\Delta x \to 0} \sum_{i=1}^{N} p_i x_i$ กับ $\sum p_i = 1 : \int_{\mathbb{R}} f(x)\,dx = 1$

$$E[X] = \int_{\mathbb{R}} x f(x)\,dx$$

<u>สมบัติของ VARIANCE</u>

1. $Var(X) \geq 0$
2. $Var(X) = 0$ if $x = constant$
3. $Var(X+a) = Var(X)$
4. $Var(aX) = a^2 Var(X)$
5. $Var(aX+bY) = a^2 Var(X) + b^2 Var(Y) + 2ab\,Cov(X,Y)$
6. $Var(aX-bY) = a^2 Var(X) + b^2 Var(Y) - 2ab\,Cov(X,Y)$

ทำไม ?

5. ผลรวมของ $Var(X_i)$ ที่ $x_i$ และ $x_j$ ซึ่งไม่ใช่ค่าเดียวกัน และไม่มีความสัมพันธ์กันเลย (เขียนในภาษาคณิตศาสตร์ได้ว่า $Cov(x_i, x_j) = 0, \forall(i \neq j)$) ได้ว่า

$$Var\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} Var(X_i)$$

หรือในรูปที่อ่านง่ายขึ้นเล็กน้อยว่า $\sigma_{SUM}^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \dots + \sigma_N^2$

## ③ COVARIANCE ($Cov(X)$, $\sigma(X,Y)$, $\sigma_{XY}$)

COVARIANCE คือ VARIANCE ในรูปแบบของข้อมูล 2 ตัวว่ามีความแปรปรวน "ร่วมกัน" เท่าใด แรก ๆ อาจยังไม่เข้าใจ หากอธิบายง่าย ๆ คือ ข้อมูล 2 ตัวเบี่ยงเบนไปจากค่าเฉลี่ยของตัวมันเองเท่าไร เขียนในรูปคณิตศาสตร์ได้ว่า

$$\begin{aligned} Cov(X,Y) &= E\left[(X-E[X])(Y-E[Y])\right] \\ &= E\left[XY - XE[Y] - YE[X] + E[X]E[Y]\right] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

หรือหากมอง X, Y เป็น DISCRETE VARIABLES :

$$Cov(X,Y) = \sum_{i=1}^{N} p_i (x_i - E[X])(y_i - E[Y])$$

ซึ่ง $p_i$ ในที่นี้ คือโอกาสที่จะเจอคู่อันดับ $(x_i, y_i)$ / อาจพูดได้ว่า $p_{XY}(x,y) = P(X \cap Y)$

ย้อนกลับไปทฤษฎีของเบย์ (*Bayes' Theorem*)!

ให้ A และ B เป็นเหตุการณ์

① หาก A และ B ไม่เกี่ยวเนื่อง (*independent*) :

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

② หาก A และ B เกี่ยวเนื่องกัน (*dependent*) :

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

$P(A|B)$ หมายถึงโอกาสที่ A จะเกิดก็ต่อเมื่อมีเหตุการณ์ B เกิดขึ้นแล้ว เช่น $\underbrace{\text{ป้อมมาทำงานสาย}}_{A}$ ขณะที่ $\underbrace{\text{ฝนตก}}_{B}$

ซึ่ง ② เป็นรูปทั่วไปของความน่าจะเป็นของความสัมพันธ์ของทั้ง 2 เหตุการณ์
เบย์ได้ใช้เงื่อนไขนั้นมาหาความสัมพันธ์ :

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

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

$$\therefore \quad P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \rightsquigarrow \text{Bayes' Theorem}$$

กลับมา COVARIANCE !

หากเรามอง Cov. ของตัวแปรเดิม , Cov(X,X) จะได้

$$\begin{aligned} Cov(X,X) &= E[X \cdot X] - E[X]E[X] \\ &= E[X^2] - E[X]^2 \\ &= Var(X) \end{aligned}$$

และเมื่อ X และ Y ไม่สัมพันธ์กัน $\xrightarrow[\text{ถ้า... แล้ว}]{}$ $Cov(X,Y) = 0$

สมบัติพิเศษบางประการ

ให้ $X, Y, W, V$ = random variables $\in \mathbb{R}$ และ $a, b, c, d$ = constants $\in \mathbb{R}$

1. $Cov(X, a) = 0$
2. $Cov(X,X) = Var(X)$
3. $Cov(X,Y) = Cov(Y,X)$
4. $Cov(aX, bY) = ab\,Cov(X,Y)$
5. $Cov(X+a, Y+b) = Cov(X,Y)$

7. $\sigma_{XY}^2 \le \sigma_X^2 \sigma_Y^2$

คล้ายๆ กับอะไร?

(ลองหาอ่าน TRIANGLE INEQUALITY)
(หาอ่านเรื่อง Vector : Cauchy-Schwarz inequality)

6. $Cov(aX+bY, cW+dV) = ac\,Cov(X,W) + ad\,Cov(X,V) + bc\,Cov(Y,W) + bd\,Cov(Y,V)$ } ทำไม?

มาลองดูเคสพิเศษของ COVARIANCE ว่าเอามาใช้อะไรได้บ้าง?

จาก $Cov(X,Y) = E[XY] - E[X]E[Y]$

$\rightsquigarrow Cov(X,Y) = 0 \rightarrow E[XY] = E[X]E[Y]$

$$\sum\sum(p_{xy}x_iy_i) = (\sum p_ix_i)(\sum p_iy_i)$$

For discrete $x_i, y_i$ ; $p_i = \frac{1}{N}$

$$\sum\sum p_xp_yx_iy_i = (\sum p_xx_i)(\sum p_yy_i)$$

$$\therefore \sum(p_xx_i)\sum(p_yy_i) = \sum(p_xx_i)\sum(p_yy_i) \blacksquare$$

หากให้ ตัวแปร x ไม่ขึ้นกับ y (y อาจเป็น constant),

$$N \cdot c\sum x_i = N \cdot c\sum x_i \rightarrow \text{เป็นจริง} \blacksquare$$

④ NORMALIZATION เบื้องต้น : ค่าสัมประสิทธิ์ความแปรปรวน

COEFFICIENT OF VARIATION ($C_V$) หรือ RELATIVE S.D. คือการนำค่าส่วนเบี่ยงเบนมาตรฐานมาคิดเทียบกับค่าเฉลี่ย หมายถึงว่า ค่าที่วัด/สำรวจได้ เบี่ยงเบนไปเป็นกี่เท่า / กี่% ของค่าที่ควรเป็น เช่น 2.52 cm ± 5% ← 0.05 เท่าของ 2.52 นั่นคือ 0.13 cm

เขียนในรูปคณิตศาสตร์ได้ว่า

$$C_V = \frac{\sigma}{\mu}$$

บางครั้งหากเรารายงาน S.D. = 950 อาจจะตกใจว่าทำไมจึงเยอะ แต่พอมารู้ทีหลังว่าค่าที่วัดได้คือ 2,500,680 ซึ่งโล่งอกไปที เพราะเพี้ยน ± 0.04% เท่านั้น (น้อยมาก) แต่ถ้าค่าที่วัดได้เป็น 1,000 แล้ว S.D. เป็น 950 คงต้องตกใจเป็นอย่างมากเพราะค่าเพี้ยนเป็นเกือบเท่าตัวของค่าจริง (<u>เลว</u>) แต่ไม่เหมาะกับค่าเล็ก ๆ เพราะ เมื่อ $\mu \to 0$ จะได้ $C_V \to \pm\infty$ ซึ่งไม่สวยเท่าไรนัก

## ⑤ NORMALIZATION เบื้องต้น : ค่า STANDARD ERROR / ${}_x\sigma_{mean}$

STANDARD ERROR (SE, $\sigma_{\bar{x}}$) มีความแพร่หลายในการใช้งานมากกว่า $C_V$ เพราะถูกหลักการการคิด *sample* (ตัวอย่าง) ของข้อมูลหลาย ๆ ชุดมากกว่า โดยคิดจากค่าคาดหวังของความแปรปรวน

POPULATION $\to \sigma \approx \sigma_x \leftarrow$ SAMPLE

$$\sigma_{\bar{x}}^2 = E[E[\sigma^2]] \quad \text{โดย } \sigma \approx \sigma_x$$

$$\sigma_{\bar{x}}^2 = \frac{1}{n}\sum\frac{\sigma^2}{N}$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

$$\therefore \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$\sigma \to \sigma_{sample}$ ใช้ $N-1$

หรือ

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + \dots + \sigma_N^2$$

$$\sigma^2 = N\sigma_{\bar{x}}^2$$

$$\therefore \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \frac{\sigma}{\sqrt{n}}$$

ซึ่งค่านี้นิยมใช้กับการวัดค่าด้วยเครื่องมือ ย้ำ ๆ หลายครั้งหาค่าที่ดีที่สุด (ค่าอะไร? อย่างไร?) สามารถใช้ $\sigma_{\bar{x}}$ เป็นค่า *uncertainty* ได้เลยหากเทียบกับค่าจากเครื่องมือแล้วเหมาะสม (if $\sigma_{\bar{x}} > \delta x$) $\delta x$ คือ UNCERTAINTY OF MEASUREMENT ซึ่งนำค่า $\sigma_{\bar{x}}$ มาใช้รายงาน 95% *confidence interval* ($\mu \pm 1.96\sigma_{\bar{x}}$) ได้ด้วย

## ⑥ NORMALIZATION ประยุกต์ : ค่าสัมประสิทธิ์สหสัมพันธ์

ค่าสัมประสิทธิ์สหสัมพันธ์ (PEARSON CORRELATION COEFFICIENT, $r$, $\rho$) คือค่าที่บอกความสัมพันธ์ของข้อมูล 2 ชุดว่าสัมพันธ์/เกี่ยวเนื่องกันหรือไม่ อย่างไร ซึ่งช่วงของค่าบ่งชี้ได้ดังนี้ :

$$r = \begin{cases} +x & ; \text{ สัมพันธ์แปรผันตาม } 0 < r \leq 1 \\ 0 & ; \text{ ไม่สัมพันธ์กันเลย, } r = 0 \\ -x & ; \text{ สัมพันธ์แปรผกผัน } -1 \leq r < 0 \end{cases}$$

โดยค่า $r$ คำนวณได้จากค่าความสัมพันธ์ของ 2 ตัวแปร (COVARIANCE) มาทำการ NORMALIZATION กับ $\sigma_x\sigma_y$ (เหมือนความสัมพันธ์ INEQUALITY ใน COV. ที่ว่า

$$\sigma_{xy}^2 \leq \sigma_x^2\sigma_y^2 \to \left|\frac{\sigma_{xy}}{\sigma_x\sigma_y}\right| \leq 1 \to -1 \leq \frac{\sigma_{xy}}{\sigma_x\sigma_y} \leq 1$$

จะได้ว่า $$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad \circledast$$

หรือ $$= \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

หรือ $$= \frac{E[XY]-E[X]E[Y]}{\sqrt{E[X^2]-E[X]^2}\sqrt{E[Y^2]-E[Y]^2}}$$

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

หรือ $$= \frac{1}{n-1}\sum z_x z_y$$ มาจากไหน ?

ซึ่งทั้งหมดที่กล่าวมาล้วนมีความคล้ายคลึงกับ COSINE'S SIMILARITY:

$$\cos\theta = \frac{\vec{A}\cdot\vec{B}}{AB} = \frac{\sum A_i B_i}{\sqrt{\sum A_i^2}\sqrt{\sum B_i^2}} \quad ; \quad A_i = NE(X), \; B_i = NE(Y)$$

$$A = |\vec{A}| = \vec{A}\cdot\vec{A} = \sum A_i^2$$

$$\vec{A} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \rightsquigarrow \vec{A}\cdot\vec{A} = \sum A^2, \; \vec{B} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

⊕ มีศาสตร์ที่น่าสนใจ น้อยคนจะหาศึกษาดู (ไม่ค่อยได้ยินใครพูดถึงเรื่องนี้สักเท่าไร) คือ เรื่อง DIRECTIONAL STATISTICS (CIRCULAR STATISTICS)

กลับมาที่การวัด

หรือ No tools are ก็ได้.

การวัดทุกอย่าง ย่อมมีความคลาดเคลื่อนเสมอ ( No tool is perfect. ) ซึ่งค่านั้นเกิดขึ้นจริง มีอยู่จริงตลอดขั้นตอน เพียงเราไม่รู้ว่ามันมีค่าเท่าไร

ทางสถิติ (STATISTICAL)

ความคลาดเคลื่อนแบ่งเป็น 2 ประเภทหลัก ๆ คือแบบสุ่ม (RANDOM) และแบบที่มีเหตุผลชัดเจน ( SYSTEMATIC ) ซึ่งสามารถลดทอนได้ด้วยวิธีต่าง ๆ แต่เมื่อรวม ๆ กันแล้ว มันก็ไม่หายไปทั้งหมด เหลือเพียงสิ่งที่เราต้องรู้ที่จะอยู่กับมัน และคำนึงถึงผลลัพธ์ที่จะเกิดจากมัน และวิธีรายงาน อย่างที่เคยกล่าวไว้ เครื่องมือที่เที่ยงตรง/แม่นยำ ( ACCURATE ) และละเอียด ( PRECISE ) ย่อมถูก PREFER เสมอ ตัวอย่างที่ชัดของความคลาดเคลื่อน คือ อัตราเร็วคลื่นแม่เหล็กไฟฟ้าในสุญญากาศ มีค่าเท่ากับ $299792458 \pm 1 \ m \cdot s^{-1}$ , $\pm 1$ คือความคลาดเคลื่อนจากการวัดค่าค่านี้ขึ้นมา มาจากไหน ?

<u>การรายงานค่าจากการวัด</u> ( หลักสำคัญที่ถูกต้อง )

อ่านว่า *delta* ถ้า $\Delta$ จะเป็น *Delta*

1. ค่าสุดท้ายท้ายกระดาษ เขียนในรูปแบบ $x \pm \delta x$ ซึ่งหมายถึงค่าของ $x$ มีได้ระหว่าง $x - \delta x$ ถึง $x + \delta x$ โดย $\delta x > 0$ เสมอ <u>ยกเว้น</u> ระบุไว้ เช่น TOLERANCE ในการเขียนแบบทางวิศวกรรม จะเขียนในรูป $50.00^{\,0}_{-0.50}$

และเป็นไปตามหลักการเขียนเลขนัยสำคัญ

01.2340 ตัวเลขนี้เป็นตัวเลขทศนิยม 4 ตำแหน่ง
1 2 3 4 5 มีเลขนัยสำคัญ 5 ตัว
1 2 3 4

สมบัติของเลขนัยสำคัญ

1. การบวก/ลบ ให้ใช้เลขทศนิยมตำแหน่งน้อยที่สุด
   เช่น $25.26 + 50.3 = 75.56$
   (2) (1) $\approx 75.6$
2. การคูณ/หาร ให้ใช้เลขนัยสำคัญน้อยที่สุด
   เช่น $0.00059 \times 20.0 = 0.0118$
   (2) (3) $\approx 0.012$

สำหรับความคลาดเคลื่อน ให้ใช้เลขนัยสำคัญเพียง 1 จำนวนเท่านั้น และทศนิยมตำแหน่งสุดท้ายตรงกัน เช่น

นาย ก. สูง $179.8 \pm 0.4$ cm หรือ
$1.798 \pm 0.004$ m หรือ
$1798 \pm 4$ mm

อย่า รายงานค่าเป็น $179.82 \pm 0.4$ cm หรือ
$179.8 \pm 0.43$ cm (*)

แต่ในบางครั้งอาจเจอ $179.82 \pm 0.45$ cm แต่ไม่นิยมเป็นอย่างมาก

(*) เป็นตัวเลขที่มีนัยสำคัญต่ำที่สุด (1) ให้ปัดเลขก่อน จึงรายงาน เช่น $\pm 0.45 \rightarrow \pm 0.5$

2. ค่าระหว่างการคำนวณ (ก่อนรายงาน) ให้เก็บไว้จนถึงตัวก่อนสุดท้าย หรือ เก็บไว้ 5-6 หลัก หรือ เก็บไว้ให้มากที่สุด ← ควรทำ เพราะอาจเกิดความคลาดเคลื่อนจากการคำนวณหยาบ ๆ (พลาดง่าย ๆ)

## การหาค่า $\delta x$ หาอย่างไร?

1. หาจากเครื่องมือ หากอ่านเพียงครั้งเดียว เช่นไม้บรรทัดทาบหาความยาว ซึ่งไม้บรรทัดมีสเกลละเอียดสุด 1 mm แต่ตามนุษย์อ่านได้ละเอียดที่สุด (แล้วยังแม่นยำอยู่) คือ ครึ่งสเกล ในที่นี้ คือ $\frac{1}{2}$ mm หรือ 0.05 cm ดังนั้นเราสามารถรายงานค่าได้เป็น 2.65 cm แต่จะเป็น 2.653 cm ไม่ได้ (ต้องมีสเกลเวอร์เนียร์) อุปกรณ์ที่มี VERNIER SCALE เช่น VERNIER CALIPER, MICROMETER, ฯลฯ

2. อ่านด้วยวิธีเดิม ซ้ำ หลายครั้ง และหากอ่านค่า N ครั้ง จะได้ค่าที่ไม่เหมือนกันออกมา N ค่า : $x_1, x_2, x_3, \dots, x_N \rightarrow X$ คือค่าที่วัดได้
ส่วน $x$ คือค่าจริง (พระเจ้าเท่านั้นที่รู้)

วิธีที่สมเหตุสมผลและง่ายที่สุด คือ พิจารณา ผลต่างจากค่าจริงกำลังสอง ว่าห่างจากความเป็นจริงขนาดไหน นั่นคือค่า $(x_i - x)^2$ เพราะ $\underbrace{\sum (x_i - x)^2}_{\text{LEAST SQUARES}} > 0$ เสมอ แต่ $\sum (x_i - x)$ อาจ $= 0$

แต่ค่า $x$ จะเป็นอย่างไร ใช้แคลคูลัสในการหาด้วยหลักการ MINIMIZE นั่นคือ

$\frac{d}{dx} f(x) = 0$ จะได้จุดวิกฤตออกมา

$$\frac{d}{dx} \sum_{i=1}^{N} (x_i - x)^2 = 0$$

$$-2 \sum_{i=1}^{N} (x_i - x) = 0$$

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x = Nx$$

$$\therefore \quad x = \frac{\sum x_i}{N}$$ → นั่นคือค่าเฉลี่ยเลขคณิต ($\bar{x}$)

ซึ่ง $x$ เป็นค่า $x$ ดีสุด ( THE BEST ESTIMATE ของ $x$ ) ดังนั้น $\frac{\sum x_i}{N}$ น่าจะ

เป็น $x$ ที่พระเจ้ารู้เห็น แต่ถ้าเราวัด N เยอะ ๆ ( $\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x_i$ )

จะเป็นค่า $x$ ที่แท้จริงในอุดมคติ แปลว่า ยิ่งวัดค่าเยอะยิ่งแม่นขึ้น (ควรจะ)

ดังนั้นความห่างของค่า $x_i$ จาก $x$ กำลังสองเฉลี่ย จะเป็น $\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 = \sigma^2$ (VARIANCE)

(แต่บางครั้งมักใช้ N - 1 เพื่อความถูกต้อง (แต่ปริมาณมาก ๆ เริ่มไม่ค่อยต่าง)

สำหรับกระบวนวิธีทางสถิติ : NORMAL DIST., VARIANCE, $\sigma_{\bar{x}}$ ได้กล่าวไปข้างต้นแล้ว

ตัวอย่าง วัดความยาว $l$ ของแท่งไม้ 5 ครั้ง ได้ผลเป็น

24.25, 24.26, 24.22, 24.28, 24.24 cm

จงรายงานผลความยาวแท่งไม้นี้ (ตอบ 24.25 ± 0.01 cm)

หรือ 24.247 ± 0.009 cm

ตามเห็นสมควร

| $i$ | $x_i$ | $(x_i - \bar{x})^2$ |
|---|---|---|
| | | คิดในเครื่องคำนวณ |

## การรวมความคลาดเคลื่อน ( *Propagation of uncertainties* )

สมมติวัดค่า X ได้ $x=\bar{x}\pm\delta x$ และ Y ได้ $y=\bar{y}\pm\delta y$

1. ผลบวก : $x+y=(\bar{x}+\bar{y})\pm(\delta x+\delta y)$

$$x+2y = x+y+y = (\bar{x}+2\bar{y})\pm(\delta x+2\delta y)$$

2. ผลต่าง : $x-y=(\bar{x}-\bar{y})\pm(\delta x+\delta y)$

บวก

3. ผลคูณและผลหาร

$$x=\bar{x}\pm\delta x=\bar{x}\left(1\pm\frac{\delta x}{\bar{x}}\right)$$

$$y=\bar{y}\pm\delta y=\bar{y}\left(1\pm\frac{\delta y}{\bar{y}}\right)$$

เล็กมาก ( $\frac{\delta x\delta y}{xy}\ll\frac{\delta x}{x}$ , $\ll\frac{\delta y}{y}$ )

$$xy=\bar{x}\cdot\bar{y}\left(1\pm\left(\frac{\delta x}{\bar{x}}+\frac{\delta y}{\bar{y}}\right)\pm\frac{\delta x\delta y}{xy}\right)$$

$$\therefore xy=\bar{x}\cdot\bar{y}\pm\left(\frac{\delta x}{x}+\frac{\delta y}{y}\right)\cdot\bar{x}\bar{y}$$

ในการคูณและการหาร ใช้ $\frac{\delta A}{A}$ มาบวกกัน ซึ่งสิ่งนี้คือ RELATIVE UNCERTAINTY

4. การยกกำลัง : ใช้นิยามของผลคูณ

อ่านว่า sigma ถ้า $\Sigma$ คือ Sigma

5. กรณีทั่วไป เช่น $\varsigma=f(x)$ ใช้การประมาณทางแคลคูลัสว่า $\frac{\delta\varsigma}{\delta x}=\frac{d}{dx}\varsigma$

ดังนั้น $\delta\varsigma=\delta x\frac{d}{dx}f(x)$

จะได้ว่า $\varsigma=f(\bar{x})\pm\delta x f'(x)\big|_{x=\bar{x}}$

6. กรณีทั่วไปจัด ๆ $\varsigma=f(x,y)\rightarrow\delta\varsigma=\frac{\partial f}{\partial x}\delta x+\frac{\partial f}{\partial y}\delta y$

$$\rightarrow\underbrace{(\delta\varsigma)^2=\left(\frac{\partial f}{\partial x}\right)^2(\delta x)^2+\left(\frac{\partial f}{\partial y}\right)^2(\delta y)^2}+\underbrace{2\left(\frac{\partial f}{\partial x}\right)\left(\frac{\partial f}{\partial y}\right)\delta x\delta y}_{=0\ \because\ \sum_x\sum_y(\dots)=0}$$

ค่าบวกเท่าค่าลบ

หากเป็นการกระจายสมมาตร เขียนแทน $\delta A$ เป็น ${}_A\sigma_{mean}$ ได้ว่า

$$ {}_{e_1}\sigma_{mean} = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 {}_x\sigma^2_{mean} + \left(\frac{\partial f}{\partial y}\right)^2 {}_y\sigma^2_{mean}} $$

$$ \therefore \quad e_1 = f(\bar{x},\bar{y}) \pm {}_{e_1}\sigma_{mean} $$

7. ผลรวมทางสถิติ

จากสมบัติผลรวมของ $Var(X_i)$ ได้ว่า $\sigma^2 = \sigma_1^2 + \sigma_2^2 + \dots + \sigma_N^2$

ใช้แทนได้ทั้งหมดว่า $\sigma = \sqrt{\Sigma \sigma^2_{mean}}$

เช่น 1. $x+y = (\bar{x}+\bar{y}) \pm \sqrt{(\delta x)^2 + (\delta y)^2}$ ซึ่งดีกว่า

3. $xy = \bar{x}\cdot\bar{y} \pm \bar{x}\cdot\bar{y}\sqrt{\left(\frac{\delta x}{\bar{x}}\right)^2 + \left(\frac{\delta y}{\bar{y}}\right)^2}$

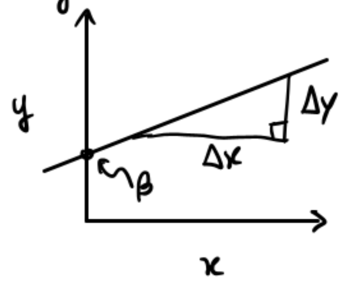
<u>การเขียนกราฟจากข้อมูลการทดลอง</u>

กราฟ มีหน้าที่ทำให้เราเห็นถึงความสัมพันธ์และพัฒนาการของข้อมูล และเราชอบ-นิยมที่จะนำข้อมูลมาเขียนในรูปของเส้นตรง เนื่องจากมีความง่าย สื่อความหมาย และมองรู้ว่าอะไรคืออะไร

เราพิจารณากราฟความสัมพันธ์ของ 2 ปริมาณ $y$ กับ $x$ ในความสัมพันธ์ $y(x)$

เช่น $y=ax+b$, $xy=c$, $y=ax^n$, $y=ae^{\lambda x}$, $y=\tan(x^a)$, ฯลฯ

1. $y=ax+b$

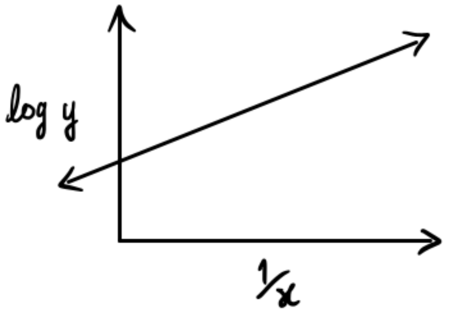



ความชัน $= \frac{\Delta y}{\Delta x}$

y-intercept $= \beta = b$

2. รูปอื่น ๆ เช่น $\log y = a\cdot\frac{1}{x} + b$

ให้ใช้แกนตั้งเป็น $\log y$ และแกนนอนเป็น $\frac{1}{x}$

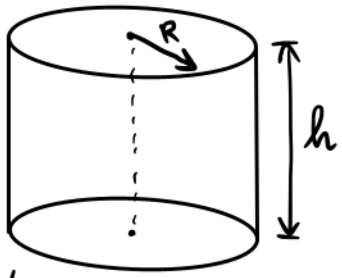



จะมีรูปเป็นเส้นตรง

โดยทั่วไปข้อมูลที่เราทำการวัดไม่ใช่เส้นต่อเนื่อง (continuous) แต่เป็นจุดข้อมูลซึ่งไม่ต่อเนื่อง (discrete) ดังนั้นเราควรวางจุดที่วัดได้ให้อยู่ในรูปของเส้นตรงให้ได้มากที่สุด ไม่ใช่โค้ง ๆ หรือ หักไปมาเบี้ยว นั่นมิใช่เส้นตรง

# ข้อมูลที่มีความสัมพันธ์แบบเส้นตรงและการถดถอย (LINEAR REGRESSION)

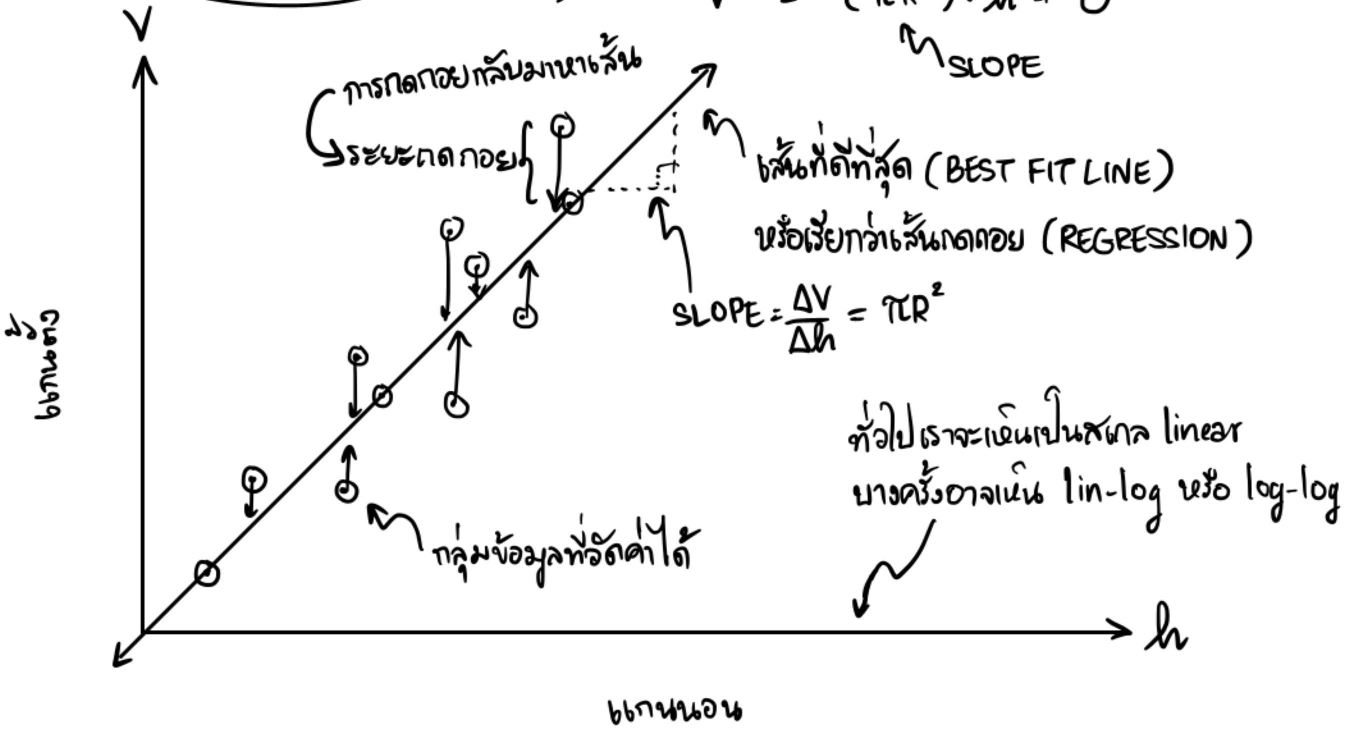
สมมติทรงกระบอก ให้หาความสัมพันธ์ของปริมาตร V เทียบกับ $h$ ว่ามีความสัมพันธ์กันอย่างไร เริ่มจากใช้ความสัมพันธ์ $V = \pi R^2 h$




เราจะเห็นได้ว่า $V \propto h$ แบบสมการเส้นตรง $y = ax + b$

โดย $a$ คือความชัน $= \pi R^2$ และ $b$ คือจุดตัดแกน $Y = 0$

→ $V = \underbrace{(\pi R^2)}_{\text{SLOPE}} \cdot h + 0$




การลากเส้น BEST FIT ด้วยมือนั้น อาศัยความชำนาญ และความสามารถในการประมาณค่าที่ดีที่สุดออกมาได้ ซึ่งเป็นพื้นฐานที่ดีของนักวิทยาศาสตร์, วิศวกร, นักวิเคราะห์, ฯลฯ โดยเราต้องการลากให้ผลรวมระยะถดถอยมีค่าน้อยที่สุด (MINIMIZATION) หรือใกล้เส้นตรงมากที่สุด แต่ถึงเวลาจริง เราก็ใช้เครื่องคำนวณหาเส้นนั้นออกมา เพราะมันแม่นกว่ามือเรามาก ๆ แต่ใครคิดสูตรพวกนั้นให้เครื่องจักร?

| V | h |
|---|---|
| ... | ... |

ตารางพล็อต

* เพิ่มเติม หากข้อมูลมีความคลาดเคลื่อนมาก จะแสดงเป็น ERROR BARS

สมมติตัวแปรเก็บค่าได้เป็นคู่อันดับ $(x_i, y_i)$ ซึ่งอาจสร้างสมการเส้นตรงได้เป็น $\boxed{\bar{Y} = aX_i + b}$

จากทฤษฎี LEAST SQUARES ซึ่งนำมาหา BEST ESTIMATE เราต้องการให้ค่า $S = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$ น้อยที่สุด,

$$S = \sum_{i=1}^{N} (Y_i - aX_i - b)^2$$

หา $b$ :

$$\frac{\partial S}{\partial b} = 0$$

$$0 = -2\sum_{i=1}^{N} (Y_i - aX_i - b)$$

$$\sum_{i=1}^{N} b = \sum_{i=1}^{N} Y_i - a\sum_{i=1}^{N} X_i$$

$$b = \frac{1}{N}\left(\sum Y_i - a\sum X_i\right)$$

$$\therefore\ b = \bar{Y} - a\bar{X}$$

หา $a$ :

$$\frac{\partial S}{\partial a} = 0$$

$$0 = -2\sum_{i=1}^{N} X_i (Y_i - aX_i - b)$$

$$0 = \sum_{i=1}^{N} (X_i Y_i - aX_i^2 - bX_i)$$

$$0 = \sum_{i=1}^{N} (X_i Y_i - (\bar{Y} - a\bar{X})X_i - aX_i^2)$$

$$a\sum_{i=1}^{N} (X_i^2 - X_i\bar{X}) = \sum_{i=1}^{N} (X_i Y_i - X_i\bar{Y})$$

$$\therefore\ a = \frac{\sum X_i (Y_i - \bar{Y})}{\sum X_i (X_i - \bar{X})}$$