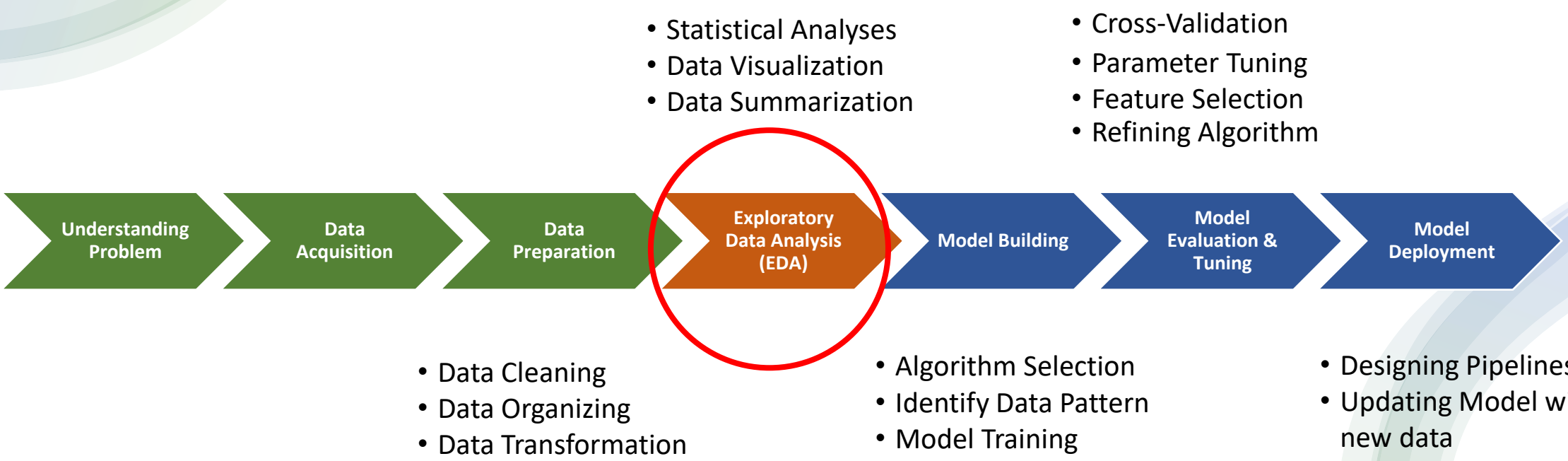
การวิเคราะห์ข้อมูลเบื้องต้น

ดร. ธรรมกร แซ่ตั้ง

Core Processes in Data Science



การวิเคราะห์ข้อมูลเบื้องต้น

ในการที่จะบ่งบอกถึงลักษณะของข้อมูลได้นั้นต้องใช้การวัด (measure)

การวัดเบื้องต้นที่นิยมใช้มี 3 แบบคือ

I. **การวัดแนวโน้มเข้าสู่ส่วนกลาง (measures of central tendency)**
II. **การวัดตำแหน่ง (measure of location)**
III. **การวัดการกระจาย (measure of dispersion)**

I. การวัดแนวโน้มเข้าสู่ส่วนกลาง

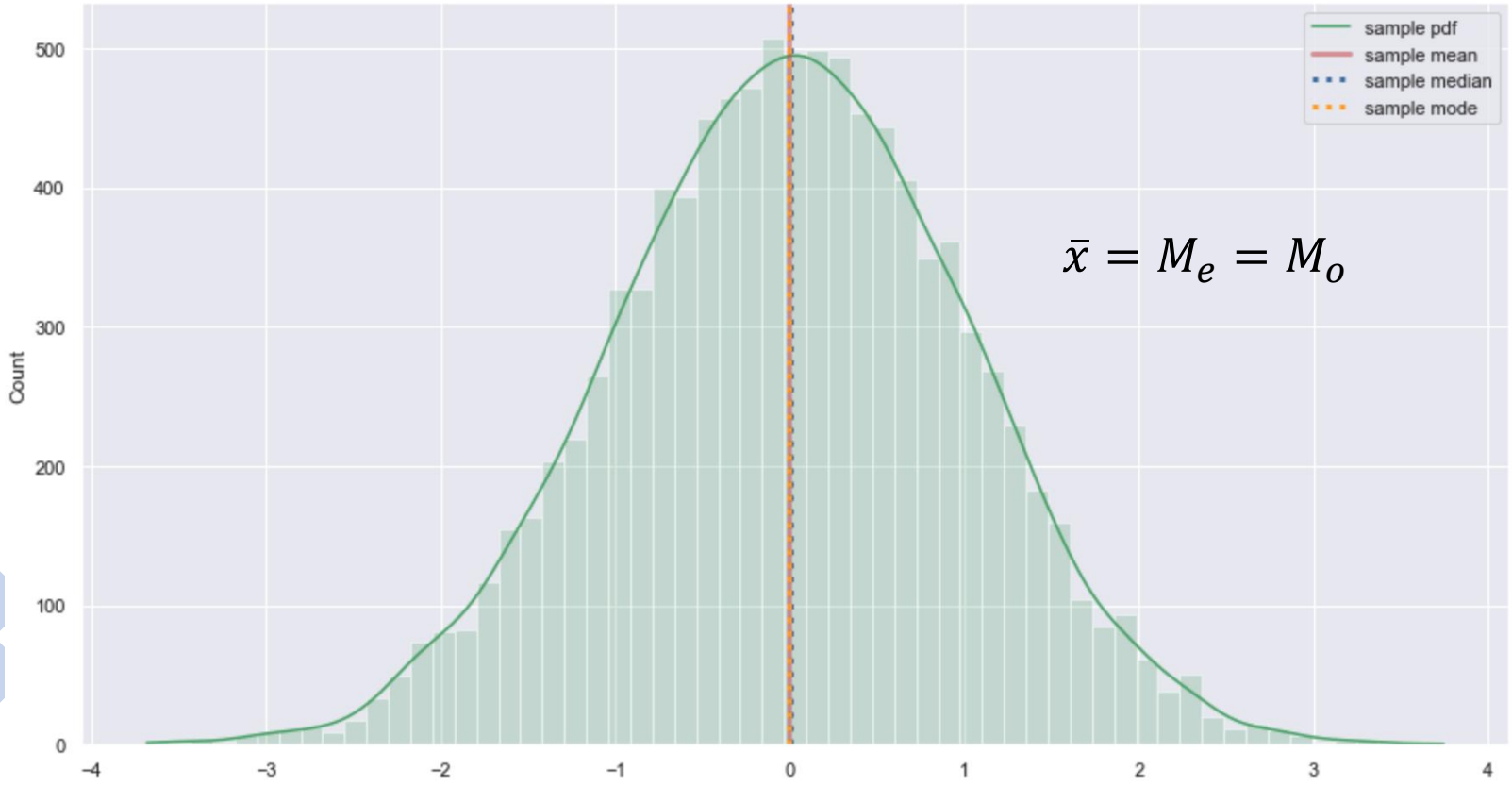
คือการคำนวณค่ากลางของข้อมูลหรือจุดกึ่งกลางของข้อมูล เพื่อมองภาพรวมของข้อมูลว่ามีลักษณะเป็นอย่างไร

ค่าที่ใช้วัดแนวโน้มเข้าสู่ส่วนกลาง ได้แก่

1. **ค่าเฉลี่ยเลขคณิต (arithmetic mean)**
2. **มัธยฐาน (median)**
3. **ฐานนิยม (mode)**

ความสัมพันธ์ระหว่างค่าเฉลี่ยเลขคณิต มัธยฐาน และฐานนิยม

1. ข้อมูลมีการแจกแจงแบบโค้งปกติ (normal distribution)
   ค่าเฉลี่ยเลขคณิต มัธยฐาน และฐานนิยม จะมีค่าเท่ากันหรือใกล้เคียงกันมาก

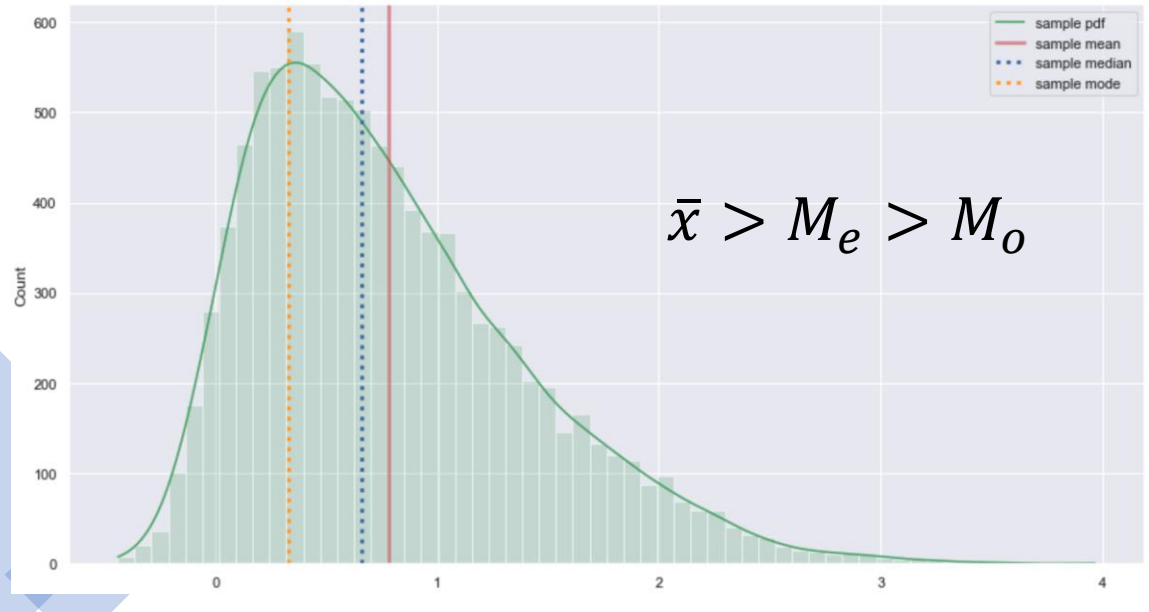


see Viz_mean_median_mode.ipynb

ความสัมพันธ์ระหว่างค่าเฉลี่ยเลขคณิต มัธยฐาน และฐานนิยม

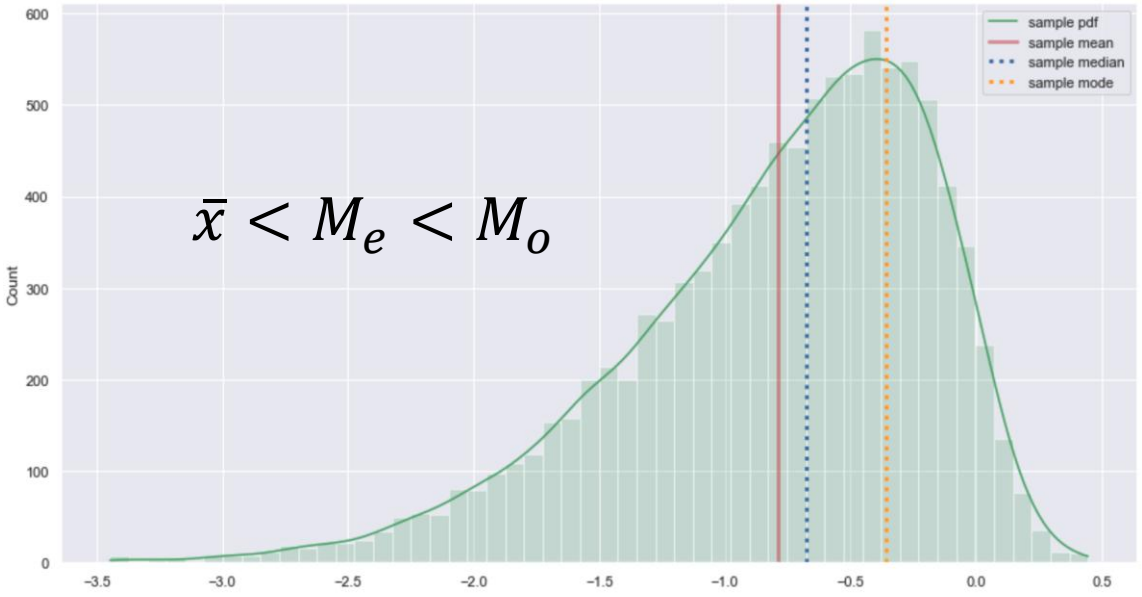
2. ข้อมูลมีการแจกแจงแบบไม่ใช่โค้งปกติมีความเบ้ของข้อมูล (skewed) เช่นมีการเบ้ขวาหรือเบ้ซ้าย

แจกแจงเบ้ขวา



$$\bar{x} > M_e > M_o$$

แจกแจงเบ้ซ้าย



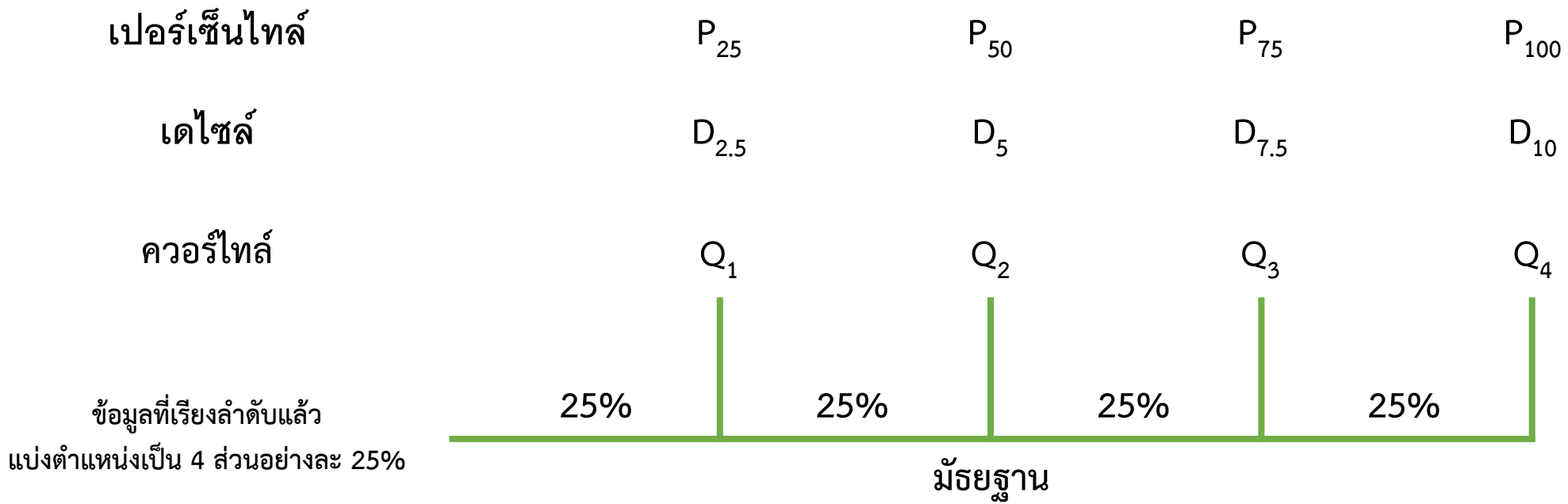
$$\bar{x} < M_e < M_o$$

II. การวัดตำแหน่งของข้อมูล (Measure of Location)

นอกจากมัธยฐานซึ่งเป็นค่าที่อยู่ ณ ตำแหน่งตรงกลางข้อมูล ค่าสถิติอื่นที่ใช้วัดตำแหน่งของข้อมูล ได้แก่ ควอร์ไทล์ (Quartiles, Q_r), เดไซล์ (Deciles, D_r) และเปอร์เซ็นไทล์ (Percentiles, P_r)

ค่าสถิติ	จำนวนการแบ่งส่วนข้อมูล	สัญลักษณ์	ช่วงของค่า r
ควอร์ไทล์	4	Q_r	1-4
เดไซล์	10	D_r	1-10
เปอร์เซ็นไทล์	100	P_r	1-100

เปรียบเทียบตำแหน่ง ควอร์ไทล์, เดไซล์ และเปอร์เซ็นไทล์



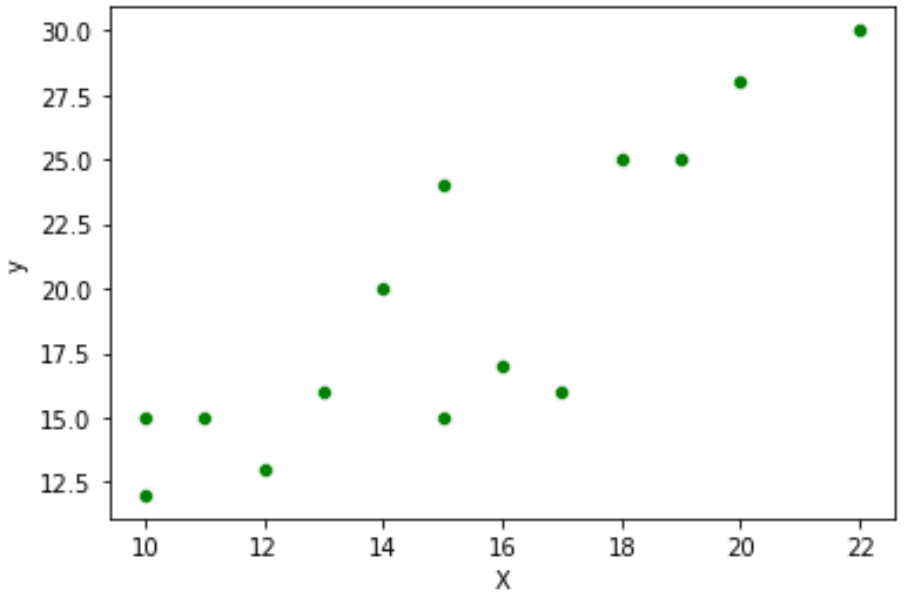
III. การวัดการกระจาย (Measure of Dispersion)

หากต้องการพิจารณาภาพรวมของข้อมูลว่ามีความแตกต่างมากน้อยแค่ไหน ค่าสถิติที่นิยมใช้วัดคือ

1. พิสัย (range)
2. ส่วนเบี่ยงเบนควอร์ไทล์ (quartiles deviation)
3. ส่วนเบี่ยงเบนมาตรฐาน (standard deviation)
4. ความแปรปรวน (variance)
5. สัมประสิทธิ์ของการแปรผัน (coefficient of variation)

ความสัมพันธ์ของตัวแปรเชิงปริมาณ

- **ตัวแปรอิสระ (Independent variable) ใช้สัญลักษณ์ X**
  - สามารถทำการวิเคราะห์โดยใช้ตัวแปรอิสระเพียงตัวเดียวหรือมากกว่า 1 ตัว
  - มีอีกชื่อว่า ตัวแปรต้น, ตัวแปรเหตุ, feature data
- **ตัวแปรตาม (Dependent variable) ใช้สัญลักษณ์ y**
  - ในการวิเคราะห์มักสนใจตัวแปรตามเพียงแค่ 1 ตัว (จึงใช้ y ตัวเล็ก)
  - มีอีกชื่อว่าตัวแปรผล, ตัวแปรตาม



การวิเคราะห์สหสัมพันธ์และการถดถอยอย่างง่าย

- ศึกษาตัวแปร 2 ตัว (X 1 ตัว, y 1 ตัว) ว่ามีรูปแบบความสัมพันธ์อย่างไร ทิศทางใด และมี ขนาดมากน้อยเพียงใด
- ศึกษาอิทธิพลของปัจจัยต่าง ๆ (X ทีละตัว) ต่อผลที่เกิดขึ้น (y)
- สามารถทำนายว่าปริมาณของตัวแปรตาม (y) มีปริมาณเท่าใด ถ้าทราบค่าของปริมาณของตัวแปรอิสระ (X) โดยพยายามให้ค่าที่ประมาณหรือค่าที่พยากรณ์ได้มีความคาดเคลื่อนน้อย หรือมีค่าใกล้เคียงกับความเป็นจริงมากที่สุด

สัมประสิทธิ์สหสัมพันธ์อย่างง่าย (Correlation Coefficient)

- เป็นการวัดว่าความสัมพันธ์ของ x และ y มีขนาดและทิศทางอย่างไร
- สำหรับ correlation coefficient ของประชากรจะใช้สัญลักษณ์ ρ เมื่อ $-1 \leq \rho \leq 1$
- สำหรับ correlation coefficient ของตัวอย่างจะใช้สัญลักษณ์ r เมื่อ $-1 \leq r \leq 1$ (โดยส่วนมากแล้วค่า correlation coefficient คำนวณจากตัวอย่าง)

การคำนวณ:

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

ใช้ corr() จาก pandas

หาก df คือ pandas dataframe จะสามารถหาสัมประสิทธิ์สหสัมพันธ์ในแนวคอลัมน์โดยใช้ .corr()

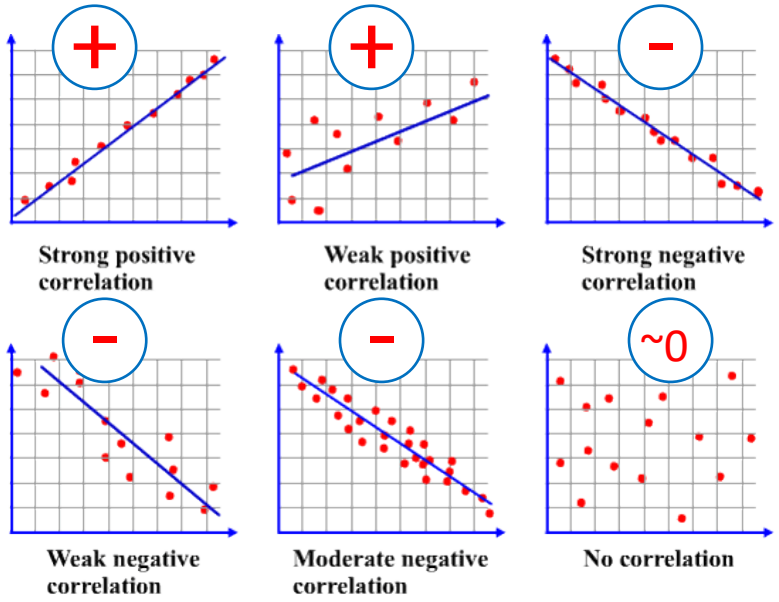
```
df.corr()
```

ใช้ pearsonr() หรือ spearmanr() จาก scipy

X และ y เป็นได้ทั้ง list หรือ np.array ผลลัพธ์จะให้ค่าสัมประสิทธิ์สหสัมพันธ์พร้อมกับค่า p-value

```
r, p_val = stats.pearsonr(x, y)
```

ค่า r ที่ได้



**pearsonr() ใช้ได้กับข้อมูลที่มีการแจกแจงแบบปกติเท่านั้น หากข้อมูลไม่ใช่การแจกแจงแบบปกติจะใช้ spearmanr()

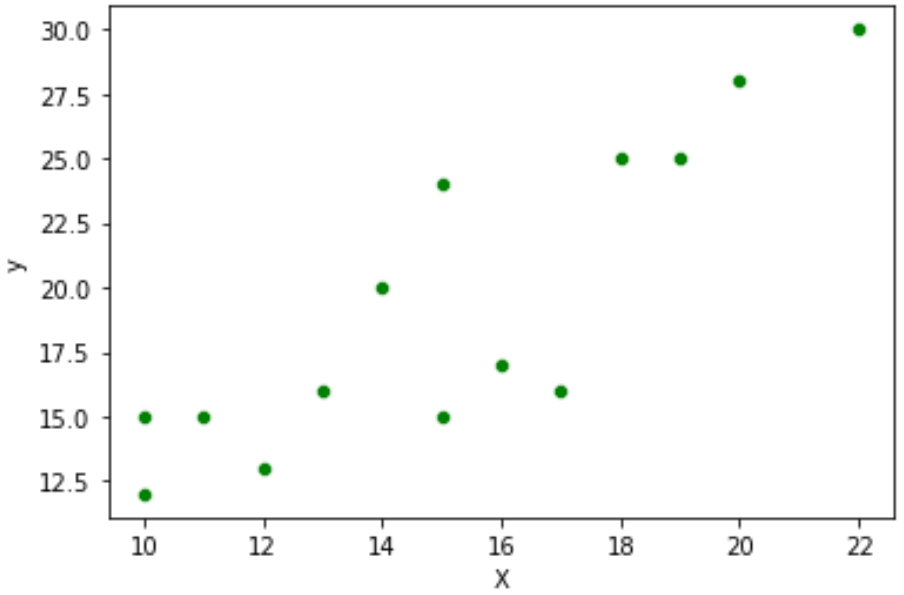
การวิเคราะห์การถดถอย (Regression Analysis)

เป็นการศึกษาและวิเคราะห์ตัวแบบความสัมพันธ์ของตัวแปรเชิงปริมาณตั้งแต่สองตัวขึ้นไป ซึ่งประกอบด้วย

- **ตัวแปรอิสระ (Independent variable) ใช้สัญลักษณ์ X**
  - สามารถทำการวิเคราะห์โดยใช้ตัวแปรอิสระเพียงตัวเดียว: การวิเคราะห์ถดถอยและสหสัมพันธ์อย่างง่าย (simple / univariate regression analysis) หรือมากกว่า 1 ตัว: การวิเคราะห์ถดถอยแบบพหุคูณ (multiple / multivariate regression analysis)
- **ตัวแปรตาม (Dependent variable) ใช้สัญลักษณ์ y**
  - ในการวิเคราะห์มักสนใจตัวแปรตามเพียงแค่ 1 ตัว (จึงใช้ y ตัวเล็ก)

สิ่งที่ศึกษา:

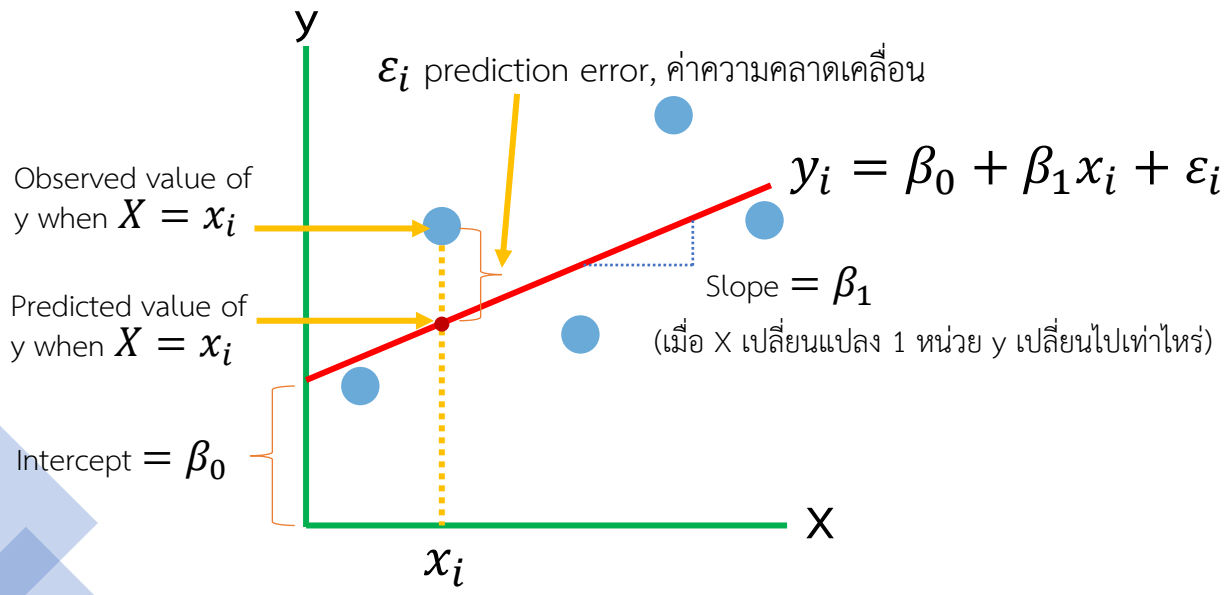
- X และ y มีความสัมพันธ์ในรูปแบบใด
- X และ y สัมพันธ์มากน้อยเพียงใด
- หาสมการเพื่อใช้สำหรับพยากรณ์หรือประมาณค่า y เมื่อทราบค่า x



การวิเคราะห์การถดถอยอย่างง่าย (Simple Regression Analysis)

เมื่อตัวแปรอิสระ (X) และตัวแปรตาม (y) มีความสัมพันธ์ลักษณะเชิงเส้นตรงแล้ว สามารถสร้างสมการถดถอยเพื่อใช้พยากรณ์ค่า y โดยใช้ค่า X

ตัวแบบความสัมพันธ์เชิงเส้นตรงสำหรับข้อมูลประชากร: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$



** β_0 และ β_1 เป็นพารามิเตอร์ของการถดถอย

หาก

$\beta_1 = 0$ แสดงว่า X และ y ไม่มีความสัมพันธ์เชิงเส้นตรงเลย

$\beta_1 > 0$ แสดงว่า X และ y มีความสัมพันธ์เชิงเส้นตรงเชิง +

$\beta_1 < 0$ แสดงว่า X และ y มีความสัมพันธ์เชิงเส้นตรงเชิง -

สมการถดถอยของตัวอย่าง

ในทางปฏิบัติไม่สามารถหาค่าพารามิเตอร์ β_0 และ β_1 ได้ จึงต้องสุ่มข้อมูลตัวอย่างเพื่อประมาณค่าพารามิเตอร์ด้วยค่าสถิติ b_0 และ b_1 ดังสมการ:

$$\hat{y} = b_0 + b_1 x_i$$

การประมาณค่าพารามิเตอร์ในที่นี้ใช้วิธีกำลังสองน้อยที่สุด (Method of Least Square)

$$b_0 = \bar{y} - b_1 \bar{x}$$ หรือ $$b_0 = \frac{\sum y - b_1 \sum x}{n}$$

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$ หรือ $$b_1 = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$ หรือ $$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

การทดสอบสมมติฐานค่า β_1 โดยการวิเคราะห์ความแปรปรวน

$H_0: \beta_1 = 0$ (ตัวแปร X และ y ไม่มีความสัมพันธ์เชิงเส้นตรง)
$H_1: \beta_1 \neq 0$ (ตัวแปร X และ y มีความสัมพันธ์เชิงเส้นตรง)

ค่าสถิติทดสอบ: $$F = \frac{MSR}{MSE}$$

โดย $MSR = \frac{SSR}{1}$, $SSR = b_0 \sum y + b_1 \sum xy - n\bar{y}^2$ (SSR: ความแปรผันจากค่าถดถอย)

$MSE = \frac{SSE}{n-2}$, $SSE = SST - SSR$, $SST = \sum y^2 - n\bar{y}^2$ (SSE: ความคลาดเคลื่อน, SST: ความแปรผันรวม)

ปฏิเสธ H_0 เมื่อ

$F > F_{\alpha,\ \nu_1=1,\ \nu_2=n-2}$ (เป็น right-tailed test)

สัมประสิทธิ์การตัดสินใจ (Coefficient of Determination)

สัมประสิทธิ์การตัดสินใจคือค่าที่ใช้อธิบายความผันแปรของ y ที่เกิดขึ้นว่าเป็นผลมาจากความผันแปรของตัวแปร X มากน้อยเพียงใด ใช้สัญลักษณ์แทนคือ r^2

$$r^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

r^2 มีค่าในช่วง 0 ถึง 1

- r^2 เข้าใกล้ 1 หมายความว่า ความผันแปรของตัวแปรตาม (y) ได้รับอิทธิพลมาจากความผันแปรของตัวแปรอิสระ (X) เท่ากับ 100%
- r^2 เข้าใกล้ 0 หมายความว่า ความผันแปรของ y ไม่ได้เกิดจากความผันแปรของ X เลย