รายวิชา 568 351 สถิติและการประยุกต์ทางเภสัชศาสตร์

สหสัมพันธ์เชิงเส้น (LINEAR CORRELATION)

รศ.ดร.ลาวัลย์ ศรัทธาพุทธ

ภาควิชาสารสนเทศศาสตร์ทางสุขภาพ คณะเภสัชศาสตร์ มหาวิทยาลัยศิลปากร

References

- **รศ.ศศิธร สุวิรัชวิทยกิจ สถิติสำหรับวิทยาศาสตร์และนักวิทยาศาสตร์ประยุกต์ เล่ม 1, เล่ม 2 มหาวิทยาลัยศิลปากร**
- Elementary Statistics: A Step by Step Approach, 8th edition, Allan G. Bluman ,McGraw-Hill, 2009.
- ผศ. ดร.ลาวัลย์ ศรัทธาพุทธ คู่มือการใช้ซอฟต์แวร์เสรีทางสถิติ PSPP สำหรับผู้เริ่มต้น, โรงพิมพ์มหาวิทยาลัยศิลปากร, 2012.

ตัวแปร (Variable)

- **ตัวแปร** คือ คุณลักษณะ (characteristics) ใดๆ ของสิ่งที่สนใจศึกษา
- สิ่งที่สนใจศึกษา = หน่วยตัวอย่าง
- เมื่อทำการวัดข้อมูลหน่วยตัวอย่างแต่ละหน่วยที่สนใจศึกษาซึ่งมีคุณลักษณะต่างๆกันไป ข้อมูลที่วัดได้จากหน่วยตัวอย่างเรียกว่า "**ค่าตัวแปร**"
- ตัวแปรถูกแบ่งตามลักษณะของค่าของตัวแปรเป็น 2 ประเภท คือ
  - ตัวแปรเชิงคุณลักษณะ (Qualitative Variable) คือ ตัวแปรที่บอกคุณลักษณะของการแบ่งกลุ่ม หรือของการจัดอันดับ - มีมาตรวัดเป็นชนิดนามบัญญัติ (nominal) หรือเป็นชนิดอันดับ (ordinal)
  - ตัวแปรเชิงปริมาณ (Quantitative Variable) คือ ตัวแปรที่มีค่าเป็นปริมาณซึ่งสามารถวัดได้เป็นตัวเลข (ที่สามารถนำไปคำนวณได้) – มีมาตรวัดเป็นชนิดอันตรภาค (interval) หรือเป็นชนิดอัตราส่วน (ratio)

ความสัมพันธ์เชิงเส้น(ตรง)

- **ความสัมพันธ์**เชิงเส้น (Linear **Correlation**) หมายถึง ความสัมพันธ์ของตัวแปรเป็นแบบเชิงเส้นตรง
- **สัมประสิทธิ์**สหสัมพันธ์เชิงเส้น (Linear Correlation **Coefficient**) หมายถึง ค่าที่ใช้วัดความสัมพันธ์เชิงเส้น

- หมายเหตุ: ความสัมพันธ์ของตัวแปรยังมีแบบอื่นๆอีก เช่น พาราโบลา เอ็กส์โปเนนท์เชียล

สัมประสิทธิ์สหสัมพันธ์เชิงเส้น

- สัมประสิทธิ์สหสัมพันธ์เชิงเส้น (Linear Correlation Coefficient) เป็นค่าที่ใช้วัด**องศาแห่งความสัมพันธ์**เชิงเส้น (degree of linear relationship) **ระหว่างตัวแปร 2 ตัว**
- ใช้สัญลักษณ์ สำหรับประชากร ρ_{xy}
- ใช้สัญลักษณ์ สำหรับตัวอย่างสุ่ม r_{xy}
- สูตรที่ใช้หาสัมประสิทธิ์สหสัมพันธ์เชิงเส้นที่นิยมมากที่สุด คือ สูตรของ Pearson ที่เรียกว่า สหสัมพันธ์ของเพียร์สัน (Pearson's correlation, Pearson's product-moment coefficient)

แนวคิดของ Pearson's Correlation

- **Pearson** (1930): แนวคิดของ Pearson's Correlation คือถ้าตัวแปร X มีความสัมพันธ์กับตัวแปร Y แล้วค่าของ X จะต้องเปลี่ยนไปตามค่า Y ซึ่งอาจเป็นไปในทิศทางเดียวกันหรือตรงข้ามกันก็ได้ **โดยใช้ค่าเฉลี่ยเป็นจุดเปรียบเทียบการเปลี่ยนไปของค่าตัวแปร**
- ความสัมพันธ์เชิงเส้นระหว่างการเปลี่ยนแปลงของค่า X และ Y แทนด้วย $(X_i - \bar{X})(Y_i - \bar{Y})$
- ทิศทางของความสัมพันธ์เชิงเส้นแสดงด้วยเครื่องหมายของ $(X_i - \bar{X})(Y_i - \bar{Y})$

สหสัมพันธ์ของเพียร์สัน

- สหสัมพันธ์ของเพียร์สัน (Pearson's correlation coefficient, Pearson's product-moment correlation coefficient) คือการนำ**ค่าความแปรปรวนร่วม**ระหว่างตัวแปรสุ่มทั้งสองมาหารด้วย**ค่าส่วนเบี่ยงเบนมาตรฐาน**ของตัวแปรทั้งสอง

นิยามของ Pearson's Correlation

- **Pearson's correlation coefficient (Pearson's product-moment correlation coefficient)** is the covariance of the two variables divided by the product of their standard deviations. (....Wikipedia)

$$corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

ข้อตกลงการใช้สูตร Pearson's Correlation

1. ตัวแปรหรือข้อมูลทั้ง 2 ชุดอยู่ในมาตรวัดชนิดอันตรภาค หรืออัตราส่วน

2. ข้อมูลทั้ง 2 ชุด มีการแจกแจงแบบปกติและมีความสัมพันธ์เชิงเส้นตรง

3. ข้อมูลในแต่ละชุดจะต้องมีความเป็นอิสระต่อกัน

การแปรปรวนร่วม (Covariation)

- **Blalock** (1972): ได้ให้นิยามของ**องศาแห่งความสัมพันธ์เชิงเส้น** (degree of linear relation) ของข้อมูลจาก ต.ย.ขนาด n ด้วยค่าของการแปรปรวนร่วม (Covariation) ดังนี้

เอสใหญ่

$$Covariation = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = S_{XY}$$

- จากหนังสือ Social statistics, 2nd ed. ของ Blalock (1972)

$$Standard\ Deviation = s_x = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

$$Variance = s_x^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

$$Sum\ of\ Square\ of\ X = \boldsymbol{S}_X^2\ (big) = \sum(X_i - \bar{X})^2$$

$$\boldsymbol{S}_X\ (big) = \sqrt{\sum(X_i - \bar{X})^2}$$

$$Covariation = Sum\ of\ SquareXY = \boldsymbol{S}_{XY}\ (big) = \sum(X_i - \bar{X})(Y_i - \bar{Y})$$

$$\text{Covariance} = Cov(X,Y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$Correlation = r_{XY} = \frac{Cov(X,Y)}{s_x s_y}$$

ความแปรปรวนร่วม (Covariance)

- ค่าความแปรปรวนร่วม (Covariance) คือค่าเฉลี่ยของ Covariation ก็คือองศาแห่งความสัมพันธ์เชิงเส้นเฉลี่ย
- ความแปรปรวนร่วมเป็นการวัดปริมาณการเปลี่ยนแปลงของสองตัวแปรว่าจะมีการเปลี่ยนแปลงตามกันมากน้อยเท่าใด
- สูตร Covariance

$$Cov(X,Y) = \frac{Covariation}{n-1} = \frac{S_{XY}}{n-1} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

ปัญหาของความแปรปรวนร่วม

- การวัดองศาแห่งความสัมพันธ์เชิงเส้นระหว่าง X และ Y โดยใช้ค่า Cov(X,Y) มีข้อเสียคือค่าความแปรปรวนร่วมที่คำนวณได้**ขึ้นกับ "หน่วย" และ "ขนาด" ของข้อมูล**ทำให้ไม่สะดวกในการตีความ การขึ้นกับขนาดของข้อมูลอาจจะ**ทำให้เข้าใจได้ยากว่าตัวแปร X และตัวแปร Y มีองศาแห่งความสัมพันธ์เชิงเส้นในระดับไหน** เช่น
  - ถ้า X_1 และ Y_1 เป็นเลขจำนวนเต็มหลายหลัก ค่าของ $Cov(X_1,Y_1)$ ก็จะเป็นหลายหลักด้วย
  - ถ้า X_2 และ Y_2 เป็นเลขทศนิยม ค่าของ $Cov(X_2,Y_2)$ ก็จะเป็นทศนิยมด้วย
  - ทั้งๆ ที่ X_1 และ Y_1 อาจมีองศาแห่งความสัมพันธ์เชิงเส้นใกล้เคียงกับ X_2 และ Y_2 แต่ค่าของ $Cov(X_2,Y_2)$ ดูต่างจากค่าของ $Cov(X_1,Y_1)$ มาก

วิธีแก้ปัญหาความแปรปรวนร่วม

- เพื่อขจัดปัญหาของการวัดองศาแห่งความสัมพันธ์เชิงเส้นระหว่างที่ขึ้นอยู่กับหน่วยและขนาดของข้อมูล จึงต้องทำ**การแปลงข้อมูลให้เป็นข้อมูลมาตรฐาน**ซึ่งมีค่าเฉลี่ยเป็น 0 และค่าความแปรปรวนเป็น 1 ดังนี้

$$Z_{X_i} = \frac{X_i-\bar{X}}{s_x} \quad \text{โดยที่} \quad s_x = \sqrt{\frac{\sum(X_i-\bar{X})^2}{n-1}}$$

$$Z_{Y_i} = \frac{Y_i-\bar{Y}}{s_y} \quad \text{โดยที่} \quad s_y = \sqrt{\frac{\sum(Y_i-\bar{Y})^2}{n-1}}$$

- s_x และ s_y **(เอสเล็ก)** = standard deviation
- ข้อมูลมาตรฐาน Zx และ Zy ไม่ขึ้นกับหน่วยและขนาดของข้อมูลเดิม

ความแปรปรวนร่วมของข้อมูลมาตรฐาน

- การวัดองศาแห่งความสัมพันธ์เชิงเส้นตรง**ของข้อมูลมาตรฐาน** ดังนี้

Cov(X,Y)

$$Cov(Z_X,Z_Y) = \frac{\sum Z_X Z_Y}{n-1} = \frac{\sum(\frac{X_i-\bar{X}}{s_x})(\frac{Y_i-\bar{Y}}{s_y})}{n-1} = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{(n-1)s_x s_y}$$

$$Cov(Z_X,Z_Y) = \frac{Cov(X,Y)}{s_x s_y} = corr(X,Y) = corr(Z_X,Z_Y)$$

- นิยมเรียก "ความแปรปรวนร่วมของข้อมูลมาตรฐาน" ว่า "สัมประสิทธิ์สหสัมพันธ์"

สัมประสิทธิ์สหสัมพันธ์เชิงเส้น

- สัมประสิทธิ์สหสัมพันธ์เชิงเส้น (Correlation coefficient)

$$corr(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

- Population Correlation Coefficient (rho)

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

- Sample Correlation Coefficient

$$r_{XY}=\frac{Cov(X,Y)}{s_x s_y}$$

สัมประสิทธิ์สหสัมพันธ์ของตัวอย่างสุ่ม

- สัมประสิทธิ์สหสัมพันธ์ของตัวอย่างสุ่ม (Sample Correlation Coefficient) หาได้จากสูตร

$$r_{XY}=\frac{Cov(X,Y)}{s_x s_y}=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})/(n-1)}{s_x s_y}$$

เอสเล็ก

$$r_{XY}=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})/(n-1)}{\sqrt{\frac{\sum(X_i-\bar{X})^2}{n-1}}\sqrt{\frac{\sum(Y_i-\bar{Y})^2}{n-1}}}$$

- ปรับสูตรใหม่โดยตัดค่า (n-1) ของเศษและส่วนออกจะได้เป็น

เอสใหญ่

$$\boxed{r_{XY}=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2}\sqrt{\sum(Y_i-\bar{Y})^2}}=\frac{S_{XY}}{S_X S_Y}}$$

สัมประสิทธิ์สหสัมพันธ์ของตัวอย่างสุ่ม

- โดยเอสใหญ่ $S_{XY}=\sum(X_i-\bar{X})(Y_i-\bar{Y})=\sum X_iY_i-\frac{(\sum X_i)(\sum Y_i)}{n}$

$$S_X^2=\sum(X_i-\bar{X})^2=\sum X_i^2-\frac{(\sum X_i)^2}{n}$$

$$S_Y^2=\sum(Y_i-\bar{Y})^2=\sum Y_i^2-\frac{(\sum Y_i)^2}{n}$$

- ปัญหาการใช้สูตร r_{XY}
  - ถ้าใช้สูตรนี้กรณี $\bar{X}$ และ $\bar{Y}$ ที่เป็น**เลขหารลงตัว** (จำนวนเต็มหรือจุดทศนิยมรู้จบ) จะ**ไม่มีความคลาดเคลื่อน**อันเนื่องมาจากการปัดจุดทศนิยม
  - ถ้าใช้สูตรนี้กรณี $\bar{X}$ และ $\bar{Y}$ ที่เป็น**เลขหารไม่ลงตัว** (จุดทศนิยมไม่รู้จบ) จะได้ค่า r_{xy} **ที่ผิดพลาด**ได้จากปัญหาอันเนื่องมาจากการปัดจุดทศนิยม

สัมประสิทธิ์สหสัมพันธ์ของตัวอย่างสุ่ม

- สูตรที่เหมาะสมในกรณี $\bar{X}$ และ $\bar{Y}$ **หารลงตัวและหารไม่ลงตัว**

$$r_{XY}=\frac{n\sum_{i=1}^{n}X_iY_i-(\sum_{i=1}^{n}X_i)(\sum_{i=1}^{n}Y_i)}{\sqrt{n\sum_{i=1}^{n}X_i^2-(\sum_{i=1}^{n}X_i)^2}\sqrt{n\sum_{i=1}^{n}Y_i^2-(\sum_{i=1}^{n}Y_i)^2}}$$

- สูตรที่เหมาะสมในกรณี $\bar{X}$ และ $\bar{Y}$ **หารลงตัว (สูตรจะง่ายลง)**

$$r_{XY}=\frac{\sum_{i=1}^{n}X_iY_i-n\bar{X}\bar{Y}}{\sqrt{\sum_{i=1}^{n}X_i^2-n\bar{X}^2}\sqrt{\sum_{i=1}^{n}Y_i^2-n\bar{Y}^2}}$$

สัมประสิทธิ์สหสัมพันธ์ของประชากร

- สูตรที่เหมาะสมในกรณี μ_x และ μ_y **หารลงตัวและหารไม่ลงตัว**

$$\rho_{XY} = \frac{N\sum_{i=1}^{N} X_i Y_i - (\sum_{i=1}^{N} X_i)(\sum_{i=1}^{N} Y_i)}{\sqrt{N\sum_{i=1}^{N} X_i^2 - (\sum_{i=1}^{N} X_i)^2}\sqrt{N\sum_{i=1}^{N} Y_i^2 - (\sum_{i=1}^{N} Y_i)^2}}$$

- สูตรที่เหมาะสมในกรณี μ_x และ μ_y **หารลงตัว**

$$\rho_{XY} = \frac{\sum_{i=1}^{N} X_i Y_i - N\mu_x\mu_y}{\sqrt{\sum_{i=1}^{N} X_i^2 - N\mu_x^2}\sqrt{\sum_{i=1}^{N} Y_i^2 - N\mu_y^2}}$$

การสรุปผลค่า r_{xy}

- ถ้าค่า r_{xy} =+1
  - สรุปได้ว่าตัวแปร 2 ตัว (X, Y) มีความ**สัมพันธ์กัน**เชิงเส้นตรงอย่างสมบูรณ์**ทางบวก** และสามารถเขียนอยู่ในรูป $Y_i = aX_i + b$ ได้
- ถ้าค่า r_{xy} =-1
  - สรุปได้ว่าตัวแปร 2 ตัว (X, Y) มีความ**สัมพันธ์กัน**เชิงเส้นตรงอย่างสมบูรณ์**ทางลบ** และสามารถเขียนอยู่ในรูป $Y_i = aX_i + b$ ได้
- ถ้าค่า r_{xy} = 0
  - สรุปได้ว่าตัวแปร 2 ตัว (X, Y) **ไม่มีความสัมพันธ์เชิงเส้นตรงต่อกัน** (**แต่ตัวแปรทั้งสองอาจจะมีความสัมพันธ์กันในลักษณะอื่น**)

ตัวอย่าง 1

- จงหาค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูลดังนี้

X	-2	-1	0	1	2
Y	-3	-1	1	3	5

$$r_{XY} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\sum X_i^2 - n\bar{X}^2}\sqrt{\sum Y_i^2 - n\bar{Y}^2}}$$

$$r_{XY} = \frac{20}{\sqrt{10}\sqrt{40}} = 1$$

- สรุปว่าตัวแปรทั้งสองตัวมีความสัมพันธ์เชิงเส้นอย่างสมบูรณ์ทางบวก

ตัวอย่าง 2

- จงหาค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูลดังนี้

X	0	-2	-4	2	4
Y	-1	0	1	-2	-3

$$r_{XY} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\sum X_i^2 - n\bar{X}^2}\sqrt{\sum Y_i^2 - n\bar{Y}^2}}$$

$$r_{XY} = \frac{-20}{\sqrt{40}\sqrt{10}} = -1$$

- สรุปว่าตัวแปรทั้งสองตัวมีความสัมพันธ์เชิงเส้นอย่างสมบูรณ์ทางลบ

ตัวอย่าง 3

- จงหาค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูลดังนี้

X	5	6	9	9	10
Y	9	11	0	15	1

$$r_{XY} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\sum X_i^2 - n\bar{X}^2}\sqrt{\sum Y_i^2 - n\bar{Y}^2}}$$

$$r_{XY} = \frac{-24.8}{\sqrt{18.8}\sqrt{168.8}} = -0.44$$

- สรุปว่าตัวแปรทั้งสองตัวมีความสัมพันธ์เชิงเส้นไม่สมบูรณ์ทางลบ

ตัวอย่าง 4

- จงหาค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูลดังนี้

X	1	2	3	4	5	6
Y	1	4	7	7	4	1

$$r_{XY} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\sum X_i^2 - n\bar{X}^2}\sqrt{\sum Y_i^2 - n\bar{Y}^2}}$$

$$r_{XY} = \frac{0}{\sqrt{17.5}\sqrt{36}} = 0$$

- สรุปว่าตัวแปรทั้งสองตัวไม่มีความสัมพันธ์เชิงเส้น

ตัวอย่าง 5

- จงหาค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูลดังนี้

X	0	-2	-4	2
Y	-3	0	1	2

$$r_{XY} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\sum X_i^2 - n\bar{X}^2}\sqrt{\sum Y_i^2 - n\bar{Y}^2}}$$

$$r_{XY} = \frac{0}{\sqrt{20}\sqrt{14}} = 0$$

- สรุปว่าตัวแปรทั้งสองตัวไม่มีความสัมพันธ์เชิงเส้น

ข้อสังเกต (การแปลงเชิงเส้น)

- ค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูลที่ผ่านการแปลงโดยการเอาค่าคงที่มาบวกหรือลบออกจากข้อมูลดิบจะมีค่าเท่าเดิม เนื่องจากเป็นการย้ายเส้นตรงเท่านั้น:

  $r_{x'y'} = r_{xy}$
- ค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูลที่ผ่านการแปลงโดยการเอาค่าคงที่มาคูณข้อมูลดิบจะมีค่าเท่าเดิม เนื่องจากเป็นการขยายเส้นตรงเท่านั้น:

  $r_{x'y'} = r_{xy}$
- ค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูลที่ผ่านการแปลงเป็นข้อมูลมาตรฐาน จะได้ค่าเท่าเดิมและเท่ากับค่าความแปรปรวนร่วมของข้อมูลมาตรฐานนั้น: $r_{ZxZy} = r_{xy}$
- **การแปลงเชิงเส้น**นี้มีประโยชน์สำหรับปรับสเกลขนาดของข้อมูลเพื่อให้สะดวกในการคำนวณ

ตัวอย่าง 6

- ให้เอา 2 บวกทุกค่าของ X และเอา 3 บวกทุกค่าของ Y แล้วหาค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูล

X	-2	-1	0	1	2
Y	-3	-1	1	3	5

$$r_{XY} = \frac{20}{\sqrt{10}\sqrt{40}} = 1$$

X'	0	1	2	3	4
Y'	0	2	4	6	8

$$r_{X'Y'} = \frac{20}{\sqrt{10}\sqrt{40}} = 1$$

ตัวอย่าง 7

- ให้เอา 2 คูณแล้วบวก 1 ทุกค่าของ X และเอา 10 คูณแล้วลบ 4 ทุกค่าของ Y แล้วหาค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูล

X	-2	-1	0	1	2
Y	-3	-1	1	3	5

$$r_{XY} = \frac{20}{\sqrt{10}\sqrt{40}} = 1$$

X'	-3	-1	1	3	5
Y'	-34	-14	6	26	46

$$r_{X'Y'} = \frac{400}{\sqrt{40}\sqrt{4000}} = 1$$

ตัวอย่าง 8

- ให้แปลงค่าของ X และ Y เป็นค่ามาตรฐาน Z แล้วหาค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูล

X	-2	-1	0	1	2
Y	-3	-1	1	3	5

$$r_{XY} = \frac{20}{\sqrt{10}\sqrt{40}} = 1$$

Zx	-1.265	-0.632	0	0.632	1.265
Zy	-1.265	-0.632	0	0.632	1.265

$$r_{ZxZy} = \frac{4}{\sqrt{4}\sqrt{4}} = 1$$

สรุปคุณสมบัติของสัมประสิทธิ์สหสัมพันธ์จากตัวอย่างสุ่ม

- r_{xy} เป็นตัวเลขที่ไม่มีหน่วยและที่ไม่ขึ้นกับขนาดของข้อมูล
- r_{xy} อยู่ในช่วงปิด [-1 1]
- การแปลงเชิงเส้นของข้อมูลดิบไม่ทำให้ค่า r_{xy} เปลี่ยนไป

- Note: Cov(X,Y) ไม่มีขอบเขตจำกัด, บอก direction ได้ ไม่นิยมใช้บอกขนาดความสัมพันธ์ (strength) เนื่องจากขึ้นกับขนาดและหน่วยของข้อมูล

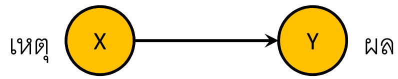
ข้อควรระวังในการตีความหมายของค่า r

- ค่าสัมประสิทธิ์สหสัมพันธ์บอกแต่เพียงองศาแห่งความสัมพันธ์เชิงเส้นเท่านั้น แต่**ไม่ได้บอกถึงความสัมพันธ์เชิงเหตุและผล**ระหว่างตัวแปรทั้งสอง
- ดังนั้นการพิจารณา **r** อย่างเดียวจะไม่สามารถสรุปความเป็นเหตุและผลของความสัมพันธ์เชิงเส้นได้
- การที่จะสรุป**ความเป็นเหตุและผล**ของความสัมพันธ์เชิงเส้นระหว่างตัวแปรทั้งสอง**จะต้องศึกษาจากทฤษฎี**ที่เกี่ยวข้องกับตัวแปรนั้นมาอธิบาย

ข้อควรระวังในการตีความหมายของค่า r

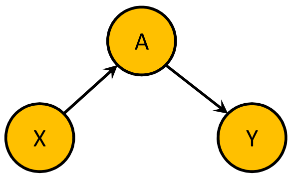
- การนำค่าสัมประสิทธิ์สหสัมพันธ์ หรือความสัมพันธ์เชิงเส้นมาเกี่ยวพันกับความเป็นเหตุและผลมีข้อควรระวังดังนี้

1. **ถ้ามีการศึกษาทฤษฎี**ที่เกี่ยวข้องกับตัวแปรอย่างละเอียด จะสามารถสรุปได้ว่า X เป็นสาเหตุของตัวแปร Y และตัวแปร Y เป็นผลจากตัวแปร X นอกจากนี้ ตัวแปร X จะมีความสัมพันธ์เชิงเส้นกับตัวแปร Y ด้วย **แต่หากไม่มีการศึกษาทฤษฎีที่เกี่ยวข้องกับตัวแปรอย่างละเอียด จะสรุปได้เพียงความสัมพันธ์เชิงเส้นของ 2 ตัวแปรเท่านั้น**



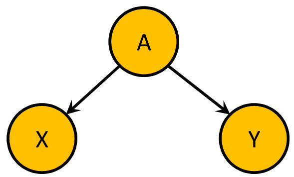
ข้อควรระวังในการตีความหมายของค่า r

2. ใน**กรณีค่า r_{xy} สูง** แต่เมื่อศึกษาทฤษฎีที่เกี่ยวข้องกับตัวแปร X และตัวแปร Y แล้วปรากฏว่า **มีตัวแปรสอดแทรก** ที่ทำให้พบความสัมพันธ์ระหว่างตัวแปร X และ Y (intervening relationship) กรณีนี้**ตัวแปร X ไม่ใช่สาเหตุโดยตรงของตัวแปร Y** แต่ส่งผ่านตัวแปรอื่นคือตัวแปรสอดแทรก **กล่าวคือถ้าไม่มีตัวแปร A แล้วตัวแปร X และ Y จะไม่มีความสัมพันธ์เชิงเส้นต่อกันเลย**



ข้อควรระวังในการตีความหมายของค่า r

3. ใน**กรณีค่า r_{xy} สูง** แต่เมื่อศึกษาทฤษฎีที่เกี่ยวข้องกับตัวแปร X และตัวแปร Y แล้วปรากฏว่า **ตัวแปร X ไม่ใช่สาเหตุของตัวแปร Y** กรณีนี้เรียกว่า **ความสัมพันธ์ลวง (spurious relationship) กล่าวคือจริงๆแล้วตัวแปรทั้งสองไม่มีความสัมพันธ์เชิงเส้นต่อกัน**



การทดสอบสมมติฐานค่าสัมประสิทธิ์สหสัมพันธ์

- ค่าสัมประสิทธิ์สหสัมพันธ์ใช้วัดองศาแห่งความสัมพันธ์เชิงเส้น การสรุปว่าตัวแปร 2 ตัวมีความสัมพันธ์เชิงเส้นกันหรือไม่ จะต้องผ่านการทดสอบสมมติฐาน

  $H_0 : \rho = 0$ (ไม่มีความสัมพันธ์เชิงเส้น)

  $H_1 : \rho \neq 0$ (มีความสัมพันธ์เชิงเส้น)

- ค่าสถิติที่ใช้ทดสอบสมมติฐานมี 3 ตัวคือ
  - ค่าสถิติ r
  - ค่าสถิติ t
  - ค่าสถิติ z

Hypothesized Population Correlation (ρ_0)

- ค่าสถิติ r และ t ใช้ทดสอบสมมติฐานกรณี $H_0 : \rho = 0$ เท่านั้นหาก กรณี $\rho \neq 0$ เช่น $H_0 : \rho = 0.5$ ต้องใช้สถิติ z

การทดสอบสมมติฐานค่าสัมประสิทธิ์สหสัมพันธ์

- ค่าสถิติ r

$$r_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}} = \frac{S_{XY}}{S_X S_Y}$$

df = n-2

- ปฏิเสธ H_0 ที่ระดับนัยสำคัญ α% ก็ต่อเมื่อ $|r| \geq r_{[\alpha, n-2]}$
  - โดยที่ α คือระดับนัยสำคัญ

    n คือขนาดตัวอย่างสุ่ม

เนื่องจากในตำราให้มาเป็นตาราง 2-tail

การทดสอบสมมติฐานค่าสัมประสิทธิ์สหสัมพันธ์

- ค่าสถิติ t

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

df = n-2

  - โดยที่ r คือค่า สปส.สหสัมพันธ์จากตัวอย่างสุ่ม

    n คือขนาดตัวอย่างสุ่ม

- ปฏิเสธ H_0 ที่ระดับนัยสำคัญ α% ก็ต่อเมื่อ $|t| \geq t_{[\alpha/2, n-2]}$

การทดสอบสมมติฐานค่าสัมประสิทธิ์สหสัมพันธ์

- ค่าสถิติ z

$$Z = \frac{Z_r - Z_\rho}{\sigma_{Z_r}}$$

  - โดยที่

$$Z_r = \frac{1}{2}[\ln(1+r) - \ln(1-r)]$$

$$Z_\rho = \frac{1}{2}[\ln(1+\rho) - \ln(1-\rho)]$$ → ค่า $Z_\rho = 0$ กรณี $\rho_0 = 0$

$$\sigma_{Z_r} = \frac{1}{\sqrt{n-3}}$$

- ปฏิเสธ H_0 ที่ระดับนัยสำคัญ α% ก็ต่อเมื่อ $|z| \geq z_{[\alpha/2]}$

ตัวอย่าง 9

- จากข้อมูลดังนี้ ได้ค่า **r =-0.44** จงทดสอบสมมติฐานว่าตัวแปร X และตัวแปร Y มีความสัมพันธ์เชิงเส้นหรือไม่ ที่ระดับนัยสำคัญ 0.05 (5%)

X	5	6	9	9	10
Y	9	11	0	15	1

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

ตัวอย่าง 9

- **กรณีใช้ค่าสถิติ r ทดสอบ**

$$r = \frac{S_{XY}}{S_X S_Y} = -0.44$$

- จากตารางสถิติ $r_{[0.05,3]}$ = 0.878 (ตาราง 2-tail) = $r_{[0.025,3]}$ (1-tail)
- พบว่า | r | < $r_{[0.05,3]}$
- สรุปว่า ไม่ปฏิเสธ H_0 ที่ α=0.05
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 5 อนุมานได้ว่า ตัวแปรทั้งสองตัวไม่มีความสัมพันธ์เชิงเส้น

การเปิดตารางหาค่า r

- $r_{0.025,3}$(1-tail) = $r_{0.05,3}$(2-tail)

Critical Values for Pearson's Correlation Coefficient

	Proportion in ONE Tail					
	.25	.10	.05	.025	.01	.005
	Proportion in TWO Tails					
DF	.50	.20	.10	.05	.02	.01
1	.7071	.9511	.9877	.9969	.9995	.9999
2	.5000	.8000	.9000	.9500	.9800	.9900
3	.4040	.6870	.8054	.8783	.9343	.9587
4	.3473	.6084	.7293	.8114	.8822	.9172
5	.3091	.5509	.6694	.7545	.8329	.8745
6	.2811	.5067	.6215	.7067	.7887	.8343
7	.2596	.4716	.5822	.6664	.7498	.7977
8	.2423	.4428	.5494	.6319	.7155	.7646
9	.2281	.4187	.5214	.6021	.6851	.7348
10	.2161	.3981	.4973	.5760	.6581	.7079

ตัวอย่าง 9

- **กรณีใช้ค่าสถิติ t ทดสอบ**

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.44\sqrt{5-2}}{\sqrt{1-(-0.44)^2}} = -0.849$$

- จากตารางสถิติ $t_{[0.025,3]}$ = -3.182
- พบว่า | t | < $t_{[0.025,3]}$
- สรุปว่า ไม่ปฏิเสธ H_0 ที่ α=0.05
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 5 อนุมานได้ว่า ตัวแปรทั้งสองตัวไม่มีความสัมพันธ์เชิงเส้น

ตัวอย่าง 9

- **กรณีใช้ค่าสถิติ z ทดสอบ:** $Z_\rho=0$, $\sigma_{z_r} = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{5-3}} = 0.7071$

$$Z_r = \frac{1}{2}[\ln(1+r) - \ln(1-r)] = \frac{1}{2}[\ln(1-0.44) - \ln(1+0.44)] = -0.472$$

$$Z = \frac{Z_r - Z_\rho}{\sigma_{z_r}} = \frac{-0.472-0}{0.7071} = -0.6675$$

- จากตารางสถิติ $z_{[0.025]}$ = -1.96
- พบว่า | z | < $z_{[0.025]}$
- สรุปว่า ไม่ปฏิเสธ H_0 ที่ α=0.05
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 5 อนุมานได้ว่า ตัวแปรทั้งสองตัวไม่มีความสัมพันธ์เชิงเส้น

ตัวอย่าง 10

- จากข้อมูลตัวอย่างสุ่มขนาด 103 มีค่า r =-0.67 จงทดสอบสมมติฐานว่า ตัวแปร X และตัวแปร Y มีความสัมพันธ์เชิงเส้นในทางลบหรือไม่ ที่ระดับนัยสำคัญ 0.01 (1%)

$$H_0 : \rho = 0$$

$$H_1 : \rho < 0$$

ตัวอย่าง 10

- **กรณีใช้ค่าสถิติ r ทดสอบ**

$$r = \frac{S_{XY}}{S_X S_Y} = -0.67$$

- จากตารางสถิติ $r_{[0.02,101]}$ = 0.230 (ตาราง 2-tail) = $r_{[0.01,101]}$ (1-tail)
- พบว่า | r | $\geq$ $r_{[0.02,101]}$
- สรุปว่า ปฏิเสธ H_0 ที่ α=0.01
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 103 อนุมานได้ว่า ตัวแปรทั้งสองตัวมีความสัมพันธ์เชิงเส้นในทางลบ

ตัวอย่าง 10

- **กรณีใช้ค่าสถิติ t ทดสอบ**

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.67\sqrt{103-2}}{\sqrt{1-(-0.67)^2}} = -9.070$$

- จากตารางสถิติ $t_{[0.01,101]}$ = -2.3638 (ตาราง 1-tail) = $t_{[0.02,101]}$ (2-tail)
- พบว่า | t | $\geq$ $t_{[0.01,101]}$
- สรุปว่า ปฏิเสธ H_0 ที่ α=0.01
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 103 อนุมานได้ว่า ตัวแปรทั้งสองตัวมีความสัมพันธ์เชิงเส้นในทางลบ

ตัวอย่าง 10

- **กรณีใช้ค่าสถิติ z ทดสอบ:** $Z_\rho=0$, $\sigma_{z_r} = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{103-3}} = 0.1$

$$Z_r = \frac{1}{2}[\ln(1+r) - \ln(1-r)] = \frac{1}{2}[\ln(1-0.67) - \ln(1+0.67)] = -0.811$$

$$Z = \frac{Z_r - Z_\rho}{\sigma_{z_r}} = \frac{-0.811-0}{0.1} = -8.11$$

- จากตารางสถิติ $z_{[0.01]} = -2.326$
- พบว่า $|z| \geq z_{[0.01]}$
- สรุปว่า ปฏิเสธ H_0 ที่ $\alpha=0.01$
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 103 อนุมานได้ว่า ตัวแปรทั้งสองตัวมีความสัมพันธ์เชิงเส้นในทางลบ

ตัวอย่าง 11

- จากข้อมูลตัวอย่างสุ่มขนาด 50 มีค่า r = 0.889 จงทดสอบสมมติฐานว่าตัวแปร X และตัวแปร Y มีความสัมพันธ์เชิงเส้นในทางบวกหรือไม่ ที่ระดับนัยสำคัญ 0.01 (1%)

$H_0 : \rho = 0$

$H_1 : \rho > 0$

ตัวอย่าง 11

- **กรณีใช้ค่าสถิติ r ทดสอบ**

$$r = \frac{S_{XY}}{S_X S_Y} = 0.889$$

- จากตารางสถิติ $r_{[0.02,48]} = 0.328$ (ตาราง 2-tail)
- พบว่า $|r| \geq r_{[0.02,101]}$
- สรุปว่า ปฏิเสธ H_0 ที่ $\alpha=0.01$
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 50 อนุมานได้ว่า ตัวแปรทั้งสองตัวมีความสัมพันธ์เชิงเส้นในทางบวก

ตัวอย่าง 11

- **กรณีใช้ค่าสถิติ t ทดสอบ**

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.889\sqrt{50-2}}{\sqrt{1-(0.889)^2}} = 13.45$$

- จากตารางสถิติ $t_{[0.01,48]} = 2.4066$ (ตาราง 1-tail)
- พบว่า $|t| \geq t_{[0.01,48]}$
- สรุปว่า ปฏิเสธ H_0 ที่ $\alpha=0.01$
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 50 อนุมานได้ว่า ตัวแปรทั้งสองตัวมีความสัมพันธ์เชิงเส้นในทางบวก

ตัวอย่าง 11

- **กรณีใช้ค่าสถิติ z ทดสอบ:** $Z_\rho=0$, $\sigma_{z_r} = \dfrac{1}{\sqrt{n-3}} = \dfrac{1}{\sqrt{50-3}} = 0.1459$

$$Z_r = \frac{1}{2}[\ln(1+r) - \ln(1-r)] = \frac{1}{2}[\ln(1+0.889) - \ln(1-0.889)] = 1.417$$

$$Z = \frac{Z_r - Z_\rho}{\sigma_{z_r}} = \frac{1.417-0}{0.1459} = 9.712$$

- จากตารางสถิติ $z_{[0.01]} = 2.326$
- พบว่า $|z| \geq z_{[0.01]}$
- สรุปว่า ปฏิเสธ H_0 ที่ $\alpha=0.01$
- แปลว่า จากข้อมูลที่ได้จาก ต.ย.ขนาด 50 อนุมานได้ว่า ตัวแปรทั้งสองตัว มีความสัมพันธ์เชิงเส้นในทางบวก

คำถาม