

Lab2: Explore Data I

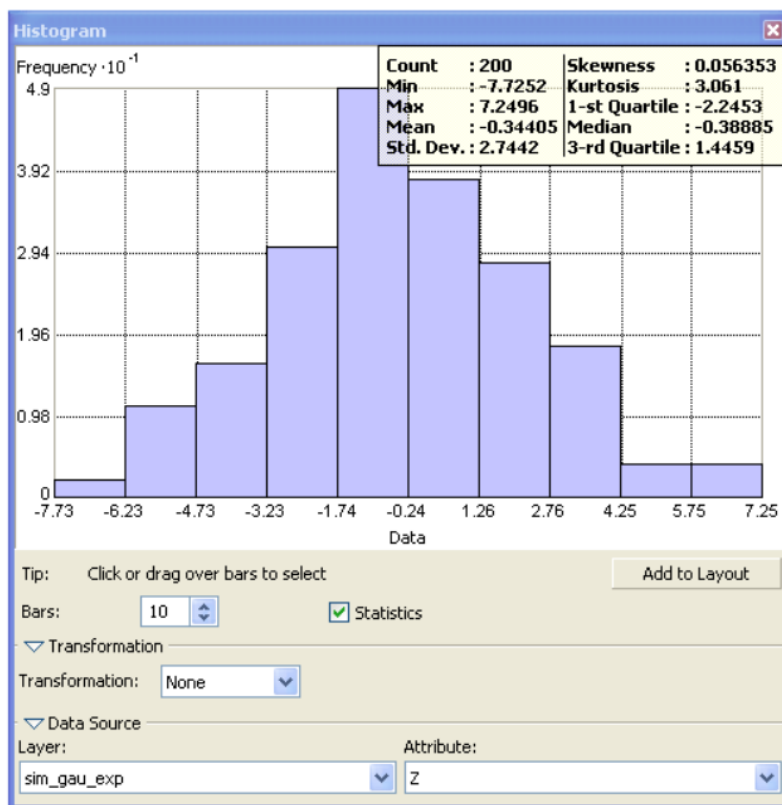
1. Histograms

Histogram เป็นเครื่องมือที่ใช้วิเคราะห์ข้อมูลแบบตัวแปรเดียว (Univariate) โดยแสดงการกระจายตัวของข้อมูลในรูปแบบกราฟแท่ง (Bar Graph) ซึ่งช่วยให้เข้าใจว่าค่าข้อมูลที่สังเกตได้ปรากฏในช่วงต่างๆ มีความถี่เป็นอย่างไร

การกระจายความถี่ (Frequency Distribution)

Histogram จะแสดงจำนวนครั้งที่ค่าของข้อมูลอยู่ในช่วงต่างๆ (Classes) ซึ่งความสูงของแท่งกราฟแทนสัดส่วนของข้อมูลในแต่ละช่วง เช่น หากแบ่งข้อมูลเป็น 10 ช่วง กราฟจะมี 10 แท่ง ความสูงของแต่ละแท่งแสดงจำนวนหรือสัดส่วนข้อมูลในช่วงนั้น

*****Class: Square Root N, Interval: (Max-Min)/Class*****



Histogram dialog box example

สถิติสรุป (Summary Statistics)

Histogram ยังช่วยสรุปคุณสมบัติที่สำคัญของการกระจายตัวของข้อมูล เช่น ตำแหน่งกระจายตัว และรูปร่างของการกระจาย ดังนี้

การวัดตำแหน่ง (Measures of Location)

มาตรวัดนี้ช่วยบอกว่าจุดศูนย์กลางหรือส่วนต่างๆ ของการกระจายข้อมูลอยู่ตรงไหน

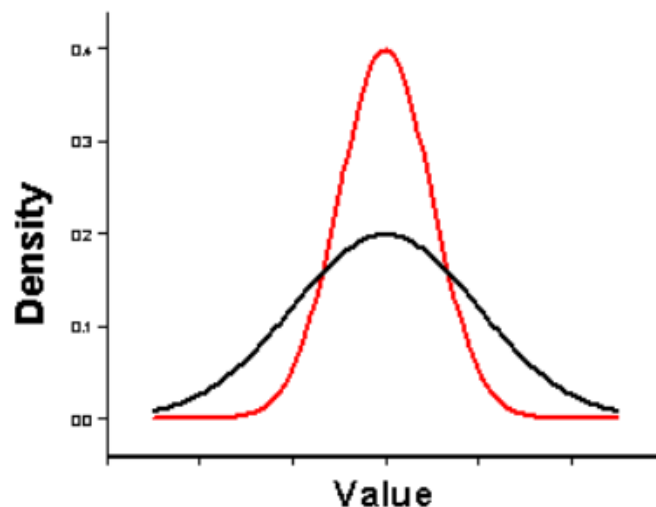
- **ค่าเฉลี่ย (Mean):** ค่าเฉลี่ยเลขคณิต เป็นค่ากลางของการกระจาย
- **ค่ามัธยฐาน (Median):** ค่ากลางของข้อมูลที่มีการเรียงลำดับแล้วแบ่งข้อมูลเป็นสองส่วนเท่าๆ กัน
- **ควอร์ไทล์ (Quartiles):** ค่าที่แบ่งข้อมูลออกเป็น 4 ส่วน ควอร์ไทล์แรก (Q1) มี 25% ของข้อมูลต่ำกว่าค่านี้ และควอร์ไทล์ที่สาม (Q3) มี 25% ของข้อมูลสูงกว่าค่านี้

การวัดการกระจายตัว (Measures of Spread)

มาตรวัดนี้ช่วยอธิบายว่าข้อมูลกระจายตัวรอบค่ากลางอย่างไร

- **ความแปรปรวน (Variance):** ค่าเฉลี่ยของความแตกต่างยกกำลังสองระหว่างข้อมูลแต่ละจุดกับค่าเฉลี่ย ใช้บอกถึงการกระจายตัว
- **ส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation):** รากที่สองของความแปรปรวน ใช้บอกถึงการกระจายตัวที่เข้าใจง่ายขึ้น

ตัวอย่าง: กราฟที่มีการกระจายตัวกว้าง (เส้นสีดำ) จะมีความแปรปรวนและส่วนเบี่ยงเบนมาตรฐานสูงกว่ากราฟที่กระจายตัวแคบ (เส้นสีแดง)

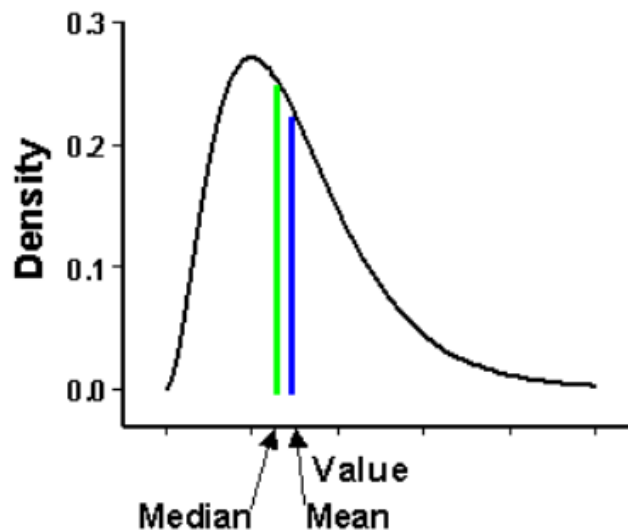


Measures of spread diagram

การวัดรูปร่าง (Measures of Shape)

กราฟ Histogram และการวิเคราะห์เหล่านี้ช่วยให้เราเข้าใจข้อมูลเชิงลึก เช่น ข้อมูลส่วนใหญ่กระจุกตัวที่ไหน กระจายอย่างไร และมีลักษณะการเบ้หรือความหนาของหางอย่างไร ซึ่งมาตรวัดนี้บอกลักษณะรูปร่างของการกระจายข้อมูล

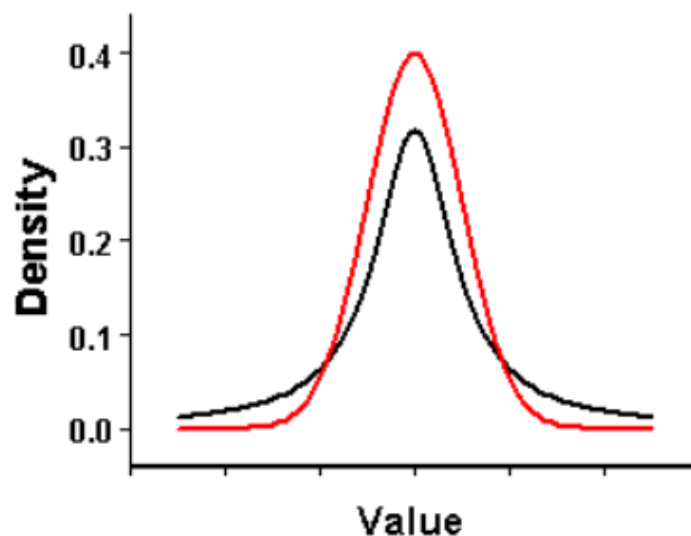
- **ความเบ้ (Skewness):** บอกความสมมาตรของข้อมูล คือ กราฟเบ้ไปทางขวา (Positive Skew) หรือเบ้ไปทางซ้าย (Negative Skew)



Positively skewed distribution example

- **ความโค้ง (Kurtosis):** เป็นค่าที่ใช้วัดลักษณะความหนาหรือบางของปลาย (หาง) ของการกระจายข้อมูล ซึ่งบ่งบอกโอกาสที่ข้อมูลจะมีค่าผิดปกติ (Outliers)
 - **Kurtosis ของการกระจายปกติ (Normal Distribution):** มีค่าเท่ากับ 3 ซึ่งหมายถึงหางของการกระจายมีลักษณะสมดุล ไม่หนาหรือบางเกินไป
 - **Leptokurtic (Thick-Tailed Distribution):** เป็นการกระจายที่มีหางหนากว่าการกระจายปกติ มีค่า Kurtosis มากกว่า 3 บ่งบอกว่าโอกาสที่จะพบ Outliers สูงขึ้น
 - **Platykurtic (Thin-Tailed Distribution):** เป็นการกระจายที่มีหางบางกว่าการกระจายปกติ มีค่า Kurtosis น้อยกว่า 3 บ่งบอกว่าโอกาสที่จะพบ Outliers ลดลง

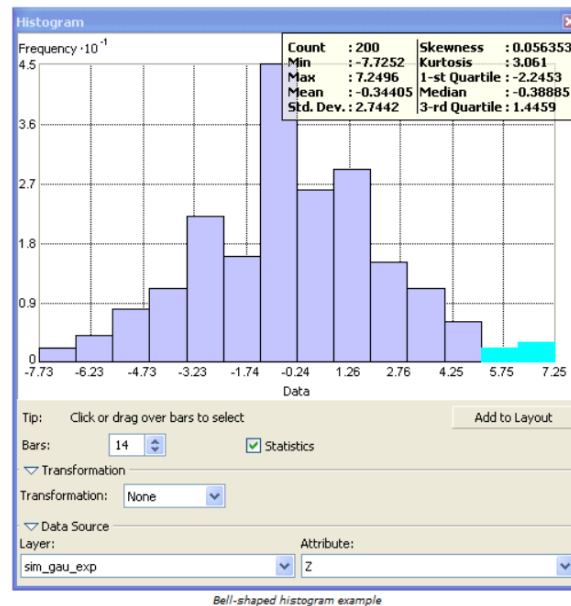
ในตัวอย่างกราฟสีแดงแสดงการกระจายปกติ (ค่า Kurtosis = 3) ส่วนกราฟสีดำแสดงการกระจายแบบ Leptokurtic ซึ่งมีหางหนาและค่า Kurtosis มากกว่า 3 การวิเคราะห์ค่า Kurtosis ช่วยให้เราประเมินความเป็นไปได้ที่ข้อมูลจะมีค่าผิดปกติ ซึ่งเป็นข้อมูลสำคัญในงานวิเคราะห์ทางสถิติและการตัดสินใจ



Normal distribution example

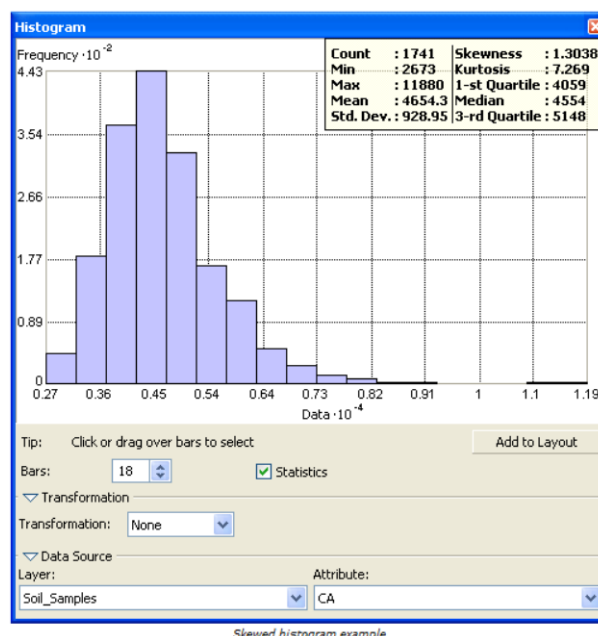
ตัวอย่างการใช้งาน Histogram Tool

จาก Histogram tool สามารถตรวจสอบรูปแบบของการกระจายข้อมูลได้โดยการสังเกตโดยตรง ด้วยการตรวจสอบค่ากลาง (mean) และค่ามัธยฐาน (median) ซึ่งสามารถกำหนดตำแหน่งของศูนย์กลางการกระจายข้อมูลได้ สังเกตได้จากรูปในภาพด้านล่าง การกระจายข้อมูลมีลักษณะเป็นรูปประฆัง และเมื่อค่าเฉลี่ยและค่ามัธยฐานมีค่าที่ใกล้เคียงกัน การกระจายนี้จึงใกล้เคียงกับการกระจายปกติ (normal distribution)

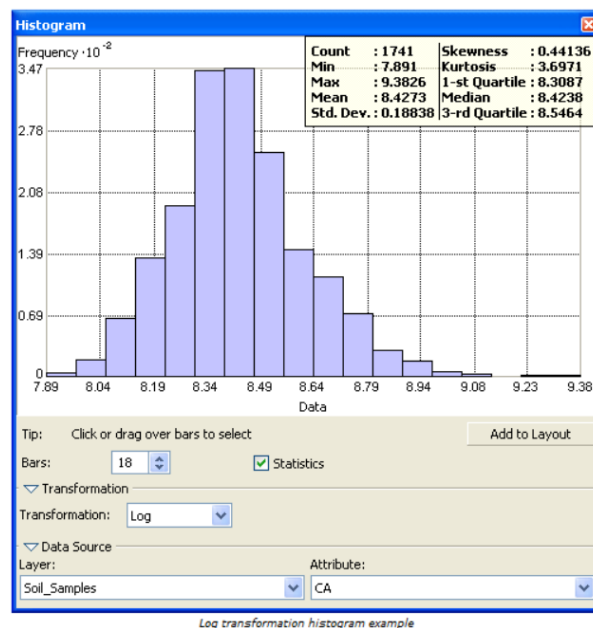


หากข้อมูลมีการกระจายที่เบ้ (skewed) อย่างมาก ก็สามารถทดสอบผลของการแปลงข้อมูล (transformation) ต่อชุดข้อมูลได้

ภาพตัวอย่างนี้แสดงการกระจายข้อมูลที่เบ้ก่อนที่จะมีการแปลงข้อมูล โดยการใช้การแปลงแบบลอการิทึม (log transformation) กับข้อมูลที่เบ้ การแปลงในกรณีนี้ทำให้การกระจายข้อมูลใกล้เคียงกับการกระจายปกติ



สำหรับข้อมูลเพิ่มเติมเกี่ยวกับการแปลงข้อมูลที่สามารถทำได้ด้วยเครื่องมือฮิสโตแกรม



การแปลงข้อมูลด้วย Box-Cox, Arcsine และ Log Transformations

ในบางกรณี การวิเคราะห์ข้อมูลจำเป็นต้องใช้ข้อมูลที่มีการกระจายตัวแบบปกติ (Normal Distribution) เช่น ในการใช้ **Geostatistical Analyst** หากข้อมูลมีการกระจายแบบเบ้ (Skewed Distribution) หรือไม่สมมาตร การแปลงข้อมูล (Data Transformation) จะช่วยปรับการกระจายให้ใกล้เคียงกับการกระจายปกติและทำให้ข้อมูลเหมาะสมสำหรับการวิเคราะห์มากขึ้น

เครื่องมือและการแปลงที่ใช้บ่อย

1. Box-Cox Transformation (Power Transformation)

- เป็นการแปลงข้อมูลด้วยฟังก์ชันที่สามารถปรับการกระจายของข้อมูลได้หลายรูปแบบ
- ใช้ได้กับข้อมูลที่มีค่าบวกทั้งหมด
- เหมาะสำหรับลดความเบ้ของข้อมูลและทำให้ความแปรปรวนมีความสม่ำเสมอทั่วพื้นที่การศึกษา

2. Arcsine Transformation

- ใช้สำหรับข้อมูลที่เป็นสัดส่วนหรือเปอร์เซ็นต์ เช่น ข้อมูลที่อยู่ในช่วง 0 ถึง 1
- ช่วยปรับการกระจายของข้อมูลให้ใกล้เคียงกับปกติ
- เหมาะสำหรับข้อมูลที่มีความหนาแน่นสูงในบางช่วง เช่น ข้อมูลเปอร์เซ็นต์ที่กระจุกตัวอยู่ใกล้ 0 หรือ 1

3. Logarithmic Transformation (Log Transformation)

- ใช้สำหรับข้อมูลที่มีค่าบวกและมีความเบ้ไปทางขวา (Right-Skewed)
- ช่วยลดผลกระทบของ Outliers และทำให้การกระจายสมมาตรมากขึ้น

ขั้นตอนการแปลงข้อมูล

- ให้สมมติว่าคุณมีข้อมูลดิบ $Z(s)$ และคุณเลือกการแปลงข้อมูล $Y(s)=t(Z(s))$
- เป้าหมายของการแปลงคือทำให้ข้อมูล $Y(s)$ มีการกระจายแบบปกติ
- นอกจากนี้ การแปลงยังช่วยทำให้ความแปรปรวนของข้อมูลคงที่ (Constant Variance) ในพื้นที่การศึกษา

ประโยชน์ของการแปลงข้อมูล

- ช่วยปรับปรุงคุณภาพของการวิเคราะห์ข้อมูล
- ทำให้ข้อมูลเหมาะสมกับสมมติฐานของแบบจำลอง
- ลดผลกระทบของ Outliers และความเบ้ของข้อมูล

การเลือกใช้ Box-Cox, Arcsine หรือ Log Transformation ขึ้นอยู่กับลักษณะของข้อมูลดิบและเป้าหมายของการวิเคราะห์.

การแปลงข้อมูลแบบ Box-Cox

สูตรการแปลง Box-Cox

$$Y(s) = (Z(s)^\lambda - 1)/\lambda,$$

สำหรับ $\lambda \neq 0$

- $Z(s)$: ข้อมูลดิบ
- λ : พารามิเตอร์ที่กำหนดรูปแบบของการแปลง

วัตถุประสงค์ของการแปลง Box-Cox

การแปลงนี้ออกแบบมาเพื่อ:

- ปรับการกระจายข้อมูลให้ใกล้เคียงกับการกระจายแบบปกติ (Normal Distribution)
- ทำให้ความแปรปรวนของข้อมูลคงที่ (Constant Variance) ทั่วพื้นที่การศึกษา

ตัวอย่างสถานการณ์การใช้งาน

สมมติว่าคุณกำลังศึกษาข้อมูลที่เกี่ยวข้องกับจำนวนการนับ (Counts) ของปรากฏการณ์บางอย่าง เช่น การนับจำนวนต้นไม้ในพื้นที่ต่าง ๆ:

- ในพื้นที่ที่มีจำนวนการนับน้อย ความแปรปรวนของข้อมูลมักจะต่ำ
- ในพื้นที่ที่มีจำนวนการนับมาก ความแปรปรวนของข้อมูลมักจะสูง
- การแปลง Box-Cox สามารถช่วยปรับข้อมูลเหล่านี้ให้มีความแปรปรวนสม่ำเสมอทั่วพื้นที่

กรณีเฉพาะ: Square-Root Transformation

- **Square-Root Transformation** เป็นกรณีเฉพาะของการแปลง Box-Cox เมื่อ $\lambda = 1/2$
- ใช้สำหรับข้อมูลที่เกี่ยวข้องกับการนับหรือข้อมูลที่มีค่าบวก
- ตัวอย่าง:
 - ข้อมูลเดิม $Z(s)$: จำนวนการนับ
 - การแปลง: $Y(s) = \sqrt{Z(s)}$

ข้อดีของ Box-Cox Transformation

1. **ความยืดหยุ่น:** สามารถปรับได้หลายรูปแบบโดยการเปลี่ยนค่าของ λ
2. **การลดความเบ้:** ปรับข้อมูลที่เบ้ (Skewed Data) ให้สมมาตรมากขึ้น
3. **การปรับความแปรปรวน:** ช่วยให้ความแปรปรวนของข้อมูลคงที่ทั่วพื้นที่

Box-Cox Transformation เป็นเครื่องมือที่ทรงพลังสำหรับการเตรียมข้อมูลในงานวิเคราะห์ทางสถิติและเหมาะสมสำหรับข้อมูลที่ต้องการความปกติและความแปรปรวนที่คงที่

การแปลงข้อมูลด้วย Log Transformation

Log Transformation เป็นกรณีพิเศษของ Box-Cox Transformation เมื่อ $\lambda = 0$;

สูตรสำหรับการแปลงข้อมูลคือ:

$$Y(s) = \ln(Z(s))$$

โดยที่ $Z(s) > 0$, และ $\ln$ คือค่า ลอการิทึมธรรมชาติ (Natural Logarithm)

วัตถุประสงค์ของ Log Transformation

1. **ลดความเบ้ (Skewness):**
 - เหมาะสำหรับข้อมูลที่มีการกระจายตัวแบบเบ้ไปทางขวา (Positively Skewed Distribution)
 - เช่น ข้อมูลที่มีค่าน้อยจำนวนมาก และมีค่าสูงบางค่าที่ห่างจากกลุ่มข้อมูลหลัก (Outliers)
2. **ปรับความแปรปรวน (Variance Stabilization):**
 - ช่วยทำให้ความแปรปรวนของข้อมูลคงที่ทั่วพื้นที่การศึกษา โดยลดผลกระทบจากค่าข้อมูลที่สูงมาก
3. **ปรับการกระจายให้ใกล้เคียงปกติ:**
 - ทำให้ข้อมูลที่มีลักษณะสมมาตรและเหมาะสมสำหรับการวิเคราะห์ทางสถิติ

ตัวอย่างการใช้งาน

- ข้อมูลดิบ $Z(s)$: จำนวนประชากรในพื้นที่
- หากข้อมูลมีบางพื้นที่ที่มีประชากรมากผิดปกติ (Outliers) การใช้ $\ln(Z(s))$ จะลดผลกระทบของค่าที่สูงมาก และช่วยปรับการกระจายให้สมมาตรมากขึ้น

การใช้กับ Kriging

1. Lognormal Kriging:

- เมื่อใช้ Log Transformation ($\lambda=0$) ร่วมกับวิธี Kriging การทำนายจะเรียกว่า **Lognormal Kriging**
- หลังการวิเคราะห์ ผลลัพธ์จะถูกแปลงกลับเป็นสเกลดั้งเดิม

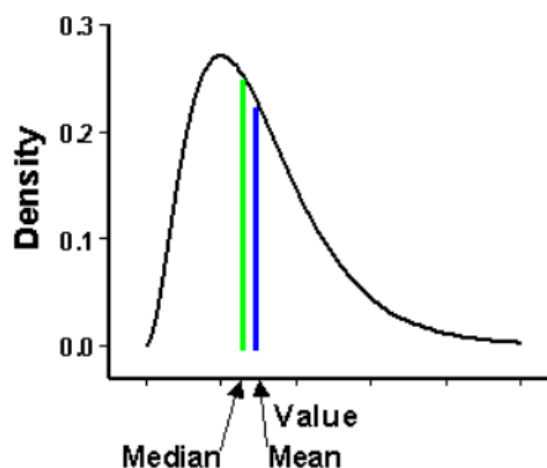
2. Trans-Gaussian Kriging:

- เมื่อใช้ค่า λ อื่น ๆ กับ Box-Cox Transformation การทำนายจะเรียกว่า **Trans-Gaussian Kriging**

ข้อดีของ Log Transformation

- ลดผลกระทบของค่าข้อมูลที่สูงมาก (Extreme Values)
- ช่วยให้การกระจายข้อมูลเหมาะสมสำหรับแบบจำลองที่ต้องการข้อมูลแบบปกติ
- ใช้ได้กับข้อมูลที่มีค่าบวกทั้งหมด

Log Transformation เป็นเครื่องมือสำคัญในงานวิเคราะห์ข้อมูล โดยเฉพาะเมื่อต้องจัดการข้อมูลที่มีการกระจายแบบเบ้และค่าผิดปกติที่สูงมาก



การแปลงข้อมูลแบบ Arcsine Transformation

สูตรการแปลง:

$$Y(s) = \sin^{-1}(Z(s)),$$

โดยที่ $Z(s)$ อยู่ในช่วงระหว่าง 0 ถึง 1

วัตถุประสงค์ของ Arcsine Transformation

1. เหมาะสำหรับข้อมูลประเภทสัดส่วนหรือเปอร์เซ็นต์:

- เช่น ข้อมูลที่แสดงสัดส่วนของพื้นที่ปกคลุมด้วยป่าไม้ หรือเปอร์เซ็นต์ของประชากรที่มีคุณสมบัติบางอย่าง

2. ปรับความแปรปรวนของข้อมูล:

- สำหรับข้อมูลสัดส่วน ความแปรปรวนมักจะเล็กที่สุดเมื่อค่าของ $Z(s)$ อยู่ใกล้ 0 หรือ 1
- ความแปรปรวนจะมีค่าสูงที่สุดเมื่อ $Z(s)$ อยู่ใกล้ 0.5
- การแปลงแบบ Arcsine ช่วยปรับความแปรปรวนให้คงที่ทั่วพื้นที่การศึกษา

3. ปรับการกระจายให้ใกล้เคียงปกติ (Normal Distribution):

- การแปลงนี้สามารถทำให้ข้อมูลที่เป็นสัดส่วนดูสมมาตรและเหมาะสมกับการวิเคราะห์ทางสถิติ

ตัวอย่างการใช้งาน

สมมติว่าคุณกำลังศึกษาสัดส่วนของพื้นที่ปกคลุมด้วยน้ำในแต่ละพื้นที่:

- ข้อมูลดิบ $Z(s)$: สัดส่วนพื้นที่ปกคลุมด้วยน้ำ (ค่าระหว่าง 0 ถึง 1)
- หลังการแปลง $Y(s) = \sin^{-1}(Z(s))$: ข้อมูลที่ได้จะมีความแปรปรวนที่สมดุลมากขึ้นและใกล้เคียงกับการกระจายปกติ

ข้อดีของ Arcsine Transformation

- **จัดการความแปรปรวนที่ไม่สม่ำเสมอ:** โดยเฉพาะในข้อมูลสัดส่วนที่มีความแปรปรวนเปลี่ยนแปลงตามค่าของ $Z(s)$
- **เหมาะสำหรับการวิเคราะห์ทางสถิติ:** ช่วยให้ข้อมูลเหมาะสมกับสมมติฐานของแบบจำลอง
- **ลดปัญหาค่าผิดปกติ:** ทำให้ข้อมูลสัดส่วนมีความสมมาตรมากขึ้น

Arcsine Transformation เป็นเครื่องมือสำคัญในการจัดการข้อมูลประเภทสัดส่วนและเปอร์เซ็นต์ โดยเฉพาะเมื่อความแปรปรวนของข้อมูลเปลี่ยนแปลงตามค่าของข้อมูล

2. Quantile-Quantile (QQ) Plot

QQ Plot เป็นกราฟที่ใช้เปรียบเทียบ **ควอนไทล์ (Quantiles)** ระหว่างการกระจายตัวของข้อมูลสองชุด

- ใช้เพื่อตรวจสอบว่าข้อมูลมีลักษณะการกระจายตัวแบบปกติหรือไม่
- ใช้เพื่อตรวจสอบว่าข้อมูลสองชุดมีลักษณะการกระจายตัวที่คล้ายกันหรือไม่
- หากจุดข้อมูลเรียงตัวเป็นเส้นตรงในกราฟ QQ Plot แสดงว่าข้อมูลทั้งสองชุดมีการกระจายตัวที่คล้ายกัน

Normal QQ Plot

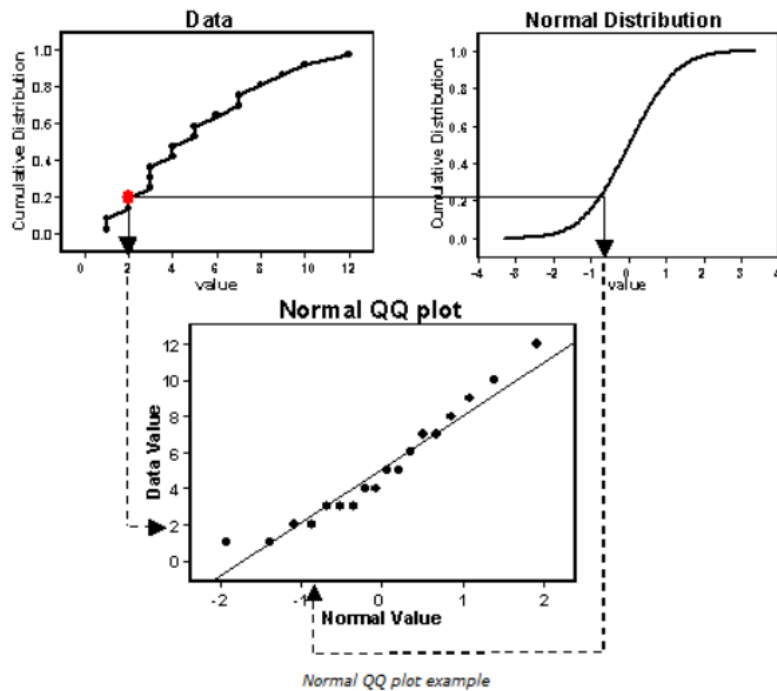
Normal QQ Plot เป็นกราฟที่ใช้เปรียบเทียบการกระจายตัวของข้อมูลกับการกระจายตัวแบบปกติ (Normal Distribution)

วิธีการสร้าง Normal QQ Plot

1. จัดเรียงข้อมูลและคำนวณค่า **Cumulative Distribution (CDF)**:
 - เรียงค่าข้อมูลจากน้อยไปมาก (Ordered Data)
 - คำนวณค่า CDF สำหรับค่าที่เรียงลำดับตามสูตร: $CDF = (i - 0.5) / n$ โดย i คืออันดับของข้อมูล และ n คือจำนวนข้อมูลทั้งหมด
 - คำนับบอกว่า **สัดส่วนของข้อมูล** ที่ต่ำกว่าค่าหนึ่ง ๆ มีค่าเท่าไร
2. สร้างกราฟการกระจายสะสม (CDF) สำหรับข้อมูลดั้งเดิม:
 - นำค่าที่เรียงลำดับไปพล็อตกราฟเทียบกับค่าที่คำนวณได้จาก CDF
3. สร้างกราฟการกระจายสะสม (CDF) สำหรับการกระจายปกติ (**Normal Distribution**):
 - ใช้ค่าการกระจายปกติมาตรฐาน (Mean = 0, Standard Deviation = 1)
 - คำนวณ CDF ตามลำดับควอนไทล์ของการกระจายปกติ
4. จับคู่ **Quantiles**:
 - นำค่าควอนไทล์จากข้อมูลดั้งเดิมมาจับคู่กับค่าควอนไทล์ที่ได้จากการกระจายปกติ
 - พล็อตค่าคู่ในกราฟ QQ Plot
5. วิเคราะห์ **QQ Plot**:
 - หากข้อมูลดั้งเดิมมีการกระจายตัวเหมือนกับการกระจายปกติ จุดในกราฟ QQ Plot จะเรียงตัวอยู่ใกล้เส้นทแยงมุม (Straight Line)
 - หากมีการกระจายที่แตกต่าง เช่น ข้อมูลเบ้หรือมี Outliers จุดจะเบี่ยงเบนจากเส้นตรง

ประโยชน์ของ Normal QQ Plot

- **ตรวจสอบความปกติของข้อมูล (Normality)**: ใช้ดูว่าข้อมูลมีการกระจายตัวใกล้เคียงกับการกระจายปกติหรือไม่
- **เปรียบเทียบการกระจายตัวของข้อมูลสองชุด**: ใช้ตรวจสอบความเหมือนหรือความแตกต่างของการกระจาย
- **ระบุ Outliers หรือข้อมูลที่ผิดปกติ**: จุดที่ห่างจากเส้นตรงอาจบ่งบอกถึงค่าผิดปกติในข้อมูล



General QQ Plot

General QQ Plot เป็นกราฟที่ใช้เปรียบเทียบการกระจายตัวของข้อมูลระหว่างชุดข้อมูลสองชุด

- จุดมุ่งหมายคือการตรวจสอบว่าข้อมูลสองชุดมีการกระจายตัวเหมือนกันหรือไม่
- หากข้อมูลทั้งสองชุดมีการกระจายตัวที่เหมือนกัน จุดในกราฟ QQ Plot จะเรียงตัวเป็นเส้นตรงทำมุม 45 องศา

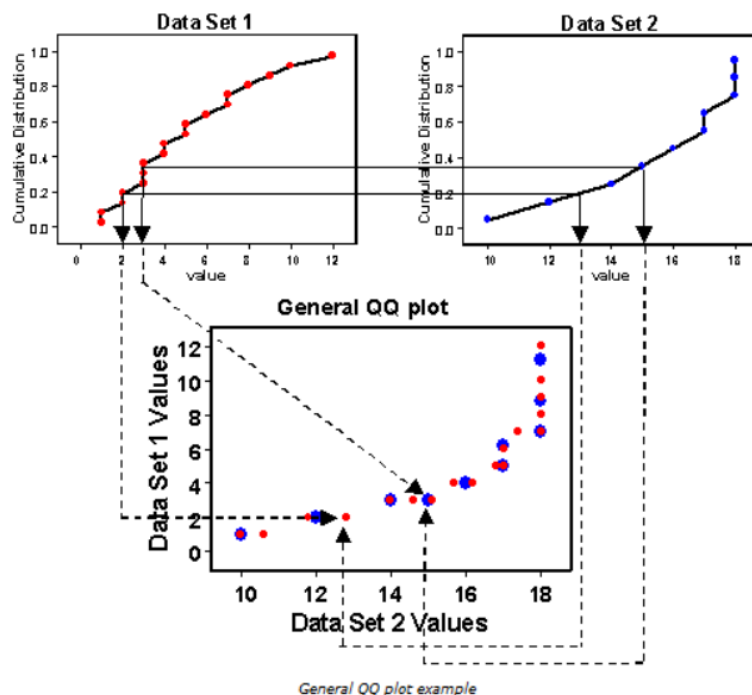
วิธีการสร้าง General QQ Plot

1. จัดเรียงข้อมูลในแต่ละชุดข้อมูล:
 - สำหรับชุดข้อมูลแรก (Data A\text{Data A}Data A) และชุดข้อมูลที่สอง (Data B\text{Data B}Data B)
 - เรียงข้อมูลในแต่ละชุดจากน้อยไปมาก (Ordered Data)
2. จับคู่ Quantiles จากทั้งสองชุดข้อมูล:
 - คำนวณตำแหน่งควอไทล์ในแต่ละชุดข้อมูล
 - คู่ควอไทล์ที่ตำแหน่งเดียวกันจากทั้งสองชุดจะถูกจับคู่กัน
3. พล็อตค่าควอไทล์:
 - พล็อตค่าควอไทล์ของ Data A\text{Data A}Data A บนแกน xxx และ Data B\text{Data B}Data B บนแกน yyy
 - กราฟที่ได้คือ General QQ Plot
4. วิเคราะห์ลักษณะกราฟ:
 - หากจุดในกราฟเรียงตัวเป็นเส้นตรงที่ทำมุม 45 องศา: หมายความว่ามีการกระจายตัวของทั้งสองชุดเหมือนกัน
 - หากจุดเบี่ยงเบนจากเส้นตรง: บ่งชี้ว่าการกระจายตัวของชุดข้อมูลทั้งสองแตกต่างกัน

ประโยชน์ของ General QQ Plot

1. **เปรียบเทียบการกระจายตัวของข้อมูลสองชุด:**
ใช้ดูว่าข้อมูลทั้งสองชุดมีลักษณะการกระจายที่คล้ายคลึงกันหรือไม่
2. **ระบุความแตกต่างของการกระจายตัว:**
ช่วยให้เข้าใจว่าเหตุใดผลลัพธ์หรือแนวโน้มในข้อมูลสองชุดจึงอาจแตกต่างกัน
3. **ตรวจสอบการแปลงข้อมูล:**
ใช้ตรวจสอบว่าการแปลงข้อมูล (Transformation) ทำให้การกระจายตัวเหมือนกันหรือไม่

General QQ Plot เป็นเครื่องมือสำคัญที่ช่วยให้คุณเข้าใจความสัมพันธ์ระหว่างการกระจายตัวของข้อมูลสองชุดอย่างละเอียดและตรงประเด็น



การตรวจสอบการกระจายตัวของข้อมูลด้วย QQ Plot

Normal QQ Plot เป็นเครื่องมือที่ใช้สำหรับตรวจสอบว่าข้อมูลชุดหนึ่งมีการกระจายตัวเป็นแบบปกติ (Normal Distribution) หรือไม่

การแปลผลจาก Normal QQ Plot

- **ข้อมูลที่มีการกระจายตัวเป็นแบบปกติ:**
 - จุดบนกราฟ QQ Plot จะเรียงตัวใกล้เคียงกับเส้นตรง 45 องศา
 - หมายความว่าข้อมูลนั้นมีลักษณะการกระจายที่เป็นปกติ (Normal Distribution)
- **ข้อมูลที่ไม่เป็นแบบปกติ:**
 - จุดบนกราฟจะเบี่ยงเบนจากเส้นตรง 45 องศา
 - การเบี่ยงเบนนี้แสดงถึงความไม่ปกติในการกระจายตัวของข้อมูล

วิธีการสร้างและการอ่านผล Normal QQ Plot

1. กราฟ Normal QQ Plot:

- ค่าควอไทล์จากการกระจายปกติ (Standard Normal Distribution) จะถูกพล็อตไว้บนแกน x
- ค่าควอไทล์ของชุดข้อมูลจริงจะถูกพล็อตไว้บนแกน y
- หากข้อมูลมีการกระจายแบบปกติ จุดจะเรียงตัวบนเส้นตรง 45 องศา

2. การเบี่ยงเบนจากเส้นตรง:

- จุดที่เบี่ยงเบนจากเส้นตรงอาจบ่งบอกถึงลักษณะการกระจายตัวที่ไม่เป็นปกติ
- ตัวอย่าง: ในกราฟที่แสดงการกระจายของค่าความเข้มข้นของโอโซน (Ozone Concentration) จุดจะเบี่ยงเบนจากเส้นตรงในค่าที่สูง

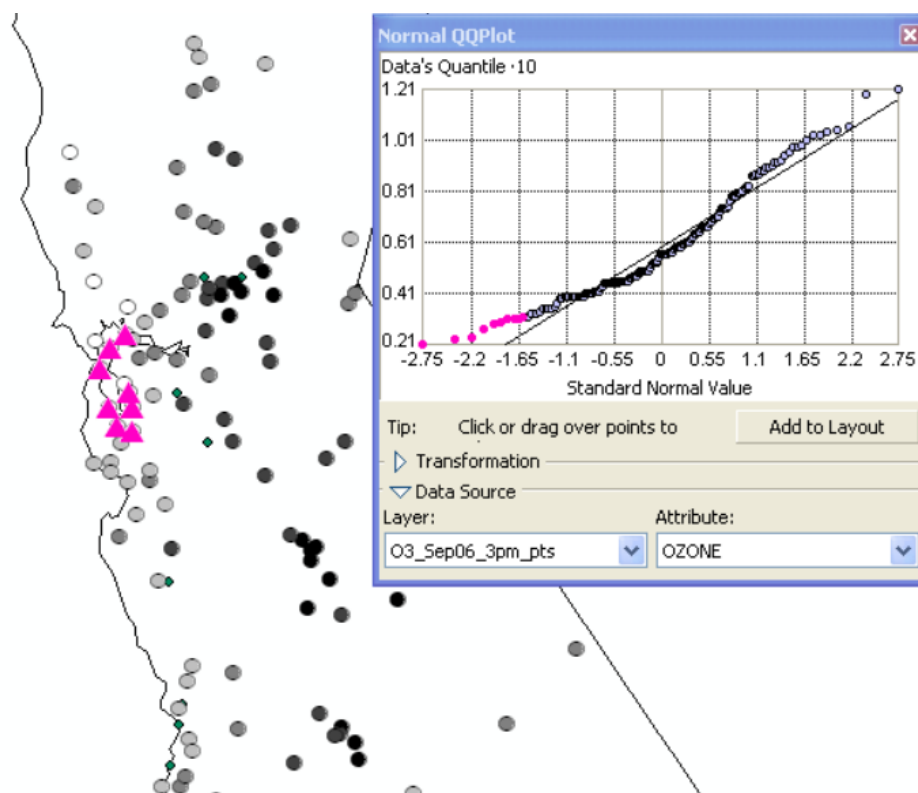
3. การระบุจุดที่เบี่ยงเบน:

- เครื่องมือ **Normal QQ Plot** ช่วยให้เลือกจุดที่ไม่อยู่ใกล้เส้นตรงได้
- จุดเหล่านี้จะถูกเน้นใน **ArcMap Data View**
- ตัวอย่างในแผนที่ด้านล่าง จุดที่เบี่ยงเบนถูกเน้นในพื้นที่ **San Francisco Bay Area** (สีชมพูในแผนที่)

การใช้งาน Normal QQ Plot

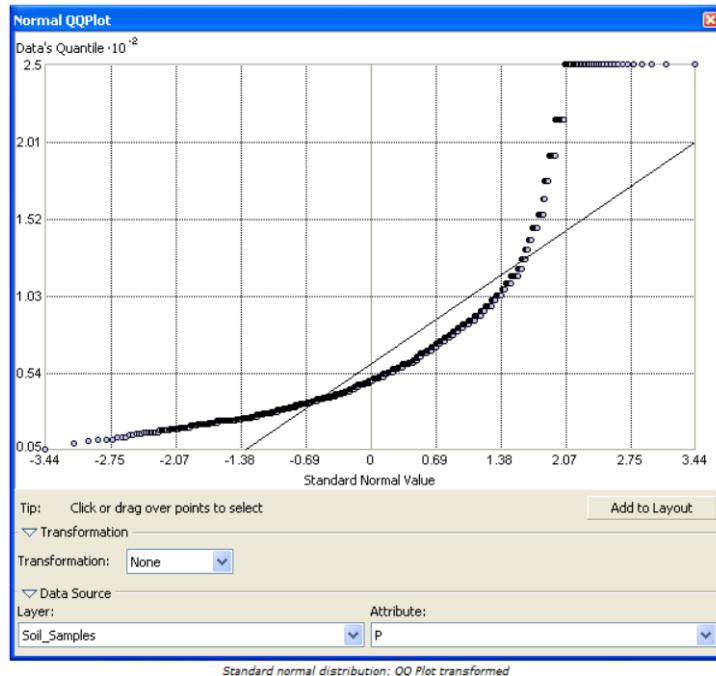
- **ตรวจสอบการกระจายตัวของข้อมูล:** ใช้เพื่อยืนยันว่าข้อมูลมีการกระจายแบบปกติหรือไม่
- **ระบุจุดที่ผิดปกติ (Outliers):** จุดที่เบี่ยงเบนจากเส้นตรงอาจบ่งบอกถึงค่า Outliers ที่ควรตรวจสอบ
- **การแสดงข้อมูลที่เป็นกลุ่ม:** การเบี่ยงเบนจากเส้นตรงอาจช่วยให้เห็นกลุ่มหรือพื้นที่ที่มีข้อมูลที่ไม่ปกติ

Normal QQ Plot เป็นเครื่องมือที่มีประสิทธิภาพในการประเมินการกระจายตัวของข้อมูลและช่วยให้คุณตรวจสอบความสมบูรณ์ของข้อมูลได้ง่ายขึ้น

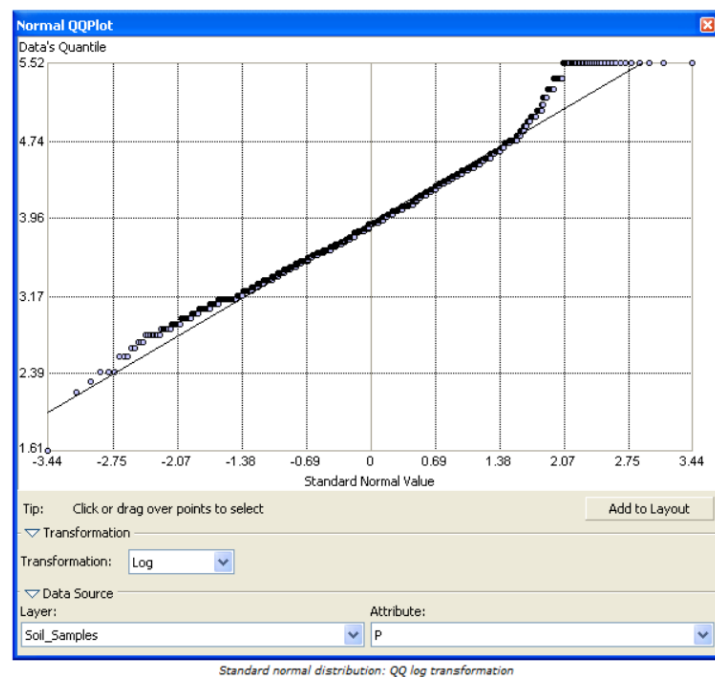


ตัวอย่างการใช้การแปลงข้อมูล

ตัวอย่างนี้คือ **Normal QQ plot** ของชุดข้อมูลตัวอย่าง: สังเกตว่า จุดในกราฟเบี่ยงเบนจากเส้นตรง



อย่างไรก็ตาม, เมื่อมีการใช้ **log transformation** กับชุดข้อมูล จุดจะอยู่ใกล้กับเส้นตรง 45 องศามากขึ้น ตามที่เห็นในภาพด้านล่าง



การแปลงข้อมูลแบบ **Box-Cox** และ **arcsine** ก็สามารถนำมาประยุกต์ใช้กับข้อมูลภายในเครื่องมือ **Normal QQ Plot** เพื่อตรวจสอบผลของการแปลงเหล่านี้ต่อการกระจายตัวที่เป็นปกติ (normality) ของข้อมูล