บทที่ 3 การวิเคราะห์ข้อมูลเชิงสถิติ
(Statistical Data Analysis)

การวิเคราะห์ข้อมูลเชิงสถิติ เป็นกระบวนการสำคัญที่ช่วยให้สามารถเข้าใจข้อมูลจำนวนมากได้อย่างเป็นระบบ สถิติเป็นเครื่องมือที่มีประสิทธิภาพในการสรุปข้อมูล ค้นหาความสัมพันธ์ และทำนายแนวโน้มในอนาคต การวิเคราะห์เชิงสถิติมีบทบาทสำคัญในการสนับสนุนการตัดสินใจทางธุรกิจและการวิจัย

1. การวิเคราะห์ข้อมูลเชิงสถิติเบื้องต้น

การวิเคราะห์ข้อมูลเชิงสถิติ (Statistical Data Analysis) หมายถึง กระบวนการในการรวบรวมข้อมูล การสรุป การตรวจสอบ และการแปลความหมายข้อมูลโดยใช้เครื่องมือและเทคนิคทางสถิติ เพื่อค้นหาความสัมพันธ์ รูปแบบ แนวโน้ม หรือสรุปข้อมูลในเชิงปริมาณ การวิเคราะห์เชิงสถิตินั้นสามารถนำไปใช้ได้ในหลายบริบท เช่น การทำนาย การประมาณค่า หรือการทดสอบสมมติฐาน

การวิเคราะห์ข้อมูลเชิงสถิติมีความสำคัญในการวิจัยและการตัดสินใจเชิงธุรกิจ เพราะช่วยให้สามารถวิเคราะห์ข้อมูลจำนวนมากได้อย่างเป็นระบบและให้ผลลัพธ์ที่สามารถนำไปใช้งานได้จริง

1.1 ความสำคัญของการวิเคราะห์ข้อมูลเชิงสถิติ

1.1.1 การสรุปข้อมูล ช่วยสรุปข้อมูลจำนวนมากให้อยู่ในรูปแบบที่ง่ายต่อการเข้าใจ เช่น การใช้ค่าเฉลี่ย (Mean), ค่ามัธยฐาน (Median), หรือการกระจายของข้อมูล (Standard Deviation)

1.1.2 การค้นหารูปแบบในข้อมูล (Pattern Identification) ช่วยให้ค้นพบแนวโน้มหรือรูปแบบที่ซ่อนอยู่ในข้อมูล เช่น การวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรต่าง ๆ หรือการหากลุ่มข้อมูลที่มีลักษณะคล้ายกัน

1.1.3 การทำนายแนวโน้ม (Trend Prediction) ใช้ในการคาดการณ์หรือทำนายสิ่งที่จะเกิดขึ้นในอนาคตจากข้อมูลในอดีต เช่น การพยากรณ์ยอดขาย หรือการทำนายผลลัพธ์ทางธุรกิจ

1.1.4 การทดสอบสมมติฐาน (Hypothesis Testing) การทดสอบสมมติฐานทางสถิติช่วยในการตรวจสอบว่าความสัมพันธ์ที่ค้นพบในข้อมูลนั้นเป็นสิ่งที่เกิดขึ้นจริงหรือเป็นเพียงเหตุการณ์ที่เกิดจากความบังเอิญ

1.1.5 การตัดสินใจเชิงข้อมูล (Data-Driven Decision Making) ช่วยให้การตัดสินใจที่มีความสำคัญสามารถอ้างอิงจากข้อมูลจริงมากกว่าการคาดเดาหรือประสบการณ์ส่วนบุคคล

1.2 ประเภทของการวิเคราะห์ข้อมูลเชิงสถิติ

1.2.1 สถิติเชิงพรรณนา (Descriptive Statistics) เป็นการสรุปข้อมูลโดยการใช้ค่ากลาง (Central Tendency) และการกระจาย (Dispersion) ได้แก่

1) ค่าเฉลี่ย (Mean) ค่าเฉลี่ยเป็นการคำนวณจากการรวมค่าทั้งหมดแล้วหารด้วยจำนวนข้อมูลทั้งหมด เช่น การคำนวณค่าเฉลี่ยรายได้ของพนักงานในบริษัท

2) ค่ามัธยฐาน (Median) ค่ากลางของข้อมูลเมื่อเรียงข้อมูลจากน้อยไปมาก เช่น การหาค่ามัธยฐานของคะแนนสอบนักเรียนในชั้นเรียน

3) ฐานนิยม (Mode) ค่าที่เกิดขึ้นบ่อยที่สุดในข้อมูล

4) ส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) เป็นการวัดการกระจายตัวของข้อมูลออกจากค่าเฉลี่ย

1.2.2 สถิติเชิงอนุมาน (Inferential Statistics) เป็นการวิเคราะห์ข้อมูลเพื่อทำการคาดการณ์หรือสรุปผลจากกลุ่มตัวอย่างไปยังกลุ่มประชากร ตัวอย่างเครื่องมือในสถิติเชิงอนุมาน

1) การทดสอบสมมติฐาน (Hypothesis Testing) เช่น การทดสอบค่าเฉลี่ยของสองกลุ่มโดยใช้ t-test หรือการวิเคราะห์ความแปรปรวนโดยใช้ ANOVA เช่น การทดสอบว่าค่าเฉลี่ยน้ำหนักของผู้ใหญ่ชายในเมืองหนึ่งมีความแตกต่างจากค่าเฉลี่ยน้ำหนักของผู้ใหญ่ชายทั่วประเทศหรือไม่

2) การสร้างช่วงความเชื่อมั่น (Confidence Interval) ใช้ในการบอกช่วงค่าที่มั่นใจว่าค่าจริงของประชากรจะอยู่ในช่วงนั้น เช่น การประเมินค่าเฉลี่ยของประชากรจากกลุ่มตัวอย่าง

1.2.3 การวิเคราะห์ความสัมพันธ์ (Correlation and Regression Analysis) เป็นการวิเคราะห์เพื่อหาความสัมพันธ์ระหว่างตัวแปรสองตัวหรือมากกว่า ตัวอย่างการวิเคราะห์ความสัมพันธ์

1) Correlation Analysis การวิเคราะห์หาค่าสหสัมพันธ์ (Correlation Coefficient) เพื่อวัดว่าความสัมพันธ์ระหว่างตัวแปรสองตัวมีความแข็งแรงและมีทิศทางอย่างไร เช่น การวิเคราะห์ว่าความสัมพันธ์ระหว่างอายุและรายได้เป็นบวกหรือลบ

2) Regression Analysis การวิเคราะห์เพื่อสร้างสมการการคาดการณ์ เช่น การใช้ข้อมูลราคาบ้านในอดีตเพื่อสร้างโมเดลทำนายราคาบ้านในอนาคต

1.2.4 การวิเคราะห์เชิงพยากรณ์ (Predictive Analytics) คือ การวิเคราะห์เพื่อทำการทำนายหรือคาดการณ์ผลลัพธ์ในอนาคต โดยใช้ข้อมูลในอดีต ซึ่งสามารถทำได้ด้วยเทคนิคต่าง ๆ เช่น การถดถอยเชิงเส้น (Linear Regression) การจำแนกประเภท (Classification) หรือการวิเคราะห์เชิงอนุกรมเวลา (Time Series Analysis)

1.3 ตัวอย่างของการวิเคราะห์ข้อมูลเชิงสถิติ

สมมติว่ามีไฟล์ข้อมูลคะแนนแบบสอบถามความพึงพอใจแต่ละหัวข้อ และต้องการทำการวิเคราะห์เชิงสถิติเพื่อหาค่าสรุป เช่น ค่าเฉลี่ย ค่ามัธยฐาน และค่ามากที่สุดในแต่ละหัวข้อ

ตัวอย่างที่ 1

โค้ด	import pandas as pd
	
	# โหลดข้อมูลจากไฟล์ CSV
	file_path = 'questionnaire.csv'
	data = pd.read_csv(file_path)
	
	# สถิติเชิงพรรณนา หาค่าเฉลี่ย, ค่ามัธยฐาน และค่ามากที่สุดในแต่ละหัวข้อ
	mean_values = data.mean()
	median_values = data.median()
	max_values = data.max()
	
	# แสดงสถิติที่ได้
	print("\nค่าเฉลี่ยในแต่ละหัวข้อ")
	print(mean_values)
	
	print("\nค่ามัธยฐานในแต่ละหัวข้อ")
	print(median_values)
	
	print("\nค่ามากที่สุดในแต่ละหัวข้อ")
	print(max_values)

ผลลัพธ์	ค่าเฉลี่ยในแต่ละหัวข้อ Q1 4.375 Q2 4.000 Q3 4.125 Q4 4.625 Q5 4.750 Q6 4.125 Q7 4.375

	Q8 4.625 Q9 4.250 Q10 4.750 dtype: float64 ค่ามัธยฐานในแต่ละหัวข้อ Q1 4.5 Q2 4.0 Q3 4.0 Q4 5.0 Q5 5.0 Q6 4.0 Q7 4.0 Q8 5.0 Q9 4.0 Q10 5.0 dtype: float64 ค่ามากที่สุดในแต่ละหัวข้อ Q1 5 Q2 5 Q3 5 Q4 5 Q5 5 Q6 5 Q7 5 Q8 5 Q9 5 Q10 5 dtype: int64

2. การทดสอบสมมติฐานทางสถิติ (Hypothesis Testing)

การทดสอบสมมติฐานทางสถิติ (Hypothesis Testing) เป็นกระบวนการในการตรวจสอบแนวคิดหรือสมมติฐานเกี่ยวกับประชากรโดยอาศัยข้อมูลจากกลุ่มตัวอย่าง (Sample) โดยใช้วิธีการทางสถิติเพื่อตัดสินว่าสมมติฐานที่ตั้งไว้นั้นมีความถูกต้องหรือไม่ หรือเป็นเพียงสิ่งที่เกิดขึ้นจากความบังเอิญ การทดสอบสมมติฐานทางสถิติเป็นพื้นฐานที่สำคัญในงานวิจัยและการวิเคราะห์ข้อมูลเพื่อตัดสินใจอย่างมีหลักฐานเชิงข้อมูล

2.1 องค์ประกอบของการทดสอบสมมติฐานทางสถิติ

2.1.1 สมมติฐานหลัก (Null Hypothesis - H_0)

สมมติฐานหลักเป็นสมมติฐานที่กล่าวว่าข้อมูลจากกลุ่มตัวอย่างไม่ได้มีความแตกต่างหรือไม่มีความสัมพันธ์กัน สมมติฐานนี้มักเขียนว่า "ไม่มีความแตกต่าง" หรือ "ไม่มีความสัมพันธ์" เช่น ค่าเฉลี่ยของน้ำหนักนักเรียนในกลุ่มนี้ไม่แตกต่างจาก 60 กิโลกรัม

2.1.2 สมมติฐานทางเลือก (Alternative Hypothesis - H_1)

สมมติฐานทางเลือกเป็นสมมติฐานที่ตรงข้ามกับสมมติฐานหลัก และแสดงว่ามีความแตกต่างหรือความสัมพันธ์บางอย่างเกิดขึ้น เช่น ค่าเฉลี่ยของน้ำหนักนักเรียนในกลุ่มนี้แตกต่างจาก 60 กิโลกรัม

2.1.3 ค่าสถิติการทดสอบ (Test Statistic)

ค่าสถิติที่ใช้ในการคำนวณความแตกต่างระหว่างข้อมูลในกลุ่มตัวอย่างกับสมมติฐานหลัก ตัวอย่างเช่น ค่า t-statistic ในการทดสอบค่าเฉลี่ย หรือ z-statistic ในการทดสอบสัดส่วน

2.1.4 ระดับนัยสำคัญ (Significance Level - α)

ระดับนัยสำคัญคือโอกาสที่จะปฏิเสธสมมติฐานหลัก (H_0) ทั้ง ๆ ที่สมมติฐานหลักเป็นจริง โดยทั่วไประดับนัยสำคัญที่ใช้มากที่สุดคือ 0.05 หรือ 5% ซึ่งหมายความว่ายอมรับโอกาสที่จะเกิดข้อผิดพลาดเพียง 5% ในการปฏิเสธสมมติฐานหลัก เช่น หากใช้ระดับนัยสำคัญที่ 0.05 หมายถึงมีโอกาส 5% ที่จะตัดสินผิดพลาด

2.1.5 ค่า p (p-value)

ค่า p-value เป็นค่าความน่าจะเป็นที่ได้จากการทดสอบสมมติฐาน หากค่า p-value ต่ำกว่าระดับนัยสำคัญที่กำหนดไว้ (เช่น 0.05) จะปฏิเสธสมมติฐานหลัก แต่หากค่า p-value สูงกว่าระดับนัยสำคัญจะไม่ปฏิเสธสมมติฐานหลัก ค่า p-value เป็นตัวช่วยบ่งชี้ว่าโอกาสที่ผลลัพธ์เกิดจากความบังเอิญมีมากน้อยเพียงใด

2.2 ขั้นตอนของการทดสอบสมมติฐานทางสถิติ

2.2.1 ตั้งสมมติฐาน

กำหนดสมมติฐานหลัก (H_0) และสมมติฐานทางเลือก (H_1) เช่น สมมติฐานหลักอาจเป็น "ค่าเฉลี่ยของกลุ่มตัวอย่างเท่ากับค่าเฉลี่ยประชากร"

2.2.2 เลือกค่าสถิติการทดสอบ (Test Statistic)

เลือกค่าสถิติที่ใช้ในการทดสอบ เช่น t-test, z-test, หรือ ANOVA ขึ้นอยู่กับลักษณะของข้อมูลและคำถามที่ต้องการทดสอบ

2.2.3 กำหนดระดับนัยสำคัญ (Significance Level - α)

เลือกระดับนัยสำคัญ เช่น 0.05 หรือ 0.01 เพื่อใช้เป็นเกณฑ์ในการตัดสินว่าควรปฏิเสธสมมติฐานหลักหรือไม่

2.2.4 คำนวณค่าสถิติการทดสอบ

ใช้ข้อมูลจากกลุ่มตัวอย่างในการคำนวณค่าสถิติการทดสอบ เช่น ค่าทดสอบ t หรือ z และคำนวณค่า p-value จากค่าสถิติดังกล่าว

2.2.5 ตัดสินใจ

เปรียบเทียบค่า p-value กับระดับนัยสำคัญ

ถ้า p-value ≤ α ปฏิเสธสมมติฐานหลัก (H_0) และยอมรับสมมติฐานทางเลือก (H_1)

ถ้า p-value > α ไม่ปฏิเสธสมมติฐานหลัก (H_0)

2.3 ประเภทของการทดสอบสมมติฐาน

2.3.1 t-test

ใช้ทดสอบค่าเฉลี่ยของกลุ่มตัวอย่างเทียบกับค่าที่คาดหวัง หรือนำค่าเฉลี่ยระหว่างสองกลุ่มมาเปรียบเทียบกัน เช่น การทดสอบว่าค่าเฉลี่ยของคะแนนนักเรียนกลุ่มหนึ่งแตกต่างจากกลุ่มอื่นหรือไม่

2.3.2 z-test

ใช้ทดสอบค่าเฉลี่ยหรือสัดส่วนของประชากรเมื่อจำนวนกลุ่มตัวอย่างมีขนาดใหญ่และข้อมูลมีการกระจายแบบปกติ

2.3.3 ANOVA (Analysis of Variance)

ใช้ทดสอบค่าเฉลี่ยของกลุ่มมากกว่าสองกลุ่มว่ามีความแตกต่างกันหรือไม่ เช่น การทดสอบว่าค่าเฉลี่ยของยอดขายในแต่ละภูมิภาคมีความแตกต่างกันหรือไม่

2.3.4 Chi-square Test

ใช้ทดสอบความสัมพันธ์ระหว่างตัวแปรเชิงหมวดหมู่ (Categorical Variables) เช่น การทดสอบว่าเพศและความพึงพอใจในการซื้อสินค้าเกี่ยวข้องกันหรือไม่

2.4 ตัวอย่างการทดสอบสมมติฐานทางสถิติ

สมมติว่ามีข้อมูลน้ำหนักนักเรียนในกลุ่มตัวอย่าง และต้องการทดสอบว่าน้ำหนักเฉลี่ยของนักเรียนในกลุ่มนี้แตกต่างจาก 60 กิโลกรัมหรือไม่ สามารถใช้ t-test เพื่อตรวจสอบสมมติฐานนี้ได้

ตัวอย่างที่ 2

โค้ด	import numpy as np
	from scipy import stats
	
	# ข้อมูลน้ำหนักของนักเรียนในกลุ่มตัวอย่าง
	weights = np.array([58, 62, 57, 59, 61, 63, 60, 64])
	
	# ตั้งสมมติฐาน
	# H0 ค่าเฉลี่ยน้ำหนักนักเรียนในกลุ่มนี้เท่ากับ 60 กิโลกรัม
	# H1 ค่าเฉลี่ยน้ำหนักนักเรียนในกลุ่มนี้ไม่เท่ากับ 60 กิโลกรัม
	
	# ใช้ t-test เพื่อทดสอบค่าเฉลี่ยน้ำหนักกับค่า 60
	t_stat, p_value = stats.ttest_1samp(weights, 60)
	
	# แสดงค่า t-statistic และ p-value
	print(f"ค่า t-statistic {t_stat}")
	print(f"ค่า p-value {p_value}")
	
	# ตัดสินใจ
	alpha = 0.05
	if p_value < alpha
	  print("ปฏิเสธสมมติฐานหลัก (H0) ค่าเฉลี่ยน้ำหนักนักเรียนแตกต่างจาก 60 กิโลกรัม")
	else
	  print("ไม่ปฏิเสธสมมติฐานหลัก (H0) ค่าเฉลี่ยน้ำหนักนักเรียนไม่แตกต่างจาก 60 กิโลกรัม")

ผลลัพธ์	ค่า t-statistic 0.5773502691896258 ค่า p-value 0.5817882345917442 ไม่ปฏิเสธสมมติฐานหลัก (H0) ค่าเฉลี่ยน้ำหนักนักเรียนไม่แตกต่างจาก 60 กิโลกรัม

การทดสอบสมมติฐานทางสถิติ (Hypothesis Testing) เป็นกระบวนการสำคัญที่ใช้ในการตรวจสอบความถูกต้องของสมมติฐานที่ตั้งขึ้น โดยอาศัยข้อมูลจากกลุ่มตัวอย่างและการคำนวณค่าสถิติ เช่น ค่า t-statistic หรือค่า p-value การตัดสินใจว่าควรปฏิเสธหรือยอมรับสมมติฐานหลักขึ้นอยู่กับค่า p-value

3. การกระจายตัวของข้อมูลและการวิเคราะห์เชิงสถิติอื่น ๆ

การกระจายตัวของข้อมูล (Data Distribution) หมายถึง การแสดงรูปแบบการกระจายตัวของค่าข้อมูลในชุดข้อมูลหนึ่ง ๆ การกระจายตัวของข้อมูลเป็นพื้นฐานสำคัญในการทำความเข้าใจลักษณะของข้อมูล เช่น ความหนาแน่น ความเป็นกลาง (Symmetry) และรูปแบบที่พบเจอ โดยการวิเคราะห์การกระจายตัวนี้จะช่วยให้ทำความเข้าใจธรรมชาติของข้อมูลและเลือกวิธีการวิเคราะห์เชิงสถิติที่เหมาะสมกับข้อมูลนั้น ๆ

3.1 ประเภทของการกระจายตัวของข้อมูล

3.1.1 การกระจายแบบปกติ

1) การกระจายแบบปกติ (Normal Distribution) หรือเรียกว่า การกระจายเบลล์ (Bell Curve) เป็นการกระจายข้อมูลที่มีลักษณะสมมาตร ทั้งสองด้านของค่าเฉลี่ยจะมีการกระจายตัวเท่า ๆ กัน และค่าข้อมูลส่วนใหญ่จะอยู่ใกล้กับค่าเฉลี่ย ตัวอย่างเช่น คะแนนสอบของนักเรียนในกลุ่มใหญ่ที่มีการกระจายตัวแบบปกติ

2) ในการกระจายแบบปกติ ค่ากลางต่าง ๆ ได้แก่ ค่าเฉลี่ย (Mean), ค่ามัธยฐาน (Median) และฐานนิยม (Mode) จะมีค่าเท่ากัน

3.1.2 การกระจายแบบเบ้

การกระจายแบบเบ้ (Skewed Distribution) เกิดขึ้นเมื่อข้อมูลมีการกระจายตัวไม่สมมาตร ข้อมูลอาจจะกระจายตัวออกไปทางด้านซ้ายหรือด้านขวาของค่าเฉลี่ยมากเกินไป

1) เบ้ขวา (Positively Skewed) มีค่าข้อมูลส่วนใหญ่กระจายตัวทางด้านซ้ายของค่าเฉลี่ย ค่าที่ออกมาจะมีค่าน้อยกว่าค่าเฉลี่ย

2) เบ้ซ้าย (Negatively Skewed) มีค่าข้อมูลส่วนใหญ่กระจายตัวทางด้านขวาของค่าเฉลี่ย ค่าที่ออกมาจะมีค่ามากกว่าค่าเฉลี่ย

3.1.3 การกระจายแบบสม่ำเสมอ

การกระจายแบบสม่ำเสมอ (Uniform Distribution) หมายถึงข้อมูลที่กระจายตัวออกมาอย่างสม่ำเสมอในทุกช่วงของข้อมูล โดยไม่มีจุดสูงสุด (peak) ตัวอย่างเช่น การสุ่มตัวเลขจากช่วงที่มีค่าคงที่

3.1.3 การกระจายแบบไบนารี

การกระจายแบบไบนารี (Binary Distribution) คือ การกระจายที่เกิดจากข้อมูลที่มีเพียงสองค่าที่เป็นไปได้ เช่น ผลการทดสอบที่มีเพียงสองผลลัพธ์คือ "ผ่าน" หรือ "ไม่ผ่าน"

3.2 การวิเคราะห์เชิงสถิติอื่น ๆ ที่เกี่ยวข้องกับการกระจายตัวของข้อมูล

3.2.1 ค่ากลาง (Central Tendency)

ค่ากลางใช้เพื่อบ่งชี้ตำแหน่งกลางของข้อมูล ซึ่งประกอบไปด้วย

1) ค่าเฉลี่ย (Mean) ผลรวมของข้อมูลทั้งหมดหารด้วยจำนวนข้อมูล เป็นการวัดค่ากลางที่ใช้บ่อยที่สุด

2) ค่ามัธยฐาน (Median) ค่ากลางของข้อมูลเมื่อเรียงจากน้อยไปมาก โดยมัธยฐานใช้เมื่อต้องการลดผลกระทบจากค่าผิดปกติ

3) ฐานนิยม (Mode) ค่าที่พบมากที่สุดในชุดข้อมูล เหมาะกับข้อมูลเชิงหมวดหมู่

3.2.2 การวัดการกระจาย (Measure of Dispersion)

การวัดการกระจายใช้เพื่อดูว่าข้อมูลมีความแปรปรวนหรือห่างไกลจากค่ากลางมากน้อยเพียงใด

1) พิสัย (Range) ค่าสูงสุดลบด้วยค่าต่ำสุด

2) ส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) ใช้วัดการกระจายตัวของข้อมูลรอบค่าเฉลี่ย โดยยิ่งมีค่าส่วนเบี่ยงเบนมาตรฐานสูง ข้อมูลก็ยิ่งกระจายตัวมาก

3) ความแปรปรวน (Variance) ค่าเฉลี่ยของความแตกต่างระหว่างแต่ละจุดข้อมูลกับค่าเฉลี่ยยกกำลังสอง

3.2.3 การวัดรูปร่างของการกระจาย (Shape of Distribution)

การวัดรูปร่างของการกระจายช่วยให้รู้ว่าการกระจายข้อมูลมีลักษณะอย่างไร

1) ความเบ้ (Skewness) วัดการสมมาตรของข้อมูล ถ้าข้อมูลเบ้ขวา แปลว่ามีข้อมูลส่วนใหญ่อยู่ทางด้านซ้าย

2) ความโหนก (Kurtosis) วัดความสูงของยอดการกระจายข้อมูล ถ้าความโหนกมากแสดงว่ามีจุดสูงสุดที่ชัดเจน ข้อมูลจะแน่นรอบค่าเฉลี่ย

3.2.4 การทดสอบการกระจายตัว (Test of Normality)

การทดสอบการกระจายตัวใช้เพื่อตรวจสอบว่าข้อมูลมีการกระจายแบบปกติหรือไม่ ตัวอย่างเช่น

1) Shapiro-Wilk Test ใช้ตรวจสอบว่าข้อมูลมีการกระจายตัวแบบปกติหรือไม่

2) Kolmogorov-Smirnov Test เป็นการทดสอบการกระจายตัวที่ใช้ในการเปรียบเทียบระหว่างข้อมูลกับการกระจายแบบปกติ

3.3 ตัวอย่างการวิเคราะห์การกระจายตัวของข้อมูล

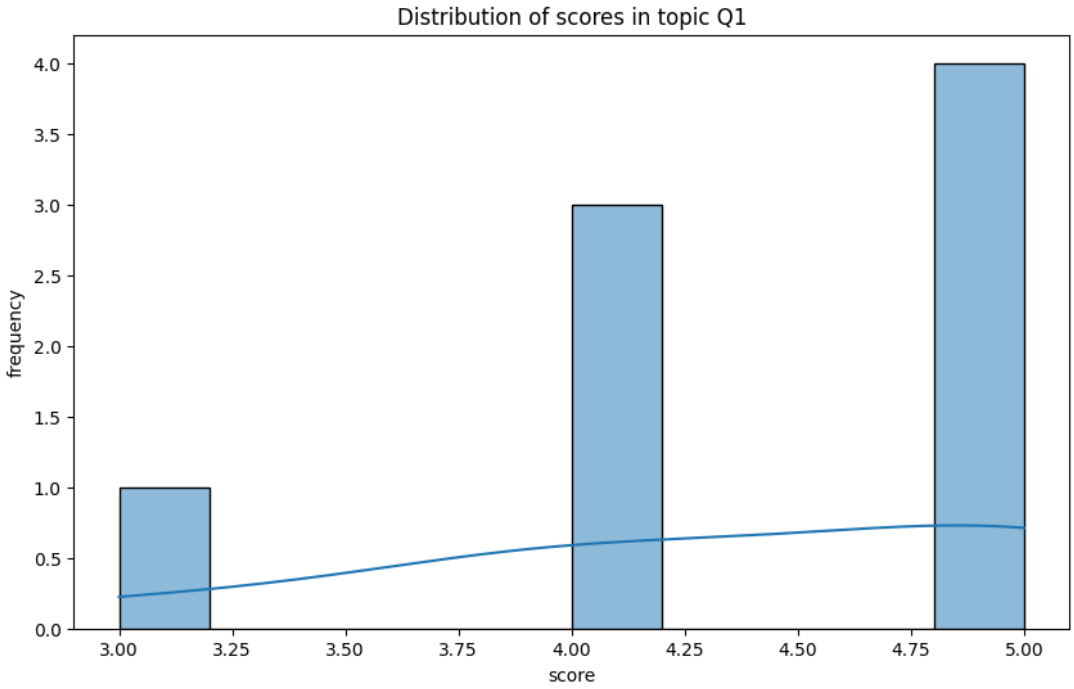
สมมติว่ามีไฟล์ข้อมูลที่แสดงคะแนนสอบของนักเรียนในหลายหัวข้อ และต้องการวิเคราะห์การกระจายตัวของข้อมูลในแต่ละหัวข้อ เพื่อทำความเข้าใจลักษณะของข้อมูล เช่น การหาค่าเฉลี่ย ค่ามัธยฐาน และส่วนเบี่ยงเบนมาตรฐาน

ตัวอย่างที่ 3

โค้ด	import pandas as pd
	import numpy as np
	import matplotlib.pyplot as plt
	import seaborn as sns
	
	# โหลดข้อมูลจากไฟล์ CSV
	file_path = 'questionnaire.csv'
	data = pd.read_csv(file_path)
	
	# ตรวจสอบค่ากลางและค่าการกระจาย
	mean_values = data.mean() # ค่าเฉลี่ย
	median_values = data.median() # ค่ามัธยฐาน
	std_values = data.std() # ส่วนเบี่ยงเบนมาตรฐาน
	
	# แสดงผลค่ากลางและค่าการกระจาย
	print("ค่าเฉลี่ยของแต่ละหัวข้อ:")
	print(mean_values)
	print("\nค่ามัธยฐานของแต่ละหัวข้อ:")
	print(median_values)
	print("\nส่วนเบี่ยงเบนมาตรฐานของแต่ละหัวข้อ:")
	print(std_values)

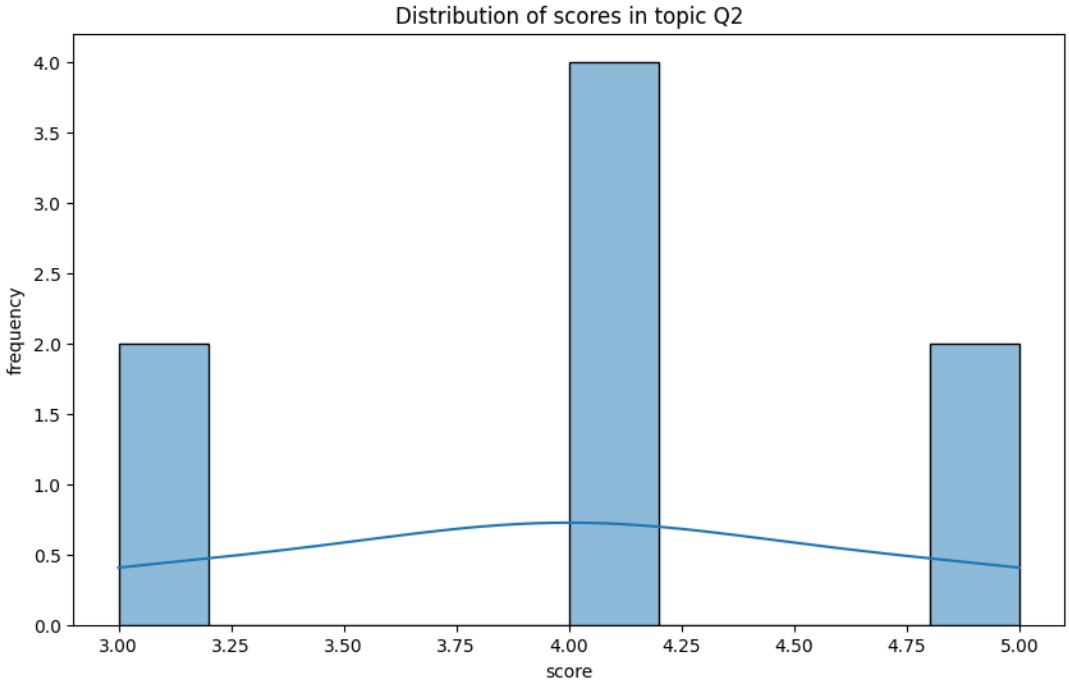
```
# การวิเคราะห์การกระจายตัวของข้อมูลด้วย Histogram Q1
plt.figure(figsize=(10, 6))
sns.histplot(data['Q1'], bins=10, kde=True)  # แสดง histogram ของ Q1
plt.title('Distribution of scores in topic Q1')
plt.xlabel('score')
plt.ylabel('frequency')
plt.show()
# การทดสอบการกระจายตัวแบบปกติด้วย Shapiro-Wilk Test
from scipy import stats
stat, p = stats.shapiro(data['Q1'])
print("\nผลการทดสอบ Shapiro-Wilk Test สำหรับ Q1:")
print(f"สถิติ: {stat}, p-value: {p}")

# การวิเคราะห์การกระจายตัวของข้อมูลด้วย Histogram Q2
plt.figure(figsize=(10, 6))
sns.histplot(data['Q2'], bins=10, kde=True)  # แสดง histogram ของ Q2
plt.title('Distribution of scores in topic Q2')
plt.xlabel('score')
plt.ylabel('frequency')
plt.show()
# การทดสอบการกระจายตัวแบบปกติด้วย Shapiro-Wilk Test
from scipy import stats
stat, p = stats.shapiro(data['Q2'])
print("\nผลการทดสอบ Shapiro-Wilk Test สำหรับ Q2:")
print(f"สถิติ: {stat}, p-value: {p}")
```

ผลลัพธ์



ผลการทดสอบ Shapiro-Wilk Test สำหรับ Q1:

สถิติ: 0.7975818786389666, p-value: 0.026967769188913724



ผลการทดสอบ Shapiro-Wilk Test สำหรับ Q2:

สถิติ: 0.8489109921172668, p-value: 0.09287781789278554

การกระจายตัวของข้อมูล และการวิเคราะห์เชิงสถิติอื่น ๆ ช่วยให้เราเข้าใจธรรมชาติของข้อมูลได้ดีขึ้น โดยการคำนวณค่ากลาง การวัดการกระจาย และการทดสอบการกระจายตัวช่วยให้เราสามารถสรุปและวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพ ข้อมูลที่มีการกระจายตัวต่างกันอาจต้องใช้เทคนิคหรือเครื่องมือในการวิเคราะห์ที่แตกต่างกัน