บทที่ ๑

บทนำ

๑.๑ คำนำ

ในการแก้ปัญหาทางสถิติทั่วไปนั้น ไม่ว่าจะเป็นการแก้ปัญหาด้านการประมาณค่า (estimation problems) หรือในการทดสอบสมมติฐาน (hypothesis testing problems) ของค่าพารามิเตอร์ (parameter) ตามที่ตำราทางสถิติทั้งขั้นต้นและขั้นกลางได้กล่าวถึงนั้น จะกล่าวถึงแต่การใช้ข้อมูลตัวอย่างที่เป็นค่าสัมบูรณ์ (absolute value)[1] เท่านั้น แต่ความจริงนั้นปัญหาทางสถิติเหล่านี้ยังสามารถจะแก้ได้ด้วยข้อมูลประเภทอื่น ๆ ได้อีก วิทยานิพนธ์ฉบับนี้มีวัตถุประสงค์ที่จะแสดงให้เห็นถึงความสำคัญทั้งทางด้านทฤษฎีและทางด้านการใช้ประโยชน์จากการใช้ข้อมูลสถิติอีกประเภทหนึ่งที่เน้นถึงความสำคัญของตำแหน่งของข้อมูล โดยมีการจัดเรียงลำดับค่าของข้อมูลตามขนาด (arranged in order of magnitude) ที่เรียกกันว่า "Order Statistics" สถิติที่จะกล่าวถึงนี้ไม่เพียงแต่จะสามารถใช้แก้ปัญหาทางสถิติได้ เช่น วิธีการทางสถิติทั่วไป (typical statistics) ที่ใช้ค่าสัมบูรณ์ แต่ยังมีคุณสมบัติอื่นที่ดีกว่าอีกหลายประการ อาทิเช่น ง่ายในการทำความเข้าใจ ใช้งานได้สะดวก (simplicity) มีประสิทธิภาพ (efficiency) ที่ดี เป็นต้น

---

[1] ค่าสัมบูรณ์ หมายถึงค่าทางพีชคณิต (algebraic value) กล่าวคือ เป็นค่าตัวเลขที่แท้จริง ซึ่งอาจเป็นได้ทั้งค่าบวกหรือค่าลบ แต่เราไม่คำนึงถึงเครื่องหมาย เช่น $-a$ และ $+a$ เราถือว่าเป็น a เท่านั้น

ด้วยความสำคัญของ order statistics ดังกล่าวมาแล้วนั้น วิทยานิพนธ์นี้มีเจตน์จำนงค์ที่จะเผยแพร่ให้เป็นที่รู้จักทั่วไป เพราะเท่าที่เป็นอยู่ในปัจจุบันนี้ ตำราทางสถิติภาคภาษาไทยทั้งในขั้นต้นและขั้นกลางยังมิได้มีการกล่าวถึงสถิติที่ได้จากข้อมูลประเภทนี้เลย หรือแม้แต่ในภาคภาษาอังกฤษซึ่งมีอยู่ไม่มากเล่มนักที่กล่าวถึงสถิติประเภทนี้ก็ยังเป็นที่ยากเย็นต้องการที่จะทำการศึกษาค้นคว้า จะมีอยู่มากก็เป็นบทความที่เขียนลงไว้ในวารสารภาษาอังกฤษด้านวิชาสถิติและคณิตศาสตร์หลายฉบับกระจัดกระจายอยู่ทั่วไป วิทยานิพนธ์นี้ได้ทำการรวบรวมทฤษฎีจากตำราและวารสารดังกล่าว โดยได้มีการเรียบเรียงเสียใหม่ให้ต่อเนื่องเพื่อความสะดวกในการศึกษาและทำความเข้าใจ พร้อมกันนี้ก็ได้นำตัวอย่างที่เป็นตัวเลขมาเสนอไว้ประกอบในการทำความเข้าใจด้วย ตัวอย่างเหล่านี้จะเป็นตัวเลขที่ได้จากการใช้ประโยชน์ของ order statistics จริง ๆ ในทางปฏิบัติ ด้วยเหตุนี้จึงหวังว่าวิทยานิพนธ์นี้จะพอเป็นประโยชน์แก่ผู้ที่ทำการศึกษาในสาขาวิชาสถิติและคณิตศาสตร์หรือสาขาอื่นใดที่มีส่วนสัมพันธ์ต่อทฤษฎีเหล่านี้ รวมทั้งผู้ที่สนใจในสถิติประเภทนี้เพื่ออาศัยเป็นบรรทัดฐานในการค้นคว้าศึกษาเพิ่มเติมต่อไป

๑.๒ Order Statistics คืออะไร

ถ้าเรามีตัวอย่างขนาด n หน่วยอยู่ชุดหนึ่ง สมมติว่าเป็น $X_1, X_2, X_3, \ldots, X_n$ โดยที่ตัวอย่างชุดนี้ถูกเลือกโดยการสุ่มมาจากประชากรที่มีฟังก์ชันของการกระจายของความน่าจะเป็น $f(x)$ จากนั้นถ้าเราจะจัดเรียงข้อมูลเสียใหม่ตามลำดับความสำคัญ เช่น เรียงจากค่าต่ำสุดไปยังค่าสูงสุดดังนี้ คือ $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ ด้วยคุณสมบัติที่ว่า

$$X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \cdots \leq X_{(n)},$$

แล้ว เราเรียกข้อมูลที่ได้รับการจัดเรียงตามขนาดของความสำคัญนี้ว่า

Order Statistics และค่า $X_{(n)}$ กับค่า $X_{(1)}$ เรียกว่า "ค่าสูงสุด" กับ "ค่าต่ำสุด" หรือโดยทั่ว ๆ ไปเราเรียก $X_{(i)}$ ว่าเป็นค่าของตัวที่มีตำแหน่งที่ i ของ order statistic ความแตกต่างระหว่างค่าสูงสุดและค่าต่ำสุด $X_{(n)} - X_{(1)} = R$ เรียกว่า "พิสัย" ของข้อมูลตัวอย่าง

ในกรณีที่ฟังก์ชันของการกระจายของความน่าจะเป็นเป็น continuous distribution แล้ว order statistics อาจเขียนได้ดังนี้

$$X_{(1)} < X_{(2)} < X_{(3)} < \cdots < X_{(n)} .$$

ตัวอย่างที่ ๑.๒.๑ จากหนังสือรายงานเศรษฐกิจประจำเดือน มกราคม พ.ศ.๒๕๑๑ ของธนาคารแห่งประเทศไทยได้รายงานไว้ว่า ราคาเฉลี่ยของข้าวชนิด ๑๐๐ % ต่อเบตริกตันในช่วงระยะเวลาจาก พ.ศ.๒๕๐๓ – ๒๕๑๐ เป็นตามตารางที่ ๑.๒.๑ ดังนี้

ตารางที่ ๑.๒.๑

แสดงการเรียงลำดับข้อมูลตามระยะเวลาที่เกิดเหตุการณ์

ปี	ราคา (บาท)
๒๕๐๓	๑,๖๔๑
๒๕๐๔	๑,๗๓๑
๒๕๐๕	๑,๘๕๖
๒๕๐๖	๑,๗๘๕
๒๕๐๗	๑,๖๘๐
๒๕๐๘	๑,๖๘๕
๒๕๐๙	๒,๑๘๙
๒๕๑๐	๒,๕๗๒

ถ้าเราจะได้จัดเรียงข้อมูลชุดนี้เสียใหม่เป็น "ราคาเฉลี่ยของข้าวชนิด ๑๐๐ % ต่อเมตริกตันในช่วงระยะเวลาจาก พ.ศ.๒๕๐๓ – ๒๕๑๐ จากราคาสูงสุดไปยังราคาต่ำสุด" จะได้ดังนี้

ตารางที่ ๑.๒.๒

แสดงการจัดเรียงข้อมูลตามลำดับขนาดของความสำคัญที่เราสนใจ

ปี	ราคา (บาท)
๒๕๑๐	๒,๕๓๒
๒๕๐๔	๒,๑๘๔
๒๕๐๕	๑,๙๙๒
๒๕๐๖	๑,๗๔๔
๒๕๐๔	๑,๗๓๑
๒๕๐๗	๑,๖๘๐
๒๕๐๘	๑,๖๔๔
๒๕๐๓	๑,๖๔๑

ข้อมูลที่ได้รับการจัดเรียงใหม่นี้เองที่เรียกว่า "Order Statistics"

ตัวอย่างของ "Order Statistics" ที่เราใช้กันอยู่อาจนำบาเสนอได้ดังนี้

<u>กึ่งพิสัย</u> (Mid-Range) หมายถึงมัชฌิมที่ได้จากค่าเฉลี่ยของคะแนนสูงสุดกับคะแนนต่ำสุด เช่น จากตัวอย่างที่ ๑.๒.๑ จะได้ค่ากึ่งพิสัยเป็น ๑/๒ (๒๕๓๒+๑๖๔๑) หรือเท่ากับ ๒๐๘๖.๕

<u>มัธยฐาน</u> (Median) หมายถึงมัชฌิมที่แสดงให้ทราบว่า ครึ่งหนึ่งของจำนวนคะแนนมีค่าสูงกว่ามัชฌิมนี้ และอีกครึ่งหนึ่งมีค่าต่ำกว่า หรืออาจกล่าวว่าเป็นค่าของคะแนนที่อยู่ในตำแหน่งที่ ๕๐ ของคะแนนทั้งหมดที่แบ่งได้เป็น ๑๐๐ ตำแหน่ง

จากตัวอย่างที่ ๑.๖.๑ มัธยฐานก็คือ $๑,๗๕๔ \pm ๑,๗๓๑$ หรือกล่าวว่ามัธยฐานของราคาเฉลี่ยของข้าวชนิด ๑๐๐ % ต่อเบอร์ริกต้นจากช่วงระยะเวลาจาก พ.ศ.๒๕๐๓ – ๒๕๑๐ เป็น ๑,๗๖๕ บาท

<u>ควอไทล์, เดไซล์, และเปอร์เซ็นไทล์</u> (Quartile, Decile, Percentile) เหล่านี้ก็ใช่มัธยฐาน แต่เป็นตัวที่แสดงให้ทราบว่ามีคะแนนที่ต่ำกว่าตัวมันเองกี่ตัวและที่สูงกว่าอีกกี่ตัว หรือกล่าวว่าเป็นการแสดงตำแหน่งของคะแนนเมื่อเทียบกับคะแนนทั้งหมด อธิบายได้ว่า

<u>ควอไทล์</u> คือคะแนนที่แสดงให้ทราบว่า จำนวนคะแนนอื่น ๆ ที่ต่ำกว่าคะแนนนี้เป็น ๑/๔ ของจำนวนคะแนนทั้งหมดเรียกว่าควอไทล์ที่หนึ่ง หรือจำนวนคะแนนอื่น ๆ ที่ต่ำกว่านี้เป็น ๓/๔ ของคะแนนทั้งหมดก็เรียกว่าควอไทล์ที่สาม

<u>เดไซล์</u> คือ คะแนนที่แสดงให้ทราบว่า จำนวนคะแนนอื่น ๆ ที่ต่ำกว่าคะแนนนี้เป็น ๑/๑๐, ๒/๑๐, ๓/๑๐,..., ๙/๑๐.

<u>เปอร์เซ็นไทล์</u> คือ คะแนนที่แสดงให้ทราบว่า จำนวนคะแนนอื่น ๆ ที่ต่ำกว่าคะแนนนี้เป็นกี่เปอร์เซ็นต์ ซึ่งก็คล้ายกับของควอไทล์และเดไซล์ เป็นแต่เพียงว่าเราแบ่งคะแนนทั้งหมดให้อยู่ในตำแหน่ง ๑๐๐ ตำแหน่งด้วยกันทั้งหมด

<u>พิสัย</u> (Range) เป็นการวัดการกระจายของข้อมูลวิธีหนึ่งโดยดูจากความแตกต่างของคะแนนที่อยู่ในตำแหน่งสูงสุดกับตำแหน่งต่ำสุด เช่น จากตัวอย่างที่ ๑.๖.๑ พิสัยของราคาเฉลี่ยของข้าวชนิด ๑๐๐ % ต่อเบอร์ริกต้นจากช่วงระยะเวลาจาก พ.ศ.๒๕๐๓ – ๒๕๑๐ เป็น ๒๕๓๒ – ๑๖๔๑ = ๘๔๑ บาท

ในการวัดการกระจายของข้อมูลโดยพิสัยนี้ เราวินิจฉัยว่า ข้อมูลชุดที่มีพิสัยมากกว่าย่อมมีการกระจายของข้อมูลมากกว่า ฉะนั้นจะเห็นว่า พิสัยเป็นการวัดการกระจายที่คร่าว ๆ เพราะใช้เพียงคะแนนตำแหน่งสูงสุดกับ

ต่ำสุดเท่านั้น โดยต้องการความรวดเร็วเป็นสำคัญ

<u>ส่วนเบี่ยงเบนควอไดล์</u> (Quartile Deviation) เป็นการวัดการกระจายของข้อมูลวิธีหนึ่งโดยดูจากส่วนต่างของควอไดล์ที่สามกับที่หนึ่ง $(Q_3 - Q_1)$ ถ้าข้อมูลใดมีส่วนเบี่ยงเบนควอไดล์มากก็ถือว่าข้อมูลมีการกระจายมาก ถ้าส่วนเบี่ยงเบนควอไดล์น้อยก็ถือว่าข้อมูลมีการกระจายน้อย

๑.๓ ความหมายของคำที่ใช้

เพื่อความสะดวกในการทำความเข้าใจกับวิทยานิพนธ์ฉบับนี้จึงได้อธิบายความหมายของคำบางคำที่จะต้องใช้อยู่เสมอ ๆ ดังต่อไปนี้

<u>ข้อมูลทางสถิติ</u> (Statistical data) อาจจำแนกได้ตามวิธีการเก็บรวบรวมเป็นสามประเภทดังนี้คือ

ก. <u>Sample data</u> เป็นข้อมูลที่ได้จากการสุ่มตัวอย่างจากประชากรที่เราสนใจ ในการรวบรวมข้อมูลเราสนใจในลักษณะของตัวอย่างเท่านั้น เช่น การสำรวจรายได้ การสำรวจการใช้จ่าย การสำรวจทัศนคติ เป็นต้น แต่ในการนี้เราไม่สนใจถึงสาเหตุหรือปัจจัยที่ทำให้เกิดลักษณะดังกล่าว กล่าวคือเราไม่สนใจว่าในการสำรวจรายได้นั้นอะไรเป็นสาเหตุทำให้คนส่วนหนึ่งมีรายได้สูงมากแต่อีกส่วนหนึ่งมีรายได้ต่ำ

ข. <u>Experimental data</u> เป็นข้อมูลที่ได้จากการทดลอง คือการแบ่งตัวอย่างที่ถูกเลือกมานั้นออกเป็นกลุ่ม ๆ ให้อยู่ภายใต้สิ่งแวดล้อมต่าง ๆ กัน เช่น การใช้ปุ๋ยต่างชนิดกันในพืชชนิดเดียวกันในอุณหภูมิที่แตกต่างกัน ข้อมูลประเภทนี้ใช้วัดถึงสาเหตุและปัจจัยที่ทำให้เกิดลักษณะประจำแก่ประชากร

๓. Observational data เป็นข้อมูลที่ได้จากการทดลองเช่นเกี่ยวกับ Experimental data แต่ต่างกันที่ว่าใน Experimental data นั้นเราสามารถที่จะกำหนดขนาดของตัวอย่างได้ เพราะสามารถที่จะหาตัวอย่างได้ตามขนาดที่ต้องการ แต่ในกรณีของ Observational data เราไม่สามารถที่จะมีหรือกำหนดขนาดของตัวอย่างได้เสมอไป ที่เป็นเช่นนั้นก็เพราะลักษณะพิเศษของประชากรนั้น ดังนั้นเราจึงใช้ตัวอย่างจากสิ่งที่เรามีอยู่ เช่น การทดสอบว่า การสูบบุหรี่เป็นสาเหตุของการเป็นมะเร็ง เราก็จะต้องศึกษาจากผู้ที่ติดบุหรี่อยู่แล้ว เราไม่สามารถที่จะเกณฑ์ให้คนที่ไม่สูบบุหรี่ให้มาสูบเพื่อเป็นเครื่องทดลองของเราได้ อีกทั้งเราก็ไม่สามารถที่จะกำหนดขนาดของตัวอย่างทดลองมากเกินกว่าจำนวนผู้สูบบุหรี่ในขณะนั้นไปได้

พาราบิเตอร์ (parameter) คือ "ลักษณะ" ของประชากร หรือ "ลักษณะ" ของการกระจายของประชากร เช่น อัตราส่วนประชากรชายต่อหญิงของจังหวัดชลบุรีเป็น ๑.๑ จำนวนครัวเรือนการเกษตรของประเทศไทยเป็น ๗๓.๕ % รายได้เฉลี่ยต่อบุคคลของประเทศไทยเป็น ๒,๔๐๐ บาทต่อปี

Statistics คือ "ลักษณะ" ของตัวอย่าง หรือ "ลักษณะ" ของการกระจายของตัวอย่าง โดยปกติจะเป็นฟังก์ชันของ observable random variables หน้าที่อันสำคัญยิ่งของ Statistics ในทางสถิติก็คือใช้เป็นตัวประมาณพาราบิเตอร์ เช่น $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ เป็น Statistic ที่ใช้ประมาณค่าของ $\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$ ซึ่งเป็นพาราบิเตอร์ที่เรายังไม่ทราบค่าที่แท้จริง

นอกจากคำเหล่านี้แล้ว คำอื่น ๆ ที่มีความหมายในทางวิชาการเมื่อถูกนำมาใช้ในวิทยานิพนธ์ฉบับนี้ก็จะได้มีคำอธิบายเพิ่มเติมเฉพาะคำไปในบทนั้น ๆ