บทที่ 4
สถิติพรรณนา (Descriptive Statistics)

สถิติพรรณนา เป็นสถิติที่มีขอบข่ายถึงระเบียบวิธีบรรยายขั้นกระบวนการต่าง ๆ ที่ใช้ในการเก็บรวบรวมข้อมูลสถิติ ระเบียบวิธีที่ใช้และคิดคำนวณค่าต่าง ๆ ของข้อมูลสถิติเหล่านั้น เพื่อที่จะสรุปและตีความหมายให้ถูกต้องตามความเป็นจริง และนำเอาความรู้เหล่านั้นมาใช้ให้เป็นประโยชน์ในการตัดสินใจ และในการดำเนินการต่าง ๆ ในบทนี้จะกล่าวถึงกำเนิดของสถิติและการอธิบายข้อมูล ดังนี้

4.1 กำเนิดของสถิติ

สถิติกำเนิดมาเป็นเวลานานหลายพันปีตั้งแต่มนุษย์รู้จักนับการรวบรวมและบันทึกข้อมูลก็เป็นที่รู้จักกันมานานหลายศตวรรษ ซึ่งในสมัยนั้นเป็นสมัยที่จักรวรรดิอียิปต์ โรมัน และกรีกรุ่งเรือง ได้มีการบันทึกข้อมูลที่เกี่ยวข้องและเป็นประโยชน์ต่อการบริหารงานของรัฐ เช่น ข้อมูลเกี่ยวกับการเกษตร ข้อมูลเกี่ยวกับจำนวนประชากร เพื่อนำข้อมูลเหล่านี้มาใช้ประโยชน์ทางด้านทหารและการเก็บภาษีอากร ซึ่งข้อมูลในสมัยนั้นเป็นการบันทึกข้อมูลอย่างง่าย ๆ ไม่สลับซับซ้อนมากนัก

คำว่าสถิติ ตรงกับคำในภาษาอังกฤษว่า Statistics ซึ่งตรงกับคำว่า Statistik ในภาษาเยอรมัน เป็นคำที่มีรากศัพท์เดียวกับคำว่า State โดยนักปราชญ์ชาวเยอรมัน ชื่อ Gottfried Achenwall (1719-1772) เป็นผู้บัญญัติขึ้นเมื่อปี ค.ศ.1749 และให้ความหมายของคำว่า State หมายถึง ข้อมูล หรือข้อเท็จจริงใด ๆ ที่เกี่ยวข้องและจะเป็นประโยชน์ต่อการบริหารของรัฐ และขณะเดียวกันก็หมายถึงศิลปศาสตร์และปรัชญาว่าด้วยการนำข้อเท็จจริงเหล่านี้มาแยกวิเคราะห์เปรียบเทียบอัตราส่วน หรือประกอบการคิดคำนวณ เพื่อนำผลลัพธ์ที่ได้มาใช้เป็นประโยชน์ต่อการบริหารงานของรัฐ

ในปี ค.ศ.1500-1800 สถิติเริ่มพัฒนาขึ้นโดย Girolamo Cardano (1501-1576) ได้ศึกษาความน่าจะเป็นคู่ไปกับการพนันและเกมโศลก ซึ่งในสมัยนั้นพวกขุนนางต่าง ๆ ได้ให้นักคณิต-ศาสตร์และนักตรรกวิทยาคำนวณโอกาสที่จะเกิดขึ้นของเกมเกี่ยวกับลูกเต๋า ไพ่ และอื่น ๆ นัก-คณิตศาสตร์ที่อุทิศตนให้กับงานทฤษฎีน่าจะเป็นเช่น De Mere, Fermat และ Blaise Pascal และในสมัยเดียวกันนี้ P.S. Laplace, A.Demoivre และ Karl F. Gauss ได้เริ่มสนใจการวิเคราะห์ด้านความคลาดเคลื่อนและการแจกแจงต่าง ๆ

ในระหว่างปี ค.ศ.1800 ถึงต้นปี ค.ศ.1900 มีการศึกษาวิชาสถิติกันอย่างกว้างขวาง และมีนักสถิติเพิ่มขึ้นมากมาย เช่น Lambert Quetelet ชาวเบลเยียมซึ่งได้ชื่อว่าเป็นบิดาแห่งสถิติได้จัดการประชุมทางสถิติขึ้นเป็นครั้งแรกและยังเป็นคนแรกที่สนใจค่าเฉลี่ย ในปี ค.ศ.1893 นักทดลองชาวอังกฤษ 2 ท่านคือ Sir Francis Galton และ Karl Pearson ได้นำเอาทฤษฎีความน่าจะเป็นและทฤษฎีความคลาดเคลื่อนมาประกอบการวิเคราะห์ข้อมูลจากผลการทดลองของเขาทางด้านพันธุ์กรรม ทางเกษตร และทางชีววิทยา จากนั้นตั้งแต่ปี ค.ศ.1895 ถึง ค.ศ. 1930 เป็นยุคที่มีวิวัฒนาการทางสถิติแผนใหม่อย่างกว้างขวางมากที่สุด เช่น การอธิบายความคลาดเคลื่อนของข้อมูล ความคลาดเคลื่อนในการทดลองในเทอมของความน่าจะเป็น หลักในการทดสอบสมมติฐาน แนวคิดเกี่ยวกับการกระจายของตัวอย่าง การใช้ค่าสหสัมพันธ์เชิงเส้นและสหสัมพันธ์เชิงพหุคูณในการอธิบายความสัมพันธ์ระหว่างข้อมูล 2 รายการ การวิเคราะห์ความแปรปรวน และการวางแผนแบบทดลอง เป็นต้น และจนกระทั่งปัจจุบัน วิวัฒนาการด้านสถิติก็ยังก้าวหน้าโดยไม่หยุดยั้ง

สถิติในปัจจุบันเป็นผลงานของ Sir Ronald Aylmer Fisher ซึ่งเป็นผู้วางรากฐานให้แก่สถิติสมัยใหม่ และท่านได้ชื่อว่าเป็นบิดาแห่งสถิตินวสมัย (Father of modern statistics) นักสถิติคนอื่น ๆ ที่ช่วยพัฒนาสถิตินวสมัย ได้แก่ William Seally Gossett, Abraham Wald, Jerzey Neyman, Egon S. Pearson และ George Waddel Snedecor

4.1.1 ความหมายของวิชาสถิติ

เป็นการยุ่งยากที่จะอธิบายความหมายของคำว่าสถิติ เพื่อที่จะให้ผู้ที่เกี่ยวข้องกับ

สถิติเพียงบางส่วนเข้าใจอย่างลึกซึ้งและครอบคลุมได้อย่างครบถ้วนและชัดเจน แต่อย่างไรก็ดี นักศึกษาคงจะเคยได้ยินได้เห็นคำว่า "สถิติ" ที่ปรากฏอยู่ตามหนังสือ บทความทางวิทยุโทรทัศน์ ฯลฯ เช่น สถิตินักศึกษาที่สมัครเข้าเรียนคณะมนุษยศาสตร์ สถิติปริมาณน้ำฝน สถิติการแข่งขันว่ายน้ำ สถิตินักศึกษามหาวิทยาลัยรามคำแหงที่จบแล้วมีงานทำ สถิติการเกษตร ภาควิชาสถิติ-คณะวิทยาศาสตร์ สำนักงานสถิติแห่งชาติ สถิติประชากร เป็นต้น ซึ่งความหมายของคำว่าสถิตินั้น มีความหมายอย่างกว้าง ๆ 2 ประการ คือ

1. สถิติ หมายถึงศาสตร์ซึ่งถือว่าเป็นทั้งวิทยาศาสตร์และศิลปศาสตร์ (Sciences and Arts) ที่ว่าด้วยวิธีการเก็บรวบรวมข้อมูลที่เป็นตัวเลข ซึ่งแสดงถึงข้อเท็จจริงต่าง ๆ และยังรวมถึงการนำเสนอข้อมูล การตีความหมาย การวิเคราะห์ การคิดคำนวณ และการสรุปข้อมูล นอกจากนี้ ยังรวมถึงการใช้ข้อมูลที่มีอยู่ในอดีตและปัจจุบันไปทำนายเหตุการณ์ในอนาคต หรือใช้ในการประกอบการตัดสินใจภายใต้ความไม่แน่นอนของเหตุการณ์บางอย่างในอนาคต

2. สถิติ หมายถึงบรรดาตัวเลขซึ่งได้จากการรวบรวมข้อเท็จจริงเกี่ยวกับเรื่องต่าง ๆ ที่เราสนใจ เช่น ปริมาณน้ำฝนที่ตกลงมา การเกิด การสมรส การย้ายที่อยู่ การตาย การศึกษา การส่งออกสินค้าประเภทสิ่งทอ เป็นต้น ซึ่งตัวเลขที่ได้มานี้มักจะอยู่ในลักษณะของยอดรวม ซึ่งประมวลมาได้จากข้อมูลเบื้องต้น หรืออาจจะเป็นตัวเลขที่ได้มาจากผลของการวิเคราะห์เปรียบเทียบ หรือจากการคิดคำนวณ ดังนั้น ตัวเลขหรือข้อมูลซึ่งเป็นข้อเท็จจริงเกี่ยวกับเรื่องต่าง ๆ ที่จะถือว่าเป็นสถิตินั้นจะต้องเป็นข้อมูลส่วนรวม ไม่ใช่ข้อมูลของคนใดคนหนึ่งหรือหน่วยใดหน่วยหนึ่งโดยเฉพาะ เพื่อให้เห็นความแตกต่างของความหมายของคำว่าสถิติทั้ง 2 ความหมาย จะเรียกสถิติในความหมายที่เป็นตัวเลขว่า ข้อมูลสถิติ (Statistical Data) และสถิติในความหมายที่เป็นศาสตร์ว่า สถิติศาสตร์ (Statistical Science)

4.1.2 ขอบข่ายของสถิติ

ขอบข่ายของสถิติในความหมายที่เป็นตัวเลขมีเนื้อหาครอบคลุมในทุกแขนงของวิชาการ และในกิจกรรมต่าง ๆ ของการบริหารงานและการดำรงชีวิตประจำวัน ซึ่งได้แก่สถิติสาขาต่าง ๆ 11 หมวด ดังนี้

1. สถิติประชากรและแรงงาน เช่น สำมะโนประชากร
2. สถิติการเกษตร เช่น สำมะโนการเกษตร สถิติจำนวนปศุสัตว์ สถิติการประมง
3. สถิติการศึกษา และสาธารณสุข
4. สถิติอุตสาหกรรม เช่น สถิติเกี่ยวกับโรงงานอุตสาหกรรม สถิติปริมาณการใช้พลังงานไฟฟ้า
5. สถิติการค้าส่ง และค้าปลีก และบริการ เช่น สถิติการจดทะเบียนการค้า
6. สถิติการคมนาคมและขนส่ง เช่น สถิติปริมาณการขนส่งทางอากาศ
7. สถิติการค้าระหว่างประเทศ เช่น สถิติการส่งสินค้าเข้า-ออก
8. สถิติการเงิน การธนาคาร การประกันภัยและการสหกรณ์ เช่น สถิติการประกันวินาศภัย สถิติสหกรณ์ในประเทศ
9. สถิติราคาสินค้า เช่น สถิติดัชนีราคา
10. สถิติรายได้รายจ่ายของครัวเรือน เช่น สถิติรายได้และรายจ่ายของครัวเรือน
11. สถิติบัญชีประชาชาติ เช่น สถิติรายได้ประชาชาติ

ส่วนสถิติในความหมายที่เป็นศาสตร์จะว่าด้วย

1. สถิติวิเคราะห์เชิงพรรณนา (Descriptive Statistical Analysis) เป็นสถิติที่ว่าด้วยระเบียบวิธีการวิเคราะห์ข้อมูลทางสถิติในรูปของการบรรยาย การนำเสนอข้อมูลในรูปบทความ ตารางสถิติ แผนภูมิ กราฟ ตลอดจนถึงรูปภาพต่าง ๆ การคิดคำนวณ และการวิเคราะห์ใช้วิธีการที่ไม่สลับซับซ้อนนัก เช่น การบวก ลบ คูณ หาร การหาเปอร์เซ็นต์ การหาอัตราส่วน ไม่ต้องใช้ทฤษฎีความน่าจะเป็นและคณิตศาสตร์ชั้นสูง สถิติวิเคราะห์เชิงพรรณนานี้มักจะทำเฉพาะกลุ่มที่ทำการศึกษาเท่านั้น ไม่สามารถนำไปกล่าวถึงกลุ่มที่กว้างขวางขึ้น

2. สถิติปฏิบัติ (Practical Statistics) เป็นสถิติที่ว่าด้วยการปฏิบัติเพื่อให้ได้มาซึ่งข้อมูลทางสถิติ หรือเรียกว่าเป็นงานสนาม ซึ่งประกอบด้วยการวางแผนการเก็บรวบรวมข้อมูล การออกแบบสอบถาม การทดสอบแบบสอบถาม การสอบถาม การควบคุมงานสนาม การบรรณาธิกรณ์ข้อมูล และการประมวลผล

3. สถิติวิเคราะห์เชิงอนุมาน (Inferential Statistical Analysis) เป็นสถิติที่ว่าด้วยการศึกษาข้อมูลจากกลุ่มตัวอย่าง แต่สามารถนำผลสรุปไปกล่าวถึงสิ่งที่ต้องการศึกษาในสภาวการณ์โดยทั่วไป โดยอาศัยทฤษฎีความน่าจะเป็นและคณิตศาสตร์ชั้นสูงมาสนับสนุนการวิเคราะห์ สถิติอนุมานนี้นับว่าเป็นสถิตินวสมัย ซึ่งมีประโยชน์แพร่หลายในสาขาวิชาต่างๆ ที่ทำการศึกษาข้อมูลเพียงบางส่วนเพื่อประโยชน์ต่างๆ ตัวอย่างของสถิติอนุมาน เช่น ทฤษฎีความน่าจะเป็น การประมาณค่า การทดสอบสมมติฐาน การวิเคราะห์ความแปรปรวน เป็นต้น

4.1.3 ประโยชน์ของสถิติ สถิติที่เป็นข้อมูลหรือตัวเลขมีจำนวนผู้ที่นำสถิติไปใช้มากมายทั้งในและนอกประเทศ ซึ่งสามารถแบ่งกลุ่มผู้นำสถิติไปใช้เป็นกลุ่มใหญ่ๆ ได้ 5 กลุ่มดังนี้ คือ

1. กลุ่มผู้ใช้ในวงการรัฐบาล ซึ่งได้แก่ กระทรวง ทบวง กรม กอง ต่างๆ ซึ่งรับผิดชอบทางด้านการวางแผนงานและการควบคุมดูแลบริหารงานด้านต่างๆ ตลอดจนบริการสาธารณะต่างๆ ของรัฐบาล กลุ่มผู้ใช้กลุ่มนี้มักจะใช้สถิติเป็นเครื่องมือช่วยในการตัดสินใจเลือกการดำเนินงาน รวมทั้งการกำหนดนโยบายต่างๆ

2. กลุ่มผู้ใช้ในวงการเอกชน ซึ่งได้แก่ ธนาคารพาณิชย์ต่างๆ นักลงทุน การโฆษณา และบริษัททำวิจัยธุรกิจต่างๆ

3. กลุ่มผู้ใช้ระดับครัวเรือนและบุคคล

4. สถาบันการศึกษาและวิจัย

5. องค์การระหว่างประเทศ

ดังนั้น ประโยชน์ของสถิติก็คือเป็นเครื่องมือสำหรับการตัดสินใจที่สำคัญอย่างหนึ่ง และมีบทบาทสำคัญในการช่วยแก้ปัญหาเกี่ยวกับการวิจัยและพัฒนาการรวมทั้งเป็นเครื่องมือที่จะใช้ในหน่วยงานทางด้านต่างๆ เกือบทุกด้าน เช่น ด้านเศรษฐกิจ สังคม แพทย์ ธุรกิจ การศึกษา ฯลฯ สถิติจัดว่ามีความสำคัญและมีประโยชน์ต่อวงการทุกระดับ โดยเฉพาะในด้านการวางแผนและการตัดสินใจในเรื่องที่สลับซับซ้อน

สำหรับในประเทศไทยนั้นการนำสถิติไปใช้ยังอยู่ในวงจำกัดทั้งในส่วนราชการและเอกชน ทั้งๆ ที่สถิติก็เป็นที่ยอมรับกันโดยทั่วไปแล้วก็ตามว่าเป็นเครื่องมือที่ช่วยในการตัดสินใจ

ที่สำคัญที่สุดอันหนึ่ง แต่ก็ยังคงใช้การตัดสินใจในเรื่องต่าง ๆ โดยไม่อาศัยข้อมูลที่เกี่ยวข้องกับสถิติ สาเหตุที่เป็นเช่นนี้อาจเนื่องมาจาก

1. ผู้บริหารยังไม่ค่อยเห็นความจำเป็นและความสำคัญที่จะต้องใช้ข้อมูลสถิติเพื่อนำมาประกอบกับการวางแผน หรือตัดสินใจในเรื่องต่าง ๆ มากนัก ซึ่งอาจจะเนื่องมาจากการขาดความรู้ในการใช้ประโยชน์จากข้อมูลสถิติ หรืออาจจะเนื่องมาจากขาดความสะดวกในการใช้ประโยชน์จากสถิติ

2. ข้อมูลสถิติที่ผลิตออกมาจากหน่วยงานทางด้านสถิติต่าง ๆ ยังมีคุณภาพไม่ดีพอ โดยเฉพาะอย่างยิ่งยังขาดคุณลักษณะที่สำคัญ ๆ ซึ่งควรจะมี เช่น ความถูกต้องแม่นยำ ความแนบนัย ความต่อเนื่อง ความสมบูรณ์ ความทันสมัย และบางครั้งก็ไม่ตรงกับความต้องการของผู้ที่จะนำไปใช้

3. ขาดการโฆษณาหรือการส่งเสริมการใช้ข้อมูลสถิติ

4.2 การอธิบายข้อมูล

ข้อมูล (Data) หมายถึงข้อเท็จจริงต่าง ๆ ซึ่งอาจจะเป็นตัวเลขหรือไม่เป็นตัวเลขก็ได้ ซึ่งข้อเท็จจริงเหล่านี้ก็คือ ค่าสังเกตที่ได้ในเรื่องใดเรื่องหนึ่งที่สนใจ เช่น จำนวนสมาชิกที่มีอยู่ในแต่ละครัวเรือน ได้แก่ 9, 3, 8, 10, 12, 6, 5, 2, 4

ข้อมูลทางสถิติ (Statistical Data) หมายถึงข้อเท็จจริงที่รวบรวมได้ จากเรื่องที่สนใจจะศึกษา เช่น คะแนนสอบของนักศึกษาที่เรียนวิชา SC 101 ภูมิลำเนาของนักศึกษาคณะมนุษยศาสตร์ เพศ ระดับการศึกษา อาชีพ อายุ ฯลฯ ซึ่งเราสามารถแบ่งข้อมูลทางสถิติตามแหล่งที่มาได้ 2 ประเภท คือ

ก. ข้อมูลปฐมภูมิ (Primary data) เป็นข้อมูลที่เก็บรวบรวมมาจากแหล่งที่ลงมือเก็บรวบรวมข้อมูลเป็นครั้งแรกด้วยตนเอง หรือจากบุคคลใดบุคคลหนึ่ง แล้วนำข้อมูลเหล่านั้นมาใช้เป็นหลักฐานอ้างอิงประกอบการศึกษาหรือวิจัยต่อไป ซึ่งข้อมูลเหล่านี้มักจะได้มาจากการทดลอง การสังเกต การส่งแบบสอบถาม การสัมภาษณ์

ข. ข้อมูลทุติยภูมิ (Secondary data) เป็นข้อมูลที่ได้มาจากแหล่งของข้อมูลที่มีการรวบรวมไว้แล้ว ซึ่งผู้ใช้ข้อมูลประเภทนี้จะสะดวกประหยัดเวลาและค่าใช้จ่าย เพราะเพียงแต่ทราบว่าข้อมูลที่ต้องการนั้นอยู่ที่หน่วยงานใดหรือที่ไหน ก็ไปติดต่อขอข้อมูลมาใช้จากหน่วยงานนั้น ซึ่งการนำเอาข้อมูลประเภทนี้มาใช้จะต้องคำนึงถึงความถูกต้องของข้อมูลที่ได้นั้นมีความน่าเชื่อถือมากน้อยแค่ไหน ตรงกับความต้องการของงานวิจัยหรือไม่ต้องมีการปรับแก้ส่วนใดของข้อมูลบ้าง และขอบเขตของข้อมูลที่ใช้เป็นอย่างไร เป็นต้น

การเก็บรวบรวมข้อมูลทางสถิตินั้นมีอยู่หลายวิธีด้วยกัน แต่ก่อนอื่นควรจะทราบความหมายของคำว่า ประชากร (Population) พารามิเตอร์ (Parameter) และตัวอย่าง (Sample) ก่อนดังนี้

ประชากร (Population) ประกอบด้วยหน่วย (unit) ต่าง ๆ ทั้งหมดที่สนใจ บางทีเรียกว่าคุ้มรวม (Coverage) ซึ่งคำว่าหน่วยในที่นี้หมายถึงสิ่งหนึ่ง หรือกลุ่มของสิ่งของต่าง ๆ ที่เราอาจจะวัดหรือสังเกตข้อเท็จจริงทางสถิติได้ และคำว่าหน่วยนี้บางทีเรียกว่า หน่วยแจงนับ (Enumeration Unit) ประชากรแบ่งออกเป็น 2 ประเภทด้วยกัน คือ

ก. ประชากรชนิดรู้จบหรือจำกัด (Finite population) เป็นประชากรที่ประกอบด้วยหน่วยแจงนับที่มีจำนวนรู้จบหรือจำกัด

ข. ประชากรชนิดไม่รู้จบ หรืออนันต์ (Infinite population) เป็นประชากรที่ประกอบด้วยหน่วยแจงนับที่มีจำนวนไม่รู้จบหรืออนันต์

พารามิเตอร์ (Parameter) เป็นค่าคงที่ซึ่งแสดงคุณลักษณะของประชากร พารามิเตอร์จะเป็นฟังก์ชันของค่าของหน่วยต่าง ๆ ทั้งหมดในประชากร เช่น ค่าเฉลี่ยประชากร (μ) สัดส่วนประชากร (π) ความแปรปรวนประชากร (σ^2) เป็นต้น

ตัวอย่าง (Sample) หมายถึงกลุ่มของบรรดาหน่วยที่เลือกได้จากประชากรตามวิธีการสำรวจด้วยตัวอย่าง (Sample Survey) เพื่อที่จะได้ตัวอย่างที่เป็นตัวแทนที่ดีของประชากร

การเก็บรวบรวมข้อมูลวิธีที่นิยมใช้กันมีอยู่ 3 วิธีดังนี้

1. การสำมะโน (Census or Complete enumeration)

หมายถึงการสำรวจหรือการแจงนับหน่วยทุกหน่วยที่อยู่ในประชากรที่สนใจ เช่น การ

ทำสำมะโนประชากร สำมะโนการเกษตร เป็นต้น ซึ่งโดยทั่วไปแล้วการทำสำมะโนต้องการเงิน เวลา และกำลังงานมาก ดังนั้น ถ้าไม่จำเป็นจริง ๆ แล้วก็จะไม่ค่อยทำกันบ่อยนัก แต่ถ้าต้องการจะได้ข้อเท็จจริงจากทุก ๆ หน่วยแล้วก็จำเป็นต้องใช้วิธีการสำมะโน สำหรับสำมะโนประชากรในประเทศไทยนั้นจะทำกันทุก ๆ 5 ปี หรือ 10 ปี การทำสำมะโนประชากรนั้นมีทั้งข้อดีและข้อเสีย ดังนี้

ข้อดี

1. ยอดข้อมูลสถิติที่รวบรวมได้ในทุกเรื่องสามารถแสดงออกในเขตบริหารหรือเขตภูมิศาสตร์ที่เล็กที่สุดได้ ทั้งนี้ เพราะมีข้อมูลจากทุกหน่วยแจงนับในคุ้มรวม เช่น ในสำมะโนเกษตรสามารถแสดงตารางข้อมูลสถิติในระดับ จังหวัด อำเภอ ตำบล หรือหมู่บ้าน

2. ข้อมูลสถิติที่ได้นั้นเป็นข้อมูลหลักซึ่งจะนำไปใช้ในการวางแผนเก็บข้อมูลอื่น ๆ ได้อีกด้วย

ข้อเสีย

1. ใช้ทรัพยากร เช่น กำลังคน เวลา และงบประมาณมาก จึงไม่สามารถทำสำมะโนได้ทุกปี ดังนั้น จึงทำให้ข้อมูลมีไม่ครบทุกปี และอาจไม่พอเพียงกับการที่ต้องใช้

2. เสียเวลามาก เนื่องจากมีหน่วยแจงนับมาก จึงทำให้ปริมาณงานมาก ไม่สามารถทำงานให้เสร็จได้ทันท่วงที

3. คุณภาพของข้อมูลที่รวบรวมได้ยังเป็นที่น่าสงสัย เพราะต้องใช้เจ้าหน้าที่ร่วมทำงานมาก ซึ่งยากแก่การควบคุม การบริหารงาน และควบคุมคุณภาพในการแจงนับ

4. ผู้ให้คำตอบในการแจงนับบางรายไม่สามารถให้ข้อมูลที่ถูกต้องแก่พนักงานแจงนับได้ จึงทำให้ยอดข้อมูลสถิติผิดไปจากความเป็นจริง

2. การสำรวจด้วยตัวอย่าง (Sample Survey) เป็นการรวบรวมข้อเท็จจริงจากหน่วยตัวอย่าง (Sampling units) ที่เลือกมาเป็นตัวแทนของประชากรที่สนใจ แล้วใช้ข้อเท็จจริงที่ได้ไปกะประมาณค่าพารามิเตอร์ที่สนใจ เหตุผลที่เราต้องใช้การสำรวจด้วยตัวอย่าง คือ ประหยัดเวลา ประหยัดเงิน รวบรวมข้อมูลได้กว้างขวางและความถูกต้องของข้อมูลมีมากกว่า แต่อย่างไรก็ตาม การสำรวจด้วยตัวอย่างก็มีข้อเสียเหมือนกันคือ

1. ไม่สามารถประมาณข้อมูลในระดับย่อย ๆ หรือท้องที่เล็ก ๆ ได้ และการวิเคราะห์ข้อมูลในรูปตารางก็ไม่สามารถจำแนกข้อมูลในรายละเอียดได้มากนัก เพราะได้ข้อมูลมาไม่เพียงพอที่จะจำแนก หรือถ้าทำได้ก็อาจจะมีความคลาดเคลื่อนสูงมาก

2. สถิติที่ประมวลได้จากตัวอย่างเป็นเพียงค่าประมาณ (Estimate) ของคุณลักษณะที่เราสนใจ ไม่ใช่ยอดสถิติจริง

วิธีการที่จะเลือกตัวอย่างให้เป็นตัวแทนที่ดีของประชากรที่สนใจนั้นมีหลายวิธีด้วยกันดังนี้ คือ

1. การสุ่มตัวอย่างแบบง่าย (Simple Random Sampling หรือ SRS) เป็นการสุ่มตัวอย่างที่ทุก ๆ หน่วยในประชากรมีโอกาสที่จะถูกเลือกเป็นตัวอย่างเท่า ๆ กัน ซึ่งวิธีการเลือกตัวอย่างดังกล่าวกระทำได้ 2 วิธีด้วยกัน คือ

ก. ใช้วิธีจับฉลาก มักจะใช้ในกรณีที่ประชากรมีจำนวนไม่มากนัก การจับฉลากทำได้ 2 วิธี คือ เลือกโดยไม่มีการแทนที่ (Sampling without replacement) และเลือกโดยมีการแทนที่ (Sampling with replacement)

ข. ใช้ตารางเลขสุ่ม (Random number tables) ซึ่งเป็นวิธีที่นิยมกันมากในทางปฏิบัติ เพราะใช้ง่ายและสะดวกรวดเร็ว การใช้ตารางเลขสุ่มขั้นแรกต้องใส่หมายเลขกำกับให้กับหน่วยประชากรทั้งหมด ขั้นต่อมาจึงใช้ตารางเลขสุ่มช่วยในการเลือกหน่วยตัวอย่าง เช่น ถ้าเรามีประชากรทั้งหมด 500 หน่วย

ขั้นแรก ใส่หมายเลขกำกับแก่หน่วยประชากรทั้งหมด ดังนี้

หน่วย	P_1	P_2	.....P_{500}
หมายเลข	001	002	 500

ขั้นที่สอง กำหนดขนาดตัวอย่างที่ต้องการสมมติว่าต้องการ 30 หน่วย ก็ต้องเลือกใช้ตารางเลขสุ่มที่มี 3 หลัก มีค่าไม่เกิน 500 แล้วทำการสุ่มในช่วง 001-500 ขึ้นมาทีละตัวโดยไม่เจาะจง เช่น ได้ 025 หน่วยที่ตกเป็นตัวอย่างก็คือ P_{025} แล้วเลือกต่อไปเรื่อย ๆ จนครบ 30 หน่วยที่ต้องการ

การสุ่มตัวอย่างแบบง่ายมีทั้งข้อดีและข้อเสีย สำหรับข้อดีนั้นก็คือ เป็นวิธีการที่ง่ายไม่สลับซับซ้อน และเข้าใจได้ง่ายกว่าวิธีอื่น ๆ และควรใช้กับประชากรที่มีลักษณะคล้ายคลึงกันมากที่สุด ส่วนข้อเสียนั้นเนื่องจากเมื่อหน่วยตัวอย่างในประชากรมีความแปรปรวนสูง ตัวอย่างที่เลือกได้จะไม่เป็นตัวแทนที่ดีของประชากร

2. การสุ่มตัวอย่างแบบแบ่งเป็นชั้นภูมิ (Stratified Random Sampling) เนื่องจากในบางครั้งประชากรประกอบด้วยหน่วยต่าง ๆ ไม่เหมือนกันทางด้านคุณลักษณะที่ต้องการศึกษา ดังนั้น จึงต้องจำแนกประชากรออกเป็นกลุ่มย่อยโดยให้หน่วยต่าง ๆ ที่เหมือนกันอยู่ในกลุ่มเดียวกัน ซึ่งเราเรียกว่า "ชั้นภูมิ" (Strata) จากนั้นจึงทำการสุ่มตัวอย่างมาจากแต่ละชั้นภูมิโดยใช้วิธีการสุ่มตัวอย่างแบบง่าย (SRS) ตัวอย่างเช่น การสำรวจค่าใช้จ่ายของนักศึกษามหาวิทยาลัยรามคำแหง โดยจำแนกนักศึกษาออกเป็นกลุ่มต่าง ๆ ตามคณะ หลังจากนั้นจึงทำการสุ่มตัวอย่างนักศึกษาจากแต่ละคณะโดยวิธี SRS สำหรับข้อดีและข้อเสียของการสุ่มตัวอย่างแบบนี้คือ

ข้อดี 1. มีประสิทธิภาพสูงกว่าแบบ SRS

2. ได้รายละเอียด แยกเป็นรายชั้นภูมิทำให้สามารถที่จะศึกษาเปรียบเทียบระหว่างกลุ่มได้

3. การควบคุมด้านบริหารสะดวกกว่าแบบ SRS

ข้อเสีย คือต้องจัดเตรียมงานล่วงหน้าเพื่อจัดแบ่งเป็นชั้นภูมิ

3. การสุ่มตัวอย่างแบบมีระบบ (Systematic Sampling) เป็นการสุ่มตัวอย่างโดยเริ่มจากจุดเริ่มต้นแบบสุ่ม (Random Start) แล้วจึงเลือกตัวอย่างต่อไปอีกทุก ๆ หน่วยที่ K จากประชากรที่เรียงลำดับไว้ โดยที่ $K = \frac{N}{n}$ เมื่อ N เป็นจำนวนประชากร n เป็นจำนวนตัวอย่าง และ K เป็น Sampling interval เช่น ต้องการสำรวจรายได้ของครัวเรือนในชุมชนหนึ่งที่มีทั้งหมด 50 ครัวเรือน โดยมีแผนที่ของครัวเรือนดังกล่าวและต้องการเลือกตัวอย่างมา 10 ครัวเรือน ดังนั้น $K = \frac{50}{10} = 5$ วิธีการเลือกตัวอย่างแบบนี้ ขั้นแรกเขียนเลขที่กำกับให้กับครัวเรือนทั้ง 50 ครัวเรือนซึ่งจะได้เป็นเลข 1, 2, ..., 50 จากนั้นก็เลือกจุดเริ่มต้นแบบสุ่ม ๆ สมมติได้เลข 6 หมายความว่า ครัวเรือนแรกที่ถูกเลือกมาเป็นตัวอย่างคือ ครัวเรือน

เลขที่ 6 ต่อไปก็คือครัวเรือนเลขที่ (6+5) = 11 และต่อไปก็คือครัวเรือนที่ (11+5) = 16, 21, 26, 31, 36, 41, 46 ซึ่งได้เพียง 9 ครัวเรือน จึงต้องขึ้นต้นใหม่จะได้เลขที่ 1 ดังนั้น ครัวเรือนทั้ง 10 ที่เลือกมาเป็นตัวอย่างคือ ครัวเรือนที่ 1, 6, 11, 16, 21, 26, 31, 36, 41 และ 46

สำหรับข้อดีและข้อเสียของการสุ่มแบบนี้มีดังนี้

ข้อดี 1. การเลือกตัวอย่างทำได้สะดวกและรวดเร็ว

2. เหมาะสำหรับงานสำรวจที่ต้องการให้หน่วยตัวอย่างกระจายโดยทั่วประชากร

3. เมื่อต้องการสุ่มตัวอย่างจากประชากรที่เรียงกันเป็นแฟ้มหรือเป็นบัตรรายการจะใช้วิธีวัดระยะเอาก็ได้

ข้อเสีย มีความเที่ยงตรงน้อย

4. การสุ่มตัวอย่างแบบกลุ่ม (Cluster Sampling) เป็นการสุ่มตัวอย่างโดยแบ่งประชากรออกเป็นกลุ่มย่อย ๆ เรียกว่า Cluster ก่อนแล้วจึงทำการสุ่มตัวอย่างกลุ่มของหน่วยตัวอย่าง (Cluster) โดยใช้วิธีการสุ่มแบบง่าย หรือแบบมีระบบจากนั้นจึงรวบรวมข้อมูลมาจากหน่วยทุกหน่วยของกลุ่มที่สุ่มได้

ลักษณะของการสุ่มแบบนี้มีลักษณะคล้ายกับการสุ่มแบบชั้นภูมิตรงที่ไม่ซ้ำซ้อนกัน แต่ต่างกันตรงที่ Cluster ประกอบด้วยหน่วยตัวอย่างที่มีลักษณะแตกต่างกัน ส่วนชั้นภูมิประกอบด้วยหน่วยตัวอย่างที่มีลักษณะคล้ายคลึงกัน ซึ่งแต่ละ Cluster จะรวมลักษณะทั้งหลายของประชากรไว้ด้วยกัน เช่น ต้องการสำรวจรายจ่ายของพนักงาน 20 คน จากแผนกต่าง ๆ 5 แผนก แต่ละแผนกมีพนักงาน 10 คน วิธีการก็คือต้องเตรียมบัญชีรายชื่อแผนกต่าง ๆ ทั้ง 5 แผนก แล้วสุ่มมาเพียง 2 แผนก จาก 5 แผนก แล้วเก็บข้อมูลจากพนักงาน 2 แผนกที่สุ่มได้นั้นก็จะได้ข้อมูลจากพนักงาน 20 คนตามที่ต้องการ ข้อดีและข้อเสียของการสุ่มแบบนี้มีดังนี้

ข้อดี 1. ไม่ต้องทำบัญชีรายชื่อหน่วยตัวอย่างทุกหน่วยในประชากรเหมือนกับวิธี SRS และวิธีสุ่มแบบแบ่งเป็นชั้นภูมิ

2. ทำให้ลดค่าใช้จ่ายในการเตรียมรายชื่อ

3. ประหยัดค่าใช้จ่ายในการเดินทาง

ข้อเสีย มีประสิทธิภาพด้อยกว่าแบบ SRS และแบบชั้นภูมิ

5. การสุ่มตัวอย่างแบบหลายขั้นตอน (Multistage Sampling) เป็นการเลือกตัวอย่างที่ต้องทำตั้งแต่ 2 ขั้นตอนขึ้นไป โดยในขั้นแรกประชากรจะถูกแบ่งเป็นหน่วยตัวอย่างในขั้นที่ 1 ก่อนแล้วสุ่มตัวอย่างมาจำนวนหนึ่ง ขั้นต่อไปก็คือการแบ่งตัวอย่างที่สุ่มได้เป็นกลุ่มย่อย ๆ อีก แล้วสุ่มมาอีกจำนวนหนึ่ง และถ้าอยากแบ่งต่อไปอีกก็สามารถทำได้ สำหรับข้อดีและข้อเสียมีดังนี้

ข้อดี 1. ประหยัดค่าใช้จ่ายในการเดินทาง

2. ลดค่าใช้จ่ายในการเตรียมบัญชีรายชื่อ

3. มีประสิทธิภาพและยืดหยุ่นได้มากกว่าการสุ่มแบบขั้นตอนเดียว

ข้อเสีย 1. เป็นวิธีการที่สลับซับซ้อนเข้าใจยาก

2. ต้องใช้การวางแผนละเอียดมากก่อนการเลือกตัวอย่าง

การสุ่มตัวอย่างทั้ง 4 วิธีข้างต้นนี้เป็นการสุ่มตัวอย่างที่ใช้กฎความน่าจะเป็นมาประยุกต์กับวิธีการเลือก ยังมีวิธีการสุ่มตัวอย่างที่นิยมใช้กันอย่างกว้างขวาง คือการสุ่มตัวอย่างแบบไม่สุ่ม (Non-random Sampling) เป็นการเลือกตัวอย่างโดยอาศัยการพิจารณาของผู้ชำนาญ ความสะดวกสบาย หรือเกณฑ์อื่น ๆ เป็นต้น ซึ่งไม่ต้องอาศัยกฎความน่าจะเป็น ซึ่งการเลือกตัวอย่างแบบไม่สุ่มมีดังนี้

ก. การเลือกตัวอย่างแบบโควต้า (Quota Sampling) เป็นการเลือกหน่วยตัวอย่างโดยไม่สนใจว่าตัวอย่างที่เลือกมานั้นจะเลือกมาโดยวิธีไหน เพียงแต่ให้มีจำนวนหน่วยครบตามที่กำหนดไว้ในแต่ละโควต้าเท่านั้น เช่น ต้องการสำรวจเกี่ยวกับขนาดของครอบครัว โดยกำหนดว่าจะทำการสอบถาม 20 ครอบครัวที่มีคุณลักษณะตามต้องการคือ แต่งงานแล้ว อยู่ด้วยกัน และมีบุตร เมื่อพบครอบครัวที่มีลักษณะดังกล่าวก็ทำการเก็บข้อมูลจากครอบครัวนั้นจนกระทั่งได้ครบตามโควต้า คือ 20 ครอบครัว

ข. การเลือกตัวอย่างเชิงพินิจพิจารณาหรือแบบมีจุดมุ่งหมาย (Judgement or Purposive Sampling) เป็นวิธีการเลือกหน่วยตัวอย่างโดยอาศัยการพิจารณาของผู้สุ่มตัวอย่าง

เองว่า จะเลือกหน่วยไหนมาเป็นตัวอย่าง การเลือกวิธีนี้อาจจะใช้สำหรับทดสอบคำถามหรือศึกษาแนวทาง

ค. การเลือกตัวอย่างแบบใช้ความสะดวก (Convenience Sampling) เป็นวิธีการเลือกหน่วยตัวอย่างโดยอาศัยความสะดวกสบายของผู้สุ่มเอง

หลักในการพิจารณาเลือกวิธีการสุ่มตัวอย่างว่าจะใช้วิธีใดนั้น ผู้สำรวจจะต้องพิจารณาลักษณะของงานสำรวจ ความถูกต้องแม่นยำของค่าประมาณ ค่าใช้จ่ายในการดำเนินงานซึ่งมักจะมีจำกัด และการควบคุมด้านบริหารจึงจำเป็นต้องพิจารณาหลายอย่างพร้อม ๆ กัน แล้วเลือกวิธีการสุ่มตัวอย่างที่ให้ผลตอบแทนที่สูงที่สุดตามข้อจำกัดที่มีอยู่

3. การทะเบียน (Registration) เป็นการเก็บรวบรวมข้อมูลจากแหล่งที่มีการบันทึกข้อมูลไว้แล้ว ข้อมูลจากทะเบียนบางประเภทสมบูรณ์และทันสมัยดี แต่บางประเภทไม่สมบูรณ์และไม่ทันสมัย ซึ่งถ้าการบันทึกลงทะเบียนไม่สมบูรณ์ผิดพลาดก็จะมีผลกระทบกระเทือนต่อการวิเคราะห์และสรุปผลได้

ข้อมูลที่เก็บมาได้จะโดยวิธีการใดก็ตามเรียกว่า ข้อมูลดิบ (Raw data) การที่จะนำข้อมูลที่รวบรวมได้ไปเสนอให้คนทั่วไปเข้าใจนั้น ทำได้โดยนำมาจัดระเบียบ และเสนอข้อมูลให้อยู่ในรูปที่น่าสนใจ เพื่อเตรียมพร้อมที่จะนำเอาข้อมูลเหล่านั้นไปวิเคราะห์ การนำเสนอข้อมูลมีหลายแบบด้วยกัน คือ การเสนอในรูปของบทความ ตาราง กราฟเส้น กราฟแท่ง แผนภาพวงกลม รูปภาพ อีสโตแกรม และตารางแจกแจงความถี่ ซึ่งจะเสนอเป็นแบบใดจึงจะเหมาะสมหรือดีนั้นมีวิธีการพิจารณาดังนี้ คือ

1. อ่านเข้าใจง่าย
2. ช่วยให้สามารถเข้าใจความหมายของข้อมูลนั้นได้ดี
3. ใช้ได้เหมาะสมกับข้อมูลแบบต่าง ๆ
4. สะดวกในการวิเคราะห์
5. ช่วยให้เข้าใจผลของการศึกษาได้ถูกต้องละเอียดและมีประสิทธิภาพ

4.2.1. ระดับการวัด (Scales of Measurement) ข้อมูลทางสถิตินั้นแบ่งออกได้เป็น 2 ประเภท คือ

1. ข้อมูลเชิงปริมาณ (Quantitative data) เป็นข้อมูลที่มีค่าเป็นตัวเลข ซึ่งตัวเลขนี้อาจจะเป็นจำนวนเต็มนับได้หรือนับไม่ได้ เช่น ข้อมูลสถิติเกี่ยวกับจำนวนนักศึกษาที่สมัครเข้าเรียนในคณะมนุษยศาสตร์ปีการศึกษา 2533 ข้อมูลสถิติเกี่ยวกับคะแนนสอบวิชา SC 101 เป็นต้น

2. ข้อมูลเชิงคุณภาพ (Qualitative data) เป็นข้อมูลสถิติที่แสดงลักษณะหนึ่ง ๆ โดยเฉพาะไม่สามารถมีค่าเป็นตัวเลขได้ เช่น ข้อมูลสถิติเกี่ยวกับเรื่องการศึกษา เพศ อาชีพ ชนิดของปุ๋ย สถานภาพสมรส ศาสนา เป็นต้น

นอกจากการแบ่งข้อมูลออกเป็น 2 ประเภทนี้แล้ว ยังแบ่งข้อมูลออกตามมาตราการวัดต่าง ๆ ซึ่งส่วนใหญ่การแบ่งวิธีนี้มักจะนำไปใช้ในการวัดตัวแปรต่าง ๆ ทางสังคม โดยที่มาตราการวัดนี้แบ่งตามประเภทของตัวแปรเชิงคุณภาพและตัวแปรเชิงปริมาณนั่นเอง คือ ถ้าเป็นตัวแปรเชิงคุณภาพก็จะใช้มาตราการวัดที่เรียกว่า มาตราการวัดแบบจำแนกประเภท (Nominal Scale) และมาตราการวัดแบบลำดับ (Ordinal Scale) แต่ถ้าเป็นตัวแปรเชิงปริมาณก็จะใช้มาตราการวัดที่เรียกว่า มาตราการวัดแบบช่วง (Interval Scale) และมาตราการวัดแบบอัตราส่วน (Ratio Scale) ดังนั้น ในทางสถิติจึงแบ่งระดับการวัดออกเป็นดังนี้

1. มาตราการวัดแบบจำแนกประเภท (Nominal Scale) เป็นตัวแปรเชิงคุณภาพที่ไม่มีลำดับ (order) ไม่มีทิศทาง (direction) ไม่มีขนาด (Magnitude) ตัวอย่าง เช่น ข้อมูลเกี่ยวกับเพศ ซึ่งจำแนกออกเป็นเพศชายและเพศหญิง ข้อมูลเกี่ยวกับระดับการศึกษา ซึ่งจำแนกออกเป็นประถมศึกษา มัธยมศึกษา และอุดมศึกษา ข้อมูลเกี่ยวกับสีของรถซึ่งจำแนกออกเป็นสีแดง ขาว เขียว น้ำตาล ฯลฯ ข้อมูลเกี่ยวกับความพอใจสินค้าบางชนิด ซึ่งอาจจำแนกออกเป็นพอใจกับไม่พอใจ หรืออาจจะจำแนกได้มากกว่านี้ ข้อมูลเกี่ยวกับการนับถือศาสนา ซึ่งอาจจำแนกออกเป็น ศาสนาพุทธ ศาสนาคริสต์ ศาสนาอิสลาม และศาสนาอื่น ๆ ข้อมูลเกี่ยวกับชนิดของสัตว์ ซึ่งอาจจะจำแนกออกเป็น สัตว์บก สัตว์น้ำ สัตว์ครึ่งน้ำครึ่งบก ข้อมูลเกี่ยวกับความสามารถในการพูดภาษา ซึ่งอาจจำแนกออกเป็น พูดภาษาอังกฤษได้ พูดภาษา

ฝรั่งเศสได้ และอื่นๆ ข้อมูลเกี่ยวกับสถานภาพสมรสซึ่งจำแนกเป็น โสด แต่งงาน หย่า หม้าย หรือข้อมูลเกี่ยวกับคำถามที่ต้องการคำตอบว่า ใช่หรือไม่ใช่ เห็นด้วย ไม่เห็นด้วย ฯลฯ เหล่านี้เป็นต้น ข้อมูลลักษณะที่กล่าวมาข้างต้นนี้จะใช้มาตราการวัดแบบจำแนกประเภททั้งสิ้น และเราสามารถกำหนดค่าของตัวเลขให้กับประเภทต่าง ๆ ของข้อมูลที่จำแนกออกมาได้ โดยใช้ Nominal scale ตัวอย่างเช่น ข้อมูลที่ได้คำตอบว่า ใช่ กับ ไม่ใช่ จะกำหนดให้เลข 0 และเลข 1 แทนคำตอบว่า ใช่ และ ไม่ใช่ ตามลำดับ หรือข้อมูลเกี่ยวกับสถานภาพสมรสอาจจะกำหนดให้เลข 1, 2, 3 และ 4 แทนคำตอบว่า โสด แต่งงาน หย่า และหม้ายได้ตามลำดับ แต่ตัวเลขที่กำหนดให้เหล่านี้ไม่มีคุณสมบัติที่แสดงถึงปริมาณ (quantitative) เลย เป็นแต่เพียงแสดงให้เห็นถึงการจำแนกข้อมูลเท่านั้น ตัวอย่างเช่น ตัวเลข 1, 2, 3 และ 4 ที่เรากำหนดให้กับสถานภาพสมรสนั้น ไม่สามารถเขียนว่า $3 > 4$ หรือ $2 < 4$ หรือ $2-1 = 4-3$ หรือ $1+3 = 4$ หรือ $4 \div 2 = 2$ ได้เลย

ข้อมูลที่ใช้มาตราการวัดแบบจำแนกประเภทนี้จะประกอบไปด้วยความถี่ในการนับหรือเป็นตารางแสดงจำนวนครั้งที่เกิดขึ้นของแต่ละประเภทของตัวแปรต่าง ๆ ที่เราทำการศึกษา เช่น ความถี่ของการนับจำนวน เพศหญิง และเพศชาย หรือการนับจำนวนผู้ที่เห็นด้วย และผู้ที่ไม่เห็นด้วย ดังนั้น ข้อมูลดังกล่าวมักจะอยู่ในรูปของ frequency data หรือ enumerative data หรือ attribute data หรือ categorical data และเครื่องหมายทางคณิตศาสตร์ที่นำมาใช้ในมาตราการวัดแบบจำแนกประเภท คือ เครื่องหมายเท่ากับ (=) และเครื่องหมายไม่เท่ากับ ($\neq$) ตัวอย่างเช่น ข้อมูลเกี่ยวกับสถานภาพสมรสของคนงานหญิง 50 คน ของบริษัทผลิตตุ๊กตาแห่งหนึ่งเป็นดังนี้

สถานภาพสมรส	จำนวนคนงานหญิง (คน)
โสด	13
สมรส	25
หย่า	7
หม้าย	5
รวม	50

หรือตัวอย่างเช่น ถ้าเราจำแนกข้อมูลเกี่ยวกับเพศและระดับการศึกษาของอาจารย์ในมหาวิทยาลัยแห่งหนึ่ง โดยให้ตัวแปรเกี่ยวกับเพศเป็นตัวแปร X ซึ่งจำแนกออกเป็นเพศหญิงและเพศชาย ตัวแปรเกี่ยวกับระดับการศึกษาเป็นตัวแปร Y ซึ่งจำแนกออกเป็นระดับปริญญาตรี ระดับปริญญาโท และระดับปริญญาเอก ก็จะได้ตารางแจกแจงความถี่เป็นดังนี้

เพศ (X)	ระดับการศึกษา (Y)			รวม
	ปริญญาตรี	ปริญญาโท	ปริญญาเอก	
ชาย	5	30	10	45
หญิง	3	45	7	55
รวม	8	75	17	100

ลักษณะของข้อมูลแบบจำแนกประเภท (Nominal data) นั้นมักจะใช้ในการคำนวณหาสัดส่วน (proportion) และเปอร์เซ็นต์ (Percentage) เช่นจะหาสัดส่วนของ

อาจารย์ที่เป็นเพศชายจากตารางข้างต้น หรือจะหาเปอร์เซ็นต์ของอาจารย์ที่จบปริญญาโทว่ามีกี่เปอร์เซ็นต์ นอกจากนี้ ในสถิติอนุมานมักจะมีการทดสอบความเป็นอิสระ ความเป็นเอกภาพและการหาความสัมพันธ์ระหว่างตัวแปร เป็นต้น

2. มาตราการวัดแบบลำดับ (Ordinal Scale) เป็นตัวแปรเชิงคุณภาพที่มีลำดับ มีทิศทาง ตัวอย่างเช่น การจัดลำดับของผู้นิยมเรียนภาษาต่างประเทศ ภาษาต่าง ๆ การจัดลำดับผลการเรียนว่า เก่งมาก เก่ง ปานกลาง อ่อน อ่อนมาก มาตราการวัดแบบนี้จะเกี่ยวข้องกับเครื่องหมายทางพีชคณิตคือเครื่องหมาย > และ < ตัวอย่างเช่น a > b หมายถึง a ใหญ่กว่า b หรือ a มีลำดับสูงกว่า b หรือชอบ a มากกว่า b เป็นต้น a < b หมายถึง a น้อยกว่า b หรือ a มีลำดับต่ำกว่า b หรือชอบ a น้อยกว่า b เป็นต้น ซึ่งจะเห็นได้ว่า เครื่องหมายทั้งสองนี้นอกจากจะหมายถึงน้อยกว่าหรือมากกว่าแล้วยังใช้ในความหมายอื่น ๆ อีกด้วย เช่น เร็วกว่า ช้ากว่า ฉลาดมากกว่า มีความพร้อมมากกว่า มีความสวยมากกว่า มีความนิยมมากกว่า ฯลฯ เป็นต้น ซึ่งสามารถกำหนดตัวเลขให้กับลำดับต่าง ๆ ของข้อมูลได้ โดยใช้มาตราการวัดแบบลำดับ (Ordinal Scale) ตัวอย่างเช่น การวัดความเร็วของรถยนต์ 4 คัน คือ A, B, C และ D โดยให้ลำดับรถที่วิ่งเร็วที่สุดเป็นลำดับที่ 1 และรองลงมาเป็นลำดับที่ 2 ลำดับที่ 3 และลำดับที่ 4 ตามลำดับ เป็นต้น ตัวเลขในมาตราการวัดแบบลำดับนี้ไม่ได้แสดงถึงปริมาณเลย เพียงแต่เป็นตัวเลขที่ชี้ให้เห็นว่าอยู่ในตำแหน่งที่เท่าใดเท่านั้น แต่ไม่ได้บอกว่าจำนวนมากน้อยเท่าใด ซึ่งจากตัวอย่างการวัดความเร็วของรถยนต์ 4 คันข้างต้นจะบอกได้แต่เพียงว่า $4 > 2$ หรือ $3 < 4$ เท่านั้น แต่จะเขียนว่า $2-1 = 4-3$ ไม่ได้หรือจะเขียนว่า $\frac{1}{2} = \frac{2}{4}$ ไม่ได้

ลักษณะของข้อมูลแบบลำดับ (Ordinal data) นั้น มักจะนำมาคำนวณหาสัดส่วน (proportion) และเปอร์เซ็นต์ (Percentage) เช่นเดียวกับข้อมูลแบบจำแนกประเภท

3. มาตราการวัดแบบช่วง (Interval Scale) ตัวแปรเชิงปริมาณที่ให้จำนวนปริมาณของสิ่งของที่ต้องการหรือที่บุคคลมีอยู่ เช่น ข้อมูลเกี่ยวกับความสูงของคน ข้อมูลเกี่ยวกับน้ำหนักของคน คะแนนสอบ จำนวนรถยี่ห้อฮอนด้า อุณหภูมิที่วัดได้ ขนาดของครอบครัว รายได้ของครอบครัว ตัวแปรเชิงปริมาณนี้สามารถที่จะนำมาบวกกัน ลบกัน คูณกัน หรือหารกันได้

มาตราการวัดแบบช่วงนั้นเป็นการวัดที่หน่วยวัดมีช่วงเท่า ๆ กัน เช่น หน่วยวัดเวลา หรือคะแนนสอบ เป็นต้น ซึ่งค่าของตัวแปรสามารถนำมาบวกกันหรือลบกันได้

4. มาตราการวัดแบบอัตราส่วน (Ratio Scale) มีความหมายคล้ายกับมาตราการวัดแบบช่วง เพียงแต่มีข้อจำกัดเพิ่มขึ้นตรงที่ว่าค่าของการวัดจะเริ่มต้นจากศูนย์เป็นต้นไปทำให้เราสามารถเปรียบเทียบค่าของการวัดเป็นรูปอัตราส่วนได้ ดังนั้น ค่าของตัวแปรสามารถนำมาหารและนำมาคูณกันได้ เช่น ข้อมูลเกี่ยวกับน้ำหนักของคน 4 คน เป็นดังนี้ 80, 92, 73, 75 กิโลกรัม สามารถหาน้ำหนักเฉลี่ยของข้อมูลชุดนี้ได้จาก

$$\bar{X} = \frac{\sum_{i=1}^{4} x_i}{n} = \frac{80+92+73+75}{4}$$

$$= \frac{320}{4} = 80 \text{ กิโลกรัม}$$

หรือจะหาพิสัย (Range) (หรือการกระจายของข้อมูลว่าค่าสูงสุดกับค่าต่ำสุดห่างกันมากน้อยแค่ไหน) ของข้อมูลชุดนี้ได้จาก

พิสัย = น้ำหนักที่มากที่สุด - น้ำหนักที่น้อยที่สุด

= 92 - 73 = 19

จากมาตราการวัดต่าง ๆ ทั้ง 4 แบบนี้ สามารถนำไปใช้ให้เหมาะสมกับลักษณะของข้อมูลที่มีอยู่ได้

4.2.2 สถิติสรุป

การสรุปข้อมูลที่รวบรวมมาได้นั้นเป็นวิธีการหนึ่งที่ช่วยให้เข้าใจคุณลักษณะบางประการของข้อมูลชุดนั้น ๆ ได้ และเพื่อจะให้มีความเข้าใจมากยิ่งขึ้นเกี่ยวกับข้อมูลชุดนั้น ๆ จำเป็นที่จะต้องอธิบายข้อมูลนั้นด้วยคุณสมบัติ 2 ประการ คือ การวัดวัดแนวโน้มเข้าสู่ส่วนกลาง และการวัดการกระจาย เพราะการวัดทั้งสอง อย่างนี้จะช่วยชี้ให้เห็นข้อแตกต่างระหว่างข้อมูลที่มีหลาย ๆ ชุด โดยการวัดแนวโน้มเข้าสู่ส่วนกลางทำให้ทราบถึงตำแหน่งของข้อมูล ส่วนการวัดการกระจายทำให้ทราบถึงขนาดหรือรูปร่างของการกระจายของข้อมูลชุดนั้น

การวัดแนวโน้มเข้าสู่ส่วนกลาง (measure of central tendency) หมายถึงการหาเลขจำนวนเดี่ยว ๆ จำนวนหนึ่งซึ่งใช้แทนค่ากลาง ๆ ของข้อมูล หรือที่เรียกกันทั่ว ๆ ไปว่าค่าเฉลี่ย (Average) เช่น คะแนนสอบวิชา SC 101 เฉลี่ยแล้วเท่ากับ 52 คะแนน ค่าใช้จ่ายของนักศึกษาชั้นปีที่ 1 เฉลี่ยแล้วเท่ากับ 1,200 บาทต่อเดือน เป็นต้น สำหรับการวัดแนวโน้มเข้าสู่ส่วนกลางที่นิยมกันมีหลายวิธี แต่ใน SC 101 จะขอกล่าวเพียง 3 วิธีที่สำคัญ ๆ ดังนี้

1. มัชฌิมเลขคณิต (Arithmetic Mean) เป็นวิธีที่นิยมใช้กันมากที่สุด และรู้จักกันแพร่หลายบางครั้งเรียกสั้น ๆ ว่า ค่าเฉลี่ย (Average หรือ Mean)

มัชฌิมเลขคณิต คือ ผลรวมของค่าสังเกตทุกค่าหารด้วยจำนวนค่าสังเกตทั้งหมด การวัดโดยวิธีนี้มีทั้งข้อดีและข้อเสียคือ

ข้อดี

1. เข้าใจและคำนวณได้ง่าย
2. การคำนวณใช้ค่าสังเกตทุกค่าที่รวบรวมได้
3. สามารถหาค่าของมัชฌิมเลขคณิตได้เสมอ และเป็นค่าที่แน่นอน
4. เหมาะสำหรับในการนำไปใช้คำนวณค่าต่าง ๆ ทางสถิติ
5. ส่วนเบี่ยงเบนของค่าสังเกตจากมัชฌิมเลขคณิตจะมีค่าน้อยที่สุด

ข้อเสีย

1. เนื่องจากมัชฌิมเลขคณิตใช้ค่าสังเกตทุกค่า ดังนั้น จึงเปลี่ยนแปลงได้ง่าย ถ้าค่าสังเกตบางค่าที่รวบรวมได้มีค่าผิดปกติก็จะทำให้มัชฌิมเลขคณิตผิดปกติไปด้วย
2. ค่ามัชฌิมเลขคณิตที่คำนวณได้จะตรงกับค่าสังเกตที่มีอยู่จริง ๆ น้อยมาก หรือไม่มีเลย

หลักในการพิจารณาว่าจะนำมัชฌิมเลขคณิตไปใช้ในการวัดแนวโน้มเข้าสู่ส่วนกลางมีดังนี้

1. เมื่อค่าสังเกตแต่ละค่ามีค่าใกล้เคียงกัน
2. เมื่อต้องการวัดการกระจายที่น้อยที่สุด
3. เมื่อต้องการมีมัชฌิมที่เชื่อถือได้มากที่สุด
4. เมื่อต้องการมัชฌิมไปใช้การคำนวณค่าต่าง ๆ ในทางสถิติต่อไป

วิธีการคำนวณหามัชฌิมเลขคณิตมีดังนี้

ถ้าให้ $\bar{x}$ แทนมัชฌิมเลขคณิต

$$\bar{x} = \frac{x_1 + x_2 + \ldots\ldots + x_n}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

เมื่อ n คือ จำนวนข้อมูลทั้งหมด และ $x_1, x_2, \ldots, x_n$ เป็นข้อมูลแต่ละข้อมูลที่รวบรวมได้

ตัวอย่างเช่น จะหามัชฌิมเลขคณิตของข้อมูลต่อไปนี้

8, 12, 15, 19 และ 24

$$\therefore \quad \bar{x} = \frac{8+12+15+19+24}{5}$$

$$= \frac{78}{5} = 15.60$$

คุณสมบัติของมัชฌิมเลขคณิต

1. จำนวนค่าสังเกตคูณด้วยค่าเฉลี่ยจะเท่ากับผลรวมทั้งสิ้นของข้อมูลนั้น นั่นคือ ถ้ามีข้อมูล n ข้อมูล

$$\sum_{i=1}^{n} x_i = n\bar{x}$$

ตัวอย่างเช่น มีข้อมูลชุดหนึ่ง คือ 2, 3, 5, 6

$$\bar{x} = \frac{2+3+5+6}{4} = \frac{16}{4} = 4$$

$$\sum_{i=1}^{4} x_i = 2+3+5+6 = 16$$

$$\therefore \quad n\bar{x} = 4 \times 4 = 16$$

ดังนั้น $\sum_{i=1}^{4} x_i = n\bar{x}$

2. ผลรวมของค่าเบี่ยงเบนจากค่าเฉลี่ยของทุกค่าสังเกตในข้อมูลหนึ่ง ๆ มีค่าเท่ากับศูนย์ นั่นคือ

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

ซึ่งแสดงให้เห็นได้ดังนี้ ถ้ามีข้อมูล n ข้อมูล

$$\because \quad \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}$$

$$= n\bar{x} - n\bar{x}$$

$$= 0$$

3. ผลรวมของกำลัง 2 ของค่าเบี่ยงเบนจากค่าเฉลี่ยจะน้อยกว่าผลรวมของกำลัง 2 ของค่าเบี่ยงเบนจากค่าสังเกตอื่น ๆ แสดงให้เห็นได้ดังตารางต่อไปนี้

x_i	$(x_i-2)^2$	$(x_i-3)^2$	$(x_i-4)^2$	$(x_i-5)^2$	$(x_i-6)^2$
2	0	1	4	9	16
3	1	0	1	4	9
4	4	1	0	1	4
5	9	4	1	0	1
6	16	9	4	1	0
รวม	30	15	10	15	30

จากข้อมูลในตารางหา $\bar{x}$ ได้เท่ากับ 4 และ $(x_i-4)^2$ คือ $(x_i-\bar{x})^2$ นั่นเอง จะมีผลรวมน้อยที่สุดคือ 10

4. มัชฌิมเลขคณิตของค่าคงที่จะมีค่าเท่ากับค่าคงที่นั้น ถ้าให้ a เป็นค่าคงที่ใดๆ จะได้มัชฌิมเลขคณิตของ a มีค่าเท่ากับ a

5. ถ้าให้ a เป็นค่าคงที่ใดๆ มัชฌิมเลขคณิตของข้อมูล (x_i) บวก (หรือลบ) กับค่าคงที่ a จะเท่ากับมัชฌิมเลขคณิตของข้อมูลเดิม (x_i) บวก (หรือลบ) ด้วยค่าคงที่ a

6. ให้ a เป็นค่าคงที่ใดๆ มัชฌิมเลขคณิตของข้อมูลเดิม (x_i) คูณ (หรือหาร) ด้วยค่าคงที่ a จะเท่ากับค่าคงที่ a คูณ (หรือหาร) ด้วยมัชฌิมเลขคณิตของข้อมูลเดิม (x_i)

มัชฌิมเลขคณิตแบบถ่วงน้ำหนัก (Weighted Arithmetic Mean)

ถ้ามีข้อมูล $x_1, x_2, \ldots, x_n$ และน้ำหนักของแต่ละข้อมูล คือ $w_1, w_2, \ldots, w_n$ ตามลำดับ มัชฌิมเลขคณิตแบบถ่วงน้ำหนัก ($\bar{x}_w$) คือ

$$\bar{x}_w = \frac{x_1w_1 + x_2w_2 + \ldots + x_nw_n}{w_1 + w_2 + \ldots + w_n}$$

หรือ $$\bar{x}_w = \frac{\Sigma xw}{\Sigma w}$$

ตัวอย่างเช่น ในการวัดผลการเรียนวิชาหนึ่งใช้ผลการสอบ 2 ครั้งด้วยกัน คือ ผลสอบกลางเทอม และสอบปลายเทอม โดยคิดผลการสอบปลายเทอมเป็น 2 เท่าของผลการสอบกลางเทอม ถ้านักเรียนคนหนึ่งได้คะแนนสอบกลางเทอมเท่ากับ 95 และคะแนนสอบปลายเทอมเท่ากับ 89 จงหาคะแนนเฉลี่ยของผลการเรียนของนักเรียนคนนี้

มัชฌิมเลขคณิตแบบถ่วงน้ำหนัก $$\bar{x}_w = \frac{95(1) + 89(2)}{1+2}$$

$$= \frac{95+178}{3} = \frac{273}{3}$$

$$= 91$$

2. มัธยฐาน (Median) เป็นค่ากลางๆ ที่ครึ่งหนึ่ง (50%) ของค่าสังเกตในข้อมูลมีค่ามากกว่าและอีกครึ่งหนึ่งของค่าสังเกตในข้อมูลมีค่าน้อยกว่า การนำมัธยฐานไปใช้วัดค่ากลางมีข้อดีข้อเสียดังนี้

ข้อดี 1. ค่าของมัธยฐานจะตรงกับค่าจริงของค่าสังเกตในข้อมูลนั้น

2. เข้าใจง่าย

3. ขจัดผลกระทบซึ่งเกิดจากข้อมูลที่มีค่าสูงหรือต่ำมากเกินไป หรือข้อมูลที่ผิดปกติ

4. ใช้ได้กับรายการซึ่งไม่สามารถหาฐานร่วมเพื่อการเปรียบเทียบได้

5. เมื่อทราบค่าของข้อมูลกลาง ๆ ก็สามารถคำนวณหาค่ามัธยฐานได้

ข้อเสีย 1. ถ้าการแจกแจงของข้อมูลไม่สม่ำเสมอ ค่ามัธยฐานที่ได้อาจจะไม่แน่นอน

2. ไม่เหมาะที่จะใช้ในการคำนวณขั้นต่อไป

นอกจากนี้ ควรจะหามัธยฐานเมื่อต้องการมัชฌิมอย่างคร่าว ๆ หรือเมื่อมีข้อมูลผิดปกติ หรือเมื่อข้อมูลบางค่ามีค่าสูง หรือต่ำมากเกินไป หรือเมื่อต้องการทราบว่าข้อมูลใดสูงกว่ามัธยฐาน ข้อมูลใดต่ำกว่ามัธยฐาน

การคำนวณหามัธยฐานทำได้โดยการเรียงลำดับข้อมูลจากมากไปหาน้อย หรือจากน้อยไปหามาก จากนั้นดูว่าข้อมูลใดอยู่ตรงกลางของบรรดาข้อมูลทั้งหมด ข้อมูลนั้นจะเป็นมัธยฐาน ซึ่งจะมีอยู่ 2 กรณี คือ

ก. เมื่อจำนวนข้อมูลเป็นเลขคี่ ข้อมูลตัวกลางจะเป็นมัธยฐาน เช่น ข้อมูลชุดหนึ่งประกอบด้วย 2, 2, 4, 5, 6, 7, 9 มัธยฐาน คือ 5 หรือหาได้จาก

ตำแหน่งมัธยฐานจะอยู่ที่ $\frac{N+1}{2} = \frac{7+1}{2} = 4$

ซึ่งตำแหน่งที่ 4 ก็คือเลข 5

$\therefore$ มัธยฐาน มีค่าเท่ากับ 5

ข. เมื่อจำนวนข้อมูลเป็นเลขคู่ จะมีข้อมูลตัวกลาง 2 ตัว มัธยฐานจะเท่ากับข้อมูลตัวกลาง 2 ตัวบวกกันแล้วหารด้วย 2 ดังนี้

มีข้อมูล 3, 4, 6, 7, 9, 10

มัธยฐานจะเท่ากับ $\frac{6+7}{2} = \frac{13}{2} = 6.5$

หรือหาตำแหน่งมัธยฐาน $= \frac{N+1}{2} = \frac{6+1}{2} = 3.5$

ตำแหน่ง 3.5 จะอยู่ระหว่างเลข 6 และ 7

$\therefore$ มัธยฐานเท่ากับค่าเฉลี่ยของเลขทั้ง 2 นั่นคือ มัธยฐาน $= \frac{6+7}{2} = 6.5$

3. ฐานนิยม (Mode)

ฐานนิยมของข้อมูลชุดหนึ่ง คือ ค่าของข้อมูล ซึ่งเกิดขึ้นด้วยความถี่สูงที่สุด ฐานนิยมอาจมีได้หลายค่า แต่ถ้ามีค่าเดียวจะเรียกว่า Unimodal ถ้ามี 2 ค่าเรียกว่า Bimodal ถ้ามีมากกว่า 2 ค่าขึ้นไปเรียกว่า Multimodal

เนื่องจากฐานนิยมไม่ค่อยจะนิยมใช้มากนัก จะใช้กันก็ต่อเมื่อต้องการมัชฌิมคร่าว ๆ และต้องการใช้อย่างรวดเร็ว หรือต้องการทราบว่าข้อมูลตัวใดมีความถี่มากที่สุด การนำฐานนิยมไปใช้วัดมีข้อดีข้อเสียดังนี้

ข้อดี 1. เข้าใจง่าย

2. ขจัดผลกระทบ ซึ่งเกิดจากข้อมูลมีค่าสูงเกินไป หรือต่ำเกินไป หรือคะแนนผิดปกติ

3. ถ้าทราบคะแนนกลาง ๆ ก็สามารถคำนวณหาฐานนิยมได้

ข้อเสีย 1. ไม่เหมาะสมในการที่จะคำนวณค่าต่าง ๆ ทางสถิติขั้นต่อไป

2. เป็นการยากที่จะคำนวณได้แน่นอน

วิธีคำนวณหาฐานนิยม หาได้โดยการเลือกข้อมูลที่มีความถี่มากที่สุด เป็นค่าของฐานนิยม ตัวอย่าง เช่น จากข้อมูลต่อไปนี้ 2, 2, 3, 4, 7, 9, 9, 9, 8, 12, 13 จะเห็นได้ว่า 9 มีความถี่มากที่สุด คือ 3 ดังนั้น ฐานนิยมของข้อมูลชุดนี้ คือ 9 ลักษณะนี้เรียกว่า Unimodal

ความสัมพันธ์ระหว่างมัชฌิมเลขคณิต มัธยฐาน และฐานนิยม

สำหรับการแจกแจงชนิดที่มีฐานนิยมค่าเดียว ซึ่งมีลักษณะเบ้เล็กน้อย (Moderately skewed) จะได้ความสัมพันธ์ของมัชฌิมเลขคณิต มัธยฐาน และฐานนิยม ดังนี้

Mean - Mode = 3(Mean - Median)

แต่สำหรับการแจกแจงความถี่ที่มีลักษณะสมมาตร (Symmetrical) ค่าของมัชฌิมเลขคณิต มัธยฐาน และฐานนิยมจะมีค่าเท่ากันนั่นคือ ค่าทั้ง 3 จะอยู่ที่เดียวกัน

การวัดการกระจาย (Measures of Dispersion)

สำหรับข้อมูลชุดหนึ่งนั้นนอกจากจะวัดแนวโน้มเข้าสู่ส่วนกลางแล้วควรจะดูความแตกต่างของค่าของข้อมูลแต่ละค่านั้นด้วยว่ามีความแตกต่างไปจากค่ากลางของข้อมูลชุดนั้น ๆ มากน้อย

แค่ไหน ซึ่งเป็นการวัดการกระจายของข้อมูล ใน SC101 จะขอกล่าวถึง 4 วิธีเท่านั้น คือ

1. พิสัย (Ranges) เป็นวิธีวัดการกระจายของข้อมูลที่ง่ายที่สุด โดยการหาความแตกต่างของข้อมูลที่สูงที่สุดและข้อมูลที่ต่ำที่สุด

$$\therefore \quad \text{พิสัย} = \text{ค่าสูงที่สุด} - \text{ค่าต่ำที่สุด}$$

จะเห็นได้ว่าพิสัยเป็นค่าวัดการกระจายของข้อมูลอย่างหยาบ ๆ เพราะนำเอาค่าสองค่าของข้อมูล (คือค่าสูงสุดกับค่าต่ำสุด) มาใช้ในการคำนวณเท่านั้น ส่วนค่าอื่น ๆ ในข้อมูลไม่ได้นำมาใช้ในการวัดการกระจายเลย เพราะฉะนั้น ถ้าข้อมูลชุดหนึ่งมีค่าใกล้เคียงกันหมดยกเว้นค่าหนึ่ง ซึ่งมีค่าสูงกว่าปกติ หรือต่ำกว่าปกติ จะทำให้พิสัยที่ได้มีค่าผิดปกติไปด้วย เช่น การชั่งน้ำหนักหมู 2 คอก ปรากฏผลดังนี้

คอกที่ 1 52, 53, 55, 57, 60, 62

คอกที่ 2 52, 54, 55, 56, 61, 95

พิสัยของน้ำหนักของหมู คอกที่ 1 = 62 - 52 = 10

พิสัยของน้ำหนักของหมู คอกที่ 2 = 95 - 52 = 43

จะเห็นได้ว่าน้ำหนักของหมู 2 คอก คล้ายคลึงกัน แต่ในคอกที่ 2 น้ำหนักหมูผิดปกติอยู่ 1 ตัว ซึ่งชั่งน้ำหนักได้ 95 จึงทำให้พิสัยของน้ำหนักของหมูทั้ง 2 คอกต่างกันมาก

ถึงแม้ว่าพิสัยเป็นค่าวัดการกระจายอย่างหยาบ แต่ก็เป็นค่าที่ใช้กันเสมอสำหรับคนทั่ว ๆ ไป เพราะค่าพิสัยเป็นค่าที่คำนวณง่าย สังเกตได้รวดเร็ว จึงเหมาะที่จะนำไปใช้ในกรณีที่ต้องการทราบการกระจายของข้อมูลโดยรวดเร็ว

2. ส่วนเบี่ยงเบนเฉลี่ย (Mean Deviation or Average Deviation) เป็นการวัดการกระจายของข้อมูล โดยวัดจากค่าเฉลี่ยของส่วนเบี่ยงเบนของแต่ละข้อมูลจากมัชฌิมเลขคณิต (ไม่คิดเครื่องหมาย) สมมติว่าข้อมูลชุดหนึ่งมีมัชฌิมเลขคณิตเท่ากับ $\bar{x}$ ฉะนั้นส่วนเบี่ยงเบนจากมัชฌิมเลขคณิต ($\bar{x}$) ของข้อมูลทั้งหมด คือ $(x_1-\bar{x})$, $(x_2-\bar{x})$, $(x_3-\bar{x})$ ....., $(x_n-\bar{x})$

ซึ่งถ้าหาผลรวมของส่วนเบี่ยงเบนจากมัชฌิมเลขคณิตของข้อมูลทั้งหมดจะได้ $\sum_{i=1}^{n}(x_i-\bar{x})$

$$\text{จาก} \quad \sum_{i=1}^{n}(x_i-\bar{x}) = \sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\bar{x}$$

$$= n\bar{x} - n\bar{x} = 0$$

ซึ่งจะได้ $\sum_{i=1}^{n}(x_i-\bar{x})$ เป็นศูนย์เสมอ ไม่ว่าข้อมูลจะกระจายมากหรือน้อย ซึ่งจะทำให้ผลรวมของส่วนเบี่ยงเบนจากมัชฌิมเลขคณิตใช้ประโยชน์ไม่ได้ ดังนั้น จึงต้องหาผลรวมของส่วนเบี่ยงเบนโดยไม่คำนึงถึงเครื่องหมายบวกและลบ นั่นคือ การหาค่าสัมบูรณ์ (Absolute value) ของผลรวมของส่วนเบี่ยงเบน

$$\therefore \quad \text{ส่วนเบี่ยงเบนเฉลี่ย (MD หรือ AD)} = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n}$$

ตัวอย่างเช่น บ้าน 5 หลัง มีอายุเป็นดังนี้ 5, 2, 4, 2 และ 2 ปี ตามลำดับ จงหาส่วนเบี่ยงเบนเฉลี่ยของอายุของบ้านทั้ง 5 หลังนี้

$$\bar{x} = \frac{5+2+4+2+2}{5} = \frac{15}{5} = 3$$

$$\text{ดังนั้น} \quad AD = \frac{\sum_{i=1}^{5}|x_i-\bar{x}|}{5} = \frac{6}{5} = 1.2 \text{ ปี}$$

เนื่องจากเครื่องหมายสัมบูรณ์ไม่สามารถเขียนให้อยู่ในรูปความสัมพันธ์กับตัวสถิติอื่นได้ ดังนั้น ส่วนเบี่ยงเบนเฉลี่ยจึงไม่ค่อยนิยมใช้

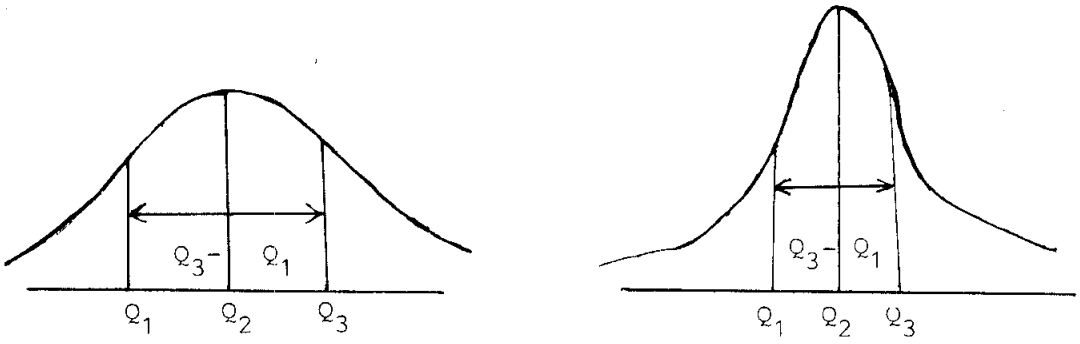
สรุปขั้นตอนในการหา AD

1. หามัชฌิมเลขคณิตของข้อมูลชุดนั้น
2. หาผลต่างของค่าสังเกตทุกตัวกับค่ามัชฌิมเลขคณิต โดยไม่คิดเครื่องหมาย
3. หาผลรวมในข้อ 2
4. หารผลรวมที่ได้ในข้อ 3 ด้วยจำนวนข้อมูลทั้งหมดก็จะได้ค่า AD ตามที่ต้องการ

3. ส่วนเบี่ยงเบนควอร์ไทล์ (Quartile Deviation) ส่วนเบี่ยงเบนควอร์ไทล์ใช้สัญลักษณ์แทนด้วย Q หรือ QD หรือมักจะเรียกกันว่า กึ่งพิสัยควอร์ไทล์ (Semi-interquartile Range) หาได้โดยพิจารณาจากค่าควอร์ไทล์ที่ 1 และ 3 ดังนี้

$$QD = \frac{Q_3 - Q_1}{2}$$

เมื่อ Q_1 และ Q_3 เป็นค่าของควอร์ไทล์ที่ 1 และควอร์ไทล์ที่ 3 ตามลำดับ จากสูตรนี้เป็นการหาความแตกต่างระหว่างค่า Q_1 และ Q_3 ทั้งนี้ เนื่องจากข้อมูลที่มีการกระจายมากค่าของ Q_3-Q_1 จะมีค่ามาก ส่วนข้อมูลที่มีการกระจายน้อยค่าของ Q_3-Q_1 จะมีค่าน้อย ดังรูป



ข้อมูลที่มีการกระจายมาก ข้อมูลที่มีการกระจายน้อย

เนื่องจากค่า QD พิจารณาจากค่า 2 ค่าเท่านั้น จึงไม่เป็นที่นิยมนำไปใช้

4. ความแปรปรวนและส่วนเบี่ยงเบนมาตรฐาน (Variance and Standard deviation) เพื่อให้ได้การวัดการกระจายที่สะดวกกว่าส่วนเบี่ยงเบนเฉลี่ย จึงใช้วิธียกกำลังสองของส่วนเบี่ยงเบนเพื่อให้เป็นบวกหมด แล้วหาค่าเฉลี่ยของค่าที่ยกกำลังสองนี้ ค่าที่ได้นี้เรียกว่า ความแปรปรวน (Variance) ค่าความแปรปรวนที่ได้จะเป็นเลขคงที่จำนวนหนึ่ง และเลขจำนวนนี้จะมีค่าเป็นบวกเสมอ ความแปรปรวนเป็นการวัดการกระจายที่สำคัญเชื่อถือได้และเป็นการวัดการกระจายที่ดีที่สุด สามารถนำไปใช้ในการคำนวณทางสถิติขั้นสูงต่อไปได้ ในการคำนวณหาความแปรปรวนอาจจะหาความแปรปรวนของประชากรทั้งหมด หรือหาความแปรปรวนของตัว

อย่างก็ได้ โดยให้ σ^2 แทนความแปรปรวนของประชากร และ s^2 แทนความแปรปรวนของตัวอย่างซึ่งหาได้ดังนี้

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

เมื่อ x_i = คือค่าของข้อมูลตัวที่ i

μ = ค่าเฉลี่ยของประชากร

N = จำนวนประชากรทั้งหมด

ซึ่งในทางปฏิบัติค่าของ σ^2 มักจะหาไม่ค่อยได้ เพราะจำนวนประชากรมักจะมีขนาดใหญ่ จึงต้องใช้การสุ่มตัวอย่างแทนแล้วนำมาหาความแปรปรวนของตัวอย่าง (s^2) จากสูตร

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

เมื่อ x_i = ค่าของข้อมูลตัวที่ i

$\bar{x}$ = ค่าเฉลี่ยของตัวอย่าง

n = จำนวนตัวอย่าง

ตัวอย่างเช่น จะหาความแปรปรวนของข้อมูลต่อไปนี้

2, 4, 6, 8, 10

จาก $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

$\bar{x} = \frac{2+4+6+8+10}{5} = 6$

$\therefore s^2 = \frac{(2-6)^2+(4-6)^2+(6-6)^2+(8-6)^2+(10-6)^2}{5-1}$

$= \frac{40}{4} = 10$

สรุปขั้นตอนในการหาความแปรปรวน

1. คำนวณหาค่าเฉลี่ย
2. นำค่าเฉลี่ยไปลบออกจากค่าสังเกตแต่ละค่า
3. หากำลังสองของส่วนเบี่ยงเบนแต่ละตัว
4. หาผลรวมของกำลัง 2 ของส่วนเบี่ยงเบน
5. หารผลรวมที่ได้ด้วย $n-1$ สำหรับตัวอย่าง และหารผลรวมที่ได้ด้วย N สำหรับประชากรก็จะได้ความแปรปรวนตามที่ต้องการ

ในกรณีที่มีข้อมูลจำนวนมาก ๆ ในทางปฏิบัติมักจะใช้คำนวณจากสูตรวิธีลัด ดังนี้

$$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}\right]$$

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

ซึ่งค่าที่คำนวณออกมาจะตรงกับสูตรแรก ดังเช่นตัวอย่างข้างต้น

หา x_i^2 ได้ 4, 16, 36, 64 และ 100

และได้ $\sum_{i=1}^{n} x_i = 2+4+6+8+10 = 30$

$$\sum_{i=1}^{n} x_i^2 = 4+16+36+64+100 = 200$$

$\therefore$ แทนค่าจะได้

$$s^2 = \frac{1}{5-1}\left[200 - \frac{(30)^2}{5}\right]$$

$$= \frac{1}{4}(200-180) = \frac{40}{4} = 10$$

ซึ่งค่าที่ได้เท่ากับที่คำนวณโดยวิธีตรง

เนื่องจากในทางปฏิบัติมีความลำบากยุ่งยากในการนำความแปรปรวนไปใช้ 2 อย่าง ด้วยกัน คือ

1. ความแปรปรวนที่ได้มักจะเป็นจำนวนเลขที่มาก เมื่อเปรียบเทียบกับค่าของข้อมูลที่ได้ เช่น ถ้าค่าของข้อมูลเป็นหลัก 1,000 ความแปรปรวนที่ได้จะเป็นหลักล้าน เป็นต้น

2. หน่วยของความแปรปรวนที่ได้จะไม่เหมือนกับหน่วยของค่าของข้อมูล เช่น หน่วยของข้อมูลเป็นฟุต หน่วยของความแปรปรวนจะเป็นฟุต^2 เป็นต้น

แต่อย่างไรก็ตาม ความแปรปรวนก็ยังมีความสำคัญในทางคณิตศาสตร์ หรือทฤษฎีทางสถิติมาก ดังนั้น เพื่อไม่ให้เกิดความยุ่งยากในการตีความหมายในทางปฏิบัติจะใช้ส่วนเบี่ยงเบนมาตรฐาน (Standard deviation) มากกว่า เพราะตีความหมายได้ดีกว่า โดยส่วนเบี่ยงเบนมาตรฐานหาได้จากการถอดรากที่สองของความแปรปรวน และหน่วยของส่วนเบี่ยงเบนมาตรฐานที่ได้ก็เป็นหน่วยเดียวกับหน่วยของข้อมูล และมีหน่วยเดียวกับค่าเฉลี่ยด้วย ส่วนเบี่ยงเบนมาตรฐานมีประโยชน์มากในทางทฤษฎีสถิติชั้นสูง และเป็นตัวที่นิยมใช้มากที่สุดในบรรดาตัวที่ใช้วัดการกระจายของข้อมูลทั้งหมด

ถ้าให้ σ = ส่วนเบี่ยงเบนมาตรฐานของประชากร

$\therefore \quad \sigma = \sqrt{\sigma^2}$ (ใช้เฉพาะค่าบวกเท่านั้น)

และถ้าให้ s = ส่วนเบี่ยงเบนมาตรฐานของตัวอย่าง

$\therefore \quad s = \sqrt{s^2}$ (ใช้เฉพาะค่าบวกเท่านั้น)

หรือเขียนสูตรเต็มๆ ได้ดังนี้

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_1 - \bar{x})^2}{n - 1}}$$

ถ้าเป็นวิธีลัดจะได้

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} x_i^2 - \frac{(\sum_{i=1}^{N} x_i)^2}{N}}{N}}$$

$$S = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1}}$$

จากตัวอย่างการหาความแปรปรวนของข้อมูลข้างต้นที่กล่าวมาแล้วได้

$$s^2 = 10$$

$$\therefore S = \sqrt{10} = 3.16$$

คุณสมบัติของความแปรปรวนและส่วนเบี่ยงเบนมาตรฐาน

ข้อมูลชุดหนึ่ง $x_1, x_2, \ldots, x_n$ ซึ่งมีส่วนเบี่ยงเบนมาตรฐานเป็น s_x^2

1. เมื่อเอาค่าคงที่ a ใดๆ ไปบวกหรือลบกับค่าของข้อมูลทุกตัว ความแปรปรวนหรือส่วนเบี่ยงเบนมาตรฐานที่ได้จะมีค่าเท่าเดิม คือ

$$s_{x \pm a}^2 = s_x^2$$

หรือ $$S_{x \pm a} = S_x$$

2. เมื่อเอาค่าคงที่ a ใดๆ ไปคูณหรือหารค่าของข้อมูลทุกตัว ความแปรปรวนหรือส่วนเบี่ยงเบนมาตรฐานที่ได้จะมีค่าเปลี่ยนแปลงไปจากเดิม คือ

$$s_{ax}^2 = a^2 s_x^2$$

และ $$s_{\frac{x}{a}}^2 = \frac{1}{a^2} s_x^2$$

หรือ $$S_{ax} = |a| S_x$$

และ $$S_{\frac{x}{a}} = \frac{1}{|a|} S_x$$

แบบฝึกหัดที่ 4

1. จากข้อความในแต่ละข้อย่อยต่อไปนี้ ข้อใดเป็นสถิติเชิงพรรณาและข้อใดเป็นสถิติเชิงอนุมาน

ก. การเพิ่มขึ้นของค่าแรงขั้นต่ำเป็นผลให้อัตราเงินเฟ้อสูงขึ้นถึงร้อยละ 5 ในปีที่ผ่านมา

ข. ครัวเรือน 1 ใน 3 ถูกสัมภาษณ์เกี่ยวกับรายได้ต่อปี

ค. ผลการศึกษาในปัจจุบันกล่าวได้ว่าคลอไรด์ที่ผสมในยาสีฟันทำให้ฟันมีคราบหินปูนน้อยลง

ง. จากคนไข้ทั้งหมดที่ได้รับการเปลี่ยนไต ปรากฏว่าร้อยละ 40 ที่อยู่รอด

จ. จากสถิติของมหาวิทยาลัยรามคำแหง พบว่าในช่วง 3 ปีที่ผ่านมา พบว่าร้อยละ 80 ของบัณฑิตภาควิชาสถิติ มีงานทำภายใน 1 ปี หลังจากที่สำเร็จการศึกษา

ฉ. จากสถิติในข้อ (จ) สามารถกล่าวได้ว่าในปีต่อไป จำนวนบัณฑิตของภาควิชาสถิติที่มีงานทำภายใน 1 ปี ภายหลังจบการศึกษามีมากกว่าร้อยละ 80

2. จากข้อมูลต่อไปนี้เป็นข้อมูลแบบจำแนกประเภทหรือข้อมูลแบบลำดับ ถ้า

ก. นับถือศาสนาของคนกลุ่มหนึ่งโดยกำหนดตัวเลขให้เป็น 1, 2, 3, 4 และ 5 แทน พุทธ อิสลาม คริสต์ อื่น ๆ และไม่ได้นับถือศาสนา

ข. นักเครื่องกลต้องการทราบว่าการเปลี่ยนชิ้นส่วนในเครื่องจักรบางชิ้นจะง่ายหรือยาก ซึ่งได้คำตอบออกมาเป็นยากมาก ยาก ธรรมดา ง่าย ง่ายมาก

ค. ลูกค้าจะต้องตอบคำถามว่า เขาชอบชนิด B มากกว่า A ชอบทั้ง 2 ชนิดเท่า ๆ กัน ชอบชนิด A มากกว่าชนิด B และไม่มีความเห็น

3. ข้อมูลต่อไปนี้เป็นข้อมูลแบบ Nominal หรือ Ordinal หรือ Interval หรือ Ratio และจงอธิบายว่าทำไม

ก. จำนวนผู้รักษาความปลอดภัยของสังคม

ข. จำนวนผู้โดยสารบนรถประจำทางที่วิ่งจากกรุงเทพฯ ไปยังขอนแก่น

ค. อุณหภูมิที่วัดเป็น Fahrenheit

ง. ลำดับของความนิยมดื่มน้ำอัดลมยี่ห้อต่าง ๆ

4. คำตอบที่ได้จากแบบสอบถามในเรื่องต่อไปนี้ใช้มาตราการวัดแบบใด

ก. ท่านสูงเท่าใด

ข. ท่านหนักเท่าใด

ค. ท่านมีอาชีพอะไร

ง. วิชานี้เป็นอย่างไร เมื่อเปรียบเทียบกับวิชาอื่น ๆ ที่เรียนมาแล้ว

จ. ท่านชื่ออะไร

ฉ. ระยะทางจากบ้านมายังมหาวิทยาลัยเท่ากับเท่าใด

ช. จำนวนของทารกที่เกิดในเวลาต่าง ๆ กันในวันหนึ่ง ๆ

5. มัชฌิมเลขคณิตมีความหมายอย่างไร

6. จงบอกผลเสียของการใช้มัชฌิมเลขคณิตในการวัดแนวโน้มเข้าสู่ส่วนกลาง

7. เมื่อใดที่มัชฌิมเลขคณิต มัธยฐาน และฐานนิยมมีค่าเท่ากัน

8. จงหาค่ามัชฌิมเลขคณิต มัธยฐาน และฐานนิยมของข้อมูลต่อไปนี้

ก. 2, 4, 5, 6, 6, 6, 9, 10, 13 และ 15

ข. 1, 3, 5, 7, 7, 7, 9, 9, 10, 10, 11 และ 12

9. จงหามัชฌิมเลขคณิตและมัธยฐานจากข้อมูลต่อไปนี้

ก. 7, 9, 2, 1, 5, 4.5, 7.5, 6, 2

ข. 1, 2, 10, 7, 7, 9, 8, 5, 2, 11

ค. 30, 2, 79, 50, 38, 17, 9

ง. 0.011, 0.032, 0.027, 0.035, 0.042

จ. 90, 87, 92, 81, 78, 85, 95, 80

ฉ. 42, 30, 27, 40, 25, 32, 33

10. ถ้า X มีค่าเฉลี่ย 200 จงหาค่าเฉลี่ยของ Y เมื่อ

ก. $Y = X + 20$ ข. $Y = 4X$ ค. $Y = 4X+20$

11. จากการวัดค่าสังเกต 100 ค่าได้ค่ามัชฌิมเลขคณิต 28.31 และจากการวัดค่าสังเกต 150 ค่า ได้ค่ามัชฌิมเลขคณิต 30.27 ถ้ารวม 250 ค่าเข้าด้วยกัน ค่ามัชฌิมเลขคณิตจะน้อยกว่าเท่ากับหรือมากกว่าค่ามัชฌิมเลขคณิตของมัชฌิมเลขคณิตทั้งสอง

12. จากข้อมูลต่อไปนี้

1, 1, 1, 2, 2, 3, 3, 3, 5, 5, 6, 6, 6, 6, 7, 7, 8, 8, 9, 9,

จงคำนวณ

ก. มัธยฐาน ข. ฐานนิยม ค. มัชฌิมเลขคณิต

ง. จงแสดงให้เห็นว่าผลรวมของค่าเบี่ยงเบนเท่ากับศูนย์

จ. ผลรวมของกำลังสองของค่าเบี่ยงเบนจากมัชฌิมเลขคณิต

ฉ. ผลรวมของกำลังสองของค่าเบี่ยงเบนจากมัธยฐาน

ช. ผลรวมของกำลังสองของค่าเบี่ยงเบนจากฐานนิยม

ซ. จงเปรียบเทียบค่าที่ได้ในข้อ จ, ฉ และ ช ว่าข้อใดน้อยที่สุด

13. ค่าสัมบูรณ์ของส่วนเบี่ยงเบนเฉลี่ยเป็นลบหรือเป็นศูนย์ได้หรือไม่จงอธิบาย

14. ค่าของส่วนเบี่ยงเบนมาตรฐานจะเป็นศูนย์ได้หรือไม่จงอธิบาย และเป็นค่าติดลบได้หรือไม่จงอธิบาย

15. มีเงื่อนไขอะไรบ้างที่ทำให้ค่าของความแปรปรวนของตัวแปรมีค่าเท่ากับศูนย์

16. จงหาส่วนเบี่ยงเบนมาตรฐานของรายได้จากการขายสินค้าต่อวันดังนี้ 81, 90, 45, 56, 76, 48, 106

17. จากข้อมูลต่อไปนี้ 2.1, 2.5, 2.7, 2.3, 2.4, 2.0, 2.7, 3.0, 1.4, 2.4 และ 2.8 จงหา

ก. พิสัย

ข. ส่วนเบี่ยงเบนเฉลี่ย

ค. ความแปรปรวนและส่วนเบี่ยงเบนมาตรฐาน

18. ถ้าข้อมูลชุดหนึ่งมีค่ามัชฌิมเลขคณิตเท่ากับ 200 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 10 จงคำนวณส่วนเบี่ยงเบนมาตรฐานของ Y เมื่อ

ก. $Y = X + 10$ ข. $Y = 5X$ ค. $Y = 7X + 3$

19. จงคำนวณหาส่วนเบี่ยงเบนมาตรฐานของข้อมูลดังต่อไปนี้

ก. 10, 8, 6, 0, 8, 3, 2, 2, 8, 0

ข. 1, 3, 3, 5, 5, 5, 7, 7, 9

ค. 20, 1, 2, 5, 4, 4, 4, 0

ง. 5, 5, 5, 5, 5, 5, 5, 5, 5, 5

20. จงคำนวณค่า s^2 และ s ของข้อมูล 3, 4, 5, 6, 7

ก. ถ้าบวกข้อมูลแต่ละตัวด้วย 2

ข. ถ้านำข้อมูลแต่ละตัวมาลบออกจาก 2

ค. ถ้านำข้อมูลแต่ละตัวคูณด้วย 2

ง. ถ้านำข้อมูลแต่ละตัวมาหารด้วย 2