บทที่ 4 ประชากร ตัวอย่าง และการแจกแจงของตัวสถิติ

The man who does not understand the laws of Sampling and Probability is intellectually crippled in today's world.

Robert Sears

ในทางสถิติเรามีวิธีเก็บรวบรวมข้อมูลสถิติกันหลายวิธี แต่ถ้าแบ่งตามลักษณะของวิธีการที่ต้องปฏิบัติแล้วจะได้เป็น 3 วิธีใหญ่ ๆ ดังนี้

(1) วิธีการเก็บรวบรวมข้อมูลจากทะเบียนและระเบียน (Registration and Record) เช่น การเก็บข้อมูลเกี่ยวกับรถยนต์หรือเรือยนต์อาจเก็บได้จากหลักฐานการจดทะเบียนซึ่งเจ้าของรถยนต์หรือเรือยนต์ต้องกรอกไว้ที่หน่วยงานที่ทำทะเบียน ข้อมูลเกี่ยวกับที่เกิดอายุ และเพศของนักศึกษาอาจเก็บได้จากหลักฐานที่ทำไว้กับแผนกทะเบียนของมหาวิทยาลัย เป็นต้น

(2) วิธีการเก็บรวบรวมข้อมูลจากการทดลอง (Experiment) การเก็บรวบรวมข้อมูลโดยวิธีนี้ต้องอาศัยการทดลอง ซึ่งอาจจะเป็นการทดลองในห้องปฏิบัติการ หรือนอกห้องปฏิบัติการก็ได้ เช่น การทดลองเปรียบเทียบพันธ์ข้าว เป็นต้น

(3) วิธีการเก็บรวบรวมข้อมูลจากการสำรวจ (Survey) วิธีการนี้แบ่งออกเป็น 2 วิธีใหญ่ ๆ คือ วิธีสำมะโน และการสำรวจด้วยตัวอย่าง การเก็บรวบรวมข้อมูลโดยวิธีการนี้อาจทำได้หลายวิธี แต่ที่นิยมใช้กันก็คือ การส่งเจ้าหน้าที่ถือแบบสำรวจไปสัมภาษณ์ผู้ที่อยู่ในข่ายที่จะต้องให้ข้อมูล โดยตรงเช่น หัวหน้าครอบครัวเจ้าของโรงงานอุตสาหกรรมเป็นต้น อีกวิธีหนึ่งที่นิยมใช้กันก็คือ การส่งแบบสำรวจหรือแบบข้อถามไปให้ผู้ที่อยู่ในข่ายที่จะสำรวจนั้นกรอกข้อมูลและส่งกลับมาเอง

4.1 วิธีสำมะโน (Census)

วิธีสำมะโนเป็นวิธีการสำรวจหรือเก็บรวบรวมข้อมูลด้วยวิธีแจงนับอย่างครบถ้วน (Complete Enumeration Method) จากแหล่งของข้อมูลที่เรียกว่าหน่วยแจงนับ (Enumeration Unit) ในการสำมะโนต้องกำหนดหน่วยแจงนับว่ามีขอบข่ายของการเก็บรวบรวมข้อมูลสถิติครอบคลุมถึงหน่วยแจงนับไหนบ้าง ขอบข่ายนั้นเรียกว่าคุ้มรวม (Coverage) ของการเก็บข้อมูล สำหรับหน่วยแจงนับทั้งหมดที่อยู่ในคุ้มรวมจะเรียกว่า ประชากรหรือจักรวาล (Population or Universe)

ดังนั้นวิธีของการสำมะโนจึงกำหนดไว้ว่า ข้อมูลสถิติที่เก็บรวบรวมได้จะต้องได้มาจากการแจงนับทุกหน่วยในคุ้มรวม เช่น ในการสำมะโนประชากรของประเทศเรากำหนด "ครัวเรือน" เป็นหน่วยแจงนับและมีคุ้มรวมไปถึงทุกครัวเรือนในประเทศ

วิธีสำมะโนมีข้อดีเด่น ดังนี้

(1) ยอดข้อมูลสถิติที่รวบรวมได้ในทุกเรื่องสามารถแสดงออกในเขตบริหาร หรือเขตภูมิศาสตร์ที่เล็กที่สุดได้ ทั้งนี้เพราะมีข้อมูลจากทุกหน่วยแจงนับในคุ้มรวม

(2) ข้อมูลสถิติที่เป็นข้อมูลหลักนั้นสามารถจะนำไปช่วยในการวางแผนการเก็บข้อมูลอื่น ๆ ได้อีก

ส่วนข้อเสียนั้นมีอยู่มาก ดังนี้

(1) ใช้ทรัพยากร (กำลังคน, งบประมาณ) มาก

(2) เสียเวลามากเพราะมีหน่วยแจงนับมาก ปริมาณงานจึงมาก ทำให้เสร็จทันท่วงทีไม่ได้

(3) คุณภาพของข้อมูลที่รวบรวมได้ยังเป็นที่น่าสงสัย เพราะคนทำมากและต้องรีบทำให้ทันคาบเวลาที่กำหนด

(4) ผู้ให้คำตอบบางรายไม่สามารถให้ข้อมูลที่ถูกต้อง

4.2 การสำรวจด้วยตัวอย่าง (Sample Survey)

การสำรวจด้วยตัวอย่าง เป็นการสำรวจโดยวิธีแจงนับเพียงบางส่วน (Partial Enumeration Method) ซึ่งมีลักษณะคล้ายคลึงกับวิธีสำมะโน แต่ผิดกันตรงที่หลังจากกำหนดข้อมูลที่จะเก็บ กำหนดคุ้มรวม และหน่วยแจงนับ แล้วการแจงนับจะทำที่หน่วยแจงนับตัวอย่างเป็นบางหน่วยเท่านั้นการสำรวจด้วยตัวอย่างนี้เป็นระเบียบวิธีสำรวจที่พัฒนาขึ้นมาเพื่อแก้ปัญหาของข้อเสียในวิธีสำมะโนทั้งนี้ก็ได้อาศัยทฤษฎีทางคณิตศาสตร์สถิติ และทฤษฎีความน่าจะเป็นช่วยในการวางแผนสำรวจและการประมาณผล

หลักปฏิบัติของการสำรวจด้วยตัวอย่าง กล่าวได้ง่าย ๆ ดังนี้ (1) สุ่มตัวอย่างหน่วยแจงนับเพียงบางส่วนจากประชากร (2) เก็บรวบรวมข้อมูลจากหน่วยแจงนับตัวอย่างนั้น และ (3) ประมาณผลหรืออนุมานยอดข้อมูลที่ต้องการ ซึ่งถือว่าเป็นยอดข้อมูลของประชากร

ประโยชน์ของการสำรวจด้วยตัวอย่างนั้นเมื่อเปรียบเทียบกับวิธีสำมะโนและเหตุผลอย่างอื่นแล้วจะมีดังนี้

(1) ลดค่าใช้จ่าย (Reduced Cost) ในการเก็บข้อมูล เพราะหน่วยแจงนับมีน้อยจึงเสียค่าใช้จ่ายน้อย

(2) ทุ่นเวลาในการทำงาน (Greater Speed) เมื่อมีข้อมูลน้อยการประมาณผลให้อยู่ในข้อมูลสำเร็จรูปก็ทำได้รวดเร็วกว่า ซึ่งจะทันต่อความต้องการใช้ข้อมูลเหล่านั้น

(3) รวบรวมข้อมูลได้กว้างขวางกว่า (Greater Scope) เมื่อหน่วยแจงนับมีน้อยและใช้พนักงานที่มีความสามารถสูง ทำให้รวบรวมข้อมูลในหลายเรื่องได้

(4) ความถูกต้องของข้อมูลมีมาก (Greater Accuracy) เมื่อปริมาณงานในการแจงนับหรือประมวลผลน้อย ก็ใช้เจ้าหน้าที่น้อยแต่มีความสามารถสูง งานจึงถูกต้อง

(5) การศึกษาหรือตรวจสอบบางอย่างต้องทำลาย (Destructive Nature) เช่นตรวจสอบ

ความแข็งแกร่งของสิ่งของบางอย่างเป็นต้น ดังนี้จึงจำเป็นต้องใช้การสำรวจด้วยตัวอย่าง

(6) ในทางธรรมชาติไม่สามารถศึกษาประชากรได้ทั้งหมด เช่น สัตว์ในป่า ปลาในน้ำ เป็นต้น จึงจำเป็นต้องใช้การสำรวจด้วยตัวอย่าง

ส่วนข้อเสียเปรียบสำหรับวิธีการสำรวจด้วยตัวอย่างนั้นก็มีอยู่ว่า

(1) ไม่สามารถประมวลข้อมูลในระดับย่อย ๆ หรือท้องที่เล็ก ๆ ได้

(2) ข้อมูลที่ประมวลขึ้นจะมีความคลาดเคลื่อนในข้อมูลที่เรียกว่า ความคลาดเคลื่อนจากการสุ่มตัวอย่าง (Sampling Error)

4.3 การวิเคราะห์ข้อมูลโดยวิธีการสรุปย่อข้อมูล (Condensation and Summarization of Data)

การวิเคราะห์ข้อมูลมีความหมายเด่นชัดยิ่งขึ้น และสามารถนำไปใช้เป็นพื้นฐานในการอธิบายและหาข้อสรุปเกี่ยวกับเรื่องหรือสภาพการณ์ต่าง ๆ ที่สามารถอธิบายจากข้อมูลซึ่งมีการเก็บรวบรวมมาได้ ในการวิเคราะห์ข้อมูลจำนวนมากนั้นเราจะต้องอาศัยเครื่องจักรประมวลผลสมัยใหม่ เช่น คอมพิวเตอร์เข้าช่วย

ในการวิเคราะห์ข้อมูลเรามีเทคนิคการวิเคราะห์หลายอย่าง ซึ่งแต่ละอย่างมีความเหมาะสมที่ใช้ไม่เหมือนกัน แล้วแต่ว่าข้อมูลที่เก็บรวบรวมมาได้นั้นเป็นข้อมูลเกี่ยวกับอะไร และเรากำลังศึกษาในเรื่องอะไร แต่เทคนิคการวิเคราะห์ข้อมูลพื้นฐานที่เราใช้กันบ่อย ๆ ก็คือ การวิเคราะห์ข้อมูลโดยวิธีการสรุปย่อข้อมูล

การวิเคราะห์ข้อมูลโดยการสรุปย่อที่สำคัญและใช้กันอย่างแพร่หลายมากที่สุดก็มี การทำการแจกแจงความถี่ และการหาค่าทางสถิติ

4.3.1 การสรุปข้อมูลโดยการทำการแจกแจงความถี่ (Frequency distribution) การแจกแจงความถี่เป็นวิธีการที่ใช้ในการจัดข้อมูลที่มีอยู่หรือที่เก็บรวบรวมมาได้ให้อยู่เป็นพวก เพื่อสะดวกในการวิเคราะห์ข้อมูลแบบอื่น เช่น การวิเคราะห์ค่าเฉลี่ย ความแปรปรวนเป็นต้น โดยทั่ว ๆ ไป การแจกแจงความถี่จะทำกันเมื่อข้อมูลที่จะทำการศึกษา หรือวิเคราะห์มีเป็นจำนวนมาก หรือมีค่าตัวเลขซ้ำกันอยู่มาก เพราะจะช่วยให้ประหยัดเวลาและสรุปผลได้รัดกุม สะดวก และเหมาะสมที่จะนำไปใช้ให้เป็นประโยชน์ต่อไป เมื่อทำการแจกแจงความถี่แล้วเราจะได้ตารางความถี่ (Frequency Table) เช่น การวิเคราะห์ต้นทุนการผลิตข้าวของชาวนาซึ่งสำรวจมาจากชาวนาจำนวนมากพอสมควรโดยระเบียบวิธีการสำรวจด้วยตัวอย่างเมื่อทำการวิเคราะห์การแจกแจงความถี่จะได้ตารางแจกแจงความถี่ดังนี้

ช่วงของต้นทุนต่อไร่ (บาท)	ความถี่
ต่ำกว่า 200	f_1
200-249	f_2
250-299	f_3
300-349	f_4
350-399	f_5
400-499	f_6
500-549	f_7
500 และสูงกว่า	f_8
รวม	$n = f_1 + f_2 + \ldots\ldots + f_8$

ในการสร้างตารางแจกแจงความถี่เรามีขั้นตอนดังนี้

(1) พิจารณาข้อมูลที่มีค่าสูงสุด $X_{(n)}$ และค่าต่ำสุด $X_{(1)}$ และหาพิสัย (Range, R) นั่นคือ

พิสัย = ข้อมูลที่มีค่าสูงสุด - ข้อมูลที่มีค่าต่ำสุด

หรือ $R = X_{(n)} - X_{(1)}$

(2) แบ่งพิสัยออกเป็นจำนวนชั้น (Class) ที่เหมาะสมซึ่งโดยทั่วไปมักจะใช้ระหว่าง 6-25 ชั้น แล้วแต่ว่าเราต้องการรายละเอียดในการวิเคราะห์ การแจกแจงความถี่มากน้อยแค่ไหน และขึ้นอยู่กับจำนวนข้อมูลที่มีอยู่อีกอย่างหนึ่ง ซึ่ง H.A.Sturges เสนอไว้ว่า

$k = 1 + 3.3 \log_{10} n$ ในเมื่อ k เป็นจำนวนชั้นที่ต้องการ และ n เป็นจำนวนข้อมูลทั้งหมด

(3) กำหนดขนาดหรือช่วง หรือความกว้างของอันตรภาคชั้น (Class interval)

ให้มีความต่อเนื่องกันจากน้อยไปหามาก โดยใช้หลักเกณฑ์ที่ว่าขนาดของอันตรภาคชั้นไม่ควรจะกว้างหรือแคบจนเกินไป เพราะถ้ากว้างเกินไปจะทำให้มองไม่เห็นภาพการแจกแจงที่ชัดเจนนัก แต่ถ้าเล็กเกินไปก็อาจจะทำให้เกิดความไม่สม่ำเสมอ หรือความไม่เป็นไปตามปกติธรรมชาติของการแจกแจงของลักษณะที่เราสนใจไปได้เช่นกัน ขนาดของอันตรภาคชั้นสามารถกำหนดได้จากสูตรที่ Sturges เสนอไว้ในเทอมของพิสัยและจำนวนชั้นดังนี้

$i = R/k$

ในเมื่อ i เป็นความกว้างของอันตรภาคชั้น R และ k เป็นพิสัยและจำนวนชั้น

(4) พิจารณาขีดจำกัดล่าง (Lower class Limit) ของชั้นแรกโดยการกะเอาหรือใช้สูตรดังนี้

ขีดจำกัดล่าง $= X_{(1)} - (k\,i - R)/2$

(5) พิจารณาจำนวนข้อมูลที่อยู่ในแต่ละชั้น หรือหาความถี่ของแต่ละชั้นนั่นเอง ใน

การหาความถี่ของชั้นมักจะอาศัยรอยขีด (tally) หรือแผ่นคะแนน (Score Sheet))

ตัวอย่าง ในการศึกษาต้นทุนของการผลิตหนังสือโดยอาศัยการสำรวจด้วยตัวอย่าง ได้ข้อมูลมาดังนี้

8.25	8.95	7.25	10.50	9.95	7.50	4.75	7.95
8.75	14.95	6.70	7.50	8.25	11.00	1.75	3.50
1.45	4.75	1.50	3.40	5.40	7.00	1.75	9.50
3.60	6.95	6.50	8.75	8.75	3.25	1.40	7.50
8.00	2.75	10.50	5.95	7.95	14.75	1.65	12.50

เมื่อเราจะวิเคราะห์ข้อมูลเหล่านี้โดยการทำการแจกแจงความถี่ เราก็จะทำได้ดังนี้

(1) ข้อมูลที่มีค่ามากสุด และต่ำสุดเป็น 14.95 และ 1.40 พิสัย (R) = 14.95 - 1.40 = 13.55

(2) จำนวนชั้น $k = 1 + 3.3 \log_{10}(40) = 6.30$ ดังนั้นชั้นทั้งหมดจะเป็น 7 ชั้น

(3) ความกว้างของอันตรภาคชั้นจะเป็น $13.55/7 \cong 2$

(4) ขีดจำกัดล่างของชั้นแรกโดยการกะเอาซึ่งจะให้เป็น 1.00-2.99

(5) พิจารณาความถี่ของแต่ละชั้นจะได้ตารางแจกแจงความถี่ ดังตารางต่อไปนี้

ราคา	รอยขีด	จุดกลาง	ราคา	จำนวนหนังสือ
1.00-2.99	𝍸 ||	1.995	1.00-2.99	7
3.00-4.99	𝍸 |	3.995	3.00-4.99	6
5.00-6.99	𝍸	5.995	5.00-6.99	5
7.00-8.99	𝍸 𝍸 |||	7.995	7.00-8.99	14
9.00-10.99	|||	9.995	9.00-10.99	4
11.00-12.99	|	11.995	11.00-12.99	2
13.00-14.99	|	13.995	13.00-14.99	2

ขอบเขตชั้น(class boundaries) ของแต่ละชั้นจะเป็น 0.995-2.995, 2.995-4.995,..... 12.995-14.995

4.3.2 การสรุปข้อมูลโดยวิธีการหาค่าทา งสถิติที่เป็นค่าวัดลักษณะสำคัญของข้อมูล

ค่าทางสถิติที่ได้จากการแจงนับทุกหน่วยในประชากรหรือจากการสำมะโนจะเรียกว่า "มาตรประชากร" หรือพารามิเตอร์ (Parameter) ซึ่งเป็นค่าคงที่ที่แสดงคุณลักษณะของประชากร ดังนั้นพารามิเตอร์ จึงเป็นฟังก์ชันของค่าสังเกตหรือข้อมูลจากหน่วยแจงนับทุกหน่วยในประชากร เราจะใช้อักษรกรีกแทนพารามิเตอร์ สำหรับพารามิเตอร์ที่สนใจกันส่วนมากก็มี

(1) ยอดรวม (Population Total) ถ้า Y เป็นคุณลักษณะที่เราต้องการศึกษาของหน่วย

แจงนับที่ i ในประชากรซึ่งมีทั้งหมด N หน่าย แล้วยอดรวมของ Y_i ทั้งหมด หรือ τ กำหนดไว้ว่า

$$\tau = {}_1 + Y_2 + \ldots\ldots + Y_n = \sum_{i=1}^{n} Yi$$

สำหรับยอดรวมที่น่าสนใจเช่น จำนวนเครื่องรับวิทยุในภาคหนึ่ง (หรือจังหวัดหนึ่ง) จำนวนผลผลิตข้าวทั้งประเทศ จำนวนข้าวเปลือกที่ใช้ในการบริโภคภายในประเทศ เป็นต้น

(2) ค่าเฉลี่ย (Population Average or Mean) ถ้า Y_i เป็นคุณลักษณะที่ต้องการศึกษาของหน่วยแจงนับที่ i ในประชากร N หน่วยแล้วค่าเฉลี่ยของ Y ทั้งหมด หรือ μ เป็นดังนี้

$$\mu = \frac{1}{N} \sum_{i=1}^{N} Y_i = \tau/N$$

ดังนั้นยอดรวม $\tau = N\mu$

(3) ความแปรปรวนและส่วนเบี่ยงเบนมาตรฐาน (Population Variance and Standard Deviation) ถ้า Y_i เป็นคุณลักษณะที่ต้องการศึกษาของหน่วยแจงนับที่ i ในประชากร N หน่วย แล้ว ความแปรปรวนและส่วนเบี่ยงเบนมาตรฐานหรือ σ^2 และ σ จะเป็น

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

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

(4) สัดส่วนหรือเปอร์เซ็นต์ (Pouplation Proportion or Percentage) ให้ $U_1, U_2, \ldots, U_N$ เป็นหน่วยแจงนับในประชากรขนาด N และให้ A เป็นจำนวนหน่วยที่มีคุณลักษณะที่เราสนใจ แล้วสัดส่วนของประชากรที่มีคุณลักษณะที่สนใจ หรือ π จะเป็น

$$\pi = A/N$$

เปอร์เซนต์หรือร้อยละเป็นค่าที่ใช้วิเคราะห์สภาพการณ์ หรือใช้เป็นเครื่องชี้หรือเครื่องวัดภาวะการณ์ต่าง ๆ จากข้อมูลได้ดีค่าหนึ่ง อย่างไรก็ดี การใช้เปอร์เซนต์สรุปเรื่องราวบางอย่างจากข้อมูลควรจะต้องระมัดระวังพอสมควร เพราะถ้าใช้อย่างฟุ่มเฟือยไม่ระมัดระวังอาจทำให้เกิดตีความผิดจากข้อมูลสถิติได้มากเช่นกัน โดยเฉพาะใช้เปอร์เซนต์สรุปสภาพการณ์จากข้อมูลจำนวนน้อย ๆ การวิเคราะห์ข้อมูลโดยใช้เปอร์เซนต์นั้นต้องใช้สรุปสภาพการณ์จากข้อมูลจำนวนมาก ๆ เช่น การวิเคราะห์ภาวะการถือครองที่ดินของเกษตรกรในจังหวัด การวิเคราะห์อัตราการว่างงานของประชากรในเมืองใหญ่ ๆ การวิเคราะห์ผลการหยั่งเสียงประชามติของคนจำนวนมาก การวิเคราะห์ข้อมูลเกี่ยวกับสาเหตุการตายของประชากร การวิเคราะห์องค์ประกอบของรายได้รายจ่ายของประเทศ หรือท้องถิ่น เป็นต้น

(5) อัตราส่วน (Population Ratio) ถ้าให้ X_i เป็นคุณลักษณะหนึ่ง และ Y_i เป็นอีกคุณลักษณะหนึ่งในหน่วยแจงนับที่ i ของประชากรที่มี N หน่วยอัตราส่วนของยอดรวมทั้งสองคุณลักษณะ

หรือ η จะกำหนดไว้ว่า

$$\eta = \tau_y / \tau_x = \mu_y / \mu_x$$

ในเมื่อ τ_y กับ τ_x เป็นยอดรวมของ Y_i และ X_i ตามลำดับ และ μ_y กับ μ_x เป็นค่าเฉลี่ยของ Y_i กับ X_i ตามลำดับ ในวงการสถิติมีอัตราส่วนที่ใช้วิเคราะห์สภาพการณ์ของส่วนรวม เช่น ภาคหรือประเทศอยู่หลายตัว เช่น

อัตราส่วนระหว่างเพศของประชากร = ประชากรชาย : ประชากรหญิง

$$\text{อัตราการเกิด} = \frac{\text{จำนวนประชากรที่เกิดใหม่ในช่วงปี}}{\text{จำนวนประชากรที่มีอยู่ทั้งหมด}} \ (1000)$$

$$\text{อัตราการตาย} = \frac{\text{จำนวนประชากรที่ตายไปในช่วงปี}}{\text{จำนวนประชากรที่มีอยู่ทั้งหมด}} \ (1000)$$

$$\text{ผลผลิตต่อไร่} = \frac{\text{ผลผลิตทั้งหมด}}{\text{เนื้อที่ที่ปลูก}}$$

สำหรับค่าทางสถิติที่ได้จากการแจงนับเพียงบางส่วนหรือได้จากการสำรวจด้วยตัวอย่าง เราจะเรียกว่า ตัวสถิติ (Statistic) ดังนั้นตัวสถิติจึงเป็นฟังก์ชั่นของค่าสังเกตจากหน่วย ตัวอย่างและไม่ขึนอยู่กับพารามิเตอร์ที่ไม่ทราบค่าแน่นอน เราจะใช้อักษรโรมันตัวใหญ่แทนตัวสถิติ ตัวสถิติบางตัวที่ใช้เป็นตัวประมาณค่าของพารามิเตอร์จะเรียกว่า ตัวประมาณค่า (Estimator) ตัวสถิติที่ใช้กันบ่อย ๆ มี

(1) ยอดรวม (Sample Total) $Y = \sum_{i=1}^{n} Y_i$

ในเมื่อ n เป็นขนาดตัวอย่าง (Sample Size)

(2) ค่าเฉลี่ย (Sample Mean) $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

(3) ความแปรปรวนและส่วนเบี่ยงเบนมาตรฐาน (Sample Variance and Standard Deviation)

$$S^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2/(n-1)$$

$$S = \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2/(n-1)}$$

(4) สัดส่วน (Sample Proportion) $P = a / n$; a เป็นจำนวนหน่วยที่มีคุณลักษณะที่สนใจจากตัวอย่าง n หน่วย

(5) อัตราส่วน (Sample Ratio) $R = \bar{Y}/\bar{X} = Y/X$

4.4 การคำนวณตัวสถิติที่สำคัญบางตัว

ในการศึกษาประชากรเพื่อที่จะค้นหาข้อมูลข่าวสารเกี่ยวกับประชากรนั้นส่วนมากเราอาศัยการสำรวจด้วยตัวอย่าง นั่นก็คือ เราจะอาศัยตัวอย่างสุ่ม (Random Sample) เป็นเครื่องมือ

ในการอนุมานหรืออ้างอิงเกี่ยวกับประชากร เช่น สมมติว่าเราต้องการจะสรุปผลเกี่ยวกับเปอร์เซนต์ของคนในประเทศไทยที่ชอบดื่มกาแฟยี่ห้อหนึ่ง ถ้าจะถามคนไทยทุกคนเพื่อหาเปอร์เซนต์ที่แท้จริงเราก็ไม่สามารถจะทำได้ ดังนั้นเราจึงต้องใช้ตัวแทนหรือตัวอย่างสุ่มขนาดใหญ่ และเปอร์เซนต์จากตัวอย่างนั้นเราจะใช้อ้างอิงหรืออนุมานเกี่ยวกับเปอร์เซนต์ที่แท้จริง สำหรับตัวอย่างสุ่มเรากำหนดไว้ว่า

สำหรับประชากร X นั้น ตัวอย่าง $X_1, X_2, \ldots, X_n$ จะเป็นตัวอย่างสุ่มขนาด n ของประชากร X ก็ต่อเมื่อการแจกแจงร่วมของตัวอย่างสุ่ม หรือ $f(x_1, x_2, \ldots, x_n)$ กำหนดไว้ดังนี้

$$f(x_1, x_2, \ldots, x_n) = f(x_1)\, f(x_2) \ldots f(x_n)$$

ค่าทางสถิติที่เป็นฟังก์ชั่นของค่าสังเกตจากตัวอย่างสุ่มนั้นเราจะเรียกว่า ตัวสถิติ โดยที่ประชากรหนึ่ง นั้นเราจะมีตัวอย่างสุ่มที่เป็นไปได้มากมาย ดังนั้นตัวสถิติจึงมีค่าผันแปรไปตามตัวอย่างสุ่มนั่นคือ ตัวสถิติจะเป็นตัวแปรเชิงสุ่มด้วย ตัวสถิติสำคัญบางตัวที่น่าจะกล่าวถึงมีดังนี้

4.4.1 ค่าเฉลี่ย (Statistical Average) ในทางสถิติถือว่าค่าเฉลี่ยเป็นค่าที่วัดสภาพปานกลางหรือแนวโน้มเข้าสู่ส่วนกลาง (Measure of Central Tendency) ของข้อมูล เพราะบางครั้งเราจำเป็นต้องสรุปหรือย่อข้อมูลจำนวนมาก ๆ โดยการหาตัวแทนของข้อมูลที่เหมาะสมเพียงบางตัวขึ้นมาซึ่งก็พอที่จะทำให้ผู้สนใจในตัวเลขหรือข้อมูลนั้นได้ข้อสรุปไปโดยไม่ต้องเอาข้อมูลทั้งหมดไปดู เช่นผลผลิตเฉลี่ย ขนาดเฉลี่ยของครัวเรือน รายได้เฉลี่ยของครัวเรือน รายจ่ายเฉลี่ยต่อเดือน เป็นต้น

ในทางสถิติเรามีวิธีการหาค่าเฉลี่ยเพื่อใช้สรุปย่อข้อมูลได้หลายวิธี แต่ที่สำคัญและใช้ได้มากกับการวิเคราะห์ค่าเฉลี่ยข้อมูลทั่ว ๆ ไป ได้แก่ค่าเฉลี่ยเลขคณิต (Arithmetic Mean) มัธยฐาน (Median) และฐานนิยม (Mode) ส่วนค่าเฉลี่ยอื่น ๆ ที่ใช้กันอยู่บ้างในบางกรณีก็มีเปอร์เซนต์ไทล์ (Percentile) กึ่งพิสัย (Midrange) ค่าเฉลี่ยเรขาคณิต G eometric Mean) ค่าเฉลี่ยกำลังสอง (Quadratic Mean) ค่าเฉลี่ยฮาร์โมนิก (Harmo nic Mean)

(1) ค่าเฉลี่ยเลขคณิตตัวอย่าง (Sample Arithematic Mean) ค่าเฉลี่ยเลขคณิตใช้สรุปข้อมูลตัวอย่างจากประชากรที่เป็นแบบปกติหรือสมมาตร และข้อมูลนั้นเป็นแบบอันตรภาคหรือแบบอัตราส่วน เราใช้ $\bar{X}$ แทนค่าเฉลี่ยเลขคณิตตัวอย่าง และกำหนดไว้ดังนี้

ก. เมื่อข้อมูลตัวอย่างไม่ได้จัดกลุ่ม หรือเป็นข้อมูลดิบ แล้วค่าเฉลี่ยเลขคณิตตัวอย่าง $\bar{X}$ จะเป็น

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

ตัวอย่าง ตัวอย่างสุ่มขนาด 6 ให้ค่าสังเกตดังนี้ 20, 27, 25, 24, 26, 22
ดังนั้นค่าเฉลี่ยเลขคณิตตัวอย่าง $\bar{X}$ จะเป็น

$\bar{X} = (20 + 27 + ... + 22)/6 = 24$

ข. เมื่อข้อมูลตัวอย่างมีความถี่เราจะหาค่าเฉลี่ยเลขคณิตได้เป็น

$$\bar{X} = \sum_{i=1}^{k} f_i x_i / n$$

ในเมื่อ $n = \sum f_i$

ตัวอย่าง จากการสอบถามราคาข้าวสารชนิดหนึ่งจากร้านขายข้าวสาร 75 ร้าน ปรากฏว่าเป็นดังนี้

X (ราคา)	f (จำนวนร้าน)	fx
50	10	500
51	15	765
53	20	1060
55	15	825
56	10	560
58	3	174
60	2	120
	รวม n = 75	4004

$\bar{X} = 4004/75 = 53.39$

(1) ในกรณีที่ข้อมูลตัวอย่างสุ่มมีน้ำหนัก (weight) มาเกี่ยวข้อง แล้วค่าเฉลี่ยเลขคณิตจะกำหนดไว้ดังนี้

$$\bar{X} = \sum_{i=1}^{k} w_i X_i / \sum_{i=1}^{k} w_i$$

ในเมื่อ w_i เป็นตัวถ่วงน้ำหนัก

ตัวอย่าง สมมติว่าข้อมูลที่เกี่ยวกับรายจ่ายเป็นเปอร์เซนต์ของครอบครัวทั่ว ๆ ไป เป็นดังนี้

อาหาร 20 % ที่พักอาศัย 25 %

เสื้อผ้า 15 % บริการ 40 %

ถ้าราคาของแต่ละประเภทเพิ่มขึ้นเป็นดังนี้

อาหาร 5% ที่พักอาศัย -1 %

เสื้อผ้า 2 % บริการ 8 %

การเพิ่มขึ้นของราคาโดยเฉลี่ยหาได้ดังนี้

$$\bar{X} = \frac{20(5) + 15(2) + 25(-1) + 40(8)}{20 + 15 + 25 + 40} = 4.25$$

สำหรับค่าสังเกตของตัวอย่างสุ่มที่มีการแจกแจงความถี่เราจะได้เฉลี่ยเลขคณิต $\bar{X}$ ดังนี้

$$\bar{X} = A + i\bar{d} \qquad = A + i(\Sigma fd / n)$$

ในเมื่อ A เป็นค่าเฉลี่ยที่สมมติ ซึ่งจะเป็นจุดกลางของช่วงที่ d = 0, d_i เป็นค่าที่แปลงไปจากค่าสังเกต นั่นคือ $d_i = (X_i - A)/i$, X_i เป็นจุดกลางของชั้น และ i เป็นความกว้างของอัตรภาคชั้น

ตัวอย่าง ตัวอย่างสุ่มขนาด 20 ของรายได้ต่อวันจากการขายสินค้าชนิดหนึ่งเป็นดังนี้

รายได้	f: จำนวนวัน	d	fd
5-14	3	2	-6
15-24	5	1	-5
25-34	6	0	-
35-44	4	1	4
45-54	2	2	4
รวม	20		-3

เมื่อกำหนด A = 29.50 เป็นจุดกลางของชั้น 25-34 และมี i = 10 เราจะได้ค่าเฉลี่ย $\bar{X}$ (โดยอาศัยตารางข้างบน) ดังนี้

$$\bar{d} = \Sigma fd/n = -3/20 = -0.15$$

$$\bar{X} = 29.50 + (10)(-0.15) = 28.00$$

ในการวิเคราะห์ข้อมูลหลาย ๆ ชุด ที่หาค่าเฉลี่ยเลขคณิตของแต่ละชุดไว้แล้ว หากต้องการทราบค่าเฉลี่ยเลขคณิตของข้อมูลทั้งหมดโดยนับรวมเป็นชุดเดียวกัน ก็จะหาได้จากสูตรที่กำหนดไว้ดังนี้

$$\bar{X} = \Sigma_{i=1}^{k} n_i \bar{X}_i / \Sigma_{i=1}^{k} n_i$$

ตัวอย่าง ในการศึกษาค่าจ้างเฉลี่ยต่อวันของกรรมกรได้ข้อมูลสรุปมาดังนี้

ประเภทของกรรมกร	ขนาดตัวอย่าง	ค่าจ้างเฉลี่ยต่อวัน
ไร้ทักษะ	100	30
กึ่งทักษะ	150	30
ทักษะ	50	90

ดังนั้นค่าจ้างเฉลี่ยของกรรมกรทุกประเภทจะเป็น

$$\bar{X} = \{100(30) + 150(60) + 50(90)\}/(100+150+50)$$

$$= 16500/300 = 55$$

(2) มัธยฐานตัวอย่าง (Sample Median) มัธยฐานตัวอย่างใช้ สรุปข้อมูลตัวอย่างจากประชากรที่ไม่เป็นแบบปกติหรือไม่สมมาตร หรือข้อมูลตัวอย่างที่เป็นแบบอันดับอันตรภาค และอัตราส่วน ให้ X เป็นมัธยฐานตัวอย่างกำหนดไว้ดังนี้

ถ้า $X_1, X_2, \ldots, X_n$ เป็นตัวอย่างสุ่มขนาด n จากประชากรซึ่งเรียงลำดับค่าสังเกตจากน้อยไปมากเป็น $X_{(1)}, X_{(2)}, X_{(n)}$ ซึ่งเรียกว่าถ้าสถิติอันดับ (Order Statistic) ที่ 1, 2, ... , n ตามลำดับ แล้วมัธยฐานตัวอย่าง $\bar{X}_m$ กำหนดไว้ดังนี้

$$\bar{X}_m = X_{(k)} \text{ ถ้า n เป็นเลขคี่ และ } k = (n+1)/2$$

$$= \frac{1}{2}(X_{(k)} + X_{(k+1)}) \text{ ถ้า n เป็นเลขคู่ และ } k = n/2$$

ดังนั้นมัธยฐานจะเป็นค่าที่แบ่งข้อมูลออกเป็นสองส่วนเท่า ๆ กัน ในเมื่อได้เรียงลำดับข้อมูลแล้ว

ตัวอย่าง (1) รายได้ต่อวันของพนักงานในบริษัทแห่งหนึ่งจากการสำรวจด้วยตัวอย่างเป็นดังนี้

30, 37, 35, 35, 34, 50, 45, 32, 35, 34

เมื่อเรียงตามขนาดจะเป็น 30, 32, 34, 34, 35, 35, 35, 37, 45, 50

ดังนั้น $\bar{X}_m = (1/2)[35 + 35] = 35$

(2) ในการศึกษาเกรดในวิชาสถิติของนักศึกษา 5 คน ได้ผลดังนี้

G P P F F

เราจะได้มัธยฐานของเกรดเป็น P

(3) ในการศึกษาถึงรายได้ของคนในเขต กขค โดยอาศัยตัวอย่าง 1075 ราย ได้ผลดังนี้

รายได้	อันดับที่	จำนวนคน
สูง	3	33
กลาง	2	472
ต่ำ	1	570
	รวม	1075

โดยที่ $n/2 = 1075/2 = 537.5$ เราจึงได้ $\bar{X}_m = 1$ (รายได้ต่ำ)

สำหรับตัวอย่างสุ่มขนาด n ที่ได้ทำการแจกแจงความถี่ของค่าสังเกตไว้แล้ว ตัวสถิติ

$\bar{X}_m$ จะกำหนดไว้ดังนี้

$$\bar{X}_m = L + i\left(\frac{n/2-F}{f_m}\right)$$

ในเมื่อ L เป็นขอบเขตล่างของชั้นที่มัธยฐานอยู่ F เป็นความถี่สะสมของชั้นที่ต่ำกว่ามัธยฐานอยู่, f_m เป็นความถี่ของชั้นที่มัธยฐานอยู่, และ i เป็นความกว้างของอันตรภาคชั้น

ตัวอย่าง รายได้ต่อวันของพนักงานในบริษัทใหญ่แห่งหนึ่ง จากการสำรวจด้วยตัวอย่าง สมมติว่าเป็นดังนี้

รายได้ต่อวัน	f (จำนวน)	F (ความถี่สะสม)
25-34	15	15
35-44	20	35
45-54	10	45
55-64	4	49
65-74	1	50
	n = 50	

$$\bar{X}m = 34.5 + 10\left(\frac{50/2-15}{20}\right)$$

$$= 34.5 + 5 = 39.5$$

ในการคำนวณหามัธยฐานของตัวอย่างนั้น เราสามารถหาได้โดยการเทียบบัญญัติไตรยางค์ (Interpolation) ตามขั้นตอน ดังนี้

ก. หา n/2

ข. รวมความถี่จากชั้นล่างสุดขึ้นไป และรวมกันไม่ให้เกิน n/2 ชั้นที่ทำให้ความถี่รวมกันเป็น n/2 นั้นจะเป็นมัธยฐานอยู่

ค. พิจารณาขอบเขตล่างของชั้นที่มัธยฐานอยู่

ง. เทียบว่าความถี่ในชั้นมัธยฐานอยู่มีเท่าไร และชั้นนั้นกว้างเท่าไร และถ้าความถี่อีก (n/2-ผลรวมความถี่ใน ข.) จะได้ความกว้างเท่าใด

จ. ดังนั้นมัธยฐานจะเป็น

X_m = ขอบเขตชั้นล่างใน ค. + ผลจาก ง.

จากตัวอย่างที่กล่าวมาแล้วเราจะหามัธยฐานได้ดังนี้

ก) หา n/2 ได้เป็น 50/2 = 25

ข) รวมความถี่ได้เป็น 15 ดังนั้นชั้น 35-44 เป็นชั้นมัธยฐานอยู่

ค) ขอบเขตล่างของชั้นที่มัธยฐานอยู่เป็น 34.5

ง) ความถี่ 20 กว้าง 10 ดังนั้น ความถี่ (25-15) กว้าง $10(25 - 15)/20 = 5$

จ) ฉะนั้น $\bar{X}_m = 34.5 + 5 = 39.5$

นั่นคือ $\bar{X}_m = 34.5 + \frac{10}{20}(10) = 39.5$

สำหรับเปอร์เซนไทล์ (Percentile) ซึ่งเป็นค่าบนสเกลของร้อยนั้นจะระบุเปอร์เซนต์ของการแจกแจงที่เท่ากับหรือต่ำกว่าค่านั้น เช่นค่าในเปอร์เซนไทล์ที่ 95 จะเป็นคะแนนที่เท่ากับหรือมากกว่าคะแนนต่าง ๆ ถึง 95% เปอร์เซนไทล์จะเป็นจุดหรือค่าซึ่งจะแทนด้วย P_r และมีอยู่ 99 ค่า คือ $P_1, P_2, \ldots, P_{99}$ เราสามารถหา P ได้จากสูตรหรือการเทียบบัญญัติไตรยางค์ในลักษณะคล้าย ๆ กับมัธยฐาน ดังนี้

$$P_r = L + i\left(\frac{rn/100 - F}{f_r}\right)$$

ในเมื่อ P_r เป็นค่าของเปอร์เซนไทล์ที่ r($r = 1, 2, \ldots, 99$); L เป็นขอบเขตล่างของชั้นที่มีค่าของ P_r อยู่ ; f_r เป็นความถี่ในชั้นที่ P_r อยู่ ; F เป็นความถี่สะสมของชั้นที่ต่ำกว่าชั้นของ P_r อยู่ ; และ i เป็นขนาดของอันตรภาคชั้นที่ P_r อยู่

ตัวอย่าง สมมติว่าข้อมูลที่ได้จากการสำรวจด้วยตัวอย่างเป็นดังนี้

ชั้น	ความถี่	F
10-19	2	2
20-29	6	8
30-39	12	20
40-49	18	38
50-59	8	46
60-69	3	49
70-79	1	50
	50	

จงหาเปอร์เซนไทล์ที่ 25 และ 70

เมื่อาศัยสูตรเราจะได้ P_{25} และ P_{70} ดังนี้

$$P_{25} = 29.5 + 10\left(\frac{25(50)/100 - 8}{12}\right)$$

$$= 33.25$$

$$P_{70} = 39.5 + 10\left(\frac{70(50)/100 - 20}{18}\right)$$

$= 47.83$

เมื่อใช้การเทียบบัญญัติไตรยางค์เราจะได้ดังนี้

$P_{25} = ?$ ก. พิจารณา $rn/100 = 25(50)/100 = 12.5$

ข. รวมความถี่จากชั้นต่ำได้เป็น $2 + 6 = 8$ ดังนั้น P_{25} อยู่ในชั้น 30 - 39 ซึ่งมีขอบเขตล่างเป็น 29.5

ค. ความถี่ 12 กว้าง 10 ดังนั้น ความถี่(12.5 - 8) กว้าง $10 (12.5 - 8)/12 = 3.75$

ง. $P_{25} = 29.5 + 3.75 = 33.25$

นั่นคือ $P_{25} = 29.5 + \frac{4.5}{12} (10) = 33.25$

และ $P_{70} = ?$ ก) พิจารณา $rn/100 = 70(50)/100 = 35$

ข) รวมความถี่จากชั้นต่ำได้เป็น $2 + 6 + 12 = 20$ ดังนั้น P_{70} อยู่ในชั้น 40-49 ซึ่งมีขอบเขตล่างเป็น 39.5

ค) ความถี่ 18 กว้าง 10 ดังนั้น ความถี่ (35 - 20) กว้าง $10 (35 - 20)/18 = 8.33$

ง) $P_{70} = 39.5 + 8.33 = 47.83$

นั่นคือ $P_{70} = 39.5 + \frac{15}{18}(10) = 47.83$

ค่าที่แบ่งค่าสังเกตออกเป็นส่วน ๆ นี้ยังมีเดไซล์ (Decile) และควอร์ไทล์(Quartile) อีก โดยที่เดไซล์จะแบ่งการแจกแจงออกเป็น 10 ส่วนเท่า ๆ กันด้วยเดไซล์ที่ 1, 2, ..., 9 หรือ D_1, D_2, D_9 ส่วนควอร์ไทล์ซึ่งมี 3 ค่า คือ Q_1, Q_2, Q_3 จะแบ่งการแจกแจงออกเป็น 4 ส่วนเท่า ๆ กัน ดังนั้นทั้งเดไซล์และควอร์ไทล์จะมีความสัมพันธ์กับเปอร์เซนไทล์ ดังนี้

$$D_1 = P_{10}, D_2 = P_2, \ldots, D_9 = P_{90}$$

$$Q_1 = P_{25}, Q_2 = P_{50} = \bar{X}_m, Q_3 = P_{75}$$

(3) ฐานนิยมตัวอย่าง (Sample Mode) ฐานนิยมตัวอย่างใช้ สรุปข้อมูลแบบนามบัญญัติ หรือแบบอื่น ๆ เราใช้ $\bar{X}_m$ แทนฐานนิยมตัวอย่าง และกำหนดไว้ว่าเป็นข้อมูลที่เกิดขึ้นบ่อยที่สุด หรือมีความถี่สูงสุดนั่นเอง

ตัวอย่าง (1) ในการศึกษาสถานภาพสมรสของคนงานในโรงงานแห่งหนึ่งโดยอาศัยตัวอย่างได้ข้อมูลมาดังนี้

สภาพสมรส	จำนวน
โสด	12
แต่งงาน	225
หม้าย	53
หย่าร้าง	10
	รวม 300

ฐานนิยมของสภาพสมรสก็คือการแต่งงาน เพราะมีความถี่สูงสุด (เท่ากับ 225)

(2) ในการศึกษาสถานภาพสังคมของคนในเขตหนึ่งโดยอาศัยตัวอย่าง ได้ข้อมูลมาดังนี้

สถานภาพสังคม	จำนวน
ชั้นสูง	227
ชั้นกลาง	665
ชั้นต่ำ	1224
รวม	2116

ฐานนิยมของสถานภาพสังคมก็คือชั้นต่ำ

(3) ในการศึกษารายได้ของคนงานทอผ้า ได้ข้อมูลมาดังนี้

รายได้ต่อวัน (บาท)	จำนวนคนงาน
45	100
50	10
60	5
	รวม 115

ฐานนิยมของรายได้ต่อวันคือ 45 บาท

กรณีที่มีการแจกแจงความถี่ของข้อมูลแบบอันตรภาคหรืออัตราส่วน เราจะได้มัธยฐานตัวอย่าง X_{mo} ดังนี้

$$\bar{X}_{mo} = L + i\left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)$$

ในเมื่อ L เป็นขอบเขตล่างของชั้นที่มีความถี่สูงสุด, i เป็นความกว้างของชั้นที่ฐานนิยมอยู่ Δ_1 และ Δ_2 เป็นผลต่างของความถี่ของชั้นที่ฐานนิยมอยู่กับชั้นที่ต่ำกว่า และชั้นที่สูงกว่าตามลำดับ

สำหรับฐานนิยมหยาบ ๆ (Crude Mode) จะเป็นจุดกลางของชั้นที่มีความถี่สูงสุด

นอกจากค่าเฉลี่ยต่าง ๆ ที่กล่าวมาแล้ว เรายังมีค่าเฉลี่ยอื่น ๆ อีกดังนิยามต่อไปนี้

นิยาม ถ้า $X_1, X_2, \ldots, X_n$ เป็นตัวอย่างสุ่มขนาด n จากประชากร X แล้วค่าเฉลี่ยเรขาคณิต $\tilde{X}_g$, ค่าเฉลี่ยฮาร์โมนิค $\tilde{X}_h$, กึ่งพิสัย $\tilde{X}_r$, และค่าเฉลี่ย กำลังสอง (Quadratic Mean) $\tilde{X}_q$ กำหนดไว้ดังนี้

$$\tilde{X}_g = \sqrt[n]{X_1 (X_2) \ldots (X_n)}$$

$$= \text{antilog}\left(\frac{1}{n}\sum_{i=1}^{n} \log X_i\right)$$

$$\tilde{X}_h = \frac{n}{\left(\sum_{i=1}^{n} 1/X\right)}$$

$$\tilde{X}_r = (1/2)(X_{(n)} + X_{(1)})$$

$$\tilde{X}_q = \sqrt{\frac{1}{n}\sum^{n} X_i^2}$$

4.4.2 การกระจาย (Statistical Dispersion) เนื่องจากการพิจารณาค่าเฉลี่ยของข้อมูลเพียงอย่างเดียวอาจไม่เพียงพอที่จะสรุปย่อข้อมูลจำนวนมาก ๆ ซึ่งมีความผันแปรในค่าตัวเลขอยู่แล้ว ฉะนั้น ในการวิเคราะห์ข้อมูลเพื่อสรุปย่อข้อมูล จึงจำเป็นต้องคำนวณค่าที่ใช้วัดความผันแปร (Variation) หรือการกระจายของข้อมูลชุดนั้น ๆ กำกับไปกับค่าเฉลี่ยด้วย

ในทางสถิติเรามีค่าที่ใช้วัดการกระจายหรือความผันแปรของข้อมูลซึ่งเรียกกันว่า มาตราวัดการกระจาย (Measure of Dispersion) หรือมาตราวัดการเกาะกลุ่ม (Measure of Concentration) นั้นอยู่ 2 ประเภท คือ มาตราวัดการกระจายแบบสัมบูรณ์ (Absolute Dispersion) และ มาตราวัดการกระจายแบบสัมพัทธ์ (Relative Dispersion)

ก. การกระจายแบบสัมบูรณ์ (Absolute Dispersion) มาตราวัดการกระจายประเภทนี้มีอยู่หลายมาตราด้วยกัน เช่น พิสัย (Range) ส่วนเบี่ยงเบนเฉลี่ย (Mean Deviation), ส่วนเบี่ยงเบนควอร์ไทล์ หรือเดไซล์ (Quartile or Decile Deviation) ความแปรปรวนและส่วนเบี่ยงเบนมาตรฐาน (Variance and Standard Deviation) และอัตราส่วนความผันแปร (Variation Ratio) หรือดัชนีของการกระจาย (Index of Dispersion)

แต่มาตรวัดการกระจายที่ใช้กันบ่อย ๆ ก็มีพิสัย และความแปรปรวนหรือส่วนเบี่ยงเบนมาตรฐาน

(1) พิสัย (Range) พิสัยเป็นมาตรวัดการกระจายของข้อมูลที่ง่ายที่สุดในเมื่อข้อมูลเป็นแบบอันตรภาค หรืออัตราส่วนพิสัย กำหนดไว้ว่า

ถ้า $X_1, X_2, \ldots, X_n$ เป็นตัวอย่างสุ่มขนาด n แล้วพิสัย R กำหนดไว้ดังนี้

$$R = X_{(n)} - X_{(1)}$$

ตัวอย่าง สมมติว่าข้อมูล 3 ชุดต่อไปนี้เป็นผลผลิตข้าวของหมู่บ้านในตำบล 3 ตำบล ซึ่งได้จากการสำรวจด้วยตัวอย่าง

ชุดที่ 1 50, 52, 54, 56, 58, 62, 67

ชุดที่ 2 55, 56, 57, 57, 57, 58, 59

ชุดที่ 3 57, 57, 57, 57, 57, 57, 57

ข้อมูลทั้ง 3 ชุดนี้มีค่าเฉลี่ยเลขคณิตเท่ากันคือ 57 ทั้ง ๆ ที่ข้อมูลทั้งสามชุดนี้ไม่อาจจะสรุปได้ว่าเหมือนกัน ดังนั้นจึงจำเป็นต้องมีค่าที่ให้ใช้วัดการกระจายกำกับอีกด้วย และเราจะใช้พิสัยเป็นค่าสรุปย่อ ซึ่งจะหาได้ดังนี้

พิสัยของข้อมูลชุดที่หนึ่ง = 37-20 = 17

พิสัยของชุดที่สอง = 29-25 = 4

พิสัยของชุดที่สาม = 27-27 = 0

เราจะเห็นได้ว่าถ้าข้อมูลมีความผันแปรในค่าตัวเลขมาก เช่น ข้อมูลชุดที่หนึ่ง แล้วค่าของพิสัยจะใหญ่และถ้าข้อมูลชุดนั้นมีความผันแปรไม่มากนักเช่นข้อมูลชุดที่สอง แล้วค่าของพิสัยจะไม่ใหญ่นัก แต่ถ้าข้อมูลไม่มีความผันแปรเลย เช่นข้อมูลชุดที่สาม แล้วค่าของพิสัยจะเป็นศูนย์

ดังนั้นในการวิเคราะห์เพื่อสรุปย่อข้อมูลจึงอาจใช้พิสัยกำกับค่าเฉลี่ยไปด้วย ซึ่งจะทำให้ค่าสถิติทั้งสองค่านั้นบอกความหมายจากข้อมูลได้ชัดเจนยิ่งขึ้น เช่นการวิเคราะห์ราคาข้าวเปลือก 2 ปี ถ้าสรุปว่า

ปี 2518 ราคาข้าวเปลือกที่ชาวนาขายได้โดยเฉลี่ยทั้งปี เท่ากับ 2,200 บาท

พิสัย เท่ากับ 400 บาท

ปี 2519 ราคาข้าวเปลือกที่ชาวนาขายได้โดยเฉลี่ยทั้งปี เท่ากับ 2,100 บาท

พิสัย เท่ากับ 600 บาท

ก็พอที่จะทำให้ผู้สนใจใช้ข้อมูลสถิติพอเข้าใจความแตกต่างในภาวะราคาข้าวเปลือกได้ดีพอสมควรโดยไม่ต้องกลับไปขอดูข้อมูลในรายละเอียด

(2) ความแปรปรวนและส่วนเบี่ยงเบนมาตรฐาน (Variance and Standard Deviation) ค่าสถิติที่ใช้วัดการกระจายของข้อมูลแบบอันตรภาคและอัตราส่วนได้ดี และใช้กันอยู่มากในวงการสถิติและวิจัยก็คือความแปรปรวนและส่วนเบี่ยงเบนมาตรฐาน ซึ่งนิยามไว้ดังนี้

ถ้า $X_1, X_2, \ldots, X_n$ เป็นตัวอย่างสุ่มขนาด n จากประชากร X แล้ว ความแปรปรวนและส่วนเบี่ยงเบนมาตรฐานของตัวอย่าง หรือ S^2 และ S กำหนดไว้ว่า

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

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

และ $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$

ตัวอย่าง รายได้ต่ออาทิตย์ของพนักงานฝ่ายผลิต 10 คน เป็นดังนี้

198, 199, 199, 200, 200, 200, 200, 201, 201, 202

และของพนักงานประจำสำนักงาน 10 คน เป็นดังนี้

190, 200, 200, 209, 191, 200, 210, 200, 191, 209

ข้อมูลทั้งสองกลุ่มนี้มีค่าเฉลี่ยเลขคณิตเท่ากันคือ 200 บาท แต่ข้อมูลทั้งสองกลุ่มมีการกระจายแตกต่างกัน เพราะจากการสังเกตก็จะเห็นว่ากลุ่มที่สองมีการกระจายมากกว่า

เมื่อเราคำนวณความแปรปรวนและส่วนเบี่ยงเบนมาตรฐานของทั้งสองกลุ่มจะได้ดังนี้

พนักงานฝ่ายผลิต $S^2 = \frac{1}{10-1}[(198-200)^2 + (199-200)^2$

$$+ \ldots\ldots + (202-200)^2] = 12/9 = 1.33$$

$$S = \sqrt{1.33} = 1.16$$

พนักงานประจำสำนักงาน $S^2 = \frac{1}{10-1}[(190-200)^2 + \ldots\ldots + (209-200)^2]$

$$= 524/9 = 58.22$$

$$S = \sqrt{58.22} = 7.63$$

ดังนั้น รายได้ของพนักงานประจำมีการกระจายมากกว่าของพนักงานฝ่ายผลิต

ในกรณีที่ค่าสังเกตของตัวอย่างสุ่มมีการแจกแจงความถี่ แล้วความแปรปรวนจะกำหนดไว้ว่า

$$S^2 = i^2 S_d^2 = i^2 \left\{\frac{\Sigma fd^2 - (\Sigma fd)^2/n}{n-1}\right\}$$

ในเมื่อ i เป็นขนาดของอัตรภาคชั้น , f เป็นความถี่ของแต่ละชั้น, และ $d = (X-A)/i$ โดยที่ A เป็นจุดกลางของชั้นที่จะทำให้ $d = 0$

ตัวอย่าง รายได้ต่อวันของกรรมการในโรงงานแห่งหนึ่งที่เลือกสุ่มมาเป็นดังนี้

รายได้	f (จำนวน)	d	fd	fd^2
25-34	10	-3	-30	90
35-44	15	-2	-30	60
45-54	20	-1	-20	20
55-64	25	0	0	0
65-74	10	1	10	10
75-84	5	2	10	20
	75		-60	200

$d = (X - A)/i = (X - 59.5)/10$

$$S^2_d = \frac{\Sigma fd^2 - (\Sigma fd)^2/n}{n-1}$$

$$= \frac{1}{75-1}\,[200 - (-60)^2/75]$$

$$= 152/74 = 2.0541$$

$$S^2 = i^2 S_d^2 = (10)^2(2.0541) = 205.41$$

$$S = \sqrt{205.41} = 14.33$$

ในการคำนวณส่วนเบี่ยงเบนมาตรฐานแบบรวดเร็วนั้น Robert L. Lathrop ได้เสนอสูตรคำนวณไว้ว่า

$$S = \frac{\text{Sum Upper (1/6) X} - \text{Sum Lower (1/6) X}}{n/2}$$

จากตัวอย่างที่ผ่านมานี้เราจะหาส่วนเบี่ยงเบนมาตรฐานได้ดังนี้

รายได้	จุดกลาง(X)	f	f		fX	
25-34	29.5	10	10	12.5	295.00	→ Sum Lower (1/6) X = 393.75
35-44	39.5	15	2.5		98.75	
45-54	49.5	20				
55-64	59.5	25				
65-74	69.5	10	7.5	12.5	521.25	→ Sum Upper (1/6) X = 918.75
75-84	79.5	5	5		397.50	
		75				

n/6 = 75/6 = 12.5

$$S = \frac{918.75 - 393.75}{75/2} = 525.00/37.5$$

$$= 14.00$$

แต่จากการคำนวณที่ผ่านมาเราได้ S = 14.33

ส่วนเบี่ยงเบนมาตรฐานจากประชากรที่แจกแจงปกติหรือเกือบปกตินั้นสามารถประมาณได้ด้วยพิสัย R ดังนี้

$$S = R/d_n$$

ในเมื่อ d_n เป็นค่าคงที่กำหนดไว้ดังนี้?

n	d_n	n	d_n	n	d_n
2	1.13			160	5.34
3	1.69	18	3.64	170	5.38
4	2.06	19	3.69	180	5.42
5	2.33	20	3.73	190	5.46
6	2.53				
7	2.70	30	4.09	200	5.49
8	2.85	40	4.32	250	5.64
9	2.97	50	4.50	300	5.76
10	3.08	60	4.64	400	5.94
11	3.17	70	4.75	450	6.01
12	3.26	80	4.85	500	6.07
13	3.34	90	4.94	600	6.18

14	3.41	110	5.08	300	6.35
15	3.47	120	5.14	900	6.42
		130	5.20	900	6.42
16	3.53	140	5.25	1000	6.48
17	3.59	150	5.30		

สำหรับขนาดตัวอย่าง $n \leq 15$ เราใช้ $\sqrt{n}$ แทน d_n ได้

จากตัวอย่างเราประมาณ S ได้เป็น

$S = (84-25)/4.8 = 12.30$

ในเมื่อ $d_n = (4.75 + 4.85)/2 = 4.80$

นอกจากนี้ M.M. Lentner ได้เสนอทฤษฎีเกี่ยวกับขอบเขตหรือช่วงของความแปรปรวนไว้ว่า

ขอบเขตต่ำสุดและสูงสุดของความแปรปรวนจะเป็น

$R^2/2(n-1) \leq S^2 \leq mR^2/4$

ในเมื่อ R เป็นพิสัย และ $m = n/(n-1)$ ถ้า n เป็นเลขคู่ หรือ $m = (n+1)/n$ ถ้า n เป็นเลขคี่

จากตัวอย่างที่ผ่านมาเราจะได้

$$(84-25)^2/2(75-1) \leq S^2 \leq [(75+1)/75](84-25)^2/4$$

$$23.52 \leq S^2 \leq 881.85$$

$$4.85 \leq S \leq 29.695$$

ซึ่งเราจะเห็นได้ $S = 14.33$ อยู่ในช่วงนี้จริง

ในบางครั้งเราต้องการที่จะรวมความแปรปรวนหรือส่วนเบี่ยงเบนมาตรฐานของข้อมูลหลาย ๆ กลุ่มเข้าด้วย เราก็ทำได้โดยอาศัยสูตร ดังนี้

$$S^2 = \frac{1}{n-1}(B - A^2/n)$$

ในเมื่อ $n = \sum_i^k n_i$, $B = \sum_i^k \left\{(n_i - 1)S_i^2 + A_i^2/n_i\right\}$

และ $A = \sum_i^k n_i \bar{X}_i$

ตัวอย่าง ตัวอย่าง 2 ชุดมีค่าเฉลี่ยเลขคณิตและความแปรปรวน ดังนี้

ชุดที่ 1 $n_1 = 20$, $\bar{X}_1 = 18$, $S_1^2 = 4$

ชุดที่ 2 $n_2 = 30$, $\bar{X}_2 = 16$, $S_2^2 = 9$

ดังนั้น $n = 20 + 30 = 50$

$A = 20(18) + 30(16) = 840$

$B = (20-1)4 + [20(18)]^2/20 + (30-1)9 + [30(16)]^2/30$

$= 6556 + 7941 = 14497$

$S^2 = \frac{1}{50-1}[14497 - (840)^2/50] = 385/49 = 7.857$

$S = 2.803$

(3) อัตราส่วนความผันแปรหรือดัชนีความผันแปร (Variation Ratio or Index of Dispersion) อัตราส่วนความผันแปรเป็นมาตราวัดการกระจายสำหรับข้อมูลแบบนามบัญญัติหรือแบบเรียงอันดับ และมาตราวัดนี้จะใช้ควบคู่กับฐานนิยม อัตราส่วนความผันแปรนี้ Freeman ได้เสนอสูตรคำนวณไว้ดังนี้

$$V = 1 - \frac{f_{mod}}{n}$$

ในเมื่อ f_{mod} เป็นนามถี่ในชั้นที่ฐานนิยมอยู่และ n เป็นขนาดตัวอย่างจะเป็นดังนี้ $0 < V < 1-1/k$ ในเมื่อ k เป็นจำนวนประเภทหรือชั้นของข้อมูล

John Galtung ได้เสนอสูตรที่มีขีดจำกัดระหว่าง 0 กับ 1 ไว้ดังนี้

$$d = \frac{n - f_{mod}}{n - n/k} \; ; \; 0 < d < 1$$

นอกจากนี้ได้มีผู้เสนอมาตราวัดการกระจายของข้อมูลแบบนามบัญญัติหรือเรียงอันดับไว้ดังนี้

$$D = \frac{k(n^2 - \Sigma f_i^2)}{n^2(k-1)} \; ; 0 < D < 1$$

ในเมื่อ f เป็นความถี่ของข้อมูลในประเภทที่ i $(i = 1, 2, \ldots, k)$

มาตราวัดนี้ Hammond และ Householder (1962) ได้เสนอไว้และใช้กันมากในทางสังคมศาสตร์ และมักเรียกกันว่าดัชนีของการกระจาย (Index of dispersion)

Mueller และ Schuessler ได้เสนอมาตราวัดการกระจายซึ่งเรียกว่าดัชนีความผันแปรเชิงคุณภาพ (Index of Qualitative Variation) ไว้ดังนี้

$$IQV = \sum_{i<j} f_i f_j / (k/2)(k-1)(n/k)^2$$

ในเมื่อ f_i หรือ f_j เป็นความถี่ในข้อมูล (เป็นเปอร์เซนต์หรือวัดส่วนก็ได้) มาตราวัดนี้มีค่าที่เป็น

ไปได้ดังนี้ $0 \leq IQV \leq 1$

M.G Kendall ได้เสนอมาตราวัดการกระจายสำหรับข้อมูลแบบคุณภาพ (Qualitative data) ไว้ดังนี้

$$K = (n^2 - \Sigma f_i^{\,2})/2n^2$$

ยังมีมาตราวัดการกระจายอื่น ๆ อีกคือส่วนเบี่ยงเบนจากฐานนิยม (Deviation from Mode, DM) กำหนดไว้ดังนี้

$$DM = 1 - \frac{\sum(f_{mod} - f_i)}{n(k-1)}, \quad 0 \leq DM \leq 1$$

และดัชนีความไม่แน่นอนสัมพัทธ์(Index of Relative Uncertainty, H)ซึ่งกำหนดไว้ดังนี

$$H = \frac{-\sum f_i/n \ln (f_i/n)}{\ln k} \quad 0 \leq H \leq 1$$

ตัวอย่าง จากการศึกษาสถานภาพสมรสของพนักงานในรามคำแหงได้ข้อมูลมาดังนี้

สถานภาพสมรส	จำนวน
โสด	15
แต่งงาน	235
หม้าย	40
หย่าร้าง	10
รวม	300

$$V = 1 - \frac{235}{300} = 0.22$$

$$d = \frac{300 - 235}{300 - 300/4} = 0.29$$

$$D = \frac{4\{300^2 - 57150\}}{300^2\ (4-1)} = 0.365$$

ในเมื่อ $\Sigma f_i^{\,2} = 15^2 + 235^2 + 40^2 + 10^2 = 57150$

(4) ส่วนเบี่ยงเบนควอร์ไทล์และเดไซล์ (Quartile and Decile Deviation QD DD) มาตราวัดทั้งสองนี้เป็นมาตราวัดการกระจายของข้อมูลแบบอันดับ หรือแบบอันตรภาค-อัตราส่วนที่ไม่สมมาตรและกำหนดไว้ดังนี้

$$QD = (Q_3 - Q_1)/2$$

$$DD = (D_9 - D_1)/2$$

ตัวอย่าง ในการศึกษาทัศนคติอย่างหนึ่งของนักศึกษาได้ข้อมูลมาดังนี้

ทัศนคติ	อันดับ	จำนวน
เห็นด้วยอย่างยิ่ง	5	25

เห็นด้วย	4	45
เฉย ๆ	3	75
ไม่เห็นด้วย	2	40
ไม่เห็นด้วยอย่างยิ่ง	1	15
	รวม	200

เนื่องจาก .25 (n) = 50 และ .75 (n) = 150 เราจึงได้

$Q_1 = 2, Q_3 = 4$ และ QD = (4−2)/2 = 1

และโดยที่ .1(n) = 20 และ .9(n) = 180 จึงได้

$D_1 = 2$, $D_9 = 5$ และ DD = (5−2)/2 = 1.5

ข. การกระจายแบบสัมพันธ์ (Relative Dispersion) ในการเปรียบเทียบข้อมูลแบบอันตรภาคหรืออัตราส่วนตั้งแต่สองชุดขึ้นไป เพื่อตัดสินว่าชุดใดมีการกระจายมากกว่า ถ้าเราใช้ค่าที่ได้จากการวัดการกระจายที่กล่าวมานั้นของแต่ละชุด มาเปรียบเทียบกันก็ย่อมตัดสินใจได้ยาก โดยเฉพาะเมื่อขนาดของข้อมูลในแต่ละชุดต่างกัน เช่นชุดหนึ่งมีค่าตั้งแต่ 0 ถึง 10 และมีส่วนเบี่ยงเบนมาตรฐาน 2.2 อีกชุดหนึ่งมีค่าตั้งแต่ 200 ถึง 800 และมีส่วนเบี่ยงเบนมาตรฐาน 60.5 ถ้าจะตัดสินว่าชุดที่หนึ่งมีการกระจายน้อยกว่าชุดที่สองก็อาจจะไม่ถูกต้องนัก เพราะค่าของข้อมูลสองชุดนี้ต่างกันมาก สค่าเฉลี่ยและค่าแสดงการกระจายย่อมจะต่างกันมากด้วย เพื่อให้การเปรียบเทียบมีความหมาย เรานิยมใช้มาตราวัดการกระจายแบบสัมพัทธ์ หรือสัมประสิทธิ์ของการกระจาย (Coefficient of Dispersion) ซึ่งกำหนดไว้ว่า

$$\text{สัมประสิทธิ์ของการกระจาย} = \frac{\text{มาตราวัดการกระจาย}}{\text{ค่าเฉลี่ยที่เหมาะสม}}$$

สำหรับสัมประสิทธิ์ของการกระจายที่สนใจ มีดังนี้

สัมประสิทธิ์ของพิสัย	$= (X_{(n)} - X_{(1)}) / (X_{(n)} + X_{(1)})$
สัมประสิทธิ์ของส่วนเบี่ยงเบนควอร์ไทล์	$= [Q_3 - Q_1] / [Q_3 + Q_1]$
สัมประสิทธิ์ของส่วนเบี่ยงเบนเดไซล์	$= \frac{D_9 - D_1}{D_9 + D_1}$
สัมประสิทธิ์ของส่วนเบี่ยงเบนเฉลี่ย	$= MD/\bar{X}_m$
สัมประสิทธิ์ความผันแปร	$= S/\bar{X}$

สำหรับสัมประสิทธิ์ของความผันแปร (CV, Coefficient of Variation) นั้นเป็นการวัดการกระจายสัมพันธ์ที่นิยมใช้กันมากที่สุด และมักเขียนในรูปเปอร์เซนต์โดยคูณด้วย 100

ตัวอย่าง ถ้าราคาข้าวเปลือกและราคาข้าวสารต่อถังของข้าวชนิดหนึ่งจากร้านค้าตัวอย่าง 10 ร้าน เป็นดังนี้

ราคาข้าวเปลือก (บาท) $\bar{X}_1 = 25$ และ $S_1 = 4$

ราคาข้าวสาร (บาท) $\bar{X}_2 = 80$ และ $S_2 = 10$

ดังนั้น $CV_1 = S_1/\bar{X}_1 = 4/25 = .16$

$CV_2 = S_2/\bar{X}_2 = 10/80 = .125$

เราจะเห็นได้ว่าราคาข้าวเปลือกมีการกระจายมากกว่าราคาข้าวสาร

เนื่องจากสัมประสิทธิ์ความผันแปรนี้ไม่มีขีดจำกัด Martin และ Gray (1971) ได้ปรับปรุงสูตรได้มีขีดจำกัดดังนี้

$$S_{(d)} = CV/\max CV$$

$$= \frac{S/\bar{X}}{\sqrt{n-1}}$$

หรือ $$S_{(d)} = \frac{MD/\bar{X}}{2(1-1/n)}$$

โดยที่ $S_{(d)}$ มีค่าที่เป็นไปได้ดังนี้ $0 \leqslant S_{(d)} \leqslant 1$

จากตัวอย่างเราได้

ข้าวเปลือก $S_{(d)} = \frac{0.16}{\sqrt{10-1}} = 0.053$

ข้าวสาร $S_{(d)} = \frac{0.125}{\sqrt{10-1}} = 0.042$

4.5 การแจกแจงของการสุ่มตัวอย่าง (Sampling Distribution)

ในทฤษฎีการสุ่มตัวอย่าง หรือสถิติอนุมาน เราจะเกี่ยวข้องกับการแจกแจงต่าง ๆ 3 ชนิด ซึ่งมีความสัมพันธ์กัน ดังนี้

(1) การแจกแจงของประชากร (Population Distribution) การแจกแจงนี้จะเป็นการแจกแจงของค่าสังเกตจากหน่วยแจงนับทุกหน่วยในประชากรที่สนใจ การแจกแจงของประชากรจะเอื้ออำนวยให้คำนวณมาตราวัดสรุปหรือบารามิเตอร์ได้ อย่างไรก็ตามการแจกแจงแบบนี้มักจะไม่ทราบ เพราะต้องศึกษาทุกหน่วยแจงนับในประชากร

(2) การแจกแจงของตัวอย่าง (Sample Distribution) เป็นการแจกแจงของค่าสังเกตจากหน่วยแจงนับทุกหน่วยในตัวอย่าง หรือเพียงบางหน่วยแจงนับในประชากร การแจกแจงนี้จะให้ค่าของตัวสถิติซึ่งเป็นค่าที่สรุปย่อของข้อมูล หรือค่าสังเกตจากตัวอย่าง อย่างไรก็ตามการแจกแจงของตัวอย่างและของประชากร จึงไม่เป็นอันเดียวกัน

(3) การแจกแจงของการสุ่มตัวอย่าง (Sampling Distribution) การเลือกตัวอย่างสุ่มจากประชากรจะเป็นการทดลองเชิงสุ่มอย่างหนึ่ง ซึ่งมีกลุ่มผลทดลองที่ประกอบด้วยตัวอย่างที่เป็นไปได้ทั้งหมด และตัวอย่างนั้นมีขนาดตามที่กำหนดไว้ เนื่องจากตัวสถิติจะกำหนดจากตัวอย่างสุ่ม ดังนั้นตัวสถิติจึงเป็นตัวแปรเชิงสุ่มด้วยและมีค่าที่เป็นไปได้ตามตัวอย่างที่เป็นไปได้เหล่านั้น เมื่อตัวสถิติเป็นตัวแปรเชิงสุ่มมันจึงมีการแจกแจงน่าจะเป็นอย่างหนึ่ง การแจกแจงของมันจะเรียกว่า "การแจกแจงของการสุ่มตัวอย่าง"

สำหรับส่วนเบี่ยงเบนมาตรฐานของตัวสถิติเราจะเรียกว่า "ความคลาดเคลื่อนมาตรฐาน (Standard Error) ของตัวสถิติ"

การแจกแจงน่าจะเป็นของค่าเฉลี่ยเลขคณิตจากตัวอย่าง $\bar{x}$ จะเรียกว่าการแจกแจงของการสุ่มตัวอย่างของค่าเฉลียเลขคณิจากตัวอย่าง $\bar{x}$ และความคลาดเคลื่อนมาตรฐานของ $\bar{x}$ ก็คือส่วนเบี่ยงเบนมาตรฐานของการแจกแจงของการสุ่มตัวอย่างของ $\bar{x}$ ในทำนองเดียวกัน การแจกแจงน่าจะเป็นของส่วนเบี่ยงเบนมาตรฐาน S จะเรียกว่าการแจกแจงของการสุ่มตัวอย่างของตัวสถิติ S และความคลาดเคลื่อนมาตรฐาน S ก็คือส่วนเบี่ยงเบนมาตรฐานของตัวสถิติ S นั่นเอง

4.6 การแจกแจงการสุ่มตัวอย่างของส่วนเฉลี่ย (Sampling distribution of Means)

การแจกแจงของการสุ่มตัวอย่างที่จะพิจารณาอันดับแรกคือการแจกแจงของตัวสถิติ $\bar{x}$ สมมติว่าตัวอย่างขนาด n สุ่มมาจากประชากรแบบปกติที่มีค่าเฉลี่ย μ และความแปรปรวน σ^2 แต่ละค่าสังเกต X_i (i = 1, 2,....., n) ของตัวอย่างสุ่มจะมีการแจกแจงแบบปกติเช่นเดียวกับประชากรที่สุ่มมา

จากคุณสมบัติของการแจกแจงแบบปกติ เราได้ว่าตัวสถิติ $\bar{x} = (X_1+X_2+.....+X_n)/n$ จะมีการแจกแจงแบบปกติที่มีค่าเฉลี่ยและความแปรปรวน ดังนี้

$E(\bar{x}) = \mu$, $V(\bar{x}) = \sigma^2/n$, และ $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

ถ้าเราสุ่มตัวอย่างจากประชากรที่ไม่ทราบการแจกแจงและไม่ว่าประชากรจะมีขนาดจำกัดหรือมีขนาดอนันต์ก็ตาม การแจกแจงของ $\bar{x}$ จะยังเป็นแบบปกติที่มีค่าเฉลี่ย μ และความแปรปรวน σ^2/n ถ้า ขนาดของตัวอย่างโต ผลอันนี้เป็นลักษณะของทฤษฎีขีดจำกัดสู่ส่วนกลาง (Central Limit Theorem) ลองพิจารณาทฤษฎีต่อไปนี้

ถ้า $\bar{x}$ เป็นค่าเฉลี่ยเลขคณิตของตัวอย่างสุ่มขนาด n ซึ่งสุ่มมาจากประชากรที่มีค่าเฉลี่ย μ และความแปรปรวนจำกัด (Finite Variance) σ^2 แล้วรูปแบบขีดจำกัด (limiting form) ของการแจกแจงของ z

$$Z = (\bar{x}-\mu)/(\sigma/\sqrt{n})$$

จะเป็นแบบปกติมาตรฐาน N(0, 1)

การแจกแจงของตัวสถิติ $\bar{x}$ จะใกล้ ๆ กับการแจกแจงแบบปกติ ถ้าตัวอย่างขนาดโต $n \geq 30$ โดยไม่คำนึงถึงรูปร่างของประชากร

ถ้าตัวอย่างขนาดเล็ก (n<30) การแจกแจงของ $\bar{x}$ จะใกล้กับแบบปกติ เมื่อประชากรไม่แตกต่างจากประชากรแบบปกติมากนัก สำหรับประชากรที่เป็นแบบปกติการแจกแจงของการสุ่มตัวอย่างของ $\bar{x}$ จะเป็นแบบปกติด้วย ถึงแม้ว่าขนาดของตัวอย่างจะเล็กสักปานใด

ตัวอย่าง อายุใช้งานของหลอดไฟจะมีการแจกแจงแบบปกติที่มีค่าเฉลี่ย 800 ชั่วโมง และส่วน

เบี่ยงเบนมาตรฐาน 40 ชั่วโมง จงหาความน่าจะเป็นที่ตัวอย่างสุ่มของหลอดไฟ 16 ดวงจะมีค่าเฉลี่ยของอายุใช้งานน้อยกว่า 775 ชั่วโมง

การแจกแจงของการสุ่มตัวอย่างของ $\bar{x}$ จะมีการแจกแจงแบบปกติที่มีค่าเฉลี่ยและความแปรปรวนดังนี้

$$E(\bar{x}) = \mu = 800$$

$$V(\bar{x}) = \sigma^2/n = 40^2/16 = 100$$

ดังนั้นความน่าจะเป็นที่ต้องการคือ

$$P(\bar{x}<775) = P(Z<\frac{775-800}{10}) = P(Z<-2.5) = 0.006$$

ตัวอย่าง ถ้าประชากรมีการแจกแจงแบบยูนิฟอร์ม ดังนี้

$$f(x) = 1/4; x = 0, 1, 2, 3$$

ความน่าจะเป็นของตัวอย่างสุ่มขนาด 36 (ที่เลือกแบบแทนที่) ที่มีค่าเฉลี่ยของตัวอย่างมากกว่า 1.4 แต่น้อยกว่า 1.8 เป็นเท่าใด ?

ค่าเฉลี่ยและความแปรปรวนของประชากร เราได้

$$\mu = 0(1/4)+1(1/4)+2(1/4)+3(1/4)$$

$$= 3/2 = 1.5$$

$$\sigma^2 = (0-1.5)^2(1/4)+....+(3-1.5)^2(1/4)$$

$$= 5/4 = 1.25$$

การแจกแจงของการสุ่มตัวอย่างของตัวสถิติ $\bar{x}$ จะเป็นแบบปกติที่มีค่าเฉลี่ยและความแปรปรวนดังนี้

$$E(\bar{X}) = 1.5$$

$$V(\bar{X}) = 1.25/36, \quad \sigma_{\bar{x}} = 0.186$$

ดังนั้น $P(1.4<\bar{X}<1.8) = P(1.45\leq\bar{X}\leq1.75) = P(-0.269\leq Z\leq1.344) = .5173$

บ่อยครั้งที่เราไม่ทราบความแปรปรวนของประชากรที่เราเลือกตัวอย่างสุ่มมา สำหรับตัวอย่างโตขนาด ($n\geq30$) ค่าประมาณที่ดีของ σ คือ S^2

ดังนั้นตัวสถิติ $(\bar{X}-\mu)/(\sigma/\sqrt{n})$ เมื่อแทน σ^2 ด้วย s^2 เราจะได้ $(\bar{X}-\mu)/(S/\sqrt{n})$ ซึ่งยังจะแจกแจงแบบปกติมาตรฐานอยู่ ทั้งนี้เพราะ S^2 ไม่ผันแปรหรือกวัดแกว่งมากนักถ้า $n\geq30$ นั่นคือไม่ผันแปรไปตามตัวอย่าง ถ้าขนาดตัวอย่างโตพอ

ถ้าขนาดตัวอย่างน้อย ($n<30$) ค่าของ S^2 จะแกว่งไปตามตัวอย่างและการแจกแจงของ $(\bar{X}-\mu)/(S/n)$ จะไม่เป็นแบบปกติมาตรฐาน แต่จะมีการแจกแจงแบบที (Student's t distribution) ดังทฤษฎีต่อไปนี้

ถ้า $\bar{X}$ และ S^2 เป็นค่าเฉลี่ยและความแปรปรวนของตัวอย่างสุ่มขนาด ซึ่งสุ่มมาจาก

ประชากรแบบปกติที่มีค่าเฉลี่ย μ แต่ไม่ทราบค่าของความแปรปรวน σ^2 แล้วตัวสถิติ T

$$T = (\bar{x} - \mu)/(S/\sqrt{n})$$

จะมีการแจกแจงแบบที ซึ่งมีองศาแห่งความเป็นอิสระ (df, degrees of freedom) $\nu = n-1$

ตามทฤษฎีนั้นประชากรต้องเป็นแบบปกติ แต่เราสามารถแสดงได้ว่าประชากรที่ไม่เป็นแบบปกติ (Nonnormal Pupulations) แต่เป็นรูประฆัง (Bell-shaped) ก็ยังคงให้ค่าของ T มีการแจกแจงใกล้กับการแจกแจงแบบที มาก

ตัวอย่าง ผู้จัดการฝ่ายผลิตหลอดไฟกล่าวว่าหลอดไฟที่ผลิตออกมาจะมีอายุใช้งานเฉลี่ย 500 ชั่วโมง เพื่อที่จะคงค่าเฉลี่ยนี้ไว้ เขาจึงทดสอบหลอดไฟ 25 หลอดทุก ๆ เดือน ถ้าคำนวณค่า T อยู่ระหว่าง $-t_{.05}$ และ $t_{.05}$ เขาก็จะพอใจ

ถ้าเดือนหนึ่งเขาสุ่มตัวอย่างมาแล้วคำนวณค่าเฉลี่ย $\bar{x}$ = 518 ชั่วโมง และส่วนเบนมาตรฐาน S = 40 ชั่วโมง เขาจะสรุปผลได้อย่าง ? สมมติว่าอายุใช้งานของของหลอดไฟแจกแจงแบบปกติ

สำหรับ t_{05} นั้นเป็นค่าของ t ที่ทำให้เกิดพื้นที่ทางขวามือเป็น .05 สำหรับ ν = 24 เราได้ $t_{.05}$ = 1.711 ดังนั้นผู้จัดการจะพอใจ ถ้าตัวอย่างขนาด 25 ให้ค่า T ระหว่าง - 1.711 และ 1.711

เมื่อ μ = 500 แล้ว T = (518 − 500)/(40/$\sqrt{25}$) = 2.25 ซึ่งโตกว่า 1.711 ความน่าจะเป็นที่ได้ค่า t (ν = 24) เท่ากับหรือมากกว่า 2.25 จะประมาณ 0.02 ซึ่งจะสรุปได้ว่าหลอดไฟจะมีอายุใช้งานมากกว่าที่กล่าวไว้ นั่นคือผลิตภัณฑ์ของเขาดีกว่าที่คิดไว้

ในกรณีที่เรามีสองประชากรโดยที่มีค่าเฉลี่ย μ_1 กับ μ_2 และความแปรปรวน σ_1^2 กับ σ_2^2 และให้ตัวสถิติ $\bar{x}_1$ แทนค่าเฉลี่ยของตัวอย่างสุ่มขนาด n_1 ที่เลือกจากประชากรแรกกับตัวสถิติ $\bar{x}_2$ แทนค่าเฉลี่ยของตัวอย่างสุ่มขนาด n_2 ที่เลือกจากประชากรที่สอง และเป็นอิสระกับตัวอย่างในประชากรแรก เราจะได้การแจกแจงของการสุ่มตัวอย่างของสถิติ $\bar{x}_1 - \bar{x}_2$ ดังนี้

ถ้าสองตัวอย่างสุ่มที่เป็นอิสระกัน มีขนาด n_1 และ n_2 ได้เลือกมาจากสองประชากรแบบปกติที่มีค่าเฉลี่ย μ_2 และ μ_2 กับความแปรปรวน σ_1^2 และ σ_2^2 แล้วการแจกแจงของการสุ่มตัวอย่างของตัวสถิติ $\bar{x}_1 - \bar{x}_2$ จะเป็นแบบปกติที่มีค่าเฉลี่ยและความแปรปรวน ดังนี้

$$E(\bar{X}_1 - \bar{X}_2) = (\mu_1 - \mu_2)$$

$$V(\bar{X}_1 - \bar{X}_2) = \sigma^2_1/n_1 + {}^2_2/n_2$$

ดังนั้น $Z = ((\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2))/\sqrt{\sigma^2_1/n_1 + \sigma^2_2/n_2}$ จะมีการแจกแจงแบบปกติมาตรฐาน N(0,1)

ถ้าประชากรทั้งสองไม่เป็นแบบปกติ แล้วการแจกแจงของตัวสถิติ $\bar{x}_1 - \bar{x}_2$ อาจจะเป็นแบบปกติ หรือไม่เป็นแบบปกติก็ได้แต่ถ้าขนาดตัวอย่างโต $(n_1, n_2 \geq 30)$ แล้วการแจกแจงของ $\bar{x}_1 - \bar{x}_2$ จะใกล้กับแบบปกติมาก

ตัวอย่าง หลอดโทรทัศน์ของโรงงาน ก.มีอายุใช้งานเฉลี่ย 6.5 ปี และส่วนเบี่ยงเบนมาตรฐาน 0.9 ปี แต่โรงงาน ข.มีค่าเฉลี่ย 6.0 ปี และส่วนเบี่ยงเบนมาตรฐาน 0.8 ปี

จงหาโอกาสที่ตัวอย่างสุ่มขนาด 36 หลอดจากโรงงาน ก.จะมีอายุใช้งานเฉลี่ยมากกว่าของโรงงาน ข.อย่างน้อย 1 ปี (เมื่อสุ่มตัวอย่างขนาด 49 หลอดมาจากโรงงาน ข.)

ประชากร 1: $\mu_1 = 6.5, \quad \sigma_1^2 = 0.9, n_1 = 36$

ประชากร 2: $\mu_2 = 6.0, \quad \sigma_2^2 = 0.8, n_2 = 49$

การแจกแจงของการสุ่มตัวอย่างของ $\bar{x}_1 - \bar{x}_2$ จะมีค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน ดังนี้

$E(\bar{x}_1 - \bar{x}_2) = 6.5 - 6.0 = 0.5$

$V(\bar{x}_1 - \bar{x}_2) = 0.81/36 + 0.64/49 = 0.036$

$\sigma(\bar{x}_1 - \bar{x}_2) = 0.189$

ดังนั้น $P(\bar{x}_1 - \bar{x}_2 \geq 1.0) = P\left(Z > \frac{1.0 - 0.5}{0.189}\right)$

$= P(Z > 2.646) = 0.0041$

ถ้าตัวอย่างสุ่มที่เป็นอิสระกันจากสองประชากร ซึ่งเป็นแบบปกติและมีค่าเฉลี่ย μ_1 กับ μ_2 ทฤษฎีที่กล่าวมานี้จะใช้ได้ก็ต่อเมื่อ σ_1^2 และ σ_2^2 นั้นทราบค่า หรือประมาณได้จากตัวอย่างขนาดโต $(n_1, n_2 \geq 30)$ เมื่อไม่ทราบค่าโดยทั่วไปเรามักไม่ทราบ σ_1^2 และ σ_2^2 จึงทำให้ไม่ทราบการแจกแจงที่แท้จริงของ $(\bar{x}_1 - \bar{x}_2)$ ถ้าเราสมมติว่า σ_1^2 และ σ_2^2 ที่ไม่ทราบค่านั้นมีค่าเท่ากันหรือใกล้เคียงกัน $\sigma_1^2 = \sigma_2^2 = \sigma^2$ แล้วเราสามารถประมาณ σ^2 ด้วย S_p^2 ดังนี้

$$S_p^2 = \left\{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2\right\} / (n_1 + n_2 - 2)$$

ในเมื่อ S_1^2 และ S_2^2 เป็นความแปรปรวนของตัวอย่างขนาด n_1 และ n_2 แล้วเราจะได้การแจกแจงของ $(\bar{x}_1 - \bar{x}_2)$ ตามทฤษฎีต่อไปนี้

ถ้า $\bar{x}_1, S_1^2$ และ $\bar{x}_2, S_2^2$ เป็นค่าเฉลี่ยและความแปรปรวนของตัวอย่างสุ่มที่เป็นอิสระกันและมีขนาด n_1 และ n_2 จากสองประชากรแบบปกติที่มีค่าเฉลี่ย μ_1, μ_2 และความแปรปรวน σ_1^2, σ_2^2 ซึ่งไม่ทราบค่า แต่ $\sigma_1^2 = \sigma_2^2 = \tau^2$ แล้วตัวสถิติ T

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{1/n_1 + 1/n_2}}$$

จะมีการแจกแจงแบบที ซึ่งมีองศาความเป็นอิสระ $= n_1 + n_2 - 2$

แต่ถ้า σ_1^2 และ σ_2^2 แตกต่างกันมาก ($\sigma_1^2 \neq \sigma_2^2$) แล้วตัวสถิติ T จะเป็น

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 - S_2^2/n_2}}$$

ซึ่งมีการแจกแจงแบบที (โดยประมาณ) แต่มี γ เป็นดังนี้

$$\gamma = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$$

ตัวอย่าง โรงงานผลิตหลอดไฟ 2 ยี่ห้อ ก. และ ข. ยี่ห้อ ก. มีอายุใช้งานเฉลี่ยมากกว่ายี่ห้อ ข. 100 ชั่วโมง ความแปรปรวนทั้งสองยี่ห้อเท่า ๆ กัน ทุกเดือนจะนำหลอดไฟยี่ห้อ ก. 16 หลอดและ ข. 9 หลอด มาทดสอบ แล้วนำมาคำนวณผลต่างของค่าเฉลี่ยจากตัวอย่างนั้น และคำนวณค่า T โรงงานจะพอใจถ้า T อยู่ระหว่าง $-t_{.025}$ และ $t_{.025}$

ถ้าเดือนหนึ่งคำนวณค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของยี่ห้อ ก. และ ข. ได้เป็น

ก. $\bar{X}_1 = 521,\ S_1 = 40$

ข. $\bar{X}_2 = 440,\ S_2 = 35$

โรงงานจะพอใจต่อการผลิตหรือไม่ ?

สำหรับ $\gamma = 16 + 9 - 2 = 23$ เราได้ $t_{.025} = 2.069$ ดังนั้นโรงงาน จะพอใจถ้าค่า T อยู่ระหว่าง - 2.069 และ 2.069

ความแปรปรวนผสม (Pooled variances) S_p^2 จะเป็น

$$S_p^2 = \frac{15(40)^2 + 8(35)^2}{16 + 9 - 2} = 1469.565$$

$$S_p = 38.34$$

เพราะฉะนั้น $T = \dfrac{(521 - 440) - 100}{38.34\sqrt{1/16 + 1/9}} = -1.2$

ค่า T นี้อยู่ระหว่าง -2.069 และ 2.069 โรงงานจึงพอใจและคุณภาพของสองยี่ห้อยังคงอยู่

ในกรณีที่ตัวอย่างไม่เป็นอิสระกัน (Dependent) ทฤษฎีที่กล่าวมาก็ใช้ไม่ได้ ถ้าให้จับคู่สังเกตของทั้งสองตัวอย่าง และพิจารณาตัวสถิติ D ซึ่งแทนผลต่างของค่าสังเกตแต่ละคู่ดังนั้น $D_i = X_{1i} - X_{2i}$ จึงเป็นค่าของตัวแปรเชิงสุ่ม D ซึ่งมีค่าเฉลี่ยเป็น $\bar{D}$ และความแปรปรวน $S_D^2 = \Sigma (D_i - \bar{D})^2/(n - 1)$ เราจะได้ทฤษฎีต่อไปนี้

ให้ $D_1, D_2, \ldots, D_n$ เป็นผลต่างของ n คู่ค่าสังเกตในตัวอย่างสุ่มที่เลือกมาจากประชากรแบบปกติของผลต่างที่มีค่าเฉลี่ย μ_D และความแปรปรวน σ_D^2 แล้วตัวสถิติ T

$T = (\bar{D} - \mu_D)/(S_D/\sqrt{n})$ จะมีการแจกแจงแบบที ซึ่งมี $\gamma = n - 1$

ตัวอย่าง สุ่มพนักงานพิมพ์ดีดมา 6 คน และให้พิมพ์ข้อความที่กำหนดด้วยพิมพ์ดีด 2 ยี่ห้อ แล้วบันทึกเวลาที่ใช้จากแต่ละยี่ห้อและหาผลต่างของเวลาที่ใช้ จะได้ดังนี้

พนักงานพิมพ์ดีด	เครื่องพิมพ์ ก.	เครื่องพิมพ์ ข.	ผลต่าง
1	23	19	4
2	18	18	0
3	29	24	5
4	22	23	-1
5	33	31	2
6	20	22	-2

ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของผลต่างของเวลาที่ใช้จะเป็นดังนี้

$\overline{D} = 4/3$, $S_D = 2.8$

ถ้าเครื่องพิมพ์ดีดมีความเร็วเท่ากัน นั่นคือ $\mu_D = 0$ ค่า T ที่คำนวณได้จะเป็น

$$T = \frac{4/3 - 0}{2.8/\sqrt{6}} = 1.166$$

4.7 การแจกแจงของการสุ่มตัวอย่างของสัดส่วน (Sampling Distribution of Proportions)

ลองพิจารณาถึงการทดลองแบบทวินาม ซึ่งแต่ละครั้ง (Trial) มีผลทดลองเป็น S (สำเร็จ) หรือ F (ไม่สำเร็จ) ด้วยความน่าจะเป็น π และ $1 - \pi$ ตามลำดับเมื่อต้องการตัวอย่างขนาด n เราก็ทำการทดลอง n ครั้ง การแจกแจงของการสุ่มตัวอย่างของ X ซึ่งแทนจำนวน **S** ในตัวอย่างขนาด n นั้น จะเป็นแบบปกติที่มีค่าเฉลี่ยและความแปรปรวน ดังนี้

$$\mu = n\pi; \quad \sigma^2 = n\pi(1 - \pi)$$

ถ้าค่าของ π ไม่ใกล้ 0 หรือ 1 มากนัก และขนาดตัวอย่างโตพอ

สำหรับแต่ละตัวอย่างขนาด n จากประชากรทวินาม ถ้าเราพิจารณาสัดส่วนของ **S** หรือ P ค่า $P = X/n$ จะผันแปรตามตัวอย่างขนาด n และ p เป็นค่าของตัวสถิติ $P = X/n$ การแจกแจงตัวอย่างของ P จะเป็นดังนี้

ถ้าตัวอย่างสุ่มขนาด n จากประชากรทวินามที่มีค่าเฉลี่ย $\mu = n\pi$ และความแปรปรวน $\sigma^2 = n\pi(1 - \pi)$ แล้วการแจกแจงตัวอย่างของ P (สัดส่วนของ **S** สำเร็จ) จะเป็นแบบปกติที่มี

$$\mu_p = \pi; \quad \sigma_p^2 = \pi(1 - \pi)/n$$

นั่นคือ $Z = \dfrac{P - \pi}{\sqrt{\pi(1 - \pi)/n}}$ จะมีการแจกแจงแบบปกติมาตรฐาน N(0, 1)

จากทฤษฎีนี้เราประยุกต์กับการทดลองแบบไฮเปอร์จีโอเมตริค ถ้าขนาดตัวอย่างเล็กเมื่อเทียบกับขนาดประชากร N โดยให้ $\pi = k/N$ เมื่อ k แทนจำนวน **S** ที่มีอยู่ในประชากร และยัง

ประยุกต์ได้กับประชากรที่มีขนาดน้อยและสุ่มแบบแทนที่ แต่ถ้าตัวอย่างมีขนาดมากกว่า 20% ของขนาดของประชากร และสุ่มแบบไม่แทนที่เราก็จะได้การแจกแจงของ P แบบเดียวกันอีก โดยมีค่าเฉลี่ยและความแปรปรวน ดังนี้

$$\mu_p = \pi, \quad \sigma_p^2 = \frac{\pi(1-\pi)}{n}\,\frac{N-n}{N-1}$$

ถ้าพิจารณาตัวอย่างสุ่มที่เป็นอิสระกันและมีขนาด n_1, n_2 ซึ่งเลือกสุ่มจากสองประชากรทวินามที่มีค่าเฉลี่ย $n_1\pi_1$ และ $n_2\pi_2$ กับความแปรปรวน $n_1\pi_1(1-\pi_1)$ และ $n_2\pi_2(1-\pi_2)$ ตามลำดับ เราให้ P_1 และ P_2 แทนสัดส่วนของ S ในแต่ละตัวอย่าง แล้วการแจกแจงของผลต่าง $(P_1 - P_2)$ สำหรับตัวอย่างต่าง ๆ ขนาด n_1 และ n_2 จะเป็นแบบปกติที่มีค่าเฉลี่ยและความแปรปรวนดังนี้

$$E(P_1 - P_2) = \pi_1 - \pi_2$$

$$V(P_1 - P_2) = \pi_1(1-\pi_1)/n_1 + \pi_2(1-\pi_2)/n_2$$ ดังนั้นเราจะได้ตัวสถิติ Z

$$Z = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{V(P_1 - P_2)}}$$

จะมีการแจกแจงแบบปกติมาตรฐาน

ถ้าทั้ง n_1 และ n_2 เท่ากับหรือมากกว่า 20 แล้วการแจกแจงตัวอย่างของ $(P_1 - P_2)$ จะใกล้กับการแจกแจงแบบปกติมากขึ้น

4.8 การแจกแจงของการสุ่มตัวอย่างของความแปรปรวน (Sampling Distribution of Sample Variances)

ถ้าตัวอย่างสุ่มขนาด n จากประชากรที่มีค่าเฉลี่ย μ และความแปรปรวน σ^2 และเมื่อคำนวณความแปรปรวนของตัวอย่างแล้วเราจะได้ตัวสถิติ S^2 มีการแจกแจงดังทฤษฎีต่อไปนี้

ถ้าตัวอย่างสุ่มขนาด n จากประชากรที่มีโมเมนต์ที่ 4 หรือ $\mu_4 = E(X-\mu)^4$ หาค่าได้แล้ว การแจกแจงของการสุ่มตัวอย่างของความแปรปรวน s^2 จะมีค่าเฉลี่ยและความแปรปรวน ดังนี้

$$E(S^2) = \sigma^2, \quad V(S^2) = (1/n)\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right)$$

ถ้าตัวอย่างจากประชากรแบบปกติ เราจะได้ $V(S^2) = 2\sigma^4/(n-1)$

ถ้า n โต $(n \geq 100)$ แล้วการแจกแจง $(S^2 - \sigma^2)/\sigma^2\sqrt{\frac{2}{n-1}}$ จะมีลักษณะใกล้กับการแจกแจงแบบปกติมาตรฐาน

สำหรับประชากรแบบปกติและตัวอย่างมีขนาดมากกว่า 100 แล้วการแจกแจงของการสุ่มตัวอย่างของส่วนเบี่ยงเบนมาตรฐานจะมีค่าเฉลี่ยและความแปรปรวนดังนี้

$$E(s) = \left(\frac{4n-4}{4n-3}\right)\sigma \simeq \sigma$$
$$V(s) = \sigma^2/2(n-1)$$

และเราจะได้ว่า $\frac{s-\sigma}{\sigma/\sqrt{2(n-1)}}$ มีการแจกแจงแบบปกติมาตรฐาน

เนื่องจากการแจกแจงตัวอย่าง s^2 นั้นในทางปฏิบัติมีการประยุกต์ในทางสถิตินัอย แต่การแจกแจงของการสุ่มตัวอย่างของตัวสถิติ $(n-1)s^2/\sigma^2 = \Sigma(x-\bar{x})^2/\sigma^2$ มีการประยุกต์กันมากกว่า และมีการแจกแจงดังทฤษฎีต่อไปนี้

ถ้า s^2 เป็นความแปรปรวนของตัวอย่างสุ่มขนาด n จากประชากรแบบปกติที่มีความแปรปรวน σ^2 แล้วตัวสถิติ χ^2

$\chi^2 = (n-1)s^2/\sigma^2$ จะมีการแจกแจงแบบไคสแควร์ที่มี $\nu = n-1$

ถ้าจำนวน ν สูงขึ้น แล้วการแจกแจงของตัวสถิติ σ^2 จะมีลักษณะใกล้กับการแจกแจงแบบปกติมากขึ้น

ตัวอย่าง โรงงานผลิตแบตเตอรี่รถยนต์รับประกันว่า แบตเตอรี่ของโรงงานจะมีอายุโดยเฉลี่ย 3 ปี และส่วนเบี่ยงเบนมาตรฐาน 1 ปี

ถ้าตัวอย่างขนาด 5 ของแบตเตอรี่ มีอายุ 1.9, 2.4, 3.0, 3.5, 4.2 ปี แล้วโรงงานยังแน่ใจหรือไม่ว่าแบตเตอรี่ยังคงมีส่วนเบี่ยงเบนมาตรฐานเป็น 1 ปี

ความแปรปรวนของตัวอย่าง $s^2 = (1/4)(48.26 - (15)^2/5) = 0.815$

ดังนั้น $\chi^2 = 4(0.815)/1 = 3.26$

จากตารางไคสแควร์ เราได้ค่าของ χ^2 ที่มี $\nu = 4$ นั้นจะมีค่าอยู่ระหว่าง 0.484 และ 11.143 ด้วยความน่าจะเป็น 95% ดังนั้นความแปรปรวนของแบตเตอรี่ยังคงเดิมเพราะ $\chi^2 = 3.26$ อยู่ในช่วงนี้

4.9 การแจกแจงของการสุ่มตัวอย่างของอัตราส่วนของความแปรปรวน (Ratio of Variances)

เราทราบกันมาแล้วว่าอัตราส่วนของสองการแจกแจงที่เป็นอิสระแบบไคสแควร์ จะเป็นการแจกแจงแบบเอฟ และจากทฤษฎีที่ผ่านมาเราจะได้ว่า $\chi_1^2 = (n_1-1)S_1^2/\sigma_1^2$ หรือ $\chi_2^2 = (n_2-1)S_2^2/\sigma_2^2$ จะมีการแจกแจงแบบไคสแควร์ ที่มี $df = n_1 - 1$ และ $n_2 - 1$ ตามลำดับ

ดังนั้น $F = \chi_1^2\nu_1 \;/\; \chi_2^2/\nu_2 = \dfrac{((n_1-1)S_1^2/\sigma_1^2)/(n_1-1)}{((n_2-1)S_2^2/\sigma_2^2)/(n_2-1)} = S_1^2/\sigma_1^2 / S_2^2/\sigma_2^2$

จะมีการแจกแจง ดังทฤษฎีต่อไปนี้

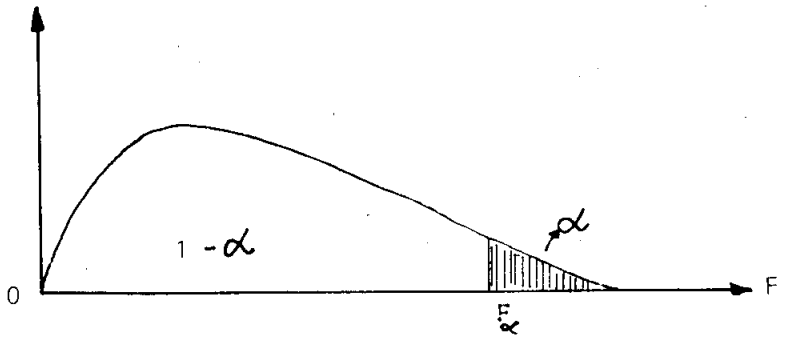
ถ้า S_1^2 และ S_2^2 เป็นความแปรปรวนของตัวอย่างสุ่มที่เป็นอิสระกัน โดยมีขนาดตัวอย่าง

n_1 และ n_2 จากประชากรแบบปกติสองประชากรที่มีความแปรปรวน σ^2_1 และ σ^2_2 ตามลำดับแล้ว

$$F = S_1^2/\sigma^2_1 \,/\, S^2_2/\sigma^2_2$$

จะมีการแจกแจงแบบเอฟที่มี $\nu = (n_1 - 1), (n_2 - 1)$

ให้ F_α เป็นค่า F ที่มีพื้นที่อยู่ทางขวาของค่านี้เป็น $\alpha = .1, .05, .025, .01, .005$ แต่ส่วนมากมักมี แต่ $\alpha = .05$ และ $.01$



แต่ถ้าเราต้องการหาค่า $F_{1-\alpha}$ ทางซ้าย เราจะหาได้กับทฤษฎีต่อไปนี้

ถ้า $F^{(\nu_1, \nu_2)}$ เป็นค่าของ F ที่มีองศาความเป็นอิสระ ν_1 ν_2 แล้ว $F_{1-\alpha}(\nu_1, \nu_2)$ ซึ่งเป็นค่าของ F จะหาได้ดังนี้

$$F_{1-\alpha}^{(\nu_1,\nu_2)} = 1/F_\alpha(\nu_2, \nu_1)$$

ตัวอย่างเช่น ถ้าค่า F ที่มี $\nu = (10, 6)$ มีพื้นที่ทางขวาของค่านี้เท่ากับ 0.05 เป็น 4.06 (ดูจากตาราง FOW) แล้วค่า F ที่มี $\nu = (6, 10)$ โดยมีพื้นที่ของขวามือเป็น 0.95 จะเป็น

$$F_{.95}^{(6,10)} = 1/F_{.05}^{(10,6)} = 1/4.06 = 0.246$$

นอกจากตัวสถิติที่กล่าวมาแล้ว ยังมีตัวสถิติอื่น ๆ อีกมากที่น่าสนใจ แต่จะได้กล่าวในเมื่อจะต้องใช้ตัวสถิตินั้น ๆ ในการอ้างอิงหรืออนุมาน