บทที่ 4
ผลการศึกษา

4.1 ความนำ

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อแสดงให้เห็นถึงแนวคิดของประชากรคงที่ และแนวคิดของอภิประชากรที่ใช้ในการอนุมานหรืออธิบายเกี่ยวกับประชากรอันตะ ซึ่งจะทำให้เกิดความเข้าใจที่ชัดเจนสำหรับเป็นรากฐานในการพัฒนาแนวคิดของทฤษฎีการสำรวจตัวอย่าง

เพื่อความสะดวกในการเสนอผลการศึกษาจึงกำหนดสัญลักษณ์ต่าง ๆ ดังนี้

แนวคิดของประชากรคงที่ ;

N	หมายถึง	ขนาดประชากร
n	หมายถึง	ขนาดตัวอย่าง
Y	หมายถึง	คุณลักษณะประชากรที่สนใจศึกษา
y	หมายถึง	ค่าของคุณลักษณะประชากรที่สนใจศึกษา
$\overline{Y}$	หมายถึง	ตัวแปรสุ่มแสดงค่าเฉลี่ยประชากร
$\bar{y}$	หมายถึง	ค่าหรือผลลัพธ์ที่ได้จากตัวแปรสุ่มแสดงค่าเฉลี่ยประชากร
f	หมายถึง	สัดส่วนการสุ่ม (sampling fraction)
S^2	หมายถึง	ความแปรปรวนประชากร

แนวคิดของอภิประชากร ;

N	หมายถึง	ขนาดประชากร
$v(s)$	หมายถึง	ขนาดตัวอย่าง
Y	หมายถึง	ตัวแปรสุ่มแสดงคุณลักษณะประชากรที่สนใจศึกษา
$\overline{Y}$	หมายถึง	ตัวแปรสุ่มแสดงค่าเฉลี่ยประชากร
$\overline{Y}_s$	หมายถึง	ตัวแปรสุ่มแสดงค่าเฉลี่ยตัวอย่าง
$\overline{Y}_{\tilde{s}}$	หมายถึง	ตัวแปรสุ่มแสดงค่าเฉลี่ยของหน่วยที่ไม่ได้ถูกเลือกเป็นตัวอย่าง
$\bar{y}$	หมายถึง	ค่าหรือผลลัพธ์ที่ได้จากตัวแปรสุ่มแสดงค่าเฉลี่ยประชากร
$\bar{y}_s$	หมายถึง	ค่าหรือผลลัพธ์ที่ได้จากตัวแปรสุ่มแสดงค่าเฉลี่ยตัวอย่าง
$\bar{y}_{\tilde{s}}$	หมายถึง	ค่าหรือผลลัพธ์ที่ได้จากตัวแปรสุ่มแสดงค่าเฉลี่ยของหน่วยที่ไม่ได้ถูกเลือกเป็นตัวอย่าง
σ^2	หมายถึง	ความแปรปรวนประชากร

ξ	หมายถึง	ฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วมของค่าสังเกตในประชากร $Y_1,\ldots,Y_N$
ξ_θ	หมายถึง	ฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วมของค่าสังเกตในประชากร $Y_1,\ldots,Y_N$ ที่ขึ้นอยู่กับพารามิเตอร์ θ ที่ไม่ทราบค่า
S	หมายถึง	เซตของหน่วยทั้งหมดในประชากร โดยที่ S เป็นตัวแปรสุ่มที่มีค่าเป็น s
s	หมายถึง	เซตของหน่วย (Unit) ในประชากรที่ถูกเลือกเป็นตัวอย่างขนาด $v(s)$
$\tilde{s}$	หมายถึง	เซตของหน่วย (Unit) ในประชากรที่ไม่ได้ถูกเลือกเป็นตัวอย่างขนาด $N-v(s)$
k	หมายถึง	ลำดับที่ (label)
k_i	หมายถึง	ลำดับที่ (label) ของตัวอย่างที่เป็นลำดับที่ i ในประชากร
r_i	หมายถึง	อันดับที่ (Rank) ของตัวแปรตัวที่ i
p	หมายถึง	ฟังก์ชันความน่าจะเป็นที่มีโดเมนเป็นตัวอย่าง เรนจ์เป็นเลขจำนวนจริงที่มีค่าอยู่ระหว่าง 0 ถึง 1 [1]
T	หมายถึง	ตัวสถิติที่ใช้ในการประมาณค่าเฉลี่ยประชากร
U	หมายถึง	ตัวสถิติที่ใช้ในการประมาณค่าเฉลี่ยของหน่วยที่ไม่ได้ถูกเลือกเป็นตัวอย่าง ($\overline{Y}_{\tilde{s}}$)
$\mathfrak{I}$	หมายถึง	ค่าคาดหวังของอภิประชากร
f_s	หมายถึง	สัดส่วนการสุ่ม (sampling fraction)

4.2 แนวคิดของประชากรคงที่กับแนวคิดของอภิประชากร

จุดมุ่งหมายที่สำคัญของการสำรวจตัวอย่างคือ การสุ่มตัวอย่างจากประชากรที่สนใจและการอนุมานทางสถิติโดยใช้รายละเอียดเกี่ยวกับข้อเท็จจริงที่เก็บรวบรวมได้จากตัวอย่าง มาทำการวิเคราะห์เพื่อหาข้อสรุปเกี่ยวกับคุณลักษณะประชากรที่สนใจนั้นให้มีคุณภาพสูงสุด ภายใต้ข้อจำกัดทางด้านทรัพยากร

แนวคิดสำหรับการสำรวจตัวอย่างที่ใช้กันอย่างแพร่หลายในปัจจุบัน เป็นแนวคิดที่พิจารณาให้ค่าของหน่วยประชากรในประชากรที่สนใจเป็นค่าคงที่ที่ไม่ทราบค่า ต้องทำการสำรวจและสังเกตค่ามาวิเคราะห์เพื่อหาข้อสรุปเกี่ยวกับคุณลักษณะประชากรที่สนใจศึกษา เรียกแนวคิดนี้ว่าแนวคิดของประชากรคงที่ สำหรับอีกแนวคิดหนึ่งที่แตกต่างจากแนวคิดข้างต้น คือแนวคิดที่

---

[1] ดูตัวอย่างฟังก์ชันความน่าจะเป็น ที่ภาคผนวก ก.1, หน้า 49.

พิจารณาให้ค่าของหน่วยประชากรในประชากรที่สนใจ เป็นตัวแปรสุ่มที่ถูกกำกับด้วยโครงสร้างเชิงสโตแคสติก หรือฟังก์ชันการแจกแจงความน่าจะเป็น โดยที่โครงสร้างหรือฟังก์ชันนี้ส่วนใหญ่จะมีข้อมูลอื่นที่เกี่ยวข้องกับตัวแปรหรือคุณลักษณะประชากรที่สนใจศึกษามาสนับสนุน เรียกแนวคิดนี้ว่า แนวคิดของอภิประชากร ตัวอย่างเช่น สนใจศึกษาปริมาณข้าวสาลีที่ปลูกได้ในปีหนึ่ง ๆ ดังนั้นหน่วยประชากรในที่นี้คือทุ่งข้าวสาลี และคุณลักษณะประชากรที่สนใจศึกษาคือปริมาณข้าวสาลีในปีที่สนใจศึกษา แนวคิดที่ใช้กันโดยทั่วไปในการอนุมานปัญหานี้คือ การพิจารณาให้ปริมาณข้าวสาลีแต่ละทุ่งข้าวสาลีในประชากรเป็นค่าคงที่แต่ไม่ทราบค่า ต้องทำการสุ่มตัวอย่างทุ่งข้าวสาลีเพื่อสำรวจหาปริมาณข้าวสาลี ดังนั้นเมื่อทุ่งข้าวสาลีหนึ่ง ๆ ถูกสุ่มขึ้นมาเป็นตัวอย่าง ทำการวัดปริมาณข้าวสาลีของทุ่งนั้น โดยที่การวัดปริมาณข้าวสาลีของทุ่งข้าวสาลีนั้นไม่มีความคลาดเคลื่อนเกิดขึ้น แต่แนวคิดที่แตกต่างอีกแนวคิดหนึ่งที่สามารถนำมาใช้ในการอนุมานปัญหานี้คือ การพิจารณาให้ปริมาณข้าวสาลีของทุ่งข้าวสาลีเป็นค่าที่ถูกสร้างขึ้น ภายใต้โครงสร้างเชิงสโตแคสติก ซึ่งโดยทั่วไปโครงสร้างนี้จะรวมข้อมูลหรือความรู้อื่นที่เกี่ยวข้องมาสนับสนุน ปริมาณข้าวสาลีที่ได้จากการสุ่มจะมีความคลาดเคลื่อนเกิดขึ้นจากโครงสร้างเชิงสโตแคสติก เหตุผลหลักที่มีการศึกษาแนวคิดของอภิประชากร คือ เมื่อทราบประชากรทั้งหมดก็ย่อมทราบการแจกแจงของประชากร และเมื่อทำการสุ่มตัวอย่างจากประชากรนั้น ตัวอย่างที่ได้ควรมีการแจกแจงเช่นเดียวกับประชากร ซึ่งแตกต่างกับแนวคิดของประชากรคงที่ที่ไม่พิจารณาการแจกแจง (Distribution Free)[2] ดังนั้นแนวคิดของการสำรวจตัวอย่างสามารถแยกพิจารณาได้เป็น 2 แนวคิด คือ

แนวคิดของประชากรคงที่ (Fixed population Approach); เป็นแนวคิดที่พิจารณาให้ค่าของตัวแปรศึกษาที่ได้จากแต่ละหน่วยประชากรเป็นค่าคงที่ที่ไม่ทราบค่า หรือกล่าวได้ว่า เป็นแนวคิดที่ใช้ในการประมาณค่าพารามิเตอร์ที่เป็นคุณลักษณะประชากรที่สนใจเพียงอย่างเดียว

แนวคิดของอภิประชากร (Superpopulation Approach); เป็นแนวคิดที่พิจารณาให้ค่าของตัวแปรศึกษาที่ได้จากแต่ละหน่วยประชากรอยู่ในรูปของผลลัพธ์ของตัวแปรสุ่ม หรือกล่าวได้ว่า เป็นแนวคิดที่ใช้ในการประมาณค่าพารามิเตอร์ที่เป็นคุณลักษณะประชากรและพารามิเตอร์ที่อยู่ในการแจกแจงของความคลาดเคลื่อนสุ่ม (Random Error)

แนวคิดทั้งสองของการสำรวจตัวอย่าง มีรายละเอียดเป็นดังนี้

---

[2] Edward L. Korn and Barry I. Graubard, "Variance Estimation for Superpopulation Parameters," Statistica Sinica 8 (1998): 1131-1151.

4.2.1 แนวคิดของประชากรคงที่

ทฤษฎีการสำรวจตัวอย่างเป็นทฤษฎีทางสถิติที่ถูกพัฒนาขึ้นมาสำหรับพิจารณา และอธิบายคุณลักษณะประชากรของประชากรอันตะที่สนใจ ซึ่งคุณลักษณะประชากรที่พิจารณาโดยทั่วไป[3] คือ ค่ารวมประชากร (Population Total) ค่าเฉลี่ยประชากร (Population Mean) ค่าสัดส่วนประชากร (Population Proportion) และค่าอัตราส่วนประชากร (Population Ratio) สำหรับการสุ่มตัวอย่างเชิงความน่าจะเป็นเพื่อใช้ในการอนุมานไปสู่ประชากรอาจแบ่งได้เป็น 4 วิธีแม่บท คือ การสุ่มตัวอย่างแบบง่าย (Simple Random Sampling) การสุ่มตัวอย่างแบบมีระบบ (Systematic Sampling) การสุ่มตัวอย่างแบบแบ่งชั้นภูมิ (Stratified Sampling) และการสุ่มตัวอย่างแบบแบ่งกลุ่ม(Cluster Sampling) การสุ่มในการศึกษาครั้งนี้จะพิจารณาเฉพาะการสุ่มตัวอย่างแบบง่าย เนื่องจากการสุ่มตัวอย่างแบบง่ายเป็นพื้นฐานของการสุ่มตัวอย่างแบบอื่น ๆ

การสำรวจตัวอย่าง เป็นการวัดและบันทึกค่าของตัวแปรศึกษาจากทุกหน่วยที่ตกเป็นตัวอย่าง ให้ค่าของคุณลักษณะประชากรที่สนใจศึกษาในประชากร คือ $y_1, y_2, \ldots, y_N$ และค่าของคุณลักษณะประชากรที่สนใจศึกษาในตัวอย่าง คือ $y_1, y_2, \ldots, y_n$ โดยที่ค่าของคุณลักษณะประชากรที่สนใจศึกษาในตัวอย่าง ไม่จำเป็นต้องเป็นค่าเดียวกับค่าของคุณลักษณะประชากรที่สนใจศึกษาในประชากร n ค่าแรก

วิธีการสุ่มตัวอย่างแบบง่าย เป็นวิธีที่กำหนดให้ตัวอย่างแต่ละตัวอย่างที่เป็นไปได้มีโอกาสเกิดขึ้นเท่า ๆ กัน ดังนั้นเมื่อทำการสุ่มตัวอย่างแบบไม่ใส่คืน (without replacement) ขนาด n จากประชากรขนาด N จำนวนตัวอย่างที่เป็นไปได้ทั้งหมดคือ ${}^{N}C_{n}$ และความน่าจะเป็นที่ตัวอย่างแต่ละตัวอย่างถูกเลือกจะเท่ากันหมด คือ $\frac{1}{{}^{N}C_{n}}$ และถ้าทำการสุ่มตัวอย่างแบบใส่คืน (with replacement) ขนาด n จากประชากรขนาด N จำนวนตัวอย่างที่เป็นไปได้ทั้งหมดคือ N^n และความน่าจะเป็นที่ตัวอย่างแต่ละตัวอย่างถูกเลือกจะเท่ากันหมด คือ $\frac{1}{N^n}$

---

[3] สุชาดา กีระนันท์, ทฤษฎีและวิธีการสำรวจตัวอย่าง (กรุงเทพมหานคร: โรงพิมพ์จุฬาลงกรณ์มหาวิทยาลัย,2542), หน้า 29.

สัญลักษณ์ที่ใช้สำหรับคุณลักษณะประชากรของการสุ่มตัวอย่างแบบง่าย เป็นดังนี้

	ประชากร	ตัวอย่าง
ขนาด	N	n
ค่ารวม	$Y = \sum_{i=1}^{N} y_i$	$y = \sum_{i=1}^{n} y_i$
ค่าเฉลี่ย	$\bar{Y} = \sum_{i=1}^{N} y_i \Big/ N$	$\bar{y} = \sum_{i=1}^{n} y_i \Big/ n$
ค่าสัดส่วน	$\bar{Y} = \sum_{i=1}^{N} y_i \Big/ N$ โดยที่ $y_i = 1$ เมื่อหน่วยที่ i ใน ประชากรมีลักษณะตามที่ต้องการ $y_i = 0$ เมื่อหน่วยที่ i ใน ประชากรไม่มีลักษณะตามที่ต้องการ	$\bar{y} = \sum_{i=1}^{n} y_i \Big/ n$ โดยที่ $y_i = 1$ เมื่อหน่วยที่ i ใน ตัวอย่างมีลักษณะตามที่ต้องการ $y_i = 0$ เมื่อหน่วยที่ i ใน ตัวอย่างไม่มีลักษณะตามที่ต้องการ
ค่าอัตราส่วน	$R = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}}$	$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{y}{x} = \frac{\bar{y}}{\bar{x}}$

จะเห็นได้ว่า การประมาณค่าของคุณลักษณะประชากรที่สนใจต่าง ๆ ขึ้นอยู่กับค่าเฉลี่ยประชากรเป็นส่วนใหญ่ และในกรณีของการสำรวจตัวอย่างด้วยวิธีการสุ่มตัวอย่างแบบง่ายแบบไม่ใส่คืน พบว่า ตัวประมาณค่าเฉลี่ยประชากรมีคุณสมบัติของตัวประมาณที่ดีหลายประการ คือ เป็นตัวประมาณที่ไม่เอนเอียง มีความคงเส้นคงวา และมีความแปรปรวนใกล้เคียงกับที่กำหนดในทฤษฎีสถิติ ดังนั้นจึงพิจารณาเฉพาะค่าเฉลี่ยประชากรด้วยวิธีการสุ่มแบบง่ายแบบไม่ใส่คืนเพียงอย่างเดียว

การประมาณค่าเฉลี่ยประชากร (Predicting the Population Mean)

คุณภาพของตัวประมาณค่าคุณลักษณะประชากรขึ้นอยู่กับวิธีการสุ่มตัวอย่าง และวิธีการคำนวณค่าที่ได้จากตัวอย่าง การพิจารณาความเหมาะสมของตัวประมาณที่จะใช้นั้น ต้องพิจารณาที่คุณสมบัติของตัวประมาณ จากคุณสมบัติของความไม่เอนเอียง จะได้ว่า ตัวประมาณที่ไม่เอนเอียง คือ ตัวประมาณที่ให้ค่าคาดหวังของตัวประมาณเท่ากับค่าประชากร ซึ่งค่าเฉลี่ยตัวอย่าง $(\bar{y})$ เป็นตัวประมาณที่ไม่เอนเอียงของค่าเฉลี่ยประชากร $(\bar{Y})$

จากค่าเฉลี่ยตัวอย่าง

$$\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n}$$

จะได้ $E(\bar{y}) = \frac{\sum \bar{y}}{{}^{N}C_n} = \frac{\sum (y_1 + y_2 + \ldots + y_n)}{n[N!/n!(N-n)!]}$

พิจารณาเฉพาะ $\sum (y_1 + y_2 + \ldots + y_n)$ จะได้ว่า

จำนวนวิธีที่ตัวอย่างขนาด n มีหน่วยที่ i ของประชากรปรากฏอยู่ คือ ${}^{N-1}C_{n-1}$

ดังนั้น $\sum (y_1 + y_2 + \ldots + y_n) = \frac{(N-1)!}{(n-1)!(N-n)!}(y_1 + y_2 + \ldots + y_N)$

$$\therefore \quad E(\bar{y}) = \frac{(N-1)!}{(n-1)!(N-n)!}\frac{n!(N-n)!}{nN!}(y_1 + y_2 + \ldots + y_N)$$

$$= \frac{(y_1 + y_2 + \ldots + y_N)}{N}$$

$$= \bar{Y}$$

เมื่อได้ตัวประมาณที่ใช้สำหรับประมาณค่าเฉลี่ยประชากรแล้ว ต้องทำการวัดคุณภาพของตัวประมาณที่ได้ด้วย เพราะคุณภาพของตัวประมาณแสดงถึงความน่าเชื่อถือของวิธีทางสถิติที่ใช้ในการเก็บข้อมูล เกณฑ์ที่ใช้ในการวัดคุณภาพของตัวประมาณคือ ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Mean Square Error ; MSE) เนื่องจากค่าเฉลี่ยตัวอย่างเป็นตัวประมาณที่ไม่เอนเอียงของค่าเฉลี่ยประชากร ดังนั้นจึงใช้ความแปรปรวนของตัวประมาณในการวัดคุณภาพ จะได้ว่า

ความแปรปรวนของค่าเฉลี่ยตัวอย่าง $\bar{y}$ จากตัวอย่างสุ่มแบบง่ายแบบไม่ใส่คืน คือ

$$V(\bar{y}) = E(\bar{y} - \bar{Y})^2 = \frac{S^2}{n}\frac{(N-n)}{N} = \frac{S^2}{n}(1-f)$$

เมื่อ $f = n/N$ คือสัดส่วนการสุ่ม (Sampling fraction)

และ $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$ คือความแปรปรวนของประชากร

เพื่อให้การพิสูจน์ $V(\bar{y}) = E(\bar{y} - \bar{Y})^2 = \frac{S^2}{n}\frac{(N-n)}{N} = \frac{S^2}{n}(1-f)$

เป็นไปด้วยความเข้าใจ จะขอแสดงก่อนว่า

$$E(y_1 + y_2 + \ldots + y_n) = \frac{n}{N}(y_1 + y_2 + \ldots + y_N)$$

เพื่อให้เห็นว่า แต่ละค่าของตัวแปรของหน่วยประชากรในประชากรมีโอกาสที่จะตกอยู่ในตัวอย่างเท่า ๆ กันคือ $\frac{n}{N}$

จาก $E(\bar{y}) = \bar{Y}$

$$E[\frac{y_1 + \ldots + y_n}{n}] = \frac{y_1 + \ldots + y_N}{N}$$

$$\therefore \quad E[y_1 + \ldots + y_n] = \frac{n}{N}[y_1 + \ldots + y_N]$$

$$= \frac{n}{N}y_1 + \frac{n}{N}y_2 + \ldots + \frac{n}{N}y_N$$

ดังนั้น เมื่อค่าของตัวแปรของแต่ละหน่วยประชากรในประชากร มีโอกาสที่จะตกอยู่ในตัวอย่างเท่า ๆ กัน คือ $\frac{n}{N}$ แล้ว ค่าของฟังก์ชันที่ขึ้นอยู่กับตัวแปรของแต่ละหน่วยประชากรในประชากรก็มีโอกาสที่จะตกอยู่ในตัวอย่างเท่า ๆ กัน คือ $\frac{n}{N}$ ด้วย

ให้ $u(y_i)$ เป็นค่าของฟังก์ชันที่ขึ้นอยู่กับตัวแปรของหน่วยประชากรที่ i

ดังนั้น

$$E\big(u(y_1) + u(y_2) + \ldots + u(y_n)\big) = \frac{n}{N}\big(u(y_1) + u(y_2) + \ldots + u(y_N)\big)$$

ถ้าให้ $u(y_i) = (y_i - \bar{Y})^2$ จะได้ว่า

$$E[(y_1 - \bar{Y})^2 + \ldots + (y_n - \bar{Y})^2] = \frac{n}{N}[(y_1 - \bar{Y})^2 + \ldots + (y_N - \bar{Y})^2]$$

และจะขอแสดงด้วยว่า

$$E(y_1y_2 + y_1y_3 + \ldots + y_{n-1}y_n)$$

$$= \frac{n(n-1)}{N(N-1)}(y_1y_2 + y_1y_3 + \ldots + y_{N-1}y_N)$$

เพื่อให้เห็นว่า แต่ละค่าของตัวแปรของสองหน่วยประชากรใด ๆ ในประชากรมีโอกาสที่จะตกอยู่ในตัวอย่างเดียวกันเท่า ๆ กันคือ $\frac{n(n-1)}{N(N-1)}$ ซึ่งคำนวณได้จาก

$$P(y_i, y_j \text{ จะตกอยู่ในตัวอย่างเดียวกัน};\ i \neq j) = \frac{{}^{N-2}C_{n-2}}{{}^{N}C_{n}}$$

$$= \frac{\frac{(N-2)!}{(n-2)!(N-2-(n-2))!}}{\frac{N!}{n!(N-n)!}}$$

$$= \frac{n(n-1)}{N(N-1)}$$

ดังนั้น เมื่อค่าของตัวแปรของสองหน่วยประชากรใด ๆ ในประชากรมีโอกาสที่จะตกอยู่ในตัวอย่างเดียวกันเท่า ๆ กัน คือ $\frac{n(n-1)}{N(N-1)}$ แล้วค่าของฟังก์ชันที่ขึ้นอยู่กับตัวแปรร่วมของสองหน่วยประชากรใด ๆ ในประชากรก็มีโอกาสที่จะตกอยู่ในตัวอย่างเดียวกันเท่า ๆ กัน คือ $\frac{n(n-1)}{N(N-1)}$ ด้วย

ให้ $u(y_i, y_j)$ เป็นค่าของฟังก์ชันที่ขึ้นอยู่กับตัวแปรร่วมของสองหน่วยประชากรใด ๆ เมื่อ $i, j = 1,\ldots,n$ และ $i \neq j$

ดังนั้น

$$E[\sum_{i\neq}^{n}\sum_{j}^{n} u(y_i, y_j)] = \frac{n(n-1)}{N(N-1)} E[\sum_{i\neq}^{N}\sum_{j}^{N} u(y_i, y_j)]$$

ถ้าให้ $u(y_i, y_j) = (y_i - \bar{Y})(y_j - \bar{Y})$ เมื่อ $i, j = 1,\ldots,n$ และ $i \neq j$

จะได้ว่า

$$E[(y_1 - \bar{Y})(y_2 - \bar{Y}) + (y_1 - \bar{Y})(y_3 - \bar{Y}) + \ldots + (y_{n-1} - \bar{Y})(y_n - \bar{Y})$$
$$= \frac{n(n-1)}{N(N-1)}\{(y_1 - \bar{Y})(y_2 - \bar{Y}) + \ldots + (y_{N-1} - \bar{Y})(y_N - \bar{Y})\}$$

เมื่อเข้าใจในแนวคิดข้างต้นแล้ว จะแสดงให้เห็นต่อว่า $V(\bar{y}) = \frac{S^2}{n}(1-f)$

จาก $(y_1 + y_2 + \ldots + y_N)^2 = y_1^2 + y_2^2 + \ldots + y_N^2$
$$+ 2[(y_1y_2) + (y_1y_3) + \ldots + (y_{N-1}y_N)]$$

จะได้ว่า $$(\sum_{i=1}^{N} y_i)^2 = \sum_{i=1}^{N} y_i^2 + 2[\sum_{i \neq j}^{N}\sum^{N} (y_i y_j)]$$

$$\therefore \quad 2[\sum_{i \neq j}^{N}\sum^{N} (y_i y_j)] = (\sum_{i=1}^{N} y_i)^2 - \sum_{i=1}^{N} y_i^2$$

ให้ $u(y_i)$ เป็นค่าของฟังก์ชันที่ขึ้นอยู่กับตัวแปรของหน่วยประชากรที่ i

$u(y_i, y_j)$ เป็นค่าของฟังก์ชันที่ขึ้นอยู่กับตัวแปรร่วมของสองหน่วยประชากรใด ๆ

ดังนั้น

$$(u(y_1) + \ldots + u(y_N))^2 = u(y_1)^2 + \ldots + u(y_N)^2$$
$$+ 2[u(y_1, y_2) + \ldots + u(y_{N-1}, y_N)]$$

$$(\sum_{i=1}^{N} u(y_i))^2 = \sum_{i=1}^{N} u(y_i)^2 + 2[\sum_{i \neq j}^{N}\sum^{N} u(y_i, y_j)]$$

$$\therefore \quad 2[\sum_{i \neq j}^{N}\sum^{N} u(y_i, y_j)] = (\sum_{i=1}^{N} u(y_i))^2 - \sum_{i=1}^{N} u(y_i)^2 \qquad \ldots(1)$$

จะเห็นได้ว่า

จาก $$n(\bar{y} - \bar{Y}) = n\frac{\sum_{i=1}^{n} y_i}{n} - n\bar{Y}$$

จะได้ $$n^2(\bar{y} - \bar{Y})^2 = \{(y_1 - \bar{Y}) + (y_2 - \bar{Y}) + \ldots + (y_n - \bar{Y})\}^2$$
$$= \{u(y_1) + u(y_2) + \ldots + u(y_n)\}^2$$
$$= \sum_{i=1}^{n} u(y_i)^2 + 2[\sum_{i \neq j}^{n}\sum^{n} u(y_i, y_j)]$$

ดังนั้น

$$E(n^2(\bar{y}-\bar{Y})^2) = E[\sum_{i=1}^{n} u(y_i)^2] + 2E[\sum_{i\neq}^{n}\sum_{j}^{n} u(y_i, y_j)]$$

$$E(n^2(\bar{y}-\bar{Y})^2) = \frac{n}{N}\{\sum_{i=1}^{N} u(y_i)^2 + 2\frac{n-1}{N-1}[\sum_{i\neq}^{n}\sum_{j}^{n} u(y_i, y_j)]\} \quad ...(2)$$

พิจารณาเฉพาะพจน์หลัง จะได้ว่า จาก (1) ;

$$2\frac{(n-1)}{(N-1)}[\sum_{i\neq}^{n}\sum_{j}^{n} u(y_i, y_j)] = \frac{(n-1)}{(N-1)}[(\sum_{i=1}^{N} u(y_i))^2 - \sum_{i=1}^{N} u(y_i)^2]$$

$$= \frac{(n-1)}{(N-1)}(\sum_{i=1}^{N} u(y_i))^2 - \frac{(n-1)}{(N-1)}\sum_{i=1}^{N} u(y_i)^2$$

แทนใน (2) จะได้ ;

$$E(n^2(\bar{y}-\bar{Y})^2) = \frac{n}{N}\{\sum_{i=1}^{N} u(y_i)^2 + \frac{(n-1)}{(N-1)}(\sum_{i=1}^{N} u(y_i))^2$$

$$-\frac{(n-1)}{(N-1)}\sum_{i=1}^{N} u(y_i)^2\}$$

$$= \frac{n}{N}\{(1-\frac{(n-1)}{(N-1)})\sum_{i=1}^{N} u(y_i)^2 + \frac{(n-1)}{(N-1)}(\sum_{i=1}^{N} u(y_i))^2\}$$

เมื่อกำหนดให้ $u(y_i) = (y_i - \bar{Y})^2$ และ $u(y_i, y_j) = (y_i - \bar{Y})(y_j - \bar{Y})$

ดังนั้น

$$E(n^2(\bar{y}-\bar{Y})^2) = \frac{n}{N}\left\{[1-\frac{(n-1)}{(N-1)}]\sum_{i=1}^{N}(y_i-\bar{Y})^2 + \frac{(n-1)}{(N-1)}[\sum_{i=1}^{N}(y_i-\bar{Y})]^2\right\}$$

ซึ่ง $\sum_{i=1}^{N}(y_i-\bar{Y})]^2 = [(y_1-\bar{Y}) + ... + (y_N-\bar{Y})]^2$

$$= [y_1 + ... + y_N - N\bar{Y}]^2$$

$$= [\sum_{i=1}^{N} y_i - N\frac{\sum_{i=1}^{N} y_i}{N}]^2$$

$$= 0$$

$$\therefore \quad E(\bar{y}-\bar{Y})^2 = \frac{1}{n^2}\frac{n}{N}\left(1-\frac{(n-1)}{(N-1)}\right)\sum_{i=1}^{N}(y_i-\bar{Y})^2$$

$$= \frac{(N-1)-(n-1)}{nN(N-1)}\sum_{i=1}^{N}(y_i-\bar{Y})^2$$

$$= \frac{N-n}{nN}S^2$$

เมื่อ $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{Y})^2$

ซึ่งจะเห็นได้ว่าความแปรปรวนของค่าเฉลี่ยตัวอย่าง $(V(\bar{y}))$ ขึ้นอยู่กับความแปรปรวนประชากร (S^2)ที่ไม่ทราบค่า จึงต้องทำการประมาณค่าความแปรปรวนประชากร(S^2)ด้วยความแปรปรวนตัวอย่าง (s^2)

เมื่อ $s^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$

โดยที่ s^2 เป็นตัวประมาณค่าที่ไม่เอนเอียงของ S^2

พิสูจน์

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

$$= \frac{1}{(n-1)}\sum_{i=1}^{n}[(y_i-\bar{Y})-(\bar{y}-\bar{Y})]^2$$

$$= \frac{1}{(n-1)}[\sum_{i=1}^{n}(y_i-\bar{Y})^2 - n(\bar{y}-\bar{Y})^2]$$

จาก $E[\sum_{i=1}^{n}(y_i-\bar{Y})^2] = \frac{n}{N}\sum_{i=1}^{N}(y_i-\bar{Y})^2$

$$= \frac{n(N-1)}{N}S^2$$

และ $E(\bar{y}-\bar{Y})^2 = \frac{(N-n)}{nN}S^2$

$$E[n(\bar{y}-\bar{Y})^2] = \frac{(N-n)}{N}S^2$$

ดังนั้น $E(s^2) = \frac{1}{(n-1)}\left[E(\sum_{i=1}^{n}(y_i - \bar{Y})^2) - E[n(\bar{y} - \bar{Y})^2] \right]$

$\therefore \quad E(s^2) = \frac{1}{(n-1)}\left[\frac{n(N-1)}{N} S^2 - \frac{(N-n)}{N} S^2 \right]$

$= S^2$

จากแนวคิดของประชากรคงที่ที่พิจารณาให้ค่าของตัวแปรศึกษาที่ได้จากหน่วยประชากรเป็นค่าคงที่ที่ไม่ทราบค่า ดังนั้นเมื่อทำการสุ่มตัวอย่างแบบง่าย ที่กำหนดให้ตัวอย่างแต่ละตัวอย่างที่เป็นไปได้มีโอกาสเกิดขึ้นเท่า ๆ กันนั้น ค่าเฉลี่ยตัวอย่างจะเป็นตัวประมาณที่ไม่เอนเอียงสำหรับค่าเฉลี่ยประชากร และคุณภาพของตัวประมาณสามารถวัดได้ด้วยความแปรปรวน ซึ่งความแปรปรวนของค่าเฉลี่ยตัวอย่างกรณีสุ่มตัวอย่างแบบง่ายแบบไม่ใส่คืน คือ

$$V(\bar{y}) = \frac{S^2}{n}(1-f)$$

เมื่อ $f = n/N$ คือสัดส่วนการสุ่ม (Sampling fraction)

และ $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$ คือความแปรปรวนของประชากร

4.2.2 แนวคิดของอภิประชากร

การสำรวจตัวอย่างด้วยแนวคิดของอภิประชากร เป็นการสำรวจที่พิจารณาให้เวคเตอร์ของค่าคุณลักษณะประชากรที่สนใจศึกษาในประชากร $y = (y_1,\ldots,y_N)$ เป็นค่าจริงที่อยู่ในรูปของผลลัพธ์ของเวคเตอร์ตัวแปรสุ่ม $\underset{\sim}{Y} = (Y_1,\ldots,Y_N)$ ที่มีการแจกแจงความน่าจะเป็น โดยที่กำหนดให้ฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วม (Joint Distribution Function) ของ $Y_1,\ldots,Y_N$ คือ ξ

หลักโดยทั่วไปของแนวคิดของอภิประชากร คือ

1. ประชากรอันตะเป็นเซตหนึ่งที่ดึงมาจากเอกภพที่ใหญ่กว่า
2. ฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วม ξ เป็นฟังก์ชันที่ใช้ในการอธิบายวิธีการสุ่มหรือกระบวนการที่เกิดขึ้นจริง
3. สำหรับการอนุมานแบบเบย์ พิจารณาให้ฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วม ξ เป็นการแจกแจงก่อน (prior distribution) และพิจารณาให้ค่าที่ไม่ได้ถูกเลือกเป็นตัวอย่างของ $y_1,\ldots,y_N$ เป็นพารามิเตอร์ที่ต้องหาการแจกแจงภายหลัง (posterior distribution)

4. ฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วม ξ สามารถหาหรือคำนวณได้จากวิธีการทางคณิตศาสตร์
5. แนวคิดของอภิประชากรมีประโยชน์ในส่วนของการลดความคลาดเคลื่อนที่ไม่ได้เกิดจากการสุ่มตัวอย่าง (Nonsampling Error)

ถ้า $y_1,...,y_N$ เป็นค่าที่ไม่คงที่ในประชากรอันตะ พิจารณาค่าเหล่านี้เป็นค่าจริงของตัวแปรสุ่ม $Y_1,...,Y_N$ ที่สามารถอธิบายความไม่แน่นอนเกี่ยวกับค่าจริงที่เกิดขึ้นได้ด้วยตัวแบบความน่าจะเป็น (Probabilistic Model)

แนวคิดของการอนุมานภายใต้ข้อสมมติเกี่ยวกับอภิประชากรสามารถแยกได้เป็น 2 แนวคิดใหญ่ ๆ ซึ่งคล้ายคลึงกับแนวคิดของการอนุมานในทฤษฎีสถิติ คือ

1. แนวคิดที่ใช้การอนุมานแบบคลาสสิค พิจารณาพารามิเตอร์ของตัวแบบในฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วม ξ เป็นพารามิเตอร์ที่ไม่ทราบค่าต้องทำการประมาณเป็นอันดับแรก

2. แนวคิดที่ใช้การอนุมานแบบเบย์ พิจารณาพารามิเตอร์ของตัวแบบในฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วม ξ ที่ไม่ทราบค่าด้วยการแจกแจงก่อน (Prior Distribution)

ธรรมชาติของสถานการณ์ต่าง ๆ จะแสดงตัวแบบหรือสูตรที่ใช้เป็นความรู้เบื้องต้น (prior knowledge) เกี่ยวกับประชากร ซึ่งตัวแบบหรือสูตรเหล่านี้ได้จากประสบการณ์หรือความเชื่อส่วนบุคคล โดยที่ตัวแบบอภิประชากร (Superpopulation Model) ต่างจากตัวแบบเบย์เซียน (Bayesian Model) ในส่วนของการพิจารณาความเชื่อส่วนบุคคล

ตัวแบบอภิประชากร (Superpopulation Model) เป็นตัวแบบที่แสดงเงื่อนไขของเซตที่เป็นตัวกำหนดกลุ่มของการแจกแจงที่ขึ้นอยู่กับฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วม ξ

ถ้า $Q = Q(Y_1,...,Y_N)$ เป็นฟังก์ชันของตัวแปรสุ่ม $Y_1,...,Y_N$;

ค่าเฉลี่ยของอภิประชากร ($\xi - \exp ected$) ของ Q คือ

$$\Im(Q) = \int Q d\xi$$

ความแปรปรวนของอภิประชากร ($\xi - \mathrm{var}\, iance$) ของ Q คือ

$$V(Q) = \int \{Q - \Im(Q)\}^2 d\xi$$

ถ้า $Q_1 = Q_1(Y_1,...,Y_N)$ และ $Q_2 = Q_2(Y_1,...,Y_N)$ เป็นสองฟังก์ชันใด ๆ ของ $Y_1,...,Y_N$ ความแปรปรวนร่วมของอภิประชากร ($\xi - \text{cov}\,ariance$) ของ Q_1 และ Q_2 คือ

$$C(Q_1,Q_2) = \int \{Q_1 - \Im(Q_1)\}\{Q_2 - \Im(Q_2)\}d\xi$$

เนื่องจากค่า y ในประชากรอันตะไม่ใช่ค่าคงที่ ดังนั้นสำหรับหน่วยใด ๆ ในประชากร $k,l = 1,...,N$ จะได้

ค่าเฉลี่ยของหน่วยที่ k ในอภิประชากรคือ $\mu_k = \Im(Y_k)$

ความแปรปรวนของหน่วยที่ k ในอภิประชากรคือ $\sigma_k^2 = V(Y_k)$

ความแปรปรวนร่วมของหน่วยที่ k และ หน่วยที่ l สำหรับ $(k \neq l)$ ในอภิประชากรคือ

$$\sigma_{kl} = C(Y_k,Y_l)$$

ดังนั้น ค่าเฉลี่ยประชากรในอภิประชากร คือ $\bar{\mu} = \dfrac{\sum_1^N \mu_k}{N}$

การประมาณค่าเฉลี่ยประชากร (Predicting the Population Mean)

วิธีการประมาณค่าเฉลี่ยประชากรสำหรับอภิประชากรนั้นเริ่มต้นจากการสุ่มค่าสังเกต y_k ขนาด $\upsilon(s)$ จากประชากรขนาด N ขึ้นมาเป็นตัวอย่าง จากนั้นทำการทำนาย (predict) ค่าสังเกต y_k ที่เหลือขนาด $N - \upsilon(s)$ เพื่อใช้ในการหาค่าเฉลี่ยประชากร $\bar{y} = \sum_1^N y_k / N$ ซึ่งอภิประชากรเกี่ยวข้องกับพารามิเตอร์ θ ที่ไม่ทราบค่าที่อยู่ในฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วม ξ ดังนั้นการทำนายค่าสังเกต y_k ขนาด $N - \upsilon(s)$ นั้นขึ้นอยู่กับการประมาณค่า (Estimation) พารามิเตอร์ θ ที่ไม่ทราบค่าโดยการใช้ค่าสังเกต y_k ที่ได้จากตัวอย่างขนาด $\upsilon(s)$ การประมาณค่าพารามิเตอร์นี้ จะคล้ายคลึงกับทฤษฎีสถิติ (statistical theory) ที่ใช้วิธีการประมาณค่าแบบจุดด้วยภาวะน่าจะเป็น (likelihood) สามารถกล่าวได้ว่า ทำการประมาณค่าพารามิเตอร์ที่อยู่ในฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วม ξ ของอภิประชากร เพื่อนำไปสู่เป้าหมายของการสำรวจตัวอย่างคือการประมาณค่าเฉลี่ยประชากรอันตะ

เมื่อพิจารณาที่ฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วม จะได้ว่า

ξ เป็นฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วมของค่าสังเกตในประชากร

ξ_θ เป็นฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วมของค่าสังเกตในประชากรที่ขึ้นอยู่กับพารามิเตอร์ θ ที่ไม่ทราบค่า

ฟังก์ชันการแจกแจงส่วนริม (marginal distribution) ของ $Y_{k1},\ldots,Y_{k_{\upsilon(s)}}$ คือ $\xi_s = \xi_{s;\theta}$ ซึ่งเป็นฟังก์ชันการแจกแจงส่วนริมสำหรับลำดับของการเพิ่มค่าของ $k \in s$ $(k_1 < \ldots < k_{\upsilon(s)})$[4] และฟังก์ชันการแจกแจงส่วนริมแบบมีเงื่อนไขของ Y_k สำหรับลำดับของการเพิ่มค่า k $(k \in \tilde{s})$ เมื่อกำหนด $Y_{k1},\ldots,Y_{k_{\upsilon(s)}}$ คือ $\xi_{\tilde{s}|s} = \xi_{\tilde{s}|s;\theta}$

เมื่อพิจารณาที่ฟังก์ชันความหนาแน่นจะได้ว่า

$g(\underset{\sim}{y}\,|\,\underset{\sim}{\theta})$ เป็นฟังก์ชันความหนาแน่นของ $Y_1,\ldots,Y_N$ เมื่อ $\underset{\sim}{y} = (y_1,\ldots,y_N)$

$g_s(\underset{\sim}{y_s}\,|\,\underset{\sim}{\theta})$ เป็นฟังก์ชันความหนาแน่นส่วนริมของ $Y_{k_1},\ldots,Y_{k_{\upsilon(s)}}$ เมื่อ $\underset{\sim}{y_s} = (y_{k_1},\ldots,y_{k_{\upsilon(s)}})$

และ $g_{\tilde{s}|s}(\underset{\sim}{y_{\tilde{s}}}\,|\,\underset{\sim}{\theta})$ เป็นฟังก์ชันความหนาแน่นแบบมีเงื่อนไขของ Y_k $(k \in \tilde{s})$ ในลำดับของการเพิ่มค่า k เมื่อกำหนด $Y_{k_1},\ldots,Y_{k_{\upsilon(s)}}$ โดยที่ $g_{\tilde{s}|s}(\underset{\sim}{y_{\tilde{s}}}\,|\,\underset{\sim}{\theta}) = g(\underset{\sim}{y}\,|\,\underset{\sim}{\theta}) / g_s(\underset{\sim}{y_s}\,|\,\underset{\sim}{\theta})$

ถ้าฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วม ξ เป็นฟังก์ชันการแจกแจงแบบเปลี่ยนแปลงได้ (exchangeable) แสดงว่า ξ_s และ $\xi_{\tilde{s}|s}$ เป็นฟังก์ชันการแจกแจงแบบเปลี่ยนแปลงได้ (exchangeable) ด้วยเช่นกัน ซึ่งการแจกแจงแบบเปลี่ยนแปลงได้ (exchangeable) เป็นการแจกแจงที่ให้ตัวแปรสุ่ม $Y_{r_1},\ldots Y_{r_N}$ มีการแจกแจงร่วมแบบเดียวกันสำหรับทุกการเรียงสับเปลี่ยนของอันดับ $r_1,\ldots,r_N$ ของ $1,\ldots,N$ หรือกล่าวได้ว่า ถ้าตัวแปรสุ่ม $Y_1,\ldots,Y_N$ มีการแจกแจงร่วมเหมือนกันทุกการเรียงสับเปลี่ยนของอันดับ สำหรับทุกเซตของตัวอย่าง s แสดงว่าตัวแปรสุ่มนั้นมีคุณสมบัติการเปลี่ยนแปลงได้ (exchangeable) ดังนั้นภายใต้คุณสมบัติการเปลี่ยนแปลงได้ (exchangeable) ฟังก์ชันความหนาแน่น $g(\cdot\,|\,\underset{\sim}{\theta})$ และฟังก์ชัน $g_{\tilde{s}|s}(\cdot\,|\,\underset{\sim}{\theta})$ จะเหมือนกันสำหรับทุกเซตของตัวอย่าง s นั่นคือ ตัวแบบอภิประชากรจะให้ข้อมูลเบื้องต้นของทุกค่า y ในเซตของตัวอย่าง s เหมือนกัน แสดงว่าลำดับที่ (labels) ของหน่วยในประชากรไม่ให้รายละเอียด (uninformation) เกี่ยวกับค่า y

สำหรับอภิประชากร สามารถเขียนค่าเฉลี่ยประชากรได้ในรูปของ

$$\bar{Y} = f_s\bar{Y}_s + (1 - f_s)\bar{Y}_{\tilde{s}}$$

---

[4] ดูตัวอย่างลำดับของการเพิ่มค่า ที่ภาคผนวก ก.2, หน้า 47.

เมื่อ $\bar{Y}_s = (\sum_s Y_k)/\upsilon(s)$

$\bar{Y}_{\tilde{s}} = (\sum_{\tilde{s}} Y_k)/(N-\upsilon(s))$

$f_s = \upsilon(s)/N$

โดยที่ค่าของ $\bar{Y}$ ที่ได้คือ

$$\bar{y} = f_s\bar{y}_s + (1-f_s)\bar{y}_{\tilde{s}}$$

ซึ่ง $\bar{y}_s$ เป็นค่าเฉลี่ยตัวอย่างที่ได้จากการสุ่มตัวอย่าง จากนั้นจึงพยายามที่จะทำนายค่าเฉลี่ย $\bar{y}_{\tilde{s}}$ ที่ไม่ได้ถูกสุ่มเป็นตัวอย่าง

การพิจารณาแนวคิดการอนุมานแบบคลาสสิค ค่าสังเกต y_k ที่ได้จากตัวอย่างจะถูกนำไปใช้ในการอนุมานเวคเตอร์ของพารามิเตอร์ θ ที่ไม่ทราบค่าที่อยู่ในฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วม ξ ของอภิประชากร จากนั้นจึงนำฟังก์ชัน ξ ที่ได้นั้นไปใช้ในการทำนายค่าเฉลี่ย $\bar{Y}_{\tilde{s}}$ ของค่าที่ไม่ได้ถูกสังเกตของประชากร การเลือกเซตตัวอย่าง s ที่ใช้เป็นตัวอย่างนั้น จะได้จากวิธีการสุ่มตัวอย่างเชิงความน่าจะเป็นเช่นเดียวกับแนวคิดของประชากรคงที่ ซึ่งการศึกษาครั้งนี้จะพิจารณาเฉพาะการสุ่มตัวอย่างแบบง่าย (Simple Random Sampling) เพื่อใช้สำหรับการพิจารณาเปรียบเทียบวิธีการประมาณค่าของแนวคิดของประชากรคงที่กับแนวคิดของอภิประชากร

คุณสมบัติที่ใช้ในการพิจารณาเลือกตัวสถิติที่เหมาะสมสำหรับการทำนายค่า (predict) ค่าเฉลี่ยประชากร $\bar{y}$ คือคุณสมบัติของค่าเฉลี่ยของอภิประชากรของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง ($\Im MSE$) สำหรับแผนแบบ (p,T)

เมื่อ p คือ ฟังก์ชันความน่าจะเป็น

T คือ ตัวสถิติที่ใช้ในการประมาณค่าเฉลี่ยประชากร

ซึ่งค่าเฉลี่ยของอภิประชากรของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง สำหรับแผนแบบ (p,T) คือ

$$\Im MSE(p,T) = \Im E(T-\bar{Y})^2$$

ถ้า ไม่มีรายละเอียด (Information) เกี่ยวกับฟังก์ชันความน่าจะเป็น p จะสามารถสลับตำแหน่งของ $\Im$และ E ได้ นั่นคือ [5]

$$\Im MSE(p,T) = E\Im(T-\bar{Y})^2 = E\Im_s\Im_{\tilde{s}|s}(T-\bar{Y})^2$$

---

[5] ดูเพิ่มเติมที่ภาคผนวก ก.3, หน้า 47.

เมื่อ $\mathfrak{I}, \mathfrak{I}_s, \mathfrak{I}_{\tilde{s}|s}$ เป็นค่าคาดหวังในอภิประชากร

วัตถุประสงค์หลักของการใช้ตัวแบบอภิประชากร คือ การเลือกตัวสถิติ T ที่ทำให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของอภิประชากร ($\mathfrak{I}(T-\bar{Y})^2$) ต่ำสุด ถ้าสามารถหาตัวสถิติ T^* ที่ทำให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของอภิประชากรต่ำสุด สำหรับเซตของตัวอย่าง $s \in S$ โดยที่ไม่มีรายละเอียดเกี่ยวกับฟังก์ชันความน่าจะเป็น p จะได้ว่าตัวสถิติ T^* นั้นมีคุณสมบัติของ $\mathfrak{I}MSE(p,T)$ ต่ำสุดสำหรับแผนแบบ (p,T) ด้วย

ดังนั้น ตัวสถิติ T ที่ใช้ในการทำนายค่าเฉลี่ยประชากร $\bar{Y}$ จะได้ว่า

$$T = f_s\bar{Y}_s + (1-f_s)U \qquad \text{..... (2)}$$

เมื่อ U เป็นตัวประมาณของ $\bar{Y}_{\tilde{s}}$

ถ้า $f_s = 1$ จะทำให้ $(1-f_s)U$ มีค่าเป็น 0 แสดงว่า $\upsilon(S) = N$ ด้วยความน่าจะเป็นเท่ากับ 1 จะได้ว่า ตัวสถิติ T คือ ค่าเฉลี่ยประชากร $\bar{Y}$ แสดงว่าประชากรทุกหน่วยถูกเลือกเป็นตัวอย่าง โดยที่ $\bar{Y}$ นั้นเป็นเป้าหมายของการประมาณค่า

สำหรับเซต s ตัวสถิติ $U = U(d)$ และ $T_d = f_s\bar{Y}_s + (1-f_s)U(d)$ ได้มาจากฟังก์ชัน ξ ด้วยหลักโครงสร้างเชิงสโตแคสติก เมื่อ $d = \{(k, Y_k); k \in s\}$

ค่าเฉลี่ยของอภิประชากรของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง $\mathfrak{I}E(T-\bar{Y})^2$ สามารถเขียนในรูปของตัวสถิติ U ได้ดังนี้

$$\mathfrak{I}E(T-\bar{Y})^2 = E\left\{(1-f_s)^2\mathfrak{I}_s\mathfrak{I}_{\tilde{s}|s}(U-\bar{Y}_{\tilde{s}})^2\right\}$$

ถ้าทราบพารามิเตอร์ θ ที่อยู่ในฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วม ξ ตัวสถิติที่ทำให้ $\mathfrak{I}MSE$ ต่ำสุด คือ $U = \mathfrak{I}_{\tilde{s}|s}(\bar{Y}_{\tilde{s}})$ สำหรับทุกเซต s แต่ถ้าไม่ทราบเวคเตอร์พารามิเตอร์ $\underset{\sim}{\theta}$ ที่อยู่ในฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วม ξ ต้องทำการประมาณ $\underset{\sim}{\theta}$ เป็นอันดับแรก

ถ้าตัวสถิติ U เป็นตัวประมาณที่ไม่เอนเอียงของ $\bar{Y}_{\tilde{s}}$ ($\xi - unbiased$)จะได้ว่า

$$\mathfrak{I}E(T-\bar{Y})^2 = E[(1-f_s)^2\{V(U)+V(\bar{Y}_{\tilde{s}})-2C(U,\bar{Y}_{\tilde{s}})\}]$$

ทฤษฎีการอนุมานสถิติเกี่ยวกับการประมาณค่าแบบคลาสสิค ใช้ในการหาตัวสถิติ U สำหรับประมาณค่าของ $\bar{Y}_{\tilde{s}}$ ถ้าตัวสถิติ T และตัวสถิติ U สัมพันธ์กันในรูป (1) จะได้ว่าตัวสถิติ

T จะเป็นตัวประมาณที่ไม่เอนเอียง ($\xi - unbiased$) สำหรับ $\bar{Y}$ ก็ต่อเมื่อตัวสถิติ U เป็นตัวประมาณที่ไม่เอนเอียง ($\xi - unbiased$) สำหรับ $\bar{Y}_{\tilde{s}}$

การทำนายค่าเฉลี่ยประชากรเมื่อเซตตัวอย่าง s ที่ใช้เป็นตัวอย่างนั้นได้มากจากวิธีการสุ่มตัวอย่างแบบง่าย จะได้ว่า

เมื่อตัวแปรสุ่ม $Y_1,...,Y_N$ ของอภิประชากร เป็นอิสระซึ่งกันและกันและมีการแจกแจงเหมือนกัน (i.i.d) แสดงว่า ตัวแปรสุ่ม Y_k แต่ละตัว เป็นตัวแปรสุ่มที่มีค่าเฉลี่ย μ และ ความแปรปรวน σ^2 ฟังก์ชันการแจกแจงความน่าจะเป็นของตัวแปรสุ่ม Y_k คือ $G(y|\underset{\sim}{\theta})$ (หรือ $G(y)$) และฟังก์ชันความหนาแน่น คือ $g(y|\underset{\sim}{\theta})$ (หรือ $g(y)$) ดังนั้น การพิจารณาคุณภาพของตัวประมาณสามารถแยกพิจารณาได้เป็น 3 กรณีตามรายละเอียดเกี่ยวกับฟังก์ชันความหนาแน่นดังนี้

กรณีที่ 1 $G(y)$ เป็นฟังก์ชันการแจกแจงความน่าจะเป็นแบบต่อเนื่องที่ไม่ทราบทั้งรูปแบบ (Shape) ของฟังก์ชัน , ค่าเฉลี่ย (μ) และ ความแปรปรวน (σ^2)

โดยที่ $\mu = \int_{-\infty}^{\infty} y dG(y)$

และ $\sigma^2 = \int_{-\infty}^{\infty} (y-\mu)^2 dG(y)$

เมื่อพิจารณา ค่าเฉลี่ยของอภิประชากรของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณในกรณีนี้ จะได้ว่า

$$\Im E(T-\bar{Y})^2 \geq \Im E(\bar{Y}_s - \bar{Y})^2 = E\{\frac{1}{v(s)} - \frac{1}{N}\}\sigma^2$$

ซึ่งจะเท่ากันก็ต่อเมื่อตัวสถิติ $T = \bar{Y}_s$ เนื่องจากตัวสถิติ U ที่ให้ความแปรปรวนต่ำสุด คือ $U = \bar{Y}_s$

พิสูจน์

จาก
$$\Im(U) = \Im\left(\frac{\sum_{\tilde{s}} y_k}{N - v(s)}\right)$$
$$= \frac{1}{N - v(s)} \sum_{\tilde{s}} \Im(y_k)$$
$$= \Im(y_k)$$
$$= \mu$$

แสดงว่า U เป็นตัวประมาณที่ไม่เอนเอียง ($\xi - unbiased$) ของ μ

เนื่องจากตัวสถิติ U เป็นฟังก์ชันของ Y_s ($U = U(Y_s)$) โดยที่ $Y_s = \{k; k \in s\}$ และ Y_k เป็นตัวแปรสุ่มที่เป็นอิสระซึ่งกันและกันและมีการแจกแจงเหมือนกัน จะได้ว่า $\Im(Y_k) = \mu$ และ $\Im(\bar{Y}_s) = \mu$ ดังนั้น ภายใต้คุณสมบัติความไม่เอนเอียงของค่าเฉลี่ยประชากร จะได้ว่า $U = \bar{Y}_s$ เป็นตัวประมาณที่ให้ความแปรปรวนต่ำสุด ดังนั้นจึงเลือก $\bar{Y}_s$ เป็นตัวประมาณของ $\bar{Y}_{\tilde{s}}$

$$\therefore \qquad T = f_s\bar{Y}_s + (1-f_s)\bar{Y}_s = \bar{Y}_s$$

ดังนั้น

$$\begin{aligned}
\Im E(\bar{Y}_s - \bar{Y})^2 &= E\Im_s\Im_{\tilde{s}|s}(\bar{Y}_s - f_s\bar{Y}_s - (1-f_s)\bar{Y}_{\tilde{s}})^2 \\
&= E\Im_s\Im_{\tilde{s}|s}\{(1-f_s)\bar{Y}_s - (1-f_s)\bar{Y}_{\tilde{s}}\}^2 \\
&= E\Im_s\Im_{\tilde{s}|s}\{(1-f_s)(\bar{Y}_s - \bar{Y}_{\tilde{s}})\}^2 \\
&= E\left\{(1-f_s)^2\Im_s\Im_{\tilde{s}|s}(\bar{Y}_s - \bar{Y}_{\tilde{s}})^2\right\} \\
&= E\{(1-f_s)^2\Im_s\Im_{\tilde{s}|s}[(\bar{Y}_s - \mu)^2 + (\bar{Y}_{\tilde{s}} - \mu)^2 \\
&\quad - 2(\bar{Y}_s - \mu)(\bar{Y}_{\tilde{s}} - \mu)]\} \\
&= E\left\{(1-f_s)^2[V(\bar{Y}_s) + V(\bar{Y}_{\tilde{s}}) - 2COV(\bar{Y}_s, \bar{Y}_{\tilde{s}})]\right\}
\end{aligned}$$

เนื่องจาก Y_k เป็นตัวแปรสุ่มที่เป็นอิสระซึ่งกันและกัน ดังนั้น $COV(\bar{Y}_s, \bar{Y}_{\tilde{s}}) = 0$

$$\begin{aligned}
\therefore \Im E(\bar{Y}_s - \bar{Y})^2 &= E\left\{(1-f_s)^2[V(\bar{Y}_s) + V(\bar{Y}_{\tilde{s}})]\right\} \\
&= E\left\{(1-f_s)^2\left[\frac{\sigma^2}{\upsilon(s)} + \frac{\sigma^2}{N-\upsilon(s)}\right]\right\} \\
&= E\left\{\frac{N-\upsilon(s)}{\upsilon(s)N}\sigma^2\right\} \\
&= E\left\{\left(\frac{1}{\upsilon(s)} - \frac{1}{N}\right)\sigma^2\right\}
\end{aligned}$$

กรณีที่ 2 $G(y)$ เป็นฟังก์ชันการแจกแจงความน่าจะเป็นแบบต่อเนื่องที่อยู่ในรูปแบบของ $G(y|\underset{\sim}{\theta}) = G_0\{(y-\theta_1)/\theta_2\}$ ที่ทราบฟังก์ชันของการแจกแจง $G_0(\cdot)$ แต่ไม่ทราบ

ค่าพารามิเตอร์ที่ตั้ง (location parameter) θ_1 และสเกลพารามิเตอร์ (scale parameter) θ_2 ของค่าเฉลี่ย (μ) และ ความแปรปรวน (σ^2)

โดยที่ $\mu = \theta_1 + c_1\theta_2$

และ $\sigma^2 = (c_2 - c_1^2)\theta_2^2$

ในที่นี้ $c_r = \int_{-\infty}^{\infty} z^r dG_0(z)$; $(r = 1,2)$ เป็นค่าคงที่ที่ทราบค่า

ค่าเฉลี่ยของอภิประชากรของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณในกรณีนี้ คือ

$$\Im E(T - \bar{Y})^2 = E\{(1 - f_s)^2 V(U) + (1 - f_s)\sigma^2 / N\}$$

โดยที่ $V(U) \geq C_\theta$

เมื่อ $$C_\theta = \frac{\{\mu'(\theta)\}^2}{\Im\left\{\frac{\partial \ln g_s(y_s|\theta)}{\partial\theta}\right\}^2}$$

$$\mu = \mu(\theta) = \int_{-\infty}^{\infty} y g(y|\theta) dy$$

และ $g_s(y_s|\theta) = \prod_s g(y_k|\theta)$

พิสูจน์

พิจารณาค่าเฉลี่ยจะได้ว่า

จาก $$\mu = \int_{-\infty}^{\infty} y\, dG(y)$$

ในกรณีนี้ $G(y|\underset{\sim}{\theta}) = G_0\{(y - \theta_1)/\theta_2\}$

ให้ $$z = \frac{y - \theta_1}{\theta_2}$$

$$z\theta_2 + \theta_1 = y$$

$$\theta_2\, dz = dy$$

จะได้ว่า $$\mu = \int_{-\infty}^{\infty} y\, dG_0\{(y - \theta_1)/\theta_2\}$$

$$= \int_{-\infty}^{\infty} (z\theta_2 + \theta_1)\, dG_0(z)$$

$$= \theta_2 \int_{-\infty}^{\infty} z\, dG_0(z) + \theta_1 \int_{-\infty}^{\infty} dG_0(z)$$

$$= \theta_2 c_1 + \theta_1 \{G_0(\infty) - G_0(-\infty)\}$$

$$= \theta_2 c_1 + \theta_1 \{1 - 0\}$$

$$\therefore \quad \mu = \theta_2 c_1 + \theta_1$$

โดยที่ $c_1 = \int_{-\infty}^{\infty} z\, dG_0(z)$

พิจารณาความแปรปรวน จะได้ว่า

จาก $$\sigma^2 = \int_{-\infty}^{\infty} (y-\mu)^2\, dG_0\{(y-\theta_1)/\theta_2\}$$

$$= \int_{-\infty}^{\infty} \{z\theta_2 + \theta_1 - (\theta_2 c_1 + \theta_1)\}^2\, dG_0(z)$$

$$= \int_{-\infty}^{\infty} \{\theta_2 (z - c_1)\}^2\, dG_0(z)$$

$$= \int_{-\infty}^{\infty} \theta_2^2 (z^2 - 2c_1 z + c_1^2)\, dG_0(z)$$

$$= \theta_2^2 \int_{-\infty}^{\infty} z^2\, dG_0(z) - 2c_1\theta_2^2 \int_{-\infty}^{\infty} z\, dG_0(z)$$

$$+ c_1^2\theta_2^2 \int_{-\infty}^{\infty} dG_0(z)$$

$$= \theta_2^2 c_2 - 2c_1\theta_2^2 c_1 + c_1^2\theta_2^2$$

$$= \theta_2^2 c_2 - \theta_2^2 c_1^2$$

$$\therefore \quad \sigma^2 = \theta_2^2 (c_2 - c_1^2)$$

โดยที่ $c_2 = \int_{-\infty}^{\infty} z^2\, dG_0(z)$

พิจารณาค่าเฉลี่ยของอภิประชากรของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณ จะได้ว่า

จาก

$$\begin{aligned}
&\Im E(T-\bar{Y})^2 \\
&\quad = E[(1-f_s)^2 \Im_s \Im_{\tilde{s}/s}(U-\bar{Y}_{\tilde{s}})^2] \\
&\quad = E[(1-f_s)^2 \{\Im_s \Im_{\tilde{s}/s}(U-\mu)^2 + \Im_s \Im_{\tilde{s}/s}(\bar{Y}_{\tilde{s}}-\mu)^2 \\
&\qquad - 2\Im_s \Im_{\tilde{s}/s}(U-\mu)(\bar{Y}_{\tilde{s}}-\mu)\}] \\
&\quad = E[(1-f_s)^2 \{V(U)+V(\bar{Y}_{\tilde{s}})-2Cov(U,\bar{Y}_{\tilde{s}})\}]
\end{aligned}$$

เมื่อทราบฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วม ξ จะได้ว่า ตัวสถิติ $U = \Im_{\tilde{s}|s}(\bar{Y}_{\tilde{s}})$ เป็นตัวสถิติที่ทำให้ความคลาดเคลื่อนกำลังสองของตัวประมาณต่ำสุด ดังนั้น $Cov(U,\bar{Y}_{\tilde{s}}) = 0$

$$\therefore \qquad \Im E(T-\bar{Y})^2 = E[(1-f_s)^2\{V(U)+V(\bar{Y}_{\tilde{s}})\}]$$

พิจารณาเฉพาะ $V(\bar{Y}_{\tilde{s}})$ จะได้ว่า;

$$\begin{aligned}
V(\bar{Y}_{\tilde{s}}) &= V\left\{\frac{\sum_{\tilde{s}} Y_k}{N-v(s)}\right\} \\
&= \frac{1}{(N-v(s))^2}\sum_{\tilde{s}} V(Y_k) \\
&= \frac{1}{(N-v(s))^2}(N-v(s))\sigma^2 \\
&= \frac{1}{(N-v(s))}\sigma^2
\end{aligned}$$

$$\therefore \qquad \Im E(T-\bar{Y})^2 = E[(1-f_s)^2 V(U) + (1-f_s)\frac{\sigma^2}{N}]$$

กรณีที่ 3 $G(y)$ เป็นฟังก์ชันการแจกแจงความน่าจะเป็นแบบต่อเนื่องที่สมบูรณ์ คือ ทราบทั้งรูปแบบของฟังก์ชันการแจกแจง $G(y)$ ค่าเฉลี่ย (μ) และ ความแปรปรวน (σ^2)

ค่าเฉลี่ยของอภิประชากรของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณในกรณีนี้ คือ

$$\Im E(T-\bar{Y})^2 = E[(1-f_s)\frac{\sigma^2}{N}]$$

พิสูจน์

จาก $$\Im E(T-\bar{Y})^2 = E[(1-f_s)^2\{V(U)+V(\bar{Y}_{\tilde{s}})\}]$$

พิจารณา $V(U)$;

$$V(U) = \Im_s \Im_{\tilde{s}/s}(U-\mu)^2$$

เนื่องจาก $G(y)$ เป็นฟังก์ชันการแจกแจงความน่าจะเป็นแบบต่อเนื่องที่สมบูรณ์ ดังนั้นตัวสถิติ $U = \Im_{\tilde{s}|s}(\bar{Y}_{\tilde{s}}) = \mu$ เป็นตัวสถิติที่ทำให้ความคลาดเคลื่อนกำลังสองของตัวประมาณต่ำสุด

$$\therefore \quad V(U) = 0$$

ดังนั้น $$\Im E(T-\bar{Y})^2 = E[(1-f_s)\frac{\sigma^2}{N}]$$

ซึ่งจะเห็นได้ว่า การทำนายค่าเฉลี่ยประชากรของแนวคิดอภิประชากร

$$\bar{Y} = f_s\bar{Y}_s + (1-f_s)\bar{Y}_{\tilde{s}}$$

โดยที่ตัวสถิติที่ใช้ในการทำนายค่าเฉลี่ยประชากร คือตัวสถิติ T ซึ่งอยู่ในรูปของ $T = f_s\bar{Y}_s + (1-f_s)U$ เมื่อ U เป็นตัวประมาณของ $\bar{Y}_{\tilde{s}}$ โดยที่คุณภาพของตัวประมาณสามารถวัดได้ด้วยค่าเฉลี่ยของอภิประชากรของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง ($\Im MSE$) สามารถแยกพิจารณาได้เป็น 3 กรณี คือ

กรณีที่ 1 $G(y)$ เป็นฟังก์ชันการแจกแจงความน่าจะเป็นแบบต่อเนื่องที่ไม่ทราบทั้งรูปแบบ (Shape) ของฟังก์ชัน , ค่าเฉลี่ย (μ) และ ความแปรปรวน (σ^2) ของประชากร จะได้

$$\Im E(T-\bar{Y})^2 \geq \Im E(\bar{Y}_s-\bar{Y})^2 = E\{\frac{1}{v(s)}-\frac{1}{N}\}\sigma^2$$

กรณีที่ 2 $G(y)$ เป็นฟังก์ชันการแจกแจงความน่าจะเป็นแบบต่อเนื่องที่ทราบรูปแบบของฟังก์ชัน แต่ไม่ทราบค่าเฉลี่ย (μ) และ ความแปรปรวน (σ^2) ของประชากร จะได้

$$\Im E(T-\bar{Y})^2 = E\{(1-f_s)^2V(U)+(1-f_s)\sigma^2/N\}$$

กรณีที่ 3 $G(y)$ เป็นฟังก์ชันการแจกแจงความน่าจะเป็นแบบต่อเนื่องที่สมบูรณ์ คือ ทราบทั้งรูปแบบของฟังก์ชันการแจกแจง $G(y)$ ค่าเฉลี่ย (μ) และ ความแปรปรวน (σ^2) ของประชากร จะได้

$$\Im E(T-\bar{Y})^2 = E[(1-f_s)\frac{\sigma^2}{N}]$$

แนวคิดของอภิประชากร เป็นแนวคิดที่พิจารณาให้ค่าของหน่วยประชากรในประชากรเป็นตัวแปรสุ่มที่ถูกกำกับด้วยโครงสร้างเชิงสโตแคสติก โดยที่โครงสร้างนี้จะทราบข้อมูลอื่นที่มีความสัมพันธ์หรือเกี่ยวข้องกับหน่วยประชากรที่สนใจ ซึ่งสามารถเขียนค่าของหน่วยประชากรในประชากรหน่วยที่ i ของตัวแปรสุ่ม Y ได้ดังนี้

$$y_i = \mu_i + e_i$$

เมื่อ μ_i เป็นค่าคงที่ที่แสดงคุณลักษณะที่สนใจของตัวแปรสุ่มหน่วยที่ i

e_i เป็นค่าคลาดเคลื่อนสุ่มที่ได้จากอภิประชากร

ดังนั้นแนวคิดของอภิประชากรจะเหมาะสำหรับการอนุมานปัญหาที่ทราบความสัมพันธ์ระหว่างหน่วยประชากรที่สนใจกับข้อมูลอื่นที่เกี่ยวข้อง ซึ่งมีผลทำให้สามารถลดความคลาดเคลื่อนที่ไม่ได้เกิดจากการสุ่มตัวอย่าง (Non-sampling Error) ได้ โดยส่วนใหญ่จะใช้ตัวแบบอภิประชากรสำหรับการหาแผนการสุ่มตัวอย่างและตัวประมาณที่เหมาะสมกับสถานการณ์ต่าง ๆ ใช้ตัวแบบอภิประชากรในการหาค่าเฉลี่ยประชากรรวมไปถึงการวัดค่าความคลาดเคลื่อนในพื้นที่เล็กๆ และใช้ตัวแบบอภิประชากรสำหรับการหาค่าข้อมูลสูญหาย (Missing Data)[6]

4.3 ความแตกต่างระหว่างแนวคิดของประชากรคงที่และแนวคิดของอภิประชากร

ความแตกต่างระหว่างแนวคิดของประชากรคงที่และแนวคิดของอภิประชากรสามารถระบุเป็นข้อ ๆ ได้ดังนี้

---

[6] Edward L. Korn and Barry I. Graubard, "Variance Estimation for Superpopulation Parameters," Statistica Sinica 8 (1998): 1131-1151.

	แนวคิดของประชากรคงที่	แนวคิดของอภิประชากร
ประชากร	ประชากรเป็นประชากรอันตะที่ทราบขนาดประชากรที่แน่นอน	ประชากรเป็นประชากรอันตะที่เป็นเซตหนึ่งที่ถูกดึงมาจากเอกภาพที่ใหญ่กว่า
ค่าของตัวแปรศึกษา	พิจารณาให้ค่าของตัวแปรศึกษาเป็นค่าคงที่ที่ไม่ทราบค่า	พิจารณาให้ค่าของตัวแปรศึกษาเป็นผลลัพธ์ของตัวแปรสุ่ม
การแจกแจงความน่าจะเป็นของหน่วยประชากร	ไม่สนใจการแจกแจงความน่าจะเป็นของหน่วยประชากร ถือว่าการแจกแจงเป็น Distribution Free	ใช้การแจกแจงความน่าจะเป็นของหน่วยประชากรในการสร้างและพัฒนาตัวแบบเชิงความน่าจะเป็น
การนำไปใช้งาน	สะดวกในการนำไปใช้งาน	ยุ่งยาก ซับซ้อน

แนวคิดของประชากรคงที่เป็นแนวคิดที่พิจารณาค่าของตัวแปรศึกษาเป็นค่าคงที่ ดังนั้นแนวคิดนี้จะเหมาะสำหรับการพิจารณาคุณลักษณะประชากรที่ไม่มีการเปลี่ยนแปลงหรือมีการเปลี่ยนแปลงได้ยาก เช่น เพศ , ระดับการศึกษา เป็นต้น ในขณะที่แนวคิดของอภิประชากรเป็นแนวคิดที่พิจารณาค่าของตัวแปรศึกษาเป็นตัวแปรสุ่มที่มีการแจกแจงที่มีความคลาดเคลื่อน และมีข้อมูลอื่นมาเกี่ยวข้อง จึงเหมาะสำหรับการพิจารณาคุณลักษณะประชากรที่มีการเปลี่ยนแปลงตลอดเวลา เช่น ความพึงพอใจ , ความคิดเห็น เป็นต้น

4.4 การเปรียบเทียบวิธีการสุ่มตัวอย่างและวิธีการประมาณค่าลักษณะประชากร

วิธีการสุ่มตัวอย่างเชิงความน่าจะเป็นที่ใช้ในการอนุมานไปสู่ประชากรสำหรับแนวคิดประชากรคงที่ แบ่งได้เป็น 4 วิธีแม่บท คือ การสุ่มตัวอย่างแบบง่าย (Simple Random Sampling) การสุ่มตัวอย่างแบบแบ่งชั้นภูมิ (Stratified Sampling) การสุ่มตัวอย่างแบบมีระบบ (Systematic Sampling) และการสุ่มตัวอย่างแบบกลุ่ม (Cluster Sampling) สำหรับวิธีการสุ่มตัวอย่างของแนวคิดของอภิประชากรนั้น จะใช้วิธีการสุ่มตัวอย่างเชิงความน่าจะเป็นเช่นเดียวกับแนวคิดของประชากรคงที่ ในการเลือกเซตตัวอย่าง s

คุณลักษณะประชากรที่พิจารณาในการศึกษาครั้งนี้ คือ ค่าเฉลี่ยประชากร (Population Mean) ซึ่งการเปรียบเทียบวิธีการประมาณค่าเฉลี่ยประชากรภายใต้กรอบแนวคิดของประชากรคงที่และแนวคิดของอภิประชากรนั้น สามารถเปรียบเทียบได้ด้วยการพิจารณาคุณภาพของตัวประมาณที่ได้จากแนวคิดทั้งสอง ซึ่งจะได้ว่า

ภายใต้กรอบแนวคิดของประชากรคงที่ คุณภาพของตัวประมาณสามารถวัดได้ด้วยความแปรปรวน เนื่องจากค่าเฉลี่ยตัวอย่างเป็นตัวประมาณที่ไม่เอนเอียงสำหรับค่าเฉลี่ยประชากร ดังนั้น ความแปรปรวนของค่าเฉลี่ยตัวอย่าง กรณีสุ่มตัวอย่างแบบไม่ใส่คืน คือ

$$V(\bar{y}) = \frac{S^2}{n}(1-f)$$

เมื่อ S^2 คือ ความแปรปรวนของประชากร

และ f คือ สัดส่วนการสุ่ม (Sampling fraction)

สำหรับกรอบแนวคิดของอภิประชากร คุณภาพของตัวประมาณสามารถวัดได้ด้วยค่าเฉลี่ยของอภิประชากรของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง ($\Im MSE$) ซึ่งแยกพิจารณาได้เป็น 3 กรณี ดังนี้

กรณีที่ 1 $G(y)$ เป็นฟังก์ชันการแจกแจงความน่าจะเป็นแบบต่อเนื่องที่ไม่ทราบทั้งรูปแบบ (Shape) ของฟังก์ชัน , ค่าเฉลี่ย (μ) และ ความแปรปรวน (σ^2) จะได้

$$\Im E(T-\bar{Y})^2 \geq \Im E(\bar{Y}_s-\bar{Y})^2 = E\{\frac{1}{\upsilon(s)}-\frac{1}{N}\}\sigma^2$$

$$= E\left\{\frac{N-\upsilon(s)}{\upsilon(s)N}\right\}\sigma^2$$

$$= E\left\{\frac{N}{N}\left(\frac{N-\upsilon(s)}{\upsilon(s)N}\right)\right\}\sigma^2$$

$$= E\left\{\frac{N}{\upsilon(s)}\left(\frac{N-\upsilon(s)}{N}\right)\right\}\frac{\sigma^2}{N}$$

เมื่อ n หมายถึง ขนาดตัวอย่างในประชากรคงที่

$\upsilon(s)$ หมายถึง ขนาดตัวอย่างในอภิประชากร

และ S^2 หมายถึง ความแปรปรวนประชากรในประชากรคงที่

σ^2 หมายถึง ความแปรปรวนประชากรในอภิประชากร

ดังนั้น เมื่อทำการสุ่มตัวอย่างแบบง่าย จะได้ว่า

$$n = \upsilon(s) \quad \text{และ} \quad S^2 = \sigma^2$$

แสดงว่า $\Im E(T-\bar{Y})^2 \geq \Im E(\bar{Y}_s-\bar{Y})^2 = E\{\frac{1}{\upsilon(s)}-\frac{1}{N}\}\sigma^2$

$$= \frac{N}{n}E[V(\bar{y})]$$

$$\geq V(\bar{y})$$

จะเท่ากันก็ต่อเมื่อ $n = N$ หรือ $\upsilon(s) = N$

เมื่อ $V(\bar{y}) = \frac{S^2}{n}(1-f)$ คือ ความแปรปรวนของค่าเฉลี่ยตัวอย่าง กรณีสุ่มตัวอย่างแบบไม่ใส่คืน ของแนวคิดของประชากรคงที่

กรณีที่ 2 $G(y)$ เป็นฟังก์ชันการแจกแจงความน่าจะเป็นแบบต่อเนื่องที่ทราบรูปแบบของฟังก์ชัน แต่ไม่ทราบค่าเฉลี่ย (μ) และ ความแปรปรวน (σ^2) จะได้

$$\Im E(T-\bar{Y})^2 = E\{(1-f_s)^2V(U)+(1-f_s)\sigma^2/N\}$$

$$= E\{(1-f_s)^2V(U)\}+E\{(1-f_s)\sigma^2/N\}$$

เนื่องจาก $f_s = \frac{\upsilon(s)}{N}$ ดังนั้น เมื่อ $n = \upsilon(s)$ จะได้ว่า $f_s = \frac{n}{N}$ และ $S^2 = \sigma^2$

แสดงว่า $\Im E(T-\bar{Y})^2 = E\{(1-f_s)^2V(U)\}+E\{V(\bar{y})\}$

$$\geq V(\bar{y})$$

จะเท่ากันก็ต่อเมื่อ $n = N$ หรือ $\upsilon(s) = N$

กรณีที่ 3 $G(y)$ เป็นฟังก์ชันการแจกแจงความน่าจะเป็นแบบต่อเนื่องที่สมบูรณ์ คือ ทราบทั้งรูปแบบของฟังก์ชันการแจกแจง $G(y)$ ค่าเฉลี่ย (μ) และ ความแปรปรวน (σ^2) จะได้

$$\Im E(T-\bar{Y})^2 = E[(1-f_s)\frac{\sigma^2}{N}]$$

$$= V(\bar{y})$$

จะเห็นได้ว่า ตัวประมาณที่ได้จากวิธีการสุ่มตัวอย่างแบบง่ายที่ใช้สำหรับการเลือกเซตตัวอย่าง s ภายใต้กรอบแนวคิดของอภิประชากร กรณีที่ 1 และกรณีที่ 2 ค่า $\Im MSE$ ของตัวประมาณภายใต้กรอบแนวคิดของอภิประชากร มากกว่าค่าความแปรปรวนของตัวประมาณภายใต้กรอบแนวคิดของประชากรคงที่ ซึ่งจะเท่ากันก็ต่อเมื่อ ประชากรทั้งหมดถูกสุ่มขึ้นมาเป็นตัวอย่าง ($\upsilon(s) = N$) ในขณะที่กรณีที่ 3 ค่า $\Im MSE$ ของตัวประมาณภายใต้กรอบแนวคิดของอภิประชากร เท่ากับค่าความแปรปรวนของตัวประมาณภายใต้กรอบแนวคิดของประชากรคงที่ ดังนั้น

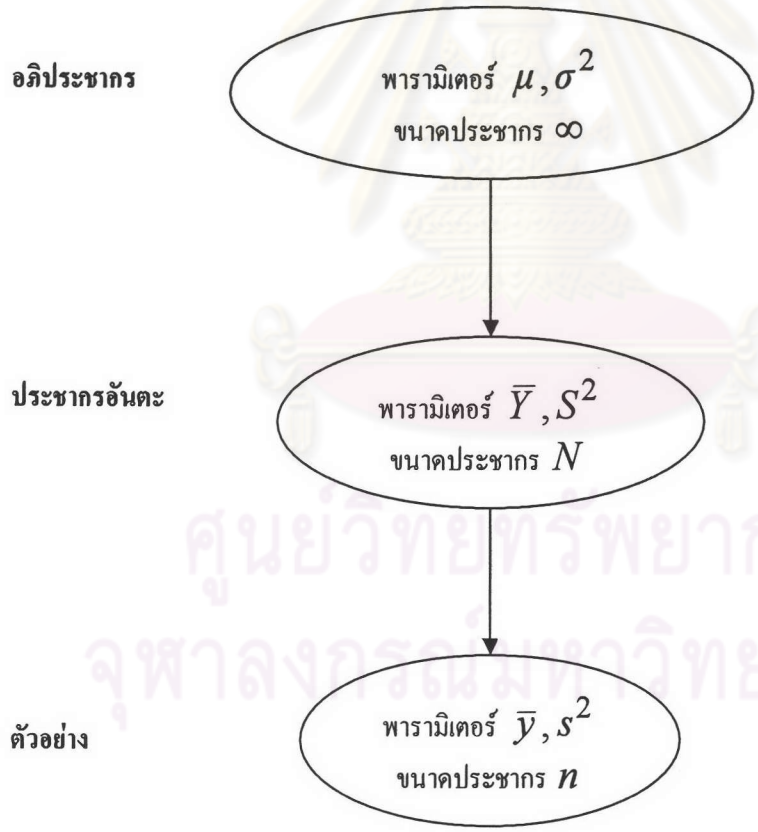
สามารถกล่าวได้ว่า เมื่อพิจารณาวิธีการสุ่มตัวอย่างแบบง่าย ตัวประมาณภายใต้กรอบแนวคิดของประชากรคงที่ จะให้คุณภาพของตัวประมาณที่ดีกว่าตัวประมาณภายใต้กรอบแนวคิดของอภิประชากร

ทฤษฎีการสำรวจตัวอย่างภายใต้กรอบแนวคิดของอภิประชากร ถือว่าประชากรในอภิประชากรมีขนาดใหญ่มาก (Infinite) การได้มาซึ่งตัวอย่างสำหรับใช้ในการพิจารณาคุณลักษณะของประชากรที่สนใจภายใต้กรอบแนวคิดนี้ สามารถแยกได้เป็น 2 ขั้นตอน คือ

ขั้นตอนที่ 1 พิจารณาประชากรอันตะขนาด N จากประชากรในอภิประชากร

ขั้นตอนที่ 2 สุ่มตัวอย่างขนาด n โดยที่ $n < N$ จากประชากรอันตะที่ได้จากขั้นตอนที่ 1

ซึ่งสามารถอธิบายเป็นแผนภาพได้ดังนี้



จะเห็นได้ว่า ทฤษฎีการสำรวจตัวอย่างที่ใช้กันโดยทั่วไปเป็นการพิจารณาประชากรในส่วนของประชากรอันตะและตัวอย่างเท่านั้น เมื่อพิจารณาคุณภาพของตัวประมาณสำหรับประมาณค่าเฉลี่ยประชากรที่ได้จากแนวคิดของอภิประชากรนั้น ในกรณีที่ 1 และกรณีที่ 2 ซึ่งเป็นกรณีที่ไม่ทราบค่าเฉลี่ยและความแปรปรวนของประชากรนั้น คุณภาพของตัวประมาณที่ได้จากแนวคิดนี้ด้อย

กว่าคุณภาพของตัวประมาณที่ได้จากแนวคิดของประชากรคงที่ ส่วนในกรณีที่ 3 ที่ทราบทั้งรูปแบบของฟังก์ชันการแจกแจง ค่าเฉลี่ย และความแปรปรวนของประชากร ตัวประมาณที่ได้นั้นให้คุณภาพเท่ากับคุณภาพของตัวประมาณที่ได้จากแนวคิดของประชากรคงที่ ซึ่งในความเป็นจริงนั้นไม่สามารถทราบค่าเฉลี่ยและความแปรปรวนของประชากร ดังนั้นการพิจารณาตัวประมาณที่ใช้สำหรับประมาณค่าเฉลี่ยประชากรภายใต้กรอบแนวคิดของประชากรคงที่ เหมาะสมกว่าการพิจารณาตัวประมาณที่ใช้สำหรับประมาณค่าเฉลี่ยประชากรภายใต้กรอบแนวคิดของอภิประชากร