الفصل الثامن عشر

المعاينة العنقودية المكيفة

Adaptive Cluster Sampling

1.18 مقدمة

المعاينة المكيفة (Apative Sampling) تعني أن الوحدات التي سيجري سحبها لتكون ضمن العينة ربما أنها ستعتمد في عملية سحبها على قيمة المتغير الذي نرغب بدراسته وتمت مشاهدته خلال المسح. فمثلاً إذا كان المسح مهتماً بتقدير كثافة نوع من الحيوانات أو النباتات النادرة في منطقة الدراسة، فسيجري ضم المواقع القريبة من الموقع الذي تم مشاهدة الحيوانات أو النبات النادر فيه. كذلك في دراسة أنواع معينة من الأمراض المعدية فإنه يجري ضم جميع أفراد العائلة إلى العينة إذا اكتشف أن أحد أفرادها مصاب بالمرض المعدي. تختلف المعاينة المكيفة عن طرق المعاينة التقليدية المعروفة مثل المعاينة العشوائية البسيطة أو غيرها، حيث إن في هذه الطرق جميع الوحدات يجري تحديدها قبل البدء بعملية السحب للوحدات، ولا علاقة لقيمة المتغير الذي نرغب في دراسته بالوحدات التي سيجرى سحبها.

تستخدم العينة العنقودية المكيفة عندما يكون المتغير الذي نرغب في دراسته على شكل عناقيد نادرة الوجود في منطقة الدراسة. لتنفيذ المسح باستخدام المعاينة العنقودية المكيفة، نقوم بسحب عينة عشوائية أولية من الوحدات، وعندما تكون قيمة المتغير الذي نرغب بدراسته محققة لبعض

الشروط نقوم بسحب الوحدات المجاورة وإضافتها إلى العينة، فمثلاً عندما يقوم الباحثون بتقدير عدد الطيور أو الحيوانات النادرة أو كمية المعادن القلية التركيز أو عدد المصابين بنوع نادر ومعدٍ من الأمراض يقومون بسحب مواضع معينة في منطقة الدراسة أو المجتمع ومن ثم يقومون بزيارة هذه المواقع ولكنهم وفي معظم هذه المواقع المختارة لا يجدون أي أثر لما يبحثون عنه، ولكن حال عثورهم على وحدة تحتوي على أحد الطيور النادرة على سبيل المثال فإنه في الغالب ستكون الوحدات المجاورة تحتوي طيوراً أخرى لأن طبيعة بعض الحيوانات أو النباتات تعيش أو توجد في مناطق متجاورة؛ لذا لابد من إضافة الوحدات المجاورة إذا أردنا الحصول على تقدير جيد لما نريد تقديره. لذا تظهر لدينا أهمية استخدام العينة العنقودية المكيفة لتقدير معدل أو مجموع أو كثافة هذه الأنواع النادرة.

يعتبر Thompson(1990) أول من اقترح استخدام العينة العنقودية المكيفة على أن تستخدم العينة العشوائية البسيطة مع الإرجاع أو دون الإرجاع لسحب الوحدات الأولية، كما قام باقتراح استخدم العينة المنتظمة والطبقية لسحب الوحدات الأولية ببحثيه Thompson(1991a,b). لقد قام كلٍ من Cormack(1988) وSeber(1986,1992) بمناقشة أهمية استخدام طرق المعاينة المكيفة للمعاينات الخاصة بالدراسات البيئية.

لقد شهدت العشر السنوات الأخيرة من القرن الماضي وبداية القرن الحالي تطورات كثيرة لاستخدام المعاينة المكيفة ومن بين أهم هذه التطورات: المعاينة العنقودية المكيفة بمرحلتين Salehi and Seber(1997b) والمعاينة العنقودية المزدوجة المكيفة لكلٍ من الباحثين Felix Medina and Thompson(1999) وFelixMedina(2000). كذلك قام كل من Pontius(1997) وRoesch(1993) وSmith et al.(1995) بدراسة العينة العنقودية المكيفة باحتمالات غير متساوية. لقد تمت دراسة المعاينة المكيفة للمربع

ألاتيني من قبل MunhollandandBrokowski(1993) و Borkowski(1999). كما تناول كل من Brown and Manly(1998) و Salehi and Seber(2002) المعاينة العنقودية المكيفة والمقيدة بحجم عينة محدد. هناك بعض النتائج التي من شأنها أن تسهل عملية الحسابات لنظرية Rao-Blackwell لتحسين المقدّرات التي تستخدم بالمعاينة المكيفة التي تمت دراستها من قبل Felix Medina(2003) و Salehi(1999). وهناك تطبيقات عديدة للمعاينة العنقودية المكيفة للمجتمعات الطبيعية والإنسانية والحيوانية وغيرها من المجتمعات المختلفة التي يهتم بها الباحثون لسبب أو لآخر تناولها العديد من الباحثين من بينهم Acharyal et al.(2000) و Boomer atel.(2000) و Clausen at el.(1999) Danahera and King(1994) و Petrucci(1998). ولمزيد من المعلومات ولمواضيع متفرقة أخرى لاستخدامات أخرى للمعاينة العنقودية المكيفة يراجع: Thompson(1993,1996,2002) و Thompson et al.(1992) و Thompson and Seber(1996) و Muttlak and Khan(2002) .

2.18العينة العشوائية البسيطة كعينة أولية

1.2.18 سحب العينة العنقودية المكيفة

إنّ من أبسط طرق المعاينة العنقودية المكيفة تكون عندما نقوم بسحب الوحدات الأولية باستخدام العينة العشوائية البسيطة، حيث نقوم بتقسيم المجتمع إلى N من الوحدات الأولية ومن ثم نختار عينة عشوائية بسيطة بحجم n من الوحدات الأولية، ونقوم بإضافة الوحدات المجاورة عندما تحقق أياً من الوحدات الأولية شرطاً معيناً وليكن C، نقول: إن الوحدة i مستوفية للشرط إذا كانت قيمة المتغير الذي نرغب في دراسته يحقق الشرط أي $y_i \in C$. فعلى سبيل المثال الوحدة i مستوفية للشرط إذا كانت y_i أكبر أو تساوي رقماً ثابتاً c أي $C = \{y : y \geq c\}$. وإذا كان أيٌّ من الوحدات المضافة تحقق الشرط

نفسه نقوم بإضافة الوحدات المجاورة لها وهكذا إلى أن تكون جميع الوحدات الخارجية أو ما تسمى المجاورة لا ينطبق عليها الشرط C.

لشرح فكرة المعاينة العنقودية المكيفة لنفترض أننا نريد أن نقدر عدد حبات الكمأة أو ما يسمى في بعض المناطق الفقع لمنطقة الدراسة، وهي عبارة عن منطقة صحراوية مساحتها 150 كم^2. نقوم بتقسيم منطقة الدراسة إلى وحدات، وهي عبارة مربعات صغيرة مساحة كل وحدة 0.5 كم^2 كما هو مبين بالشكل 1. كما هو معروف فإن وجود الفقع يكون على شكل عناقيد ونادراً ما تجد حبة واحدة بمفردها. لنفترض أن الأرقام الموجودة داخل المربعات تمثل عدد حبات الفقع داخل كل مربع لتقدير عدد حبات الفقع في منطقة الدراسة نقوم بسحب عينة عشوائية بسيطة أولية بحجم n=10 من المربعات التي ميزناها بوضع علامة x في هذه المربعات. الآن نبدأ بالبحث في المربعات عن الفقع، فإذا عثرنا على فقع في أحد المربعات فإننا نقوم بالنظر في المربعات المجاورة لهذا المربع التي تكون إلى اليمين والشمال والأعلى والأسفل فإن عثرنا في أيٍّ من هذه المربعات المجاورة على فقع فإننا نقوم بإضافتها إلى العينة.

					14		6							
					17	13	12	x						
					12	15	18				7	16		
											12	18	6	
		X									x 10	7		
		12	x 4								5			
										2	3			
												x		
		15	10			x								
		8	11	13										
		10x	15	4						11	15			
		7	8						16	15	8			
						x				10	20			
													x	
	X													

ومن الشكل1: المعاينة العنقودية المكيفة، لتقدير عدد حبات الفقع لمنطقة الدراسة التي حجمها 300 وحدة سحبنا عينة عشوائية بسيطة أولية بحجم 10 وحدات موضحة بالعلامة x.

ثم نقوم بفحص المربعات المجاورة للمربع الذي وجدنا فيه فقعاً ونضيفها إلى العينة، وهكذا إلى أن نصل إلى أن جميع المربعات الخارجية خالية من الفقع، عند ذلك نتوقف عن إضافة مربعات جديدة كما هو موضح بالشكل 2. وتكون هذه هي المعاينة العنقودية المكيفة والوحدات المضللة تمثل العينة العنقودية المكيفة.

على الرغم من كون العنقود يُعَدُّ مجموعة طبيعية، فإنه لا يمكن استخدام العناقيد أو العنقود للقيام بالاشتقاقات النظرية؛ وذلك للدور المزدوج الذي تلعبه الوحدات الموجودة على حافة العناقيد التي تسمى وحدات الحافة التي هي عبارة عن الوحدات الخالية كما هو موضح بالشكل 2، إذا تم اختيار إحدى وحدات الحافة بالعينة فهي تمثل عنقوداً بحجم 1، ولكن إذا لم يتم اختيار الوحدة في العينة الأولية فإنه لا يزال يمكن اختيارها كوحدة حافة في أحد العناقيد؛ لهذا سنقوم باستعمال فكرة الشبكة A_i للوحدة i وهي عبارة عن العنقود الذي تم الحصول عليه مع الوحدة i ولكن من دون وحدات الحافة، لذا فإن اختيار أي وحدة بالشبكة i سيؤدي إلى اختيار كل الشبكة A_i. إذا كانت i هي الوحدة الوحيدة التي ينطبق عليها الشرط $C = \{y : y \geq 1\}$ فسيكون لدينا شبكة بحجم 1. كذلك نعرف الوحدة التي جرى سحبها في العينة الأولية ولا ينطبق عليها الشرط C بالشبكة وبحجم 1. وهذا يعني أن جميع العناقيد وبحجم وحدة واحدة عبارة عن شبكات بحجم 1. كذلك أي وحدة من وحدات الحافة.

					14		6							
					17	13	12	x						
					12	15	18				7	16		
											12	18	6	
		x									x 10	7		
		12	x 4								5			
							X			2	3			
												x		
		15	10			x								
		8	11	13										
		x 10	15	4							11	15		
		7	8						16	15	8			
						x				10	20			
													x	
	x													

الشكل2. المعاينة العنقودية المكيفة لتقدير عدد حبات الفقع لمنطقة الدراسة. المربعات المظللة التي عددها 54 مربعاً (وحدة) تمثل حجم العينة العنقودية المكيفة.

عبارة عن شبكة بحجم 1. لذا فإن أي عنقود يحتوي على أكثر من وحدة يمكن تقسيمه إلى شبكة أو شبكات بحجم 1 (شبكة لكل وحدة من وحدات الحافة). يمكننا أن نلاحظ أن العناقيد يمكن أن يكون بينها عناصر مشتركة وذلك من خلال وحدات الحافة، ولكن الشبكات ستكون منفصلة عن بعضها واتحادها يساوي عدد وحدات المجتمع.

2.2.18 تقدير الوسط الحسابي والمجموع الكلي للمجتمع

عندما يتم اختيار وحدات العينة الأولية دون إرجاع سيؤدي ذلك إلى أن تكون جميع الوحدات متميزة وذلك بسبب عدم الإرجاع، ولكن البيانات النهائية يمكن أن تحتوي على مشاهدات معادة وذلك بسبب أن بعض العناقيد ممكن أن تحتوي على أكثر من وحدة أولية؛ لذا ستكون الوحدة i في العينة

إما لأنها جزء من شبكة جرى سحب إحدى وحداتها بما فيها الوحدة i في العينة الأولية، أو لكونها إحدى وحدات الحافة لإحدى الشبكات لنفترض أن m_i تمثل عدد الوحدات في الشبكة التي تحتوي الوحدة i و a_i تمثل عدد الوحدات في الشبكات التي تكون الوحدة i وحدة حافة. إذا كانت الوحدة i ينطبق عليها الشرط C فإن $a_i = 0$ ، ولكن إذا كانت الوحدة i لا ينطبق عليها الشرط فإن $m_i = 1$. إن احتمال سحب الوحدة i في أي سحبة من السحبات المتعاقبة التي عددها n هو

$$p_i = \frac{m_i + a_i}{N}$$

أما احتمال أن تتضمن العينة الوحدة i فسيكون

$$\pi_i = 1 - \binom{N - m_i - a_i}{n} \Big/ \binom{N}{n}$$

عندما يجري سحب الوحدات الأولية مع الإرجاع، فسنحصل على مشاهدات معادة إما في العينة الأولية وذلك بسبب عدم الإرجاع أو بسبب سحب أكثر من وحدة في العنقود. وباستخدام هذا التصميم سيكون احتمال سحب الوحدة i في أي سحبة من السحبات المتعاقبة هو

$$p_i = \frac{m_i + a_i}{N}$$

أما احتمال أن تتضمن العينة الوحدة i فسيكون

$$\pi_i = 1 - (1 - p_i)^n .$$

باستخدام أي التصميمين (دون إرجاع أو مع الإرجاع) فإنه لا يمكن حساب جميع الاحتمالات p_i أو π_i باستخدام المعلومات المتوافر في العينة وذلك بسبب عدم معرفة جميع قيم a_i.

لا يمكن استخدام المقدّرات المعتادة التي مرت معنا في المعاينة العشوائية البسيطة أو المعاينة العنقودية في المعاينة العنقودية المكيفة دون إدخال بعض التعديلات عليها؛ وذلك لكونها ستكون مقدّرات منحازة. ولكن إذا استخدمنا العينة الأولية على اعتبار أن الوحدات التي ينطبق عليها الشرط C تمثل عنقوداً بحجم 1 فإن الوسط الحسابي سيكون تقديراً غير متحيز لتقدير الوسط الحسابي للمجتمع، ولكنه أهمل جميع البيانات التي تمت إضافتها باستخدام المعاينة العنقودية المعدلة وعليه سيكون أقل فاعلية من المقدّرات التي سنتناولها في هذا الفصل.

سوف نكتفي بطريقة واحدة لتقدير الوسط الحسابي للمجتمع أو المجموع الكلي باستخدام المعاينة العنقودية المكيفة، وذلك عندما يتم سحب الوحدات بشكل متعاقب. ولمزيد من المعلومات يراجع(Thompson(2002. لنفترض أن A_i تمثل الشبكة التي تحتوي على الوحدة i و m_i يمثل عدد الوحدات بالشبكة i والوحدة التي لا تستوفي الشرط يمكن اعتبارها شبكة بحجم 1. ولنفترض أن w_i يمثل متوسط المشاهدات للشبكة التي تحتوي على الوحدة i للعينة الأولية أي:

$$w_i = \frac{1}{m_i} \sum_{j \in A_i} y_j$$

لذا فإن المقدّر غير المتحيز للوسط الحسابي للمجتمع سيكون

$$\hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^{n} w_i$$

أما التباين فسيكون

$$\text{var}(\hat{\mu}_1) = \frac{N-n}{n(N-1)} \sum_{i=1}^{N} \frac{(w_i - \mu)^2}{N}$$

إذا سحبت العينة الأولية من دون إرجاع، أما إذا سحبت العينة الأولية مع الإرجاع فسيكون التباين

$$\text{var}(\hat{\mu}_1) = \frac{1}{n} \sum_{i=1}^{N} \frac{(w_i - \mu)^2}{N}$$

أما تقدير التباين فسيكون

$$s^2(\hat{\mu}_1) = \frac{N-n}{nN} \sum_{i=1}^{n} \frac{(w_i - \hat{\mu}_1)^2}{n-1}$$

إذا سحبت العينة الأولية من دون إرجاع، أما إذا سحبت العينة الأولية مع الإرجاع فسيكون التباين

$$s^2(\hat{\mu}_1) = \frac{1}{n} \sum_{i=1}^{n} \frac{(w_i - \hat{\mu}_1)^2}{n-1}$$

مثال 1: العينة الموضحة بالشكل 1 و2 تم سحبها باستخدام العينة العشوائية البسيطة ومن دون إرجاع وبحجم n=10. قَدِّر عدد الفقع في منطقة الدراسة والخطأ المعياري للتقدير، علماً بأن الأرقام الموجودة داخل المربعات تمثل عدد الفقع في كل مربع أو وحدة.

الحل:

لتقدر الوسط الحسابي لعدد حبات الفقع في منطقة الدراسة والخطأ المعياري للتقدير لابد من حساب w_i

$w_1 = \frac{76}{7} = 10.857$ و $w_2 = \frac{16}{2} = 8$ و $w_3 = \frac{101}{10} = 10.1$ و $i=1,2,...,10$ و $w_4 = w_5 = ... = w_{10} = 0$. وعليه سيكون تقدير الوسط الحسابي هو

$$\hat{\mu}_1 = \frac{1}{10}(10.857 + 8 + 10.1 + 0 + L\,0) = 2.896$$

أما تقدير التباين إلى $\hat{\mu}_1$ فهو

$$s^2(\hat{\mu}_1) = \frac{N-n}{Nn(n-1)}\sum_{i=1}^{n}(w_i - \hat{\mu}_1)^2 = \frac{300-10}{300(10)(10-1)}\{(10.857-2.896)^2 +$$

$$+(8-2.896)^2 + (10.1-2.896)^2 + (0-2.896)^2 + ... + (0-2.896)^2 = 2.149$$

لذا فإن تقدير عدد الفقع في منطقة الدراسة (المجموع الكلي) سيكون

$$N\hat{\mu}_1 = 300(2.896) = 869$$

وأخيراً فإن الخطأ المعياري للمجموع الكلي هو

$$\sqrt{N^2 s^2(\hat{\mu}_1)} = \sqrt{(300)^2(2.149)} = 439.784$$

ولكن إذا استخدمنا الوسط الحسابي العادي للعينة النهائية فإننا سنحصل على $\bar{y} = 193/54 = 3.574$: أما تقدير للمجموع الكلي للفقع فهو $N\bar{y} = 300(3.574) = 1072$. في الواقع إن الوسط الحسابي العادي سيؤدي إلى تقديرات أعلى مما يجب.

3.2.18 المقارنة بين المعاينة العنقودية المكيفة والمعاينة العشوائية البسيطة

لابد من الإشارة هنا إلى أن عدم التحيز للتصميم المكيف في هذا الفصل لا يعتمد على نوع المجتمع الذي نتعامل معه أو نرغب في دراسته؛ لأن عدم التحيز يعتمد على التصميم أو ما يسمىdesign-based؛ لذا فإن كون المعاينة

العنقودية المكيفة أكثر فاعلية من المعاينة العشوائية البسيطة، وهذا لا يعتمد على نوع المجتمع الذي نرغب في دراسته أو نسحب العينة منه.

لنفترض أن لدينا المعاينة العنقودية المكيفة مع عينة أولية بحجم n، جرى سحبها باستخدام المعاينة العشوائية البسيطة مع الإرجاع، وتم الحصول على المقدّر $\hat{\mu}_1$ لتقدير الوسط الحسابي للمجتمع μ. لنفترض أن تباين المجتمع المحدود

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

لذا فإن تباين الوسط الحسابي للعينة العشوائية البسيطة بحجم عينة ثابت n^* هو

$$\sigma^2(N-n^*)/Nn^*$$

و بمقارنة هذا بتباين $\hat{\mu}_1$ $var(\hat{\mu}_1)$ سنحصل على أن تباين المعاينة العنقودية المكيفة سيكون أصغر من تباين الوسط الحسابي للعينة العشوائية البسيطة بحجم n^* إذا كان

$$(\frac{1}{n} - \frac{1}{n^*})\sigma^2 < \frac{N-n}{Nn(N-1)}\sum_{k=1}^{K}\sum_{i\hat{I}A_k}(y_i - w_i)^2$$

حيث إن k يمثل عدد الشبكات. نلاحظ أن الحد الأيمن من المتباينة يحتوي على التباين داخل الشبكة. لذا فإن المعاينة العنقودية المكيفة ستكون أفضل من المعاينة العشوائية البسيطة إذا كان التباين داخل الشبكة كبيراً. يراجع Thompson(2002) لمزيد من المعلومات.

3.18 العينة الطبقية كعينة أولية

بعد تقسيم المجتمع أو منطقة الدراسة إلى طبقات، نقوم بسحب عينة عشوائية طبقية أولية من المجتمع أو منطقة الدراسة، عندما تكون قيمة المتغير الذي نرغب في دراسته في أي وحدة من الوحدات التي جرى سحبها منطبقاً عليه الشرط نقوم بإضافة الوحدات المجاورة إلى العينة، كذلك نقوم بإضافة أي من الوحدات المجاورة للوحدات التي تم إضافتها حديثاً للعينة إذا تحقق الشرط وهكذا، إلى أن تكون جميع الوحدات الموجودة على الحافة لا ينطبق عليها الشرط، أو بعبارة أخرى خالية من المتغير الذي نرغب في دراسته.

يُعَدُّ تصميم المعاينة العنقودية الطبقية المكيفة من التصاميم المهمة من الناحية العملية؛ لأن كثيراً من المجتمعات تتوافر عنها معلومات يمكن أن تستخدم لتقسيم المجتمع إلى طبقات، ولكن توزيع التركيزات أو العناقيد في المجتمع قد لا يمكن التنبؤ به. نقوم في المعاينة الطبقية التقليدية بجمع الوحدات المتقاربة أو المشابهة لبعضها في طبقة بناءً على معلومات سابقة. أما في المعاينة العنقودية المكيفة فنقوم بالاستفادة من الحالة العنقودية التي تكون عليها الوحدات في المجتمع عندما يكون حجم وتوزيع العناقيد في المجتمع لا يمكن التنبؤ بها قبل بدء المسح بالعينة.

باستخدام المعاينة العنقودية الطبقية المكيفة تكون المقدّرات التقليدية (كالوسط الحسابي الطبقي للعينة) متحيزةً، لذا سنقترح مقدّرات تكون غير متحيزةً باستخدام هذا النوع من المعاينة. كذلك ستكون لدينا بعض التعقيدات مثل الاختيار في طبقة معينة قد يؤدي بسبب إضافة الوحدات للعينة إلى التجاوز إلى طبقة أخرى. أي سيكون لدينا في عينة واحدة وحدات من أكثر من طبقة وهذا يعني أن المشاهدات في الطبقات المختلفة ليست مستقلة كما هو معلوم عن المعاينة الطبقية التقليدية. ولقد جرى اقتراح مقدّرات معينة

لحل هذه الإشكالية. وأخيراً لابد من الإشارة إلى أن معظم ما هو موجود في هذا الجزء مأخوذ من Thompson(1991b).

1.3.18 تصميم المعاينة العنقودية الطبقية المكيفة

لتنفيذ المعاينة العنقودية الطبقية المكيفة لابد أن يكون المجتمع مقسماً إلى L طبقة، وتحتوي الطبقة h وL-1,2,...h على N_h وحدة، ومجموع وحدات المجتمع نرمز لها N. ترمز u_{hi} إلى الوحدة i في الطبقة h ومع كل وحدة نعرف المتغير y_{hi} الذي نرغب في دراسته. لكل وحدة من وحدات المجتمع u_{hi} نعرف المجاور أو الجوار لهذه الوحدة بأنه عبارة عن مجموعة من الوحدات بما فيها الوحدة u_{hi} ينطبق على هذه الوحدات خاصية التماثل. التي تعني إذا كانت الوحدة $u_{h'i'}$ مجاورة للوحدة u_{hi} فإن الوحدة u_{hi} مجاورة للوحدة $u_{h'i'}$. ربما يكون المجاور لإحدى الوحدات يحتوي على وحدات من أكثر من طبقة. نقول إن الوحدة u_{hi} ينطبق عليها الشرط إذا كانت قيمة المتغير y لهذه الوحدة يقع في المجموعة المحددة C. على سبيل المثال يمكن أن تكون $C = \{y : y \geq c\}$.

					14		6							
					x17	13	12							
					12	15	18				7	16		
											x12	18	6	
											10	7		
		x12	4								5			
										x2	3			
		x												
							x							
		15	10											
		8	11	13										
		10	15	4						11	15			
		7	8						16	15	8		X	
					x					10	20			
										x				
	x													

الشكل3: المعاينة العنقودية الطبقية المكيفة لتقدير عدد حبات الفقع لمنطقة الدراسة التي يوجد فيها طبقتان سحبنا عينة عشوائية بسيطة أولية بحجم 5 وحدات من كل طبقة وهي الموضح بالعلامة x.

نقوم بسحب عينة عشوائية أولية من مجتمع أو منطقة الدراسة باستخدام المعاينة العشوائية الطبقية، وهذا يعني أننا سنقوم بسحب عينة عشوائية بسيطة دون إرجاع من كل طبقة h من طبقات المجتمع وبحجم n_h ويكون السحب من كل طبقة بشكل مستقل عن الطبقات الأخرى. وعندما ينطبق الشرط على أي وحدة فإن الوحدات المجاورة لها يتم إضافتها إلى العينة إن لم تكن ضمن العينة الأولية، وأيضًا نقوم بإضافة أي من الوحدات المجاورة للوحدات التي جرى إضافتها إذا انطبق الشرط على الوحدات الجديدة، وهكذا حتى تكون العينة تحتوي على جميع الوحدات التي ينطبق عليها الشرط.

لتوضيح فكرة سحب الوحدات للعينة العنقودية الطبقية المكيفة لنفترض أن في مثالنا الأول حول الفقع الموضح في الشكل 1 أن منطقة الدراسة كانت مقسمة إلى طبقتين كما هو موضح في الشكل 3 أعلاه. وسحبنا عينة عشوائية بسيطة بحجم وحدات 5 من كل طبقة، والمربعات التي تحتوي على x تمثل العينة الطبقية الأولية التي جرى سحبها بصورة مستقلة من كل طبقة. تكون أي من الوحدات التي جري سحبها بالعينة الطبقية الأولية مستوفية للشرط إذا كان $C = \{y : y \geq 1\}$ ، أي إذا كانت أي من الوحدات تحتوي

على حبة فقع واحدة أو أكثر فإنها مستوفية للشرط. وكذلك جرى إضافة الوحدات

					14		6							
					x17	13	12							
					12	15	18				7	16		
											x12	18	6	
											10	7		
		x12	4								5			
										X 2	3			
		x												
							x							
		15	10											
		8	11	13										
		10	15	4						11	15			
		7	8						16	15	8		X	
					x					10	20			
										x				
	x													

الشكل4: المربعات المظللة تمثل العينة العنقودية الطبقية المكيفة النهائية.

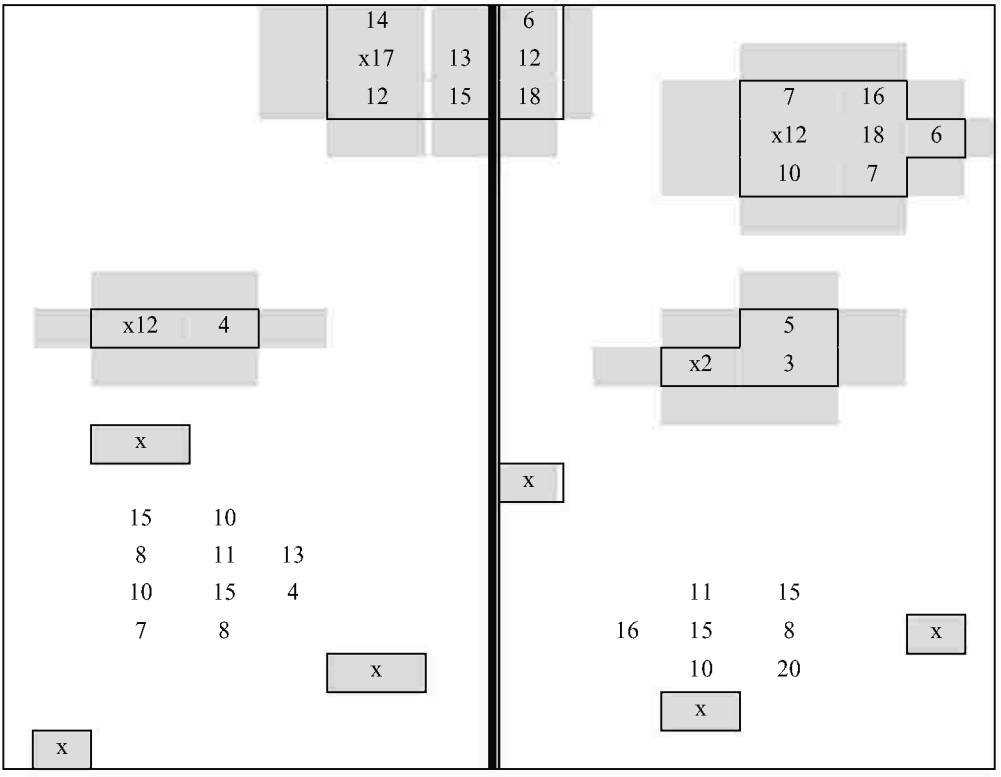
ملاحظة: بعض الوحدات من الطبقة الثانية في اليمين جرى إضافتها للعينة كنتيجة لاختيار إحدى وحدات الطبقة الأولى التي جرى سحبها في العينة الأولية من هذه الطبقة.

المجاورة للوحدات التي ينطبق عليها الشرط في المرحلة السابقة وهكذا. الشكل4 يوضح العينة العنقودية الطبقية المكيفة بشكلها النهائي.

يمكن تقسيم المجتمع إلى K من المجموعات ذات الوحدات والتي نسميها شبكات، بحيث إن اختيار أي وحدة في العينة الأولية سيؤدي إلى اختيار جميع الوحدات في الشبكة. الوحدة التي جرى سحبها بالعينة الأولية ولا تستوفي الشرط $C = \{y : y \geq 1\}$ هي عبارة عن شبكة بحجم 1. الشبكات المتميزة التي

تم الحصول عليها من خلال العينة الطبقية الأولية موضحة بالشكل 5 من خلال وضع خطوط حول وحدات كل شبكة.

لنفترض أن r_{hi} يمثل عدد المرات التي يمكن أن نختار بها الوحدة u_{hi}. ولنفترض أن m_{khi} يمثل عدد الوحدات التي تقع في تقاطع الطبقة k مع الشبكة التي تحتوي الوحدة u_{hi}. للوحدة u_{hi} والتي لا تستوفي الشرط ثم لنفترض أن a_{khi} تمثل مجموع الوحدات في تقاطع الطبقة k مع مجموعة من الشبكات المتميزة التي جرى تميزها بوضع خطوط حولها.



الشكل5: الشبكات المتميزة التي جرى الحصول عليها من العينة الأولية.

لا تحتوي الوحدة u_{hi} التي تتقاطع مع جيران الوحدة u_{hi}. الاختيار الأولي إلى أي من a_{khi} سيؤدي إلى إضافة الوحدة u_{hi} إلى العينة. لنعرف $a_{khi} = 0$ لأي وحدة ينطبق عليها الشرط.

إن توقع عدد المرات التي نختار بها الوحدة u_{hi} سيكون

$$E(r_{hi}) = \sum_{k=1}^{L} n_k \frac{m_{khi} + a_{khi}}{N_k}$$

2.3.18 تقدير الوسط الحسابي والمجموع الكلي للمجتمع

كما أشرنا سابقاً فإن المقدّرات التقليدية كالوسط الحسابي الطبقي سيكون متحيزاً في حالة استخدام المعاينة العنقودية الطبقية المكيفة، ويمكن أن نحصل على مقدّر غير متحيز لتقدير الوسط الحسابي للمجتمع ولو أنه غير فاعل وذلك باستخدام العينة العشوائية الطبقية، مع إهمال جميع الوحدات التي تمت إضافتها باستخدام المعاينة العنقودية الطبقية المكيفة.

للوحدة u_{hi} لنعرف متغير جديد نرمز له ب w_{hi}

$$w_{hi} = \frac{n_h}{N_h} \sum_{k=1}^{L} \xi_{khi} \Big/ \sum_{k=1}^{L} \frac{n_k}{N_k} m_{khi}$$

حيث إن ξ_{khi} يمثل مجموع قيم المتغير y الواقعة في الطبقة k والمتقاطعة مع الشبكة التي تحتوي على الوحدة u_{hi} و m_{khi} يمثل عدد الوحدات الواقعة في هذا التقاطع. لذا سيكون المقدّر للوسط الحسابي للمجتمع هو

$$\hat{\mu}_1 = \frac{1}{N} \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} w_{hi}$$

أما التباين للمقدّر $\hat{\mu}_1$ فسيكون

$$\text{var}(\hat{\mu}_1) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{\sigma_h^2}{n_h}$$

حيث إن

$$\sigma_h^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(w_{hi} - \overline{W}_h)^2$$

و

$$\overline{W}_h = (1/N_h)\sum_{i=1}^{N_h} w_{hi}$$

للحصول على تقدير غير متحيز للتباين $\text{var}(\hat{\mu}_1)$ فيمكن إيجاده باستبدال σ_h^2 بتباين العينة s_h^2 لنحصل

$$s^2(\hat{\mu}_1) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{s_h^2}{n_h}$$

حيث إن

$$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(w_{hi} - \overline{w}_h)^2$$

و

$$\overline{w}_h = (1/n_h)\sum_{i=1}^{n_h} w_{hi}$$

أما تقدير المجموع الكلي للمجتمع وتقدير تباينه فيمكن إيجادهما باستخدام $N\hat{\mu}_1$ و $N^2s^2(\hat{\mu}_1)$ على التوالي.

يمكننا أن نحصل على مقدّر جديد $\hat{\mu}_1'$ لتقدير الوسط الحسابي للمجتمع وذلك بتعديل لمقدّر $\hat{\mu}_1$ باتباع الطريقة التي اقترحها كل من

Birnbaum and Sirken(1965) و Levy(1977) و Sirken(1972) والتي يعتمد فيها الوزن على الطبقة التي توجد بها الوحدة التي سحبت في العينة الأولية وتتقاطع مع الشبكة. سنعرف للوحدة u_{hi} المتغير w'_{hi} الذي يمثل مجموع المتغير y في الشبكة التي تحتوي على الوحدة u_{hi} مقسوماً على عدد الوحدات في تلك الشبكة. لذا فإن

$$w'_{hi}=\sum_{k=1}^{L}\xi_{khi}/\sum_{k=1}^{L}m_{khi}$$

يمكننا أن نستخدم w'_{hi} للحصول على مقدّر غير متحيز للوسط الحسابي للمجتمع

$$\hat{\mu}'_1=\frac{1}{N}\sum_{h=1}^{L}\frac{N_h}{n_h}\sum_{i=1}^{n_h}w'_{hi}$$

والآن يمكننا أن نحصل على تباين للمقدّر $\hat{\mu}'_1$ وكذلك مقدّراً لتباينه باستخدام w'_{hi} بدلاً من w_{hi} في المعادلات الخاصة بالتباين وتقديره أعلاه.

مثال2: العينة العشوائية الطبقية والموضحة بالشكل 3 و4 و5 تم سحبها بعد تقسيم المجتمع إلى طبقتين حجمهما $N_1=140$ و $N_2=160$ ، حيث تم سحب خمس وحدات من كل طبقة كعينة أولية أي $n_1=5$ و $n_2=5$. قدر معدل عدد الفقع في منطقة الدراسة والخطأ المعياري للتقدير، علماً أنَّ الأرقام الموجودة داخل المربعات تمثل عدد الفقع في كل مربع أو وحدة.

الحل:

لنبدأ أولاً بإيجاد قيم w_{hi}

$$w_{11}=\frac{n_1}{N_1}\sum_{k=1}^{2}\xi_{khi}/\sum_{k=1}^{2}\frac{n_k}{N_k}m_{khi}=\frac{5}{140}(107)/[\frac{5}{140}(5)+\frac{5}{160}(3)]=14.033$$

$$w_{12}=\frac{n_1}{N_1}(\xi_{12})/\frac{n_1}{N_1}m_{12}=\frac{5}{140}(16)/\frac{5}{140}(2)=8$$

وبالطريقة نفسها نجد $w_{21}=76/7=10.857$ و $w_{22}=10/3=3.333$. نستطيع أن نحسب قيمة $\hat{\mu}_1$ لتكون

$$\hat{\mu}_1=(1/300)\{[140/5(14.033+8+0+0+0)]$$

$$+[160/5(10.857+3.33+0+0+0)]\}=3.57$$

لحساب تقدير التباين إلى $\hat{\mu}_1$ نريد أن نحسب $\overline{w}_h$ حيث $h=1, 2$

$$\overline{w}_1=\frac{1}{n_h}\sum_{i=1}^{n_h} w_{hi}=\frac{1}{5}(17.125+8)=4.41$$

وبالطريقة نفسها نحسب $\overline{w}_2=2.838$. الآن نستطيع أن نحسب s_h^2 و $h=1, 2$

$$s_1^2=[(14.033-4.41)^2+(8-4.41)^2+(0-4.41)^2+...+(0-4.41)^2=40.96$$

وبالطريقة نفسها نحسب $s_2^2=22.178$. الآن نستطيع أن نحسب تقدير التباين إلى $\hat{\mu}_1$

$$s^2(\hat{\mu}_1)=\frac{1}{300^2}[140(140-5)\frac{40.96}{5}+160(160-5)\frac{22.178}{5}]=2.943$$

أما الخطأ المعياري للمقدّر فهو

$$\sqrt{s^2(\hat{\mu}_1)}=\sqrt{2.943}=1.72 .$$

References

1. Acharyal, B., Bhattarai, G., de Gier, A. and Stein, A. (2000). Systematic Adaptive Cluster Sampling for the Assessment of Rare Tree Species in Nepal, *Forest Ecology and Management*, 137, 65-73.
2. Battista, T. D. (2003). Resampling Methods foe Estimating Dispersion Indices in Random and Adaptive Design, *Environmental and Ecological Statistics*, 10, 83-93.
3. Birnbaum, Z. W. and Sirken, M. G. (1965). Design of Sample Survey to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimators, *Vital and Health Statistics*, Ser. 2, No. 11, Washington, DC: US Government Printing Office.
4. Boomer, K., Werner, C. and Brantley, S. (2000). CO_2 Estimation Related to the Yellowstone Volcanic System: 1. Developing a Stratified Adaptive Cluster Sampling Plan, *Journal of Geophysycal Research*, 105, 817-830.
5. Borkowski, J. J. (1999). Network Inclusion Probabilities and Horvitz-Thompson Estimation for Adaptive Simple Latin Square Sampling, *Environmental and Ecological Statistics*, 6, 291-311.
6. Brown, J. A. (2003). Design and Efficient Adaptive Cluster Sample, *Environmental and Ecological Statistics*, 10, 95-105
7. Brown, J. A. and Manly, B. F. J. (1998). Restricted Adaptive Cluster Sampling, *Environmental and Ecological Statistics*, 5, 47-62.
8. Christman, M. C. (2003). Adaptive Two-Stage One-Per-Stratum Sampling, *Environmental and Ecological Statistics*, 10, 43-60.
9. Clausen, D., Hanselman, D., Lunsford, C., Quinn, T. and Heifetz, J. (1999). Rockfish Adaptive Sampling Experiment in the Central Gulf of Alaska, 1998. AFSC Processed Report 99-04., Juneau, AK: Alaska Fisheries Science Center.
10. Cormack, R. M. (1988). Statistical challenges in the environmental sciences: A personal view, *Journal of the Royal Statistical Society*, Series A, 151, 201-210.
11. Danaher, P. J. and King, M. (1994). Estimating Rare Household Characteristics using Adaptive Sampling, *The New Zealand Statistician*, 29, 14-23.
12. Dryver, A. L. (2003). Performance of Adaptive Cluster Sampling Estimators in s Multivariate Setting, *Environmental and Ecological Statistics*, 10, 107-113.
13. Felix Medina, M.H. (2000). Analytical Expression for Rao-Blackwell Estimators in Adaptive Cluster Sampling, *Journal of Planning and Inference*, 84, 221-236.
14. Felix Medina, M.H. (2003). Asymptotics in Adaptive Cluster Sampling, *Environmental and Ecological Statistics*, 10, 61-82.
15. Felix Medina, M. H. and Thompson, S. K. (1999). Adaptive Cluster Double Sampling, Proceeding of the Survey Research Section, American Statistical Association, 86, 445-449.
16. Levy, P. S. (1977). Optimum Allocation in Stratified Random Network Sampling for Estimating the Prevalence Attributes in Rare Populations, *journal of the American Statistical Association*, 72, 758-763.
17. Munholland, P. L. and Borkowski, J. J. (1993). Adaptive Latin Square Sampling +1 Designs, Technical Report No. 3-23-93, Department of Mathematical Sciences, Montana State University, Bozeman.
18. Muttlak, H. A and Khan, A., (2002). Adjusted Two-sage Adaptive Cluster Sampling, *Environmental and Ecological Statistics*, 9, 111-120.
19. Petrucci, A. (1998). Adaptive Sampling for Environmental Pollution Data: Some Simulation Results, *Statistica Applicata*, 103.
20. Pontius, J. S. (1997). Strip Adaptive Cluster Sampling: Probability Proportion to Size Selection of Primary Units, *Biometrics*, 53, 1092-1096.
21. Roesch, F. A., Jr. (1993). Adaptive Cluster Sampling for Forest Inventories, *Forest Science*, 39, 655-669.

22. Salehi, M. M. (1999). Rao-Blackwell Versions of the Hansen-Hurwitz and Horvitz-Thompson Estimators in Adaptive Cluster Sampling, *Environmental and Ecological Statistics*, 6, 183-1195.
23. Salehi, M. M. and Seber, G. F. A. (1997a). Adaptive Cluster Sampling with Networks Selected without Replacement, *Biometrika*, 84, 209-219.
24. Salehi, M. M. and Seber, G. F. A. (1997b). Two-Stage Adaptive Cluster Sampling, *Biometrics*, 53, 959-970.
25. Salehi, M. M. and Seber, G. F. A. (2002). Unbiased Estimators for Restricted Adaptive Cluster Sampling, *Australian and New Zealand Journal of Statistics*, 44, 63-75.
26. Salehi, M. M. (2003). Comparison Between Hansen-Hurwitz and Horvitz-Thompson Estimators for Adaptive Cluster Sampling, *Environmental and Ecological Statistics*, 10, 115-127.
27. Seber, G. F. A. (1986). A Review of Estimating Animal Abundance, *Biometrics*, 42, 267-292.
28. Seber, G. F. A. (1992). A Review of Estimating Animal Abundance: II *International Statistical Review*, 60, 129-166.
29. Sirken, M. G. (1972). Stratified Sample Survey with Multiplicity, *journal of the American Statistical Association*, 67, 257-266.
30. Smith, D. R., Conroy, M. J., and Brakhage, D. H. (1995). Efficiency of Adaptive Cluster Sampling for Estimating Density of Wintering Waterfowl, *Biometrics*, 51, 777-778.
31. Smith, D. R., Villella, R. F. and Lemarie, D. P. (2003). Application of Adaptive Cluster Sampling to Low-Density Populations of Freshwater Mussels, *Environmental and Ecological Statistics*, 10, 7-15.
32. Su Z. and Quinn, T. J. (2003). Estimators Bias and Efficiency for Adaptive Cluster Sampling with Order Statistics and a Stopping Rule, *Environmental and Ecological Statistics*, 10, 17-41.
33. Thompson, S. K. (1990). Adapive Cluster Sampling, *journal of the American Statistical Association*, 85, 1050-1059.
34. Thompson, S. K. (1991a). Adaptive Cluster Sampling: Design with Primary and Secondary units, *Biometrics*, 47, 1103-1115.
35. Thompson, S. K. (1991b). Stratified Adaptive Cluster Sampling, *Biometrika*, 78, 378-397.
36. Thompson, S. K. (1996). Adaptive Cluster Sampling based on Order Statistics, *Enviornmetrics*, 7, 123-133.
37. Thompson, S. K (2002). Sampling, 2end Wiley, New York.
38. Thompson, S. K., Ramsey, F. L. and Seber A. F. G. (1992). An Adaptive Procedure for Sampling Animal Populations, *Biometrics*, 48, 1195-1199.
39. Thompson, S. K. and Seber A. F. G. (1996). *Adaptive Sampling*, Wiley: New York.

الفصل التاسع عشر

المعاينة الشبكية

Network Sampling

1.19 مقدمة

لتقدير انتشار أحد الأمراض النادرة، نقوم بسحب عينة عشوائية من المستشفيات أو المراكز الطبية في مجتمع أو منطقة الدراسة، ونقوم بفحص سجلات هذه المستشفيات للبحث عن المرضى الذين عولجوا في هذه المستشفيات بسبب هذا المرض النادر، ونستخرج سجلاتهم ونقوم بدراستها أو جمع معلومات منها. ولكن بعض المرضى لديهم سجلات في أكثر من مستشفى أو مركز طبي لنفس المرض أي عولجوا في أكثر من مستشفى. من الواضح أنه كلما عولج أحد المرضى في أكثر من مستشفى فإن احتمال أن يسحب سجله في العينة يكون أكبر.

في دراسة أخرى لنفرض أننا نريد أن نقدر انتشار بعض الصفات النادرة في المجتمع، لتحقيق ذلك نقوم بسحب عينة عشوائية بسيطة من بعض المنازل في مجتمع الدراسة، ونسأل ساكنيها البالغين عن حدوث هذه الصفة التي نرغب بدراستها ولكن ليس فقط عن حدوثها معهم بل مع جميع أشقائهم أو شقيقاتهم. لذا من الواضح أن الشخص الذي يسكن في منزل آخر ولديه عدد كبير من الأشقاء والشقيقات سيكون احتمال سحبه في العينة أكبر من الشخص الذي ليس لديه أشقاء يسكنون في منازل مستقلة، حتى في البيت

الواحد ليس من الضرورة أن تكون احتمالات ضم جميع الأشخاص البالغين والساكنين في البيت نفسه متساوية.

يطلق على هذا التصميم المعاينة الشبكية أو المعاينة المتعددة. في هذا النوع من المعاينة نقوم بسحب عينة عشوائية بسيطة أو عينة عشوائية طبقية من الوحدات تسمى الوحدات المختارة، ومن ثم نقوم بإضافة جميع الوحدات المشاهدة التي لها علاقة بالوحدات المختارة والتي سحبت بالعينة. على سبيل المثال إذا قمنا بسحب عينة عشوائية بسيطة من المنازل في مدينة ما وقمنا بسؤال الأشخاص البالغين والساكنين في هذه المنازل عن صفة معينة نادرة يتصفون بها هم وأشقاؤهم الساكنون في المدينة، وليس بالضرورة في المنزل نفسه الذي يسكنون فيه، فالمنازل في هذه الدراسة تمثل الوحدات المختارة وأما الأشخاص فيمثلون وحدات المشاهدة. تعرف تعددية (Multip licity) الشخص بأنها عبارة عن عدد الوحدات المختارة (مستشفى أو منزل كما في أمثلتنا أعلاه) وهي التي يكون للشخص علاقة بها، ونعرف الشبكة على أنها عبارة عن مجموعة من الوحدات التي تمت مشاهدتها والتي ترتبط ببعض ها الآخر برباط معين. يمكن أن ترتبط الشبكة بأكثر من وحدة مختارة (أشقاء يعيشون في أكثر من منزل). كذلك فإن أي وحدة مختارة يمكن أن ترتبط بأكثر من شبكة (أشخاص بالغين يعيشون في بيت واحد وليسوا بأشقاء). إذا كان المجتمع يحتوي على طبقات يمكن أن تتقاطع الشبكة مع أكثر من طبقة.

بما أن الوحدات يجري سحبها باحتمالات غير متساوية، فإن الوسط الحسابي للعينة سيكون مقدّراً متحيزاً للوسط الحسابي للمجتمع باستخدام هذا النوع من المعاينة. قام كلٌّ من Birnbaum and Sirken (1965) باقتراح مجموعة من المقدّرات غير المتحيزة. هنالك مجموعة من البحوث ركزت على المقدّر المتعدد (Multiplicity Estimator) قام بها الباحثون: Nathan(1976)

و Sirken(1970,1972a,b) أما Levy(1977) و Sirken and Levy وقد ركزوا على نسب المقدّرات المتعددة التي يمكن أن تستعمل لتقدير نسبة مجموعة عرقية مصابة بنوع نادر من الأمراض. بينما ركز Czaja et al.(1986) على تأثير الأخطاء الناتجة من جمع المعلومات عن طريق الترابط بين الشبكات، على سبيل المثال يمكن الاعتماد على المريض بصورة أكبر من الاعتماد على قريبه في المنزل للحصول على المعلومات المطلوبة. هنالك تطبيقات كثيرة ومتعددة للعينة الشبكية. يراجع كلٌّ من: Kalton and Anderson(1986) و Faulkenberry and Garoui(1991) و Sudman et al.(1988).

2.19 تقدير الوسط الحسابي والمجموع الكلي

لنفترض أن قيمة المتغير الذي نرغب في دراسته للوحدة i في المجتمع هو y_i. في المسوحات التي نرغب في تقدير عدد المرضى المصابين بنوع نادر من الأمراض تكون قيمة $y_i = 1$ إذا كانت الوحدة تحمل الصفة، وهي أن يكون المريض مصاباً بالمرض النادر، وخلاف ذلك تكون قيمة المتغير $y_i = 0$. يمكن أن يأخذ المتغير y_i أي قيمة وليس فقط 1 و0. فعلى سبيل المثال يمكن أن تكون قيمة تمثل تكلفة العلاج للمريض i المصاب بالمرض النادر. لنفترض أن N تمثل عدد وحدات المجتمع التي يمكن مشاهدتها. أما المجموع الكلي للمجتمع فيعرف

$$\tau = \sum_{i=1}^{N} y_i$$

لنفترض أن m_i تمثل تعددية الوحدة المشاهدة i، أي عدد الوحدات المختارة التي ترتبط بوحدة المشاهدة. سوف نرمز لعدد الوحدات المختارة من المجتمع بـM. لذا فإن الوسط الحسابي للمجتمع للوحدات المختارة سيكون

$$\tau = t/M$$

سوف نستخدم العينة العشوائية البسيطة دون إرجاع لاختيار n من الوحدات ذالمختارة وكل وحدة مشاهدة ترتبط بأي وحدة من الوحدات المختارة سيجري ضمها إلى العينة.

سوف نستخدم المقدّر المتعدد لتقدير مجموع أو الوسط الحسابي للمجتمع. ولمزيد من المعلومات ولمقدّرات أخرى يراجع Birnbaum and Sirken(1965) أو Thompson(2002) إن احتمال سحب وحدة المشاهدة i في أي سحبة من السحبات المتعاقبة هو

$$p_i = \frac{m_i}{M}$$

يمكننا أن نجد المقدّر غير المتحيز للمجموع الكلي للمجتمع τ بقسمة قيمة المشاهدة للمتغير y على احتمال اختيار الوحدة المصاحبة، لذا فإن المقدّر المتعدد للمجموع الكلي يمكن تعريفه بما يأي:

$$\hat{\tau}_m = \frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i} = \frac{M}{n}\sum_{i\in s}\frac{y_i}{m_i}$$

حيث إن s تمثل وحدات المشاهدة المتعاقبة في العينة بما فيها الوحدات المعادة. لذا فإن وحدات المشاهدة يمكن سحبها أكثر من مرة بالرغم من أننا نسحب الوحدات دون إرجاع؛ لأن وحدات المشاهدة يمكن أن ترتبط بأكثر من وحدة من الوحدات المختارة؛ لذا فإن توقع عدد المرات التي يمكن أن تختار بها وحدة المشاهدة i هو np_i.

يمكننا أن نعيد تعريف المقدّر المتعدد للمجموع الكلي للمجتمع بشكل أكثر سهولة، للوحدة المختارة j في المجتمع نعرف المتغير w_j ليكون مجموع y_i/m_i لجميع وحدات المشاهدة المرتبطة بالوحدة j أي

$$w_j = \sum_{i\in A_j}\frac{y_i}{m_i}$$

حيث إن A_j عبارة عن وحدات المشاهدة المرتبطة بالوحدة المختارة j. وباستخدام الرموز الجديدة يمكننا أن نعيد كتابة المقدّر المتعدد للمجموع الكلي كما يأتي:

$$\hat{\tau}_m = \frac{M}{n}\sum_{j=1}^{n} w_j$$

يمكننا أن نفكر بالمتغير w_j بصفته متغيرًا جديدًا متعلقًا بالوحدة المختارة j ونرغب في دراسته. لذا سيكون المقدّر المتعدد عبارة عن $M\,\overline{w}$ ، حيث إن $\overline{w} = \frac{1}{n}\sum_{j=1}^{n} w_j$ عبارة عن الوسط الحسابي للعينة العشوائية البسيطة وبحجم n. لذا وباستخدام النتائج الأساسية للعينة العشوائية البسيطة نحصل على تباين $\hat{\tau}_m$ كما يأتي:

$$\text{var}(\hat{\tau}_m) = \frac{M(M-n)}{n}\sigma_w^2$$

حيث إن

$$\sigma_w^2 = \frac{1}{M-1}\sum_{j=1}^{M}(w_j-\mu)^2$$

حيث إن $m = \tau / M$ تمثل الوسط الحسابي للمجتمع لكل وحدة اختيار. أما المقدّر غير المتحيز لتباين المجتمع $\text{var}(\hat{\tau}_m)$ فسيكون

$$s^2(\hat{\tau}_m) = \frac{M(M-n)}{n}s_w^2$$

حيث إن

$$s_w^2 = \frac{1}{n-1}\sum_{j=1}^{n}(w_j-\overline{w})^2$$

لتقدير الوسط الحسابي لمجتمع الوحدة المختارة فسيكون

$$\hat{\mu}_m = \hat{\tau}_m / M$$

و تباين

$$\text{var}(\hat{\mu}_m) = \text{var}(\hat{\tau}_m) / M^2$$

و تقدير للتباين

$$s^2(\hat{\mu}_m) = s^2(\hat{\tau}_m) / M^2$$

مثال:

لشرح طريقة حسابات مقدّرات العينة الشبكية سوف نستخدم المثال الآتي: لتقدير شيوع أو انتشار أحد الأمراض في إحدى المدن والمؤلفة من M=5000 منزل، سحبنا عينة عشوائية بسيطة وبحجم n=100 منزل. البالغون في المنازل المختارة سيقومون بالتبليغ فيما إذا كانوا مصابين بالمرض أو أياً من أشقائهم أو شقيقاتهم في المدينة، هنا المنازل تمثل وحدات الاختيار أو الوحدات المختارة بينما الناس البالغون يمثلون وحدات المشاهدة، المتغير الذي نرغب في دراسته $y_i = 1$ إذا كان الشخص مصاباً بالمرض و $y_i = 0$ إذا كان الشخص غير مصاب بالمرض.

نقوم بترتيب المنازل والبالغ عددها 100 في العينة لنضع المنازل التي تحتوي على إصابات في المقدمة أي المنازل التي تكون فيها قيمة المتغير y لا تساوي صفراً بالمقدمة. يعيش في البيت الأول شخصان بالغان رجل وامرأة، لدى الرجل شقيق يسكن في دار منفصل، الرجل لا يحمل المرض $(y_1 = 0)$ ولكن شقيقه مصابٌ $(y_2 = 1)$، هؤلاء الشقيقان يشكلان الشبكة الأولى بتعددية تساوي 2 أي $m_1 = 2$، أما المرأة في البيت الأول فهي مصابة بالمرض $(y_3 = 1)$ ولديها شقيقان يسكنان في منزل مستقل، الأول مصاب بالمرض $(y_4 = 1)$ والثاني غير مصاب بالمرض $(y_5 = 0)$. هؤلاء الأشقاء الثلاثة يشكلون الشبكة الثانية بتعددية مقدارها 3 أي $m_2 = 3$، ولكن المنزل الذي يسكن فيه الشقيق الخامس جرى سحبه في العينة أي من بين المئة منزل، لذا فإن المنزلين الذين يسكن فيهما الأشقاء الثلاثة جرى سحبهما مرتين، أما المنزل الثاني الذي تم

اختياره بالعينة فيسكن فيه رجل وزوجته، أحدهما شقيق الزوجة في البيت الأول والثاني غير مصاب بالمرض($y_6 = 0$) ولا يوجد لديه أشقاء في المدينة؛ لذا نرى أن الزوج (الشخص السادس) يشكل شبكة بحجم وحدة واحدة أي $m_3 = 1$. أما المنزل الثالث يسكن فيه شخص واحد بالغ ويحمل المرض ($y_7 = 1$) ولا يوجد لديه أشقاء في المدينة، وهذا يشكل الشبكة الرابعة بتعددية تساوي 1 أي $m_4 = 1$. جميع المنازل الباقية التي عددها 97 منزل لا يوجد فيها مصابون ولا يوجد لديهم أشقاء مصابون بالمرض. لذا فإن قيم المتغير y=0.

الحل:

لحساب المقدّر المتعدد سنقوم بحساب w_j في جميع المنازل. بالنسبة للمنزل الأول فإن $w_1 = \frac{1}{2}+\frac{2}{3} = \frac{7}{6}$. أما بالنسبة للمنزل الثاني فإن $w_2 = \frac{2}{3}+\frac{0}{1} = \frac{2}{3}$. وأما بالنسبة للمنزل الثالث فإن $w_3 = \frac{1}{1} = 1$. لكل منزل من المنازل الباقية التي عددها 97 فإن $w_j = 0$. الآن نستطيع أن نحسب قيمة المقدّر المتعدد للمجمع الكلي τ

$$\hat{\tau}_m = \frac{5000}{100}\left(\frac{7}{6}+\frac{2}{3}+1+0+L+0\right)=141.7$$

لكي نحسب تقدير التباين إلى $\hat{\tau}_m$ لابد من حساب الوسط الحسابي وتباين العينة للمتغير W، لذا فإن $\overline{w} = 0.02833$ و $s_w^2 = 0.02753$. الآن نستطيع حساب تقدير التباين إلى $\hat{\tau}_m$ فهو

$$s^2(\hat{\tau}_m) = \frac{5000(500-100)}{100}(0.02753) = 6745.$$

أما تقدير الخطأ المعياري فسيكون

$$\sqrt{s^2(\hat{\tau}_m)} = \sqrt{6745} = 82.13.$$

3.19 المعاينة الشبكية الطبقية

قد تنشأ تعقيدات عندما يتم سحب الوحدات المختارة من مجتمع مؤلف من طبقات وذلك بسبب أن بعض وحدات المشاهدة يمكن أن ترتبط بالوحدات المختارة في أكثر من طبقة؛ لذا فإن المشاهدات في الطبقات المختلفة ليست مستقلة كما هو متعارف عليه في العينة العشوائية الطبقية المعروفة.

لنفترض أن الوحدات المختارة في المجتمع التي عددها M مقسمة إلى L من الطبقات بحجم M_h وحدة لكل طبقة h، ولنفرض أنه جرى سحب عينة عشوائية بسيطة بحجم n_h من كل طبقة حيث إن h=1,2,...,L. لكل وحدة اختيار في العينة، جميع وحدات المشاهدة المرتبطة بها يجري ضمها للعينة بغض النظر عن الطبقات التي تنتمي لها هذه الوحدات. لنفترض أن A_{hj} عبارة عن مجموعة من الوحدات المشاهدة التي ترتبط بالوحدة j في الطبقة h. لوحدة المشاهدة i نفرض أن m_i تمثل عدد وحدات الاختيار التي يمكن أن تكون جاءت من أكثر من طبقة ومرتبطة بالوحدة i. لنعرف وحدة الاختيار j في الطبقة h المتغير الجديد الذي نرغب في دراسته

$$w_{hj} = \sum_{i \in A_{hj}} \frac{y_i}{m_i}$$

لنعرف الوسط الحسابي للمتغير w في الطبقة h بما يأتي:

$$\overline{w}_h = \frac{1}{n_h} \sum_{j=1}^{n_h} w_{hj}$$

المقدّر الطبقي المتعدد للمجموع الكلي للمجتمع كما اقترحه BirnbaumandSirken(1965) هو

$$\hat{\tau}_m = \sum_{h=1}^{L} M_h \overline{w}_h$$

حيث إنه مقدّر غير متحيز للمجموع الكلي للمجتمع وبتباينه

$$\text{var}(\hat{\tau}_m) = \sum_{h=1}^{L} \frac{M_h(M_h - n_h)}{n_h} \sigma^2_{w_h}$$

حيث إن $\sigma^2_{w_h}$ يمثل التباين للمجتمع المحدود لقيم المتغيرw في الطبقة h. يمكننا الحصول على تقدير غير متحيز لتباين $\hat{\tau}_m$ أعلاه باستبدال $\sigma^2_{w_h}$ بتباين العينة للمتغير w في الطبقة h وهو s^2_{wh}.

لابد من ملاحظة أنه بالرغم من كون $\hat{\tau}_m$ مقدّراً غير متحيز للمجموع الكلي للمجتمع τ فإن $M_h \overline{w}_h$ قد لا يكون بصورة عامة مقدّراً غير متحيز إلى المجموع الكلي لكل طبقة. وذلك ربما يكون $\overline{w}_h$ متحيزاً بصورة جزئية لقيم المتغير y لوحدات المشاهدة المرتبطة مع طبقات أخرى، على سبيل المثال إذا كانت وحدات الاختيار هي المنازل في منطقة سكنية، والمنطقة مقسمة إلى طبقات حسب الموقع الجغرافي، ووحدات المشاهدة عبارة عن أشخاص بالغين ومرتبطين بعلاقات أخوية مع آخرين في المنطقة السكنية؛ لذا فإن اختيار منزل في طبقة ما قد يؤدي إلى الإبلاغ عن أشقاء موجودين في أكثر من طبقة. نقوم بجمع قيم المتغير y للأشقاء في الطبقات المختلفة في w_{hj} لهذا المنزل. لمعالجة هذه المشكلة ولمزيد من المعلومات يراجع كلٌّ من BirnbaumandSirken(1965) و Thompson(2002).

References

1. Birnbaum, Z. W. and Sirken, M. G. (1965). Design of Sample Survey to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimators, *Vital and Health Statistics*, Ser. 2, No. 11, Washington, DC: US Government Printing Office.
2. Czaja, R. F. Snowdon, C. B. and Casady, R. J. (1986). Reporting Biased and Sampling Errors in a Survey of a Rare Population using Multiplicity Counting Rules, *Journal of the American Statistical Association*, 81, 411-419.
3. Faulkenberry, G. D. and Garoui, A. (1991). Estimating a Population Total using an Area Frame, *Journal of the American Statistical Association*, 86, 445-449.
4. Kalton, G. and Anderson, D. W. (1986). Sampling Rare Populations, *Journal of the Royal Statistical Society* A, 149, 65-82.
5. Levy, P. S. (1977). Optimum Allocation in Stratified Random Network Sampling for Estimating the Prevalence Attributes in Rare Populations, *journal of the American Statistical Association*, 72, 758-763.
6. Nathan, G. (1976). An Empirical Study of Response and Sampling Errors for Multiplicity Estimates with Different Counting Rules, *Journal of the American Statistical Association*, 71, 808-815.
7. Sirken, M. G. (1970). Household Survey with Multiplicity, *Journal of the American Statistical Association*, 63, 257-266.
8. Sirken, M. G. (1972a). Stratified Sample Survey with Multiplicity, *Journal of the American Statistical Association*, 67, 224-227.
9. Sirken, M. G. (1972b). Variance Components of Multiplicity Estimators, *Biometrics*, 28, 869-873.
10. Sirken, M. G. and Levy, P. S. (1974). Multiplicity Estimation of Proportions Based on Ratios of Random Variables, *Journal of the American Statistical Association*, 69, 68-73.
11. Sudman, S., Sirken, M. G. and Cowan, C. D. (1988). Sampling Rare and Elusive Populations, *Science*, 240, 991-996.
12. Thompson, S. K. (2002). Sampling, 2end Wiley, New York.

الفصل العشرون

معاينة المجموعات المرتبة

Ranked Set Sampling

1.20 مقدمة

تُعَدُّ تكلفة جمع البيانات من أهم أسباب استخدام طرق المعاينة المختلفة، خصوصاً إذا كانت تكاليف قياس صفات معينة للوحدات التي نرغب في دراستها مرتفعةً أو تحتاج وقتاً طويلاً لقياسها. لقد كان McIntyre(1952) أول من اقترح طريقة أكثر فاعلية لتقدير إنتاج حقول الرعي في أستراليا والتي أصبحت تعرف فيما بعد بطريقة معاينة المجموعات المرتبة. تُعَدُّ طريقة معاينة المجموعات المرتبة من الطرق الفاعلة في تقليل كلفة جمع البيانات؛ وذلك من خلال تخفيض حجم العينة إذا توافرت بعض الشروط. وبالرغم من كون معاينة المجموعات المرتبة قديمة إلا أنها لم تستخدم بشكل واسع وتنتشر إلا في العقد الأخير من القرن الماضي على الرغم من فاعليتها في خفض تكاليف جمع البيانات.

إن أول من اقترح فكرة معاينة المجموعات المرتبة هو McIntyre(1952) ضمن جهوده المتميزة لإيجاد مقدّر يكون أكثر فاعلية لتقدير إنتاج حقول الرعي الواسعة في أستراليا؛ لأن قياس إنتاجية الحقول يتطلب قطع ووزن الحشيش الموجود في هذه الحقول وهذا يتطلب جهداً ووقتاً كبيرين، ولكن شخصاً خبيراً يستطيع أن يرتب مجموعة من القطع المحتوية على الحشيش بالعين المجردة حسب كمية إنتاجها من الأدنى إلى الأعلى من غير أن تكون

هناك حاجة إلى قطع ووزن الحشيش في هذه القطع؛ لذا فإن McIntyre تبنى طريقة المعاينة الآتية: نقوم بسحب n من المجموعات بطريقة عشوائية حجم كل مجموعة n من الوحدات (القطع)، ومن ثم نقوم بترتيب كل مجموعة من القطع التي عددها n بالعين المجردة من دون قياس الوحدات من دون تكاليف حسب إنتاجيتها من الحشيش من الأدنى إلى الأعلى. من المجموعة الأولى والمرتبة من الأدنى إلى الأعلى نقوم بقطع ووزن الحشيش، من القطعة التي تحتوي على الحد الأدنى (القطعة بالمرتبة الدنيا) من الحشيش. ومن المجموعة الثانية نقوم بعد ترتيبها من حيث الإنتاج من الأدنى إلى الأعلى بقطع الحشيش من القطعة التي تحتوي على المرتبة الثانية (القطعة بالمرتبة الثانية) من حيث كمية الحشيش. وهكذا إلى أن نقطع الحشيش من القطعة ذات الإنتاج الأعلى ضمن المجموعة الأخيرة (القطعة بالمرتبة العليا). يمكننا أن نكرر العملية r من الدورات أو المرات لنحصل على عينة بحجم nr من الوحدات.

يبدو أن فكرة العينة المرتبة التي اقترحها McIntyre لم تلاقي رواجاً، وبقيت منسية حتى عام 1966 حيث قام Halls and Dell بتطبيق هذه الفكرة لتقدير إنتاج علف الماشية في غابات أشجار الصنوبر. في الحقيقة إن أول من استخدم مصطلح معاينة المجموعات المرتبة هما Halls and Dell. أما أول من قدم البراهين الرياضية لهذا النوع من المعاينة فهما العالمان اليابانيان Takahasi and Wakimoto(1968) حيث برهنا على أن الوسط الحسابي لهذا النوع من المعاينة مقدّر غير متحيز إلى الوسط الحسابي للمجتمع، وتباينه أقل من تباين الوسط الحسابي للعينة العشوائية البسيطة، وذلك إذا كان ترتيب الوحدات يتم بصورة كاملة، أي دون أخطاء في ترتيب وحدات المجموعة من الأدنى إلى الأعلى. أما Dell and Clutter(1972) فقد توصلا إلى نفس النتيجة أعلاه ولكن مع إسقاط الشرط الذي يفترض كون الوحدات داخل المجموعة ترتب بصورة كاملة، أي سمحا للخطأ في ترتيب الوحدات داخل المجموعة أو

عينة من الأصغر إلى الأكبر. وهذا مهم في الحياة العملية حيث يصعب التأكد من عدم وجود أخطاء في ترتيب وحدات المجموعة أو العينة الواحدة. يُعَدُّ Dell and Clutter(1972) و David and Levine(1972) أول من أعطى معالجة نظرية لمعاينة المجموعات المرتبة مع الخطأ في ترتيب الوحدات داخل كل مجموعة أو عينة. اقترحت Stokes(1977) باستخدام المتغير الملازم أو المصاحب لتقدير رتب المتغير الذي نرغب في دراسته والذي يصعب ترتيب وحداته بالعين المجردة. قامت Stokes(1980a, b) بتقدير تباين المجتمع ومعامل الارتباط بين متغيرين إذا كان المجتمع يتبع التوزيع الطبيعي الثنائي باستخدام عينة المجموعات المرتبة. قام كلٌّ من Muttlak and McDonald(1990a,b,1992) باستخدام معاينة المجموعات المرتبة إذا تم سحب الوحدات من المجتمع باحتمالات متناسبة مع أحجامها، واقترحا مقدّرات للوسط الحسابي للمجتمع، واثبتا أن تباين هذه المقدّرات أقل من تباين العينة العشوائية البسيطة، كما استخدما الطريقة –أول مرة– مع طريقة المعاينة بخط التقاطع. وأخيراً قاما بتطبيق هذه الطريقة بتقدير التغطية والكثافة لنوع معين من الشجيرات في حقل الدراسة.

شهد العقد الأخير من القرن العشرين تطوراً كبيراً في استخدام معاينة المجموعات المرتبة؛ وذلك من خلال تناول أوجه متعددة لاستخدام هذا النوع من المعاينة. سنشير إلى بعض هذه البحوث بعجالة ولكننا ننصح بمراجعة Patil et al.(1999) و Muttlak and Al-Saleh(2000) للحصول على مزيد من البحوث في هذه الطريقة حتى عام 2000. لتقدير دالة التوزيع الاحتمالي يراجع كلٌّ من Stoke and Sager(1988) و Kvam and Samanigeo(1994) و Chen(2000a). أما فيما يختص الطرق غير المعلماتية يراجع على سبيل المثال كلٌّ من Bohn and Wolfe(1992,1994) و Bohn (1996,1998) و Omer(1999,2002). فيما يخص الطرق المعلماتية باستخدام معاينة المجموعات المرتبة فهي أكثر مما نستطيع ذكره هنا، نذكر منها: Fei et. Al.(1994)

و Bohj(1997a,b,c,1999a,b) و Abu-DayyehandMuttlak(1996) و BohjandAhsanullah(1996) و Shen(1994) و Stokes(1995) و Chen(2000b) و Sinha et al.(1996) لمزيد من المعلومات ومراجعة حول الطرق المعلماتية يراجع Ni and Sinha(1998). لقد لقي تحليل الانحدار باستخدام طريقة معاينة المجموعات المرتبة لجمع البيانات اهتمام مجموعة من الباحثين من بينهم Muttlak(1995,1996b,1998) و Patil et al.(1993) و Yu and Lam(1997) و Chen(2001). أما التصميم الأمثل في سياق معاينة المجموعات المرتبة في حالة كون المجموعات غير متساوية فقد تناولها Kaur et al.(1997) و Chen(2001). أما استخدام معاينة المجموعات المرتبة في مراقبة وضبط الجودة فقد تناولها كلٌّ من الباحثين Salazar and Sinha(1997) و MuttlakandAl-Sabah(2003a,b).

2.20 تقدير الوسط الحسابي للمجتمع

يمكن تلخيص طريقة معاينة المجموعات المرتبة كما هو موضح في الشكل 1 أدناه، ففي الخطوة الأولى نقوم بسحب 5 مجموعات (عينات) بطريقة عشوائية من المجتمع، حجم كل مجموعة 5 وحدات. في الخطوة الثانية نقوم بترتيب كل مجموعة من المجموعات الخمسة من الأدنى إلى الأعلى، وفي الخطوة الثالثة نقوم بقياس الوحدات الواقعة على القطر، وبذلك نكون حصلنا على العينة العشوائية المرتبة وبحجم 5 وحدات. بإمكاننا إعادة العملية من جديد للحصول على 5 وحدات جديدة وهكذا.

1.2.20 تقدير الوسط الحسابي للمجتمع إذا كان ترتيب الوحدات داخل لكل مجموعة دون أخطاء

لنفترض أن $X_{(i:n)j}$ القيمة الإحصائية للمتغير برتبة i في العينة التي حجمها n في الدورة j حيث إن i=1,2,....n و j=1,2,...,r. لذا فإن المقدّر غير المتحيز لتقدير

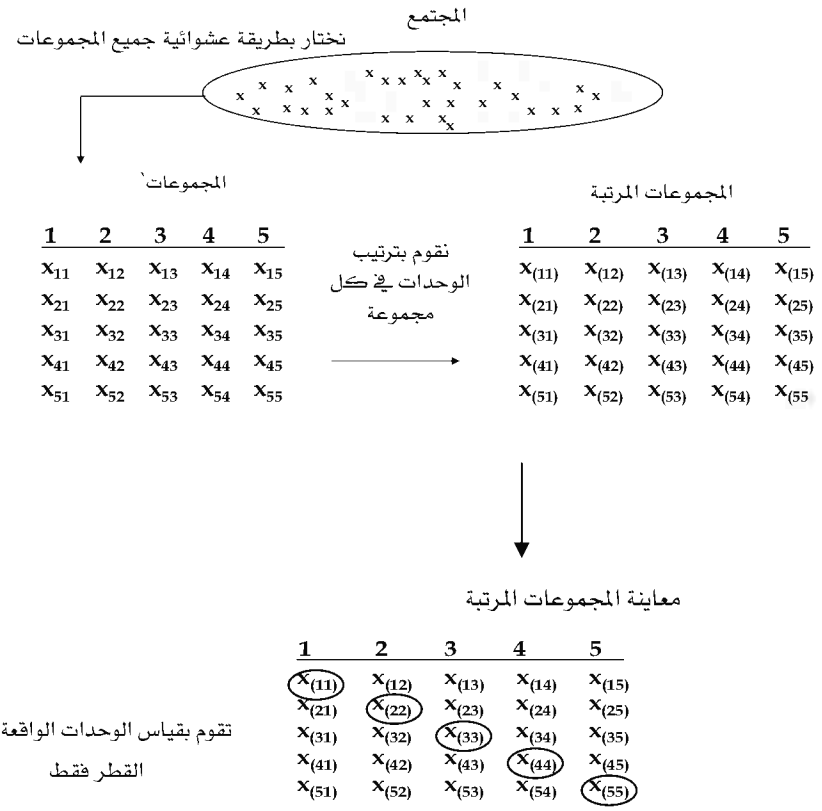
متوسط المجتمع إذا كان ترتيب الوحدات داخل كل مجموعة كاملاً أي دون أخطاء هو

$$\overline{X}_{rss} = \frac{1}{nr}\sum_{i=1}^{n}\sum_{j=1}^{r} X_{(i:n)\,j}$$

أما تباين $\overline{X}_{rss}$ فهو

$$var\left(\overline{X}_{rss}\right) = \frac{1}{n^2 r}\sum_{i=1}^{n}\sigma^2_{(i:n)}$$

الشكل 1 : خطوات اختيار عينة المجموعات المرتبة



حيث إن

$$\sigma^2_{(i:n)} = E\left[X_{(i:n)} - E\left(X_{(i:n)}\right)\right]^2$$

لقد برهنا Takahasi and Wakimoto (1968) على أن $\overline{X}_{rss}$ هو مقدّر غير متحيز لتقدير متوسط المجتمع وعلى أن

$$var(\overline{X}_{rss}) \le var(\overline{X}_{srs})$$

حيث إن $\overline{X}_{srs}$ يمثل متوسط العينة العشوائية البسيطة وبالحجم نفسه، أما تقدير التباين فيمكن الحصول على مقدّر غير متحيز إلى $var(\overline{X}_{rss})$ وهو

$$s^2_{\overline{x}_{rss}} = \frac{1}{n^2 r(r-1)} \sum_{j=1}^{r} \sum_{i=1}^{n} (X_{(i:n)\,j} - \overline{X}_{(i:n)})^2$$

حيث إن

$$\overline{X}_{(i:n)} = \frac{1}{r} \sum_{j=1}^{r} X_{(i:n)j};\ i = 1,2,\ldots,n$$

يراجع كلٌّ من Stokes(1980) Johnson et al. (1993) وللبراهين الرياضية ومزيد من المعلومات.

أما إذا كانت عدد الدورات r=1 فإن $\overline{X}_{rss}$ سيعرف كما يأتي:

$$\overline{X}_{rss} = \frac{1}{n} \sum_{i=1}^{n} X_{(i:n)}$$

و أما تباينه فسيكون

$$var\left(\overline{X}_{rss}\right) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2_{(i:n)}$$

2.2.20 تقدير الوسط الحسابي إذا كان ترتيب الوحدات ضمن كل مجموعة فيها أخطاء

لنفترض أن $X_{[i:n]j}$ القيمة الإحصائية للمتغير برتبة المقدرة i في العينة التي حجمها n في الدورة j حيث إن i=1, 2, …,n و j=1,2,…r. ونعني بالرتبة المقدرة هنا أن هذه الرتبة يمكن أن تكون خطأ؛ لذا فإن المقدّر غير المتحيز لتقدير لمتوسط المجتمع إذا كان ترتيب الوحدات داخل كل مجموعة فيها أخطاء هو

$$\overline{X}_{rsse} = \frac{1}{nr}\sum_{i=1}^{n}\sum_{j=1}^{r} X_{[i:n]j}$$

أما تباين $\overline{X}_{rsse}$ فهو

$$var\left(\overline{X}_{rsse}\right) = \frac{1}{n^2 r}\sum_{i=1}^{n}\sigma^2_{[i:n]}$$

حيث إن

$$\sigma^2_{[i:n]} = E\left[X_{[i:n]} - E\left(X_{[i:n]}\right)\right]^2$$

لقد برهنا Dell and Clutter(1972) على أن $\overline{X}_{rss}$ هو مقدّر غير متحيز لتقدير متوسط المجتمع، وعلى أن

$$var(\overline{X}_{rsse}) \le var(\overline{X}_{srs})$$

بغض النظر فيما إذا كان هناك أخطاء بترتيب وحدات المجموعة أم لا.

أما تقدير التباين فيمكن الحصول على مقدّر غير متحيز إلى $var(\overline{X}_{rsse})$ وهو

$$s^2_{\overline{x}_{rsse}} = \frac{1}{n^2 r(r-1)}\sum_{j=1}^{r}\sum_{i=1}^{n}(X_{[i:n]j} - \overline{X}_{[i:n]})^2$$

حيث إن

$$\overline{X}_{[i:n]} = \frac{1}{r}\sum_{j=1}^{r} X_{[i:n]j} \; ; i = 1,2,\ldots,n$$

مثال: لتقدير الجزء غير المملوء من القناني الزجاجية لمشروب البيبسي كولا، قمنا باستخدام معاينة المجموعات المرتبة، حيث تم سحب 16 زجاجة بصورة عشوائية من خط إنتاج المصنع الكائن في مدينة الخبر في المملكة العربية السعودية، ومن ثم قمنا بتقسيم هذه القناني الست عشرة إلى أربع مجموعات، كل مجموعة تحتوي على n=4 قنينات، وبالعين المجردة حُدِدَت القنينة التي تحتوي على أقل ارتفاع من الجزء الفارغ وقياسه من المجموعة الأولى. من المجموعة الثانية حُددت، القنينة التي تحتوي على ثاني أقل ارتفاع فارغ وقياسه. وهكذا حتى وصلنا إلى المجموعة الرابعة حيث قمنا بقياس أعلى ارتفاع فارغ من بين قناني المجموعة الرابعة، ثم قمنا بإعادة الدورة r=8 مرات للحصول على عينة بحجم nr=32 قنينة. قَدّر متوسط ارتفاع الجزء الفارغ لإنتاج المصنع من قناني الببسي كولا وخطئه المعياري.

	المجموعات			
الدورات	1	2	3	4
1	5.7	5.8	6.1	6.0
2	5.6	5.8	5.8	6.1
3	5.6	5.8	6.1	5.9
4	5.8	6.0	6.1	6.7
5	5.8	6.0	6.0	6.0
6	5.8	5.9	6.1	6.0
7	5.9	5.8	5.9	6.1
8	5.9	5.9	6.1	6.0

الحل: متوسط ارتفاع الجزء الفارغ هو

$$\overline{X}_{rss} = \frac{1}{4(8)} \sum_{i=1}^{4} \sum_{j=1}^{8} X_{(i:n)j} = 5.94$$

لحساب تقدير تباين $\overline{X}_{rss}$ ، نحتاج أن نحسب الوسط الحسابي $\overline{X}_{(i:n)}$ لكل مجموعة، وكذلك مجموع المربعات الفروق داخل كل مجموعة، الجدول الآتي يعطينا هذه المعلومات لكل مجموعة

المجموعة	1	2	3	4
$\overline{X}_{(i:n)}$	5.763	5.875	6.025	6.10
$\sum_{j=1}^{8}(X_{(i:n)j} - \overline{X}_{(i:n)})^2$	0.09875	0.05500	0.09499	0.43999

$$s^2_{\overline{x}_{rss}} = \frac{1}{4^2 8(8-1)} \sum_{j=1}^{8} \sum_{i=1}^{4} (X_{(i:n)j} - \overline{X}_{(i:n)})^2 = \frac{0.688748}{896} = 0.00076869$$

لذا فإن الخطأ المعياري إلى $\overline{X}_{rss}$ هو

$$s_{\overline{x}_{rss}} = \sqrt{s^2_{\overline{x}_{rss}}} = \sqrt{0.00076869} = 0.0277$$

3.20 طرق معدلة لمعاينة المجموعات المرتبة

إن ترتيب الوحدات داخل المجموعات من الأصغر إلى الأكبر بالنسبة للمتغير الذي نرغب في دراسته بالعين المجردة غالباً ما يكون تنفيذه صعباً إذا ما كان حجم المجموعة كبيراً (أكثر من خمس وحدات)، وإذا أنجز بالغالب ستعتري عملية الترتيب أخطاء، وهذا سيؤدي إلى خفض فاعلية معاينة المجموعات المرتبة؛ لذا أصبح من الضرورة البحث عن بدائل لعملية ترتيب

الوحدات داخل المجموعة لتلافي الأخطاء في الترتيب. سنتناول هنا ثلاث طرق بديلة من بين الطرق المقترحة لتعديل عملية ترتيب الوحدات داخل المجموعات.

1.3.20 معاينة المجموعات المرتبة وسطياً

تُعَدُّ معاينة المجموعات المرتبة وسطياً من الطرق الفاعلة في تقليل الأخطاء في عملية ترتيب الوحدات داخل المجموعات، ويمكن تلخيص الطريقة فيما يأتي: نقوم بسحب n من المجموعات (العينات) بطريقة عشوائية من المجتمع حجم كل مجموعة n من الوحدات. ثم نقوم بترتيب المجموعات من الأصغر إلى الأكبر، إذا كان حجم المجموعةn عبارة عن عدد فردي، فإننا نختار الوسيط من كل مجموعة للقياس، أي الوحدة ذي الرتبة (n+1)/2، أما إذا كان حجم المجموعة أو المجموعات عدداً زوجياً، فإننا نقوم بسحب الوحدات ذات الرتبة n/2 للقياس من نصف المجموعات، ومن النصف الثاني نقوم بسحب الوحدات ذات الرتبة (n/2)+1 وقياسها، في كلا الحالتين سنحصل على عينة بحجم n من الوحدات. يمكننا إعادة الدورة r من المرات للحصول على عينة بحجم nr.

لنفترض أن $X_{(i:m)j}$ القيمة الإحصائية للمتغير للوسيط في العينة i والتي حجمها n في الدورة j حيث إن i=1,2,....n وj=1,2,...r إذا كان حجم المجموعة أو العينة عدداً فردياً. أما إذا كان حجم المجموعة عبارة عن عدد زوجي فإن i تمثل المرتبة الإحصائية n/2 لقيمة المتغير في المجموعة والتي بحجم n من الوحدات حيث إن i=1,2,...L=n/2. ويمثل المرتبة الإحصائية (n/2)+1 لقيمة المتغير في لمجموعة حيث إن i=L+1,L+2,...,n. المقدّر لتقدير متوسط المجتمع باستخدام معاينة المجموعات المرتبة وسطياً هو

$$\overline{X}_{mrss} = \frac{1}{nr}\sum_{i=1}^{n}\sum_{j=1}^{r} X_{(i:m)j}$$

أما تباين $\overline{X}_{mrss}$ فهو

$$var\left(\overline{X}_{mrss}\right)=\frac{1}{n^2r}\sum_{i=1}^{n}\sigma^2_{(i:m)}$$

حيث إن

$$.\sigma^2_{(i:m)}=E\left[X_{(i:m)}-E\left(X_{(i:m)}\right)\right]^2$$

إذا كان المجتمع متماثلاً حول الوسط الحسابي $\overline{X}_{mrss}$ هو مقدّر غير متحيز لتقدير متوسط المجتمع، وكذلك فإن

$$var(\overline{X}_{mrss})\le var(\overline{X}_{rss})\le var(\overline{X}_{srs})$$

لمزيد من المعلومات يراجع Muttlak(1997).

2.3.20 معاينة المجموعات المرتبة تطرفياً

لقد اقترح أكثر من باحث استخدام معاينة المجموعات المرتبة تطرفياً لتسهيل عملية ترتيب الوحدات داخل المجموعات، ولتقليل أخطاء ترتيب الوحدات الذي يؤدي إلى تقليل فاعلية معاينة المجموعات المرتبة. يمكننا تلخيص الطريقة فيما يأتي: نقوم بسحب n من المجموعات (العينات) بطريقة عشوائية من المجتمع حجم كل مجموعة n من الوحدات. ثم نقوم بترتيب المجموعات من الأصغر إلى الأكبر، إذا كان حجم المجموعة أو المجموعات عدداً زوجياً، فإننا نقوم بسحب الوحدة ذي الرتبة الأصغر من n/2 مجموعة للقياس أي نصف المجموعات، ومن النصف الثاني تختار الوحدات ذات الرتبة العليا ونقيسها، أما إذا كان حجم المجموعة n عبارة عن عدد فردي، فإننا نختار (n-1)/2 من المجموعات ونقوم بقياس الوحدات ذات الرتبة الصغرى، ومن (n-1)/2 نختار الوحدات ذات الرتبة العليا ونقيسها. وأخيراً نختار إحدى المجموعات ونقيس الوسيط من هذه المجموعة بعد ترتيب وحداتها. في كلتا

الحالتين سنحصل على عينة بحجم n من الوحدات، ويمكننا إعادة الدورة r من المرات للحصول على عينة بحجم nr.

لنفترض أن $X_{(i:e)j}$ القيمة الإحصائية للمتغير بالرتبة الصغرى i من العينة والتي حجمها n في الدورة j حيث $i=1,2,\ldots,L1=n/2$ و $j=1,2,\ldots r$ إذا كان حجم المجموعة أو العينة عدداً زوجياً، كذلك تمثل القيمة الإحصائية للمتغير بالرتبة العليا من العينة التي حجمها n في الدورة j حيث $i = L_1+1, L_1+1,\ldots,n$. أما إذا كان حجم المجموعة عبارة عن عدد فردي فإن i تمثل المرتبة الإحصائية الصغرى لقيمة المتغير من المجموعة بحجم n و $i=1,2,\ldots,L_2=(n-1)/2$ والوسيط من إحدى المجموعات والمرتبة العليا لقيم المتغير من المجموعة بحجم n و $i = L_2+2, L_2+3,\ldots,n$. يمكننا إعادة الدورة r من الدورات للحصول على عينة بحجمnr. المقدّر لتقدير متوسط المجتمع باستخدام معاينة المجموعات المتطرفة هو

$$\bar{X}_{erss}=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{r}X_{(i:e)\,j}$$

أما تباين $\bar{X}_{erss}$ فهو

$$var(\bar{X}_{erss})=\frac{1}{n^2 r}\sum_{i=1}^{n}\sigma^2_{(i:e)}$$

حيث إن

$$\sigma^2_{(i:e)}=E\ [X_{(i:e)}-\ E(X_{(i:e)})]^2$$

هناك أكثر من طريقة لاختيار الوحدات باستخدام معاينة المجموعات المرتبة تطرفياً، كذلك هنالك أكثر من صيغة لمقدّر متوسط المجتمع. يراجع Stkose(1980) و Samawi et al.(1996) و Muttlak and Al-Sabah(2003a) لمزيد من المعلومات.

3.3.20 معاينة المجموعات المرتبة باستخدام المتغير المصاحب

لنفترض أن المتغير الذي نرغب في دراسته X يصعب قياسه وترتيبه أي ترتيب مجموعة من الوحدات حسب قيمة المتغير X ولكن هناك متغير Y مصاحب ومرتبط مع المتغير X يسهل ترتيبه. يمكننا أن نستخدم المتغير Y لتقدير رتب المتغير X على النحو الآتي: نقوم بسحب n من المجموعات حجم كل مجموعة n من الوحدات الثنائية أي تحتوي المتغيرين X و Y. نرتب المجموعة الأولى باستخدام المتغير Y ونقوم بقياس المتغير X المصاحب للرتبة الصغرى للمتغير Y. من المجموعة الثانية نقوم بقياس المتغير X المصاحب للرتبة الثانية للمتغير Y. وهكذا إلى أن نقيس المتغير X المصاحب للرتبة العليا للمتغير Y من المجموعة الأخيرة. يمكننا إعادة الدورة r من المرات لنحصل على عينة بحجم nr من الوحدات. لابد من ملاحظة أن رتب المتغير X ستكون مع أخطاء بالترتيب أي $X_{[i:n]j}$ حيث تمثل القيمة الإحصائية للمتغير برتبة المقدرة i في العينة والتي حجمها n في الدورة j، حيث إن i=1,2,...n و j=1,2,....r.

لنفترض أن (Y,X) تتبع التوزيع الثنائي الطبيعي وعلاقة الانحدار بين X و Y خطية. لذا يمكننا أن نتبع طريقة Stokes(1977) ونكتب

$$X = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(Y - \mu_y) + \varepsilon$$

حيث إن Y و ε مستقلين والمتوسط والتباين للمتغير ε هما على النحو الآتي 0 و $\sigma_x^2(1-\rho^2)$ ، أما ρ فيمثل معامل الارتباط بين X و Y و $\mu_x, \mu_y, \sigma_x, \sigma_y$ يمثلون متوسط والانحراف المعياري للمجتمع للمتغيرين X و Y. لذا فإن المقدّر غير المتحيز لمتوسط المجتمع للمتغير X هو

$$\overline{X}_{rssc} = \frac{1}{nr}\sum_{i=1}^{n}\sum_{j=1}^{r} X_{[i:n]j}$$

أما التباين للمقدّر $\overline{X}_{rssc}$ فيمكن الحصول عليه من Stokes(1977) وهو

$$var(\overline{X}_{rssc}) = \frac{\sigma_x^2}{nr}[(1-\rho^2)+\frac{\rho^2}{n\sigma_y^2}\sum_{i=1}^{n}\sigma_{y(i:n)}^2]$$

حيث إن

$$\sigma_{y(i:n)}^2 = E\ [Y_{(i:n)} - E(Y_{(i:n)})]^2$$

4.20 تقدير الوسط الحسابي باستخدام الانحدار لمعاينة المجموعات المرتبة

لقد حاول أكثر من باحث الجمع بين معاينة المجموعات المرتبة وخط الانحدار نذكر منهم Barreto and Barnett(1999) و Barnett and Moore (1977) و Muttlak(1995,1998) و Patil et al.(1993) و Yu and Lam(1997).

سوف نقتصر في بحثنا هنا على بعض النتائج التي اقترحها Yu and Lam (1997) والتي لها علاقة مباشرة بتقدير الوسط الحسابي للمتغير الذي نرغب في دراسته.

سوف نتبع نفس الطريقة التي اقترحها Stokes (1977) والتي تناولناها أعلاه. ونكتب

$$X = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(Y-\mu_y)+\varepsilon$$

لنفترض أن $Y_{(i:n)j}$ و $X_{[i:n]j}$ تمثل القيمة الإحصائية للمتغير برتبة i في العينة والتي حجمها n في الدورة j للمتغير Y والقيمة الإحصائية للمتغير برتبة المقدرة i في العينة والتي حجمها n في الدورة j للمتغير X، نستطيع أن نكتب

$$X_{[i:n],j} = \mu_x + \frac{\rho\sigma_x}{\sigma_y}(Y_{(i:n),j}-\mu_y)+\varepsilon_{ij}\ ;\quad i=1,2,\ldots n;\ j=1,2,\ldots,r$$

لقد اقترح Yu and Lam (1997) مقدّراً غير متحيز لتقدير متوسط للمتغير X بافتراض أن العلاقة

بين X و Y خطية وهو

$$\overline{X}_{reg} = \overline{X}_{rsse} + \hat{\beta}(\mu_y - \overline{Y}_{rss})$$

حيث إن

$$\hat{\beta} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{r}(Y_{(i:n)j} - \overline{Y}_{rss})(X_{[i:n]j} - \overline{X}_{rsse})}{\sum_{i=1}^{n}\sum_{j=1}^{r}(Y_{(i:n)j} - \overline{Y}_{rss})^2}$$

أما تباين $\overline{X}_{reg}$ فهو

$$var(\overline{X}_{reg}) = \frac{\sigma_y^2}{nr}(1-\rho^2)[1+E(\frac{\overline{Z}_{rss}}{S_z^2})]$$

حيث إن

$$Z_{(i:n)j} = \frac{Y_{(i:n)} - \mu_y}{\sigma_y}$$

و

$$\overline{Z}_{rss} = \frac{1}{nr}\sum_{i=1}^{n}\sum_{j=1}^{r}Z_{(i:n)j}$$

و

$$S_z^2 = \frac{1}{nr}\sum_{i=1}^{n}\sum_{j=1}^{r}(Z_{(i:n)j} - \overline{Z}_{rss})^2 .$$

نلاحظ أن $\overline{X}_{reg}$ دائماً يكون مقدّراً غير متحيز إلى الوسط الحسابي للمجتمع بغض النظر عن التوزيع الذي يتبعه المتغير X ولكن التباين يعتمد على التوزيع الاحتمالي للمتغير X. لمزيد من المعلومات ولمقارنة المقدّر $\overline{X}_{reg}$ ببعض المقدّرات المعروفة يراجع Yu and Lam(1997).

5.20 تقدير التباين باستخدام معاينة المجموعات المرتبة

لم تكن معاينة المجموعات المرتبة فاعلة بتقدير تباين المجتمع σ^2 كما هو الحال إلى تقدير الوسط الحسابي للمجتمع، لقد أثبتت الدراسة التي قامت بها Stokes (1980 b) أن استخدام معاينة المجموعات المرتبة غير فاعلٍ بتقدير تباين المجتمع بالمقارنة إلى المعاينة التقليدية وهي المعاينة العشوائية البسيطة. وكذلك المقدّر الذي يستخدم معاينة المجموعات المرتبة فإنه مقدّر متحيز، على خلاف المقدّر الذي يستخدم العينة العشوائية البسيطة وهو تباين العينة.

لقد أثبتت Stokes (1980 b) أن المقدّر الذي اقترحته والذي يستخدم معاينة المجموعات المرتبة يمكن أن يكون مقدّراً غير متحيز إذا كان حجم العينة كبيراً وبالتحديد إذا كان بالإمكان إعادة الدورة مرات كثيرة. كذلك يكون هذا المقدّر أكثر فاعلية من تباين العينة العشوائية البسيطة المقدّرة ولكن هذا ليس كبيرًا.

لقد اقترح كلٌّ من Abu-Dayyeh and Muttlak (2002) مجموعة من المقدّرات لتقدير تباين المجتمع، باستخدام البيانات من معاينة المجموعات المرتبة إذا كان التوزيع الاحتمالي للمجتمع الذي نرغب في دراسته يكون من عائلة التوزيعات الاحتمالية $f(x, \theta, \lambda) = g((x- \theta)/\lambda)/\lambda$. ولقد برهنوا أن معظم هذه المقدّرات هي غير متحيزة لتقدير تباين المجتمع وأكثر فاعلية من تباين العينة العشوائية البسيطة لمعظم التوزيعات التي دُرست من قبل الباحثين.

References

1. Abu-Dayyeh, W. and Muttlak, H. A. (1996). Using Ranked Set Sampling for Hypothesis Tests on the Scale Parameter for the Exponential and Uniform Distributions, *Pakistan Journal of Statistics*, 12, 131-138.
2. Al-Saleh, M. F. and Muttlak, H. A. (1998). A Note on Bayesian Estimation using Ranked Set Sample, *Pakistan Journal of Statistics*, 14, 49-56.
3. Al-Saleh, M. F. and Al-Kadiri, M. A. (1999). Double Ranked Set Sampling, *Statistics and Probability Letters*, 48, 205-212.
4. Al-Saleh, F. M., Al-Shrafat, K. and Muttlak, H. A. (2000). Bayesian Estimation using Ranked Set Sampling, *Biometrical Journal*, 42, 3, 1-12.
5. Barabesi, L. (1998). The Computation of the Distribution of the Sign Test Statistic for Ranked-Set Sampling, *Commun. Statist. Simula.* 27, 833-842.
6. Barnett, V. and Moore, K. (1997). Best Linear Unbiased Estimates in Ranked-Set Sampling with Particular Reference to Imperfect Ordering, *Journal of Applied Statistics*, 24, 697-710.
7. Barnett, V. (1999). Ranked Set Design for Environmental Investigations, *Environmental and Ecological Statistics*, 6, 59-74.
8. Barreto, M. C. M. and Barnett, V. (1999). Best Linear Unbiased Estimators for the Simple Linear Regression Model using Ranked Set Sampling, *Environmental and Ecological Statistics*, 6, 119-133.
9. Bhoj, D. S. and Ahsanullah, M. (1996). Estimation of Parameters of the Generalized Geometric Distribution using Ranked Set Sampling, *Biometrics*, 52, 685-694.
10. Bhoj, D. S. (1997a). Estimation of Parameters of the Extreme Value Distribution using Ranked Set Sampling, *Commun. Statist. Theory Meth.*, 26, 653-667.
11. Bhoj, D. S. (1997b). New Parametric Ranked Set Sampling, *Applied Statistical Sciences*, 6), 275-289.
12. Bhoj, D. S. (1997c). Estimation of Parameters using Modified Ranked Set Sampling, *Applied Statistical Sciences*, 6, 145-163.
13. Bhoj, D. S. (1999a). Estimation of Parameters of the Exponential Distribution using Three Ranked Set Sampling Procedures, *Applied Statistical Sciences*, 8, 155-172.
14. Bhoj, D. S. (1999b). Minimum Variance Linear Unbiased Estimators of the Rayleigh Parameter based on Ranked Set Sampling Procedures, *Applied Statistical Sciences*, 8, 269-277.
15. Bohn, L. L. and Wolfe, D. A. (1992). Nonparametric Two-Sample Procedures for Ranked Set Samples Data, *Journal of American Statistical Association*, 87, 552-561.
16. Bohn, L. L. and Wolfe, D. A. (1992). The Test of Imperfect Rankings on Properties of Procedures based on the Ranked-Set Samples Along with Mann-Whiteny-Wlicoxon Statistics, *Journal of American Statistical Association*, 89, 168-176.
17. Bohn, L. L. (1996). A Review of Nonparametric Ranked-Set Sampling Methodology, *Commun. Statist. Theory Meth.* 25, 2675-2685.
18. Bohn, L. L. (1998). A Ranked-Set Sample Signed-Ranked Statistic, *J Nonparametric Statistics*, 9, 295-306.
19. Chen, Z (1999). Density Estimation using Ranked-Set Sampling Data, *Environmental and Ecological Statistics*, 6, 135-146.
20. Chen, Z (2000a). On Ranked Set Sampling Quantiles and their Applications, *Journal of Statistical Planning and Inference*, 83, 125-135.
21. Chen, Z (2000b). The Efficiency of Ranked-Set Sampling Relative to Simple Random Sampling under Multi-Parameter Families, *Statistica Sinica*, 10, 125-135.
22. Chen, Z (2001). Ranked-Set Sampling with Regression Type Estimators, *Journal of Statistical Planning and Inference*, 92, 181-192.
23. David, H. A. and Levine, D. N. (1972). Ranked Set Sampling in the Present of the Judgment Error, *Biometrics*, 28, 553-55
24. Dell, D. R. and Clutter, J. L. (1972). Ranked Set Sampling Theory with Order Statistics Background, *Biometrics*, 28, 545-53.

25. Fei, H., Sinha, B. K. and Wu, Z. (1994). Estimation of Parameters in Two-Parameter Weibull and Extreme-Value Distributions using Ranked Set Sampling, *Journal of Statistical Research*, 28, 149-161.
26. Halls, L.S. and Dell, T. R. (1966). Trial of Rank Set Sampling for Forage Yields, *Forest Science*, 12, 22-26.
27. Hettmansperger, T. (1995). The Ranked-set Sample Sign Test, *Nonparametric Statistics,* 4, 263-270.
28. Hossain, A. S. and Muttlak, H. A. (1999). Paired Ranked Set Sampling: A More Efficient Procedure, *Enviornmetrics*, 10, 195-212.
29. Johnson, G. D., Patil, G. P. and Sinha, A. K. (1993). Ranked Set Sampling for Vegetation Research, *Abstracta Botanica*, 17, 87-102.
30. Kaur, A., Patil, G. P., Sinha, B. K. and Taillie, C. (1995). Ranked Set Sampling: an Annotated Bibliography, *Environmental and Ecological Statistics*, 2, 25-54.
31. Kaur, A, Patil, G. P. and Taillie, C. (1997). Unequal Allocation Models for Ranked Set Sampling with Skew Distributions, *Biometrics*, 53, 123-130.
32. Kim, Y. H and Arnold, B. C. (1999). Parameter Estimation under Generalized Ranked Set Sampling, *Statistics & Probability Letters*, 42, 353-360.
33. Kvam, P. H. and Samaniego, F. J. (1994). Nonparametric Maximum Likelihood Estimation Based on Ranked Set Samples, *Journal of American Statistical Association*, 89, 526-537.
34. Lam, K., Sinha, B. K. and Wu, Z. (1994). Estimation of Parameters in Two-Parameter Exponential Distribution using Ranked Set Sample, *Annals of the Institute of Statistical Mathematics,* 46, 723-736.
35. Lam, K., Sinha, B. K. and Wu, Z. (1995). Estimation of Location and Scale parameters of a Logistic Distribution using Ranked Set Sample. In: Nagaraja, Sen and Morrison, ed., *Papers in Honor of Herbert A. David*, 187-197.
36. Li, D. Sinha, B. K. and Perron, F. (1999). Random selection in Ranked Set Sampling and its Applications, *Journal of Statistical Planning and Inference*, 76, 185-201.
37. McIntyre, G. A. (1952). A Method of Unbiased Selective Sampling, Using Ranked sets, *Australian Journal of Agricultural Research*, 3, 385-390.
38. Muttlak, H. A. and McDonald, L. L. (1990a). Ranked Set Sampling with Respect to Concomitant Variables and with Size Biased Probability of Selection, *Communication in Statistics Theory and Methods*, 19, 205-219.
39. Muttlak, H. A. and McDonald, L. L. (1990b). Ranked Set Sampling with Size Biased Probability of Selection, *Biometrics*, 46, 435-445.
40. Muttlak, H. A. and McDonald, L. L. (1992). Ranked Set Sampling and Line Intercept Method: A More Efficient Procedure, *The Biometrical Journal*, 34, 329-346.
41. Muttlak, H. A. (1995). Parameters Estimation in a Simple Linear Regression using Ranked Set Sampling, *Biometrical Journal*, 37, 799-810.
42. Muttlak, H. A. (1996a). Estimation of Parameters for One-Way Layout with Ranked Set Sampling, *Biometrical Journal*, 38, 507-515.
43. Muttlak, H. A. (1996b). Estimation of Parameters in a Multiple Regression Model Using Rank Set Sampling, *Information & Optimization Sciences*, 17, 521-533.
44. Muttlak, H. A. (1996c). Pair Ranked Set Sampling, *Biometrical Journal*, 38, 897-885.
45. Muttlak, H. A. (1997). Median Ranked Set Sampling, *Journal of Applied Statistical Science*, 6, 245-55.
46. Muttlak, H. A. (1998). Median Ranked Set Sampling with Concomitant Variables and Comparison with Ranked Set Sampling and Regression Estimators, *Enviornmetrics*, 9, 255-267.
47. Muttlak, H. A and Abu-Dayyeh, W. (1998). Testing Some Hypotheses about the Normal Distribution Using Ranked Set Sampling: A More Powerful Test, *Journal of Information & Optimization Sciences*, 19, 1-11.
48. Muttlak, H. A. (1999a). Median Ranked Set Sampling with Size Biased Probability of Selection, *Biometrical Journal*, 40, 455-465.
49. Muttlak, H. A. (1999b). On Extreme Ranked Set Sampling with Size Biased Probability of Selection. *Far East Journal Theory of Statistics*, 3, 319-329.

50. Muttlak, H. A. and Al-Saleh, M. F. (2000). Recent Developments on Ranked Set Sampling, *Pakistan Journal of Statistics*, 16, 269-290.
51. Abu-Dayyeh, W. A. and Muttlak, H. A. (2002). Variance Estimation for the Locatio Scale Family Distributions using Ranked Set Sampling
52. , *Pakistan Journal of Statistics*, 18.
53. Muttlak, H. A. and Al-Sabah, W. (2003a). Statistical Quality Control using Ranked Set Sampling, *Journal of Applied Statistics*, 30, 1055-1078.
54. Muttlak, H. A. Al-Sabah, W. (2003b). Statistical Quality Control based on Pair and Selected Ranked Set Sampling, *Pakistan Journal of Statistics*, 19, 107-128.
55. Ni Chuiv, N. N. and Sinha, B. K. (1998). On Some Aspects of Ranked Set Sampling in Parametric Estimation, Handbook of statistics, 17, 337-377.
56. Özturk, Ö. (1999). One-and Two-Sample Sign Tests for Ranked Set Sample Selective Designs, *Commun. Statist. Theory Meth.* 28, 1231-1245.
57. Özturk, Ö. (2002). Rank Regression in Ranked-Set Samples, *Journal of American Statistical Association*, 97, 1180-1191.
58. Özturk, Ö. and Wolfe, D. A (2000). Alternative Ranked Set Sampling Protocols for Sign Test, *Statistics & Probability Letters*, 47, 15-23.
59. Patil, G. P., Sinha, A. K. and Taillie, C. (1993). Relative Precision of Ranked set Sampling: A Comparison with the Regression Estimator, *Enviornmetrics*, 4, 399-412.
60. Patil, G. P., Sinha, A. K. and Taillie, C. (1994). Ranked Set Sampling, *A Handbook of Statistics*, 12,167-200.
61. Patil, G. P., Sinha, A. K. and Taillie, C. (1995). Finite Population Corrections for Ranked Set Sampling, *Annals of the Institute of Statistical Mathematics*, 47, 621-636.
62. Patil, G. P., Sinha, A. K. and Taillie, C. (1999). Ranked Set Sampling: a Bibliography, *Environmental and Ecological Statistics*, 6, 91-98.
63. Salazar, R. D. and Sinha, A. K. (1997). Control chart X bar based on ranked set sampling. Comunicacion Tecica No. 1-97-09(PE/CIMAT), Department of Probability and Statistics, Centro de Investigation en Matematicas (CIMAT), Apdo. Postal. Gto. 36000, Mexico.
64. Samawi, H. M., Ahmed, M. S. and Abu-Dayyeh. Estimating the Population Mean using Extreme Raked Set Sampling, *The Biometrical Journal*, 38, 577-86.
65. Samawi, H. M. and Muttlak, H. A. (1996). Estimation of Ratio Using Ranked Set Sampling, *Biometrical Journal*, 38, 753-764.
66. Shen, W. H. and Yuan, W. (1996). A Test for a Normal Mean Based on a Modified Partial Ranked Set Sampling, *Pakistan Journal of statistics*, 11, 227-233.
67. Shen, W. H. (1994). Use the Ranked Set Sampling for Testing a Normal Mean, *Calcutta Statistical Association Bulletin*, 44, 183-193.
68. Sinha, B. K., Wu, Z. and Fei, H. (1994). Estimation of a Gamma Mean Based on Ranked Set Sample, *Pakistan Journal of Statistics*, 10, 235-249.
69. Sinha, B. K., A. K. Sinha, and Purkayastha, S. (1996). On some Aspects of Ranked Set Sampling for Estimation of Normal and Exponential Parameters, *Statistics and Decisions*, 14, 223-240.
70. Stokes, S. L. (1977). Ranked Set Sampling with Concomitant Variables, *Commun. Statist. Theory Meth.* A6, 1207-1211.
71. Stokes, S. L. (1980a). Inference on the Correlation coefficient in Bivariate Normal Distribution from Raked Set Samples, *Journal of American Statistical Association*, 75, 989-995.
72. Stokes, S. L. (1980b). Estimation of Variance Using Judgment Ordered Ranked Set Samples, *Biometrics*, 36, 35-42.
73. Stokes, S. L. and Sager, T. W. (1988). Characterization of a Ranked-Set Sample with Application to Estimating Distribution Function, *Journal of American Statistical Association*, 83, 374-381.
74. Stokes, S. L. (1995). Parametric Ranked Set Sampling, *Annals of the Institute of Statistical Mathematics*, 47, 465-482.

75. Takahasi K. and Wakimoto K. (1968). On Unbiased Estimates of the Population Mean Based on the Sample Stratified by Means of Ordering, *Annals of the Institute of Statistical Mathematics*, 21, 249-55.
76. Tam, C. Y. C., Yu, P. L. H. and Fung, T. W. K. (1998). Sensitivity Analysis of BLUE for the Population Mean based on a Ranked Set Sample, *Commun. Statist. Theory Meth.*, 27, 1075-1091.
77. Yu, P. L. H. and Lam, K. (1997). Regression Estimator in Ranked Set Sampling, *Biometrics*, 53, 1070-1080.
78. Yu, P. L. H., Lam, K. and Sinha, B. K. (1997). Estimation of Mean Based on Unbalanced Ranked Set Sample, Applied Statistical Science **II**, *87-97*
79. Yu, P. L. H., Lam, K. and Sinha, B. K. (1999). Estimation of Normal Variance Based on Balanced and Unbalanced Ranked Set Samples, *Environmental and Ecological Statistics*, 6, 23-4

الفصل الحادي والعشرون

طريقة السكين الحادة لإعادة المعاينة

Jackknife Resampling Method

1.21 مقدمة

إن أول من صمم طريقة السكين الحادة (Jackknife) هو Quenouille (1949,1956) لتقليل التحيز للمقدّر، ومن المميزات الجذابة لهذه الطريقة أنه يمكن استخدامها للحالات المعقدة حيث يستحيل تطبيق النماذج المعلماتية أو المعالجات النظرية.

2.21 الطريقة العامة

لقد أعطى Quenouille (1949) وصفاً لتقنية تقليل التحيز لمقدّر الترابط المتسلسل، الذي يعتمد على تقسيم العينة إلى مجموعتين. ثم قام في بحثه المنشور في عام 1956 بتعميم هذه الطريقة على الوجه الآتي: نقوم بتقسيم العينة g من المجاميع المتساوية الحجم وليكن h، أي إن حجم العينة $n=gh$. لنفرض أن $X_1, X_2, \ldots, X_n$ عبارة عن عينة عشوائية، ولنفترض أن $\hat{\theta}$ عبارة عن المقدّر للمعلمة غير المعلومة θ الذي يعتمد على العينة بحجم n. لنفرض $\hat{\theta}_{-i}$ عبارة عن تقدير θ بالاعتماد على عينة بحجم $(g-1)h$، حيث تم حذف المجموعة i والتي حجمها h.

لنفترض

$$\tilde{\theta}_i = g\hat{\theta} - (g-1)\hat{\theta}_{-i} \; ; (i = 1, ..., g)$$

لذا فإن المقدّر الآتي يكون تحيزه بالرتبة أو المرتبة n^{-2}

$$\tilde{\theta} = \frac{1}{g}\sum_{i=1}^{g}\tilde{\theta}_i = g\hat{\theta} - \frac{g-1}{g}\sum_{i=1}^{g}\hat{\theta}_{-i}$$

لقد سمى Tukey (1958) المقدّر $\tilde{\theta}$ بالسكين الحادة (jackknife)، التي يمكن أن تكون أداة إحصائية قاسية وجاهزة للاستعمال عند الحاجة.

قام Miller(1974) بمسح جميع النتائج المنشورة حول طريقة السكين الحادة، وهذا البحث يُعَدُّ أساسًا لما تبعه من بحوث في هذا المجال، حيث إن كثيراً من البحوث اللاحقة في هذا المجال تفترض أن g=n و h=1.

لقد اقترح Quenouille(1956) مقدّراً آخر يمكن استخدامه لتقديرات حد $O(n^{-2})$ من التحيز، يسمى مقدّر الدرجة الثانية، وهو

$$\tilde{\theta}^{(2)} = \frac{n^2\tilde{\theta} - (n-1)^2\sum_{j=1}^{n}\tilde{\theta}_{-j}/n}{n^2-(n-1)^2}$$

حيث إن $\tilde{\theta}_{-j}$ هو عبارة $\tilde{\theta}_i$ ولكن بحجم عينة هو (n-1) حيث تم حذف الوحدة j. ولكن إذا جرى كتابة المقدّر $\tilde{\theta}^{(2)}$ بدلالة المقدّر الأصلي $\hat{\theta}$ فسنحصل على

$$\tilde{\theta}^{(2)} = (2n-1)^{-1}[n^3\hat{\theta} - (2n^2-2n+1)(n-1)(\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{-i})$$

$$+ (n-1)^2(n-2)\{\frac{2}{n(n-1)}\sum_{i<j}^{n}\hat{\theta}_{-ij}\}]$$

حيث إن $\hat{\theta}_{ij}$ عبارة عن المقدّر الأصلي عندما يكون حجم العينة $(n-2)$ حيث جرى حذف المشاهدتين i وj. لذا فإن

$$E(\tilde{\theta}^{(2)} = q + O(n^{-3})$$

لقد اقترح كل من Gray and Owen (1971) تعديلاً للوزن لكي نحصل على مقدّر غير متحيز عندما يكون التحيز يحتوي على الدرجة الأولى والثانية بدلالة 1/n أي

$$E(\hat{\theta}) = q + \frac{a_1}{n} + \frac{a_2}{n^2}$$

لذا فإن المقدّر الذي اقترحاه هو

$$\tilde{\theta}^{(2)^*} = \frac{1}{2}[n^2\hat{\theta} - 2(n-1)^2(\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{-i}) + (n-2)^2\{\frac{2}{n(n-1)}\sum_{i<j}\hat{\theta}_{-ij}\}]$$

مثال: لنفترض أن θ تمثل متوسط المجتمع و $\hat{\theta} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. حيث إن h=1 و n=g لذا فإن

$$\hat{\theta}_{-i} = (n\hat{\theta} - x_i)/(n-1)$$

لنفترض أن

$$\hat{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{-i}$$

لذا فإن

$$\hat{\theta}_{(.)} = \hat{\theta}, \quad \hat{\theta}_{-i} - \hat{\theta}_{(.)} = (\overline{x} - x_i)/(n-1)$$

لذا فإن مقدّر التباين $\hat{\theta}_{(.)} = \hat{\theta}$ هو

$$s^2_{\hat{\theta}_{(.)}} = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{-i} - \hat{\theta}_{(.)})^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

وهو نفس المقدّر لتباين متوسط العينة $\overline{x}$.

3.21 التطبيقات الأساسية

يمكن استخدام طريقة السكين الحادة بشكل مثمر في تقديرات النسبة والانحدار . للعينة ثنائية المتغيرات $(Y_i, X_i), i = 1, 2, ..., n$ ، حيث إن $E(X_i) = \mu_x$ و $E(Y_i) = \mu_y$ ، ونرغب في تقدير $\theta = \mu_y/\mu_x$. في مسوح العينات نفترض أن متوسط المجتمع μ_x للمتغير الإضافي أو الاحتياطي يكون معلوماً، لذا فإن تقدير μ_y هو $\hat{\mu}_y = \hat{\theta}\mu_x$ ، حيث إن $\hat{\theta} = \overline{Y}/\overline{X}$ وهو ما يسمى تقدير النسبة باستخدام عينة بحجم n. وما هو متعارف عليه أن $\hat{\mu}_y$ أكثر دقة في تقدير μ_y من $\overline{Y}$.

قام Durbin (1959) بتقدير β باستخدام النسبة من خلال طريقة السكين الحادة بالنموذج

$$Y_i = \alpha + \beta X_i + e_i$$

حيث إن e_i تتوزع بشكل مستقل ومتماثل وتتبع التوزيع الطبيعي أو توزيع جاما، ولقد قام بدراسة مقدّر السكين الحادة β باستخدام العلاقة

$$\tilde{\theta} = \frac{1}{g}\sum_{i=1}^{g}\tilde{\theta}_i = g\hat{\theta} - \frac{g-1}{g}\sum_{i=1}^{g}\hat{\theta}_{-i}$$

مع g=2. لقد أثبت أن التحيز بالرتبة أو المرتبة n^{-4} ويمكن إهماله؛ لذا فإن مقدّر السكين الحادة لديه تحيز أقل وتباين أقل من $\hat{\theta} = \overline{Y}/\overline{X}$ إذا كانت e_i تتبع التوزيع الطبيعي، أما إذا كانت e_i تتبع توزيع جاما ومع معامل تغير أقل من 1/4 ، فإن مقدّر السكين الحادة سوف يقلل التحيز، ويزيد التباين قليلاً، لذا سيؤدي إلى خفض متوسط مربعات الخطأ عند مقارنته بالمقدّر $\hat{\theta} = \overline{Y}/\overline{X}$. ولقد قام كل من Rao(1965) و Rao and Webster(1966) بإعطاء الاختيار الأمثل g و n للتوزيعين الطبيعي وجاما، ولابد من الأخذ في الحسبان أن هناك مقدّرات بديلة بالإضافة إلى مقدّر السكين الحادة، وأن مقدّر السكين

الحادة لا يكون دائماً الأفضل، ولكنه ليس متخلفاً كثيراً عن المقدّر الأمثل، وأخيراً فقد قدم Brillinger (1966) تطبيقات متعددة لاستخدام طريقة السكين الحادة في مسوحات المعاينة المختلفة.

4.21 التقدير بفترة

اقترح Tukey (1958) أن قيم المجموعات التي عددها g وهي $\tilde{\theta}_i ; i = 1, ..., g$ تكون تقريبا متغيرات مستقلة عن بعضها الآخر وتتبع التوزيع نفسه في حالات كثيرة؛ لذا فإن

$$\sqrt{g}(\tilde{\theta}-\theta)/s_{\tilde{\theta}}$$

حيث إن

$$(g-1)s_{\tilde{\theta}}^2 = \sum_{i=1}^{g}(\tilde{\theta}_i-\tilde{\theta})^2$$

تتبع توزيع t التقريبي مع درجات حرية (g-1). يمكن استخدام هذا الإحصاء لإيجاد فترة ثقة θ. ولقد تم إثبات ما ذهب إليه Tukey من أن $\sqrt{g}(\tilde{\theta}-\theta)/s_{\tilde{\theta}}$ تتبع توزيع t أو التوزيع الطبيعي إذا كانت قيمة g كبيرة.

5.21 التحويل

لقد اقترح الباحثون الذين يدافعون عن طريقة السكين الحادة بأن نقوم بمحاولة تثبيت التباين وذلك بالقيام بتحويل المقدّر قبل تطبيق طريقة السكين الحادة. على سبيل المثال يكون أفضل استخدام طريقة السكين الحادة على $\log(s^2)$ و $\tanh^{-1}(r)$ بدلاً من s^2 و r حيث إن r يمثل معامل ارتباط العينة. في بعض الأحيان يكون التحويل واجباً لتفادي تشويه أو تحريف النتائج. على سبيل المثال إذا كنا نرغب في تقدير σ^2 ، فإنه من دون التحويل ستكون بعض القيم $ns^2-(n-1)s_{-i}^2$ سالبة، وبما أن طريقة السكين الحادة لا ترى القيم

السالبة، لذا لابد من التحويل باستخدام $\log(s^2)$ للتخلص من هذه المشكلة. لمزيد من المعلومات يراجع Govindarajulu(1999).

6.21 التحيز بتقدير التباين

لقد بحث كل من Efron and Stein (1981) تقدير التباين $S(X_1, ..., X_n)$ وهي عبارة عن دالة متماثلة للمتغيرات (X_i) العشوائية المستقلة والمتماثلة. لقد برهنا على أن طريقة السكين الحادة لتقدير التباين دائماً تعطينا تقديرات متحيزة نحو الأعلى.

لقد قدم لنا كل من Quenouille and Tukey تقديرات مفيدة للتباين والتحيز ذات الطبيعة غير المعلماتية، نفترض أن $S(X_1, ..., X_n)$ عبارة عن مقدّر مصمم لتقدير معلمة ما نرغب في دراستها، نفترض أن S متماثلة $X_i, i = 1, ..., n$. لنفترض

$$S_{(i)} = (X_1, X_2, ..., X_{i-1}, X_{i+1}, ..., X_n)$$

وهي عبارة عن قيمة S تم حسابها بعد حذف X_i. لذا فإن تقدير التباين $var[S(X_1, ..., X_n)]$ باستخدام طريقة السكين الحادة هو

$$s^2(S(X_1, ..., X_n)) = \frac{n-1}{n}\sum_{i=1}^{n}(S_{(i)} - \bar{S})^2$$

حيث إن

$$\bar{S} = \frac{1}{n}\sum_{i=1}^{n} S_{(i)}$$

نلاحظ أن هذا التقدير هو عبارة عن تقدير التباين باستخدام طريقة السكين الحادة S معطى بالشكل $nS-(n-1)\bar{S}$، مع ذلك يمكن أن يكون تقدير التباين S نفسها أو $S_{(i)}$. لمزيد من المعلومات يراجع Efron and Stein (1981) و Govindaragulu(1999).

References

1. Brillinger, D. R. (1966). The Application of Jackknife to the Analysis of Sample Surveys, *Commentary*, 8, 74-80.
2. Booth, J. G. and Hall, P. (1993). An Improvement of the Jackknife Distribution Function Estimator, *Ann. Statist.*, 21, 1476-1485.
3. Durbin, J. (1959). A Note in the Application of Quenouille's Method of Bias Reduction to the Estimation of Ratios, *Biometrika* 46, 477-480.
4. Eforn, B. (1982). The Jackknife, Bootstrap and other Resampling Plans, Siam Publication No. 38, Philadelphia, PA.
5. Eforn, B. (1992). Jackknife-after-Bootstrap Standard Errors and Influence Functions (with Discussions), *J. R. Statist. Soc.*, B54, 83-127.
6. Eforn, B. and Stein, C. (1981). The Jackknife Estimate of Variance, *Ann. Statist.*, 9, 586-596.
7. Frangos, C. C. and Schucany, W. R. (1990). Jackknife Estimation of the Bootstrap Acceleration Constant, *Compu. Statist. Data Anal.*, 9, 271-282.
8. Good, P. H. (2001). Resampling Methods: A Practical Guide to Data Analysis, 2end., Birkhauser, Boston.
9. Govindarajulu, Z. (1999). Elements of Sampling Theory and Methods, Prentice Hall.
10. Gray, H. L. Watkins, T. A. Adams, J. E. (1972). On the Jackknife Statistic, its Extensions, and its Relations to e_n-Transformations, *Ann. Math. Statist.* 43, 1-30.
11. Kunsch, H. R. (1989). The Jackknife and Bootstrap for General Stationary Observations. *Ann. Statist*, 17, 1217-1241.
12. Miller, R. G. (1964). A Trustworthy Jackknife, *Ann. Math. Statist.* 35, 1594-1605.
13. Miller, R. G. (1974). The Jackknife- a Review, *Biometrika* 61, 1-15.
14. Nagao, H. (1988). On the Jackknife Statistics for Eigenvalues and Eigenvectors of a Correlation Matrix, *Ann. Inst. Statist. Math.*, 40, 477-489.
15. Parr. W. C. (1983). A Note on the Jackknife, Bootstrap, and the Delta Method Estimates of Bias and Variance, *Biometrika* 70, 719-722.
16. Parr. W. C. (1985). Jackknifing Differentiable Statistical Functionals, *J. R. Statist. Soc.*, B47, 56-66.
17. Parr. W. C. and Schucany, W. R. (1982). Jackknifing L-Statistics with Smooth Weight Functions, *J. Amer. Statist. Assoc.* **77**, 629-638.
18. Quenuouille, M. H. (1949). Approximate Test of Correlation in Time Series, *J. R. Statist. Soc.*, B11, 68-84.
19. Quenuouille, M. H. (1956). Notes on Bias in Estimation, *Biometrika* 52, 647-649.
20. Rao, J. N. K. (1965). A Note on the Estimation of Ratios by Quenouille's Method, *Biometrika* 52, 647-649.
21. Rao, J. N. K. and Shao, J. (1992). Jackknife Variance Estimation with Survey Data under Hot Deck Imputation, *Biometrika* 79, 811-822.
22. Reeds, J. A. (1978). Jackknifing the Maximum Likelihood Estimates, *Ann. Statist.*, 6, 727-739.
23. Schucany, W. R. and Sheather, S. J. (1989). Jackknifing R-Estimators, *Biometrika* 76, 393-398.
24. Rao, J. N. K. and Webster, J. (1966). On Two Methods of Bias Reduction in the Estimation of Ratios, *Biometrika* 53, 571-577.
25. Shao, J. (1988). Consistency of Jackknife Estimators of the Variance of Sample Quantiles, *Comm. Statist. A*, 17, 3017-3028.
26. Shao, J. (1989). Jackknifing Weighted Least Square Estimators, *J. R. Statist. Soc.*, B51, 139-156.
27. Shao, J. (1992a). One-Step Jackknife for M-Estimators Computed using Newton's Method, *Ann. Inst. Statist. Math.*, 44, 687-701.

28. Shao, J. (1992b). Jackknifing Generalized Linear Models, *Ann. Inst. Statist. Math.*, 44, 673-686.
29. Shao, J. and Tu, D. (1995). The Jackknife and the Bootstrap, Springer, New York.
30. Simonoff, J. S. and Tasi, C. (1988). Jackknifing and Bootstrapping Quasi-Likelihood Estimators, *J. Statist. Compu. Simul.*, 30, 213-232.
31. Stute, W. and Wang, J. (1994). The Jackknife Estimate of a Kaplan-Meire Integral, *Biometrika* 81, 602-606.
32. Tukey, J. W. (1958). Bias and Confidence in Not-Quite Large Samples (Abstract), *Ann. Math. Statist.* 29, 614.
33. Wu, C. F. (1986). Jackknife, Bootstrap and other Resampling Methods in Regression Analysis (with Discussions), *Ann. Statist.*, 14, 1261-1350.

الفصل الثاني والعشرون

طريقة البوتستراب لإعادة المعاينة

Bootstrap Resampling Method

1.22 مقدمة

إن من أهم مغريات استخدام طريقة السكين الحادة والبوتستراب أنه يمكن استخدامهما عندما تكون الحالات التي نرغب في دراستها معقدة ولا تتوافر النماذج المعلماتية ولا المعالجات النظرية. يُعَدُّ Eforn (1979) الرائد لطريقة البوتستراب، ومن ثم توسيع الطريقة لحالات عديدة، على سبيل المثال تقدير وسيط المجتمع حيث أثبت أن البوتستراب أفضل من طريقة السكين الحادة، لمزيد من المعلومات يراجع Eforn and Tibshirani(1993).

2.22 طريقة البوتستراب (Bootstrap Method)

لنفترض أن $X_1, X_2, ..., X_n$ عبارة عن عينة عشوائية من مجتمع لديه f(x) و f(x) دالّتا التوزيع الاحتمالي والكثافة الاحتمالية على التوالي. نفترض أن θ المعلمة التي نرغب في دراستها، ولنفترض أن $\hat{\theta}(X_1, ..., X_n)$ هو تقدير إلى θ والتي تكون متماثلة في $X_1, X_2, ..., X_n$، نستطيع أن نكتب الانحراف المعياري إلى $\hat{\theta}$ كما يأتي

$$\sigma_{\hat{\theta}} = \sigma(F, n, \hat{\theta}) = \sigma(F)$$

إن استخدام $\sigma(F)$ في العلاقة أعلاه للتأكيد على أن الانحراف المعياري عبارة عن دالة بدلالة التوزيع الاحتمالي غير المعروف F. إن تقدير البوتستراب إلى الانحراف المعياري يمكن الحصول عليه باستبدال F غير المعروفة بـ $\hat{F}$ وهي عبارة عن تقدير الاحتمالية العظمى إلى F لنحصل على

$$\hat{\sigma}_{\hat{\theta}} = \sigma(\hat{F})$$

لتوضح طريقة البوتستراب لنفترض أننا سحبنا عينة عشوائية بحجم n، متوسط العينة $\bar{x}$ يُعَدُّ تقديراً لمتوسط المجتمع μ. الانحراف المعياري إلى $\bar{x}$ هو

$$(\mu_2/n)^{1/2}$$

حيث إن

$$\mu = E(X) \text{ و } \mu_2 = E(X-\mu)^2$$

لذا فإن

$$\hat{\sigma}_{\bar{x}} = (\hat{\mu}_2/n)^{1/2}$$

حيث إن

$$\hat{\mu}_2 = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

حيث إن $x_1, ..., x_n$ و $\bar{x}$ يمثلون قيم المشاهدات ومتوسطها للمتغيرات $X_1, ..., X_n$ و $\bar{X}$ على التوالي. بما أن $\hat{\mu}_2$ متحيز نحو الأسفل أي $E(\hat{\mu}_2) < \mu_2$ ، يمكننا ضرب μ_2 بـ $n(n-1)$ لنحصل على

$$SD = \hat{\sigma}_{\bar{x}} = \{(n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\}^{1/2}$$

في العادة $\sigma(F)$ لا يوجد لها صيغة أو علاقة محددة أو صريحة؛ لذا من أجل حساب قيمة SD لابد من اتباع الخطوات الآتية:

1- لنفترض أن $\hat{F}$ عبارة عن تقدير الاحتمالية العظمى إلى F أي تعطي احتمال 1/n لكل مشاهدة x_i.

2- نقوم بسحب عينة البوتستراب من $\hat{F}$ وبالتحديد $X^*_1, ..., X^*_n$ والتي تتوزع بصورة متماثلة كما $\hat{F}$ ومن ثم نحسب $\hat{\theta}^* = \hat{\theta}(X^*_1, ..., X^*_n)$.

3- نقوم بإعادة الخطوة 2 عدة مرات وبالتحديد B مرة (وتكون قيمة B كبيرة) لنجد قيم البوتستراب $\hat{\theta}^*_1, ..., \hat{\theta}^*_B$ ومن ثم نحسب

$$SD = \{\sum_{j=1}^{B}(\hat{\theta}^*_j - \bar{\hat{\theta}}^*)^2/(B-1)\}^{1/2}$$

حيث إن

$$\bar{\hat{\theta}}^* = \sum_{j=1}^{B}\hat{\theta}^*_j/B$$

إذا افترضنا أن $B \to \infty$ فإننا سنحصل على

$$SD = \{\sum_{j=1}^{B}(\hat{\theta}^*_j - \bar{\hat{\theta}}^*)^2/(B-1)\}^{1/2} = \sigma_{\hat{\theta}}$$

أي أن القيمة التقديرية باستخدام طريقة البوتستراب ستساوي القيمة التقديرية فيما لو كنا نعرف F. ولكن في الحياة العملية نريد أن تكون قيمة B محدودة بسبب تكاليف الحسابات، بالاعتماد على بعض الحسابات التي قام بها Efron (1982) فإن B=100 تعطينا الدقة نفسها بالتقدير فيما لو كانت قيم B تساوي 200 أو512 أو1000.

مثال 1: سحبنا عينة عشوائية بحجم 5 من مواليد الماعز ووجدنا أوزانها (1,2,3,4,5) كلغم.

أولاً نحسب

$$\overline{x} = (1+2+3+4+5)/5 = 3$$

و

$$s^2 = \frac{1}{4}\sum_{i=1}^{5}(x_i - 3)^2 = 2.5$$

لذا فإن الانحراف المعياري إلى $\overline{X}$ هو

$$\hat{\sigma}_{\overline{x}} = \frac{s}{\sqrt{n}} = (2.5/5)^{1/2} = 0.707$$

الآن لنطبق طريقة البوتستراب، لنسحب B=10 عينة بوتستراب من(1,2,3,4,5) وهي عبارة عن عينة عشوائية مع الإرجاع باستخدام الكمبيوتر أو جداول الأعداد العشوائية، حصلنا على العينات الآتية:

(1, 2, 2, 3, 5) و (3, 4, 5, 5, 2) و (1, 3, 4, 1, 2) و (5, 3, 1, 4, 2) و (5, 4, 4, 2, 5)
و (3, 3, 4, 1, 5) و (4, 1, 2, 2, 5) و (4, 4, 2, 4, 2) و (2, 5, 3, 2, 4) و (5, 4, 1, 5, 5)

والتي تعطينا

$\hat{\theta}_1^* = 2.6$ و $\hat{\theta}_2^* = 3.8$ و $\hat{\theta}_3^* = 2.2$ و $\hat{\theta}_4^* = 3$ و $\hat{\theta}_5^* = 4$ و $\hat{\theta}_6^* = 3.2$ و $\hat{\theta}_7^* = 2.8$
و $\hat{\theta}_8^* = 3.2$ و $\hat{\theta}_9^* = 3.2$ و $\hat{\theta}_{10}^* = 4$ وأخيراً $\overline{\hat{\theta}}^* = 3.2$.

و

$$\sum_{j=1}^{10}(\hat{\theta}_j^* - \overline{\hat{\theta}}^*)^2 = 3.2$$

لذا فإن تقدير الانحراف المعياري باستخدام طريقة البوتستراب هو

$$SD = (3.2/9)^2 = 0.596$$

نلاحظ أن قيمة SD = 0.596 ليست قريبة من $\hat{\sigma}_{\bar{x}}$ = 0.707 وهذا يعود لكون قيمة B=10 صغيرة جداً.

3.22 طرق البوتستراب للمشكلات العامة

لنفترض أن $\underline{X} = (X_1, ..., X_n)$ و $R(\underline{X},F)$ عبارة عن دالة إلى $\underline{X}$ ونرغب في دراستها، لنفترض أننا نرغب في تقدير بعض صفات التوزيع الاحتمالي إلى R مثل $E_F(R)$ أو $P_F(R<a)$ لقيمة محددة إلى a. ومن ثم نتبع الخطوات الثلاث السابقة نفسها، ولكن في الخطوة الثانية نحسب

$$R^* = R(\underline{X}^*, \hat{F})$$

بدلاً من $\hat{\theta}^*$. وفي الخطوة الثالثة نحسب صفات R التي نحن بصدد دراستها، على سبيل المثال إذا كنا نرغب في تقدير $E_F(R)$ نحسب

$$E_*(R^*) = B^{-1}\sum_{j=1}^{B} R_j^*$$

و إذا كنا نرغب في تقدير $P_F(R<a)$ نحسب

$$P_*(R^* < a) = \{\#(R_j^* < a)\}/B$$

4.22 تقدير التحيز بالبوتستراب

إذا كنا نرغب في تقدير التحيز لتقدير دالي $\theta(\hat{F})$ إلى $\theta(F)$ وبالتحديد

$$\theta(\hat{F}) - \theta(F)$$

نأخذ

$$R(\underline{X},F) = \theta(\hat{F}) = \theta(\hat{F}^*) - \theta(\hat{F}) = \hat{\theta}^* - \hat{\theta}$$

حيث إن $\hat{\theta}^* = \theta(\hat{F}^*)$ و $\hat{F}^*$ عبارة عن التوزيع التجريبي (empirical) الذي يعتمد على عينة البوتستراب $\underline{X}^*$ ، أي أن $\hat{F}^*$ تقوم بإعطاء احتمال مقداره M^*/n لكل x_i ، حيث إن M^* تمثل عدد المرات التي تظهر فيها x_i في عينة البوتستراب.

تقدير البوتستراب للتحيز هو

$$\text{Bias} = E_*(R_*) = B^{-1}\sum_{j=1}^{B}\hat{\theta}^*_j - \hat{\theta} = \bar{\hat{\theta}}^* - \hat{\theta}$$

في المثال أعلاه $\bar{\hat{\theta}}^*=3.2$ و $\hat{\theta}=3.0$ لذا فإن التحيز

$$\text{Bias} = \bar{\hat{\theta}}^* - \hat{\theta} = 3.2-3 = 0.2$$

لمزيد من المعلومات يراجع Eforn (1982, 1992) أو Govindarajulu(1999).

5.22 مشكلة الانحدار

لقد تناولنا إلى الآن طريقة البوتستراب لعينة واحدة أو ما يسمى مشكلة العينة الواحدة التي يكون فيها المتغير العشوائي X_i لديه نفس التوزيع الاحتمالي F. ولكن طرق البوتستراب يمكن تطبيقها إلى حالات أكثر تعقيداً، فيما يلي سوف نستخدم طرق البوتستراب على مشكلة الانحدار.

لنأخذ النموذج الآتي:

$$Y_i = g_i(\beta) + \varepsilon_i \; ; \; i = 1,...,n.$$

حيث y_i تمثل المشاهدات للمتغير $Y_i (i = 1, ..., n)$. نفترض الدوال $g_i(.)$ معلومة الشكل وتعتمد على موجه معين ثابت c_i . بينما β_{px1} موجه للمعلمات غير المعروفة، أما حدود الخطأ ε_i فتتوزع بشكل مستقل ومتماثل كما F، حيث $E_F(\varepsilon_i) = 0$ أو الوسيط إلى ε_i يساوي صفر. بعد مشاهدة موجه المتغير

$\underline{Y} = (Y_1, ..., Y_n)'$ ولـيكن $\underline{y} = (y_1, ..., y_n)'$ نرغـب في تقـدير الموجـه β باستخدام أحد المعايير مثل تصغير $D(\underline{y}, \eta)$ المسافة بين $\underline{y}$ وموجه المتنبئات

$$\eta(\beta) = (g_1(\beta), ..., g_n(\beta))'$$

لذا فإن

$$\hat{\beta}: \min_{\beta} D(\underline{y}, \eta(\beta))$$

إن الاختيار المناسب إلى D هو

$$D(\underline{y}, \eta) = \sum_{i=1}^{n} (y_i - \eta_i)^2$$

حيث إن $\eta_i = g_i(\beta)$.

نفترض أن النموذج فيه من التعقيد ما يجعل تحليله بالطرق المعروفة صعباً جـداً، علـى سـبيل المثـال $g_i(\beta) = \exp(\underline{c}_i\beta)$ و F غـير معروفـة و $D(\underline{y}, \eta) = \sum_{i=1}^{n} |y_i - \eta_i|$. لـذا فـإن الخطـوات الـثلاث السـابقة يمكـن تعديلـها لتصبح

1- ابنِ $\hat{F}$ التي تعطي احتمال 1/n لكل مشاهدة من مشاهدات المتبقي $\hat{\varepsilon}_i$ والذي يعرف

$$\hat{\varepsilon}_i = y_i - g_i(\hat{\beta})$$

2- اسحب عينة البوتستراب من $\hat{F}$ وبالتحديد $\varepsilon^*_1, ..., \varepsilon^*_n$ ، لذا فإن عينة البوتستراب تعرف

$$Y^*_i = g_i(\hat{\beta}) + \varepsilon^*_i \; ; i = 1, ..., n.$$

حيث إن ε^*_i تتوزع بشكل مستقل ومتماثل كما $\hat{F}$ ، ومن ثم نحسب

$$\hat{\beta}^*: \min_{\beta} D(\underline{Y}^*, \eta(\beta))$$

3- نقوم بإعادة الخطوة الثانية B مرة (وتكون قيمة B كبيرة) لنجد قيم البوتستراب $\hat{\beta}_1^*, ..., \hat{\beta}_B^*$.

نفترض أننا نرغب في تقدير مصفوفة التغاير إلى $\hat{\beta}$ ، لذا فإن تقدير البوتستراب إلى مصفوفة التغاير هو

$$Cov = \frac{1}{B-1}\sum_{j=1}^{B}(\hat{\beta}_j^* - \bar{\hat{\beta}}^*)(\hat{\beta}_j^* - \bar{\hat{\beta}}^*)'$$

مثال 2: لنفترض أن Y تمثل عمر الشخص وc دخل الشخص بآلاف الدولارات. ونفترض أن نموذج الانحدار هو

$$Y_i = \beta c_i + \varepsilon_i, i = 1, ..., n$$

حيث إن $E(\varepsilon_i) = 0$ و $var(\varepsilon_i) = \sigma^2$ ، وقد حصلنا على البيانات الآتية:

(20, 0.8), (25, 1.1), (30, 1.2), (35, 1.3), (40, 1.5)

نرغب في تقدير تباين $\hat{\beta}$. نقدر β باستخدام طريقة المربعات الصغرى

$$\hat{\beta} = \sum_{i=1}^{5} c_i y_i / \sum_{i=1}^{5} c_i^2 = 185/4750 = 0.039$$

$$\hat{y}_i = \hat{\beta} c_i$$

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta} c_i$$

لذا فإن

$\hat{\varepsilon}_1 = 0.8 - 0.78 = 0.02, \hat{\varepsilon}_2 = 1.1 - 0.975 = 0.125, \hat{\varepsilon}_3 = 0.03, \hat{\varepsilon}_4 = -0.065, \hat{\varepsilon}_5 = -0.06$

أما تباين $\hat{\beta}$ فيمكن حسابه باستخدام

$$var(\hat{\beta}) = \frac{\hat{\sigma}^2}{\sum c_i^2}$$

حيث

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{5} \hat{\varepsilon}_i}{5-1} = \frac{0.02475}{4} = 0.0062$$

لذا فإن تباين $\hat{\beta}$ هو

$$\text{var}(\hat{\beta}) = \frac{\hat{\sigma}^2}{\sum c_i^2} = \frac{0.0062}{7.23} = 0.0008575$$

لنستخدم طريقة البوتستراب لتقدير تباين $\hat{\beta}$ سنطبق المثال بسحب عشر عينات للبوتستراب أي B=10، أولاً نسحب عشر عينات حجم كل عينة 5 وحدات من الأرقام العشوائية بين 1 و5 ولنفترض أن الأرقام التي حصلنا عليها في المثال الأول تمثل هذه العينات، على سبيل المثال العينة الأولى (1,2,2,3,5) والعينة الثانية (3,4,5,5,2) وهكذا؛ لذا فإن عينة البوتستراب الأولى للمتبقي هي: $\varepsilon_1^* = \hat{\varepsilon}_1 = 0.02$ و $\varepsilon_2^* = \hat{\varepsilon}_2 = 1.1$ و $\varepsilon_3^* = \hat{\varepsilon}_2 = 1.1$ و $\varepsilon_4^* = \hat{\varepsilon}_3 = 0.03$ و أخيراً $\varepsilon_5^* = \hat{\varepsilon}_5 = -0.06$. وهكذا لجميع العينات. الآن نستخدم ε_i^* , i = 1, ..., 5 للعينة الأولى لحساب Y_i^* باستخدام

$$Y_i^* = \hat{\beta} c_i + \varepsilon_i^* \; ; i = 1,...,n.$$

الآن لحساب Y_1^* نقوم بالتعويض عن قيم ε_1^* في

$$Y_1^* = 0.039 c_1 + \varepsilon_1^* = 0.039(20) + 0.02 = 0.8$$

لذا فإن الزوج الأول من قيم عينة البوتستراب هو: $(Y_1^*, c_1) = (0.8, 20)$. لتوضيح الفكرة نحسب (Y_3^*, c_3)

$$Y_3^* = 0.039 c_3 + \varepsilon_3^* = 0.039(20) + 1.1 = 1.295$$

لذا فإن الزوج الثالث من قيم عينة البوتستراب هو: $(Y_3^*, c_3) = (1.295, 30)$. وهكذا لجميع قيم العينة لنجد:

(0.8, 20), (1.1, 25), (1.295, 30), (1.395, 35), (1.56, 40)

نستخدم هذه العينة لحساب قيمة β^* للعينة الأولى باستخدام طريقة المربعات الصغرى لنحصل على $\hat{\beta}^* = 0.0408$.

الجدول الآتي يعطينا قيم جميع عينات البوتستراب العشر وقيمة $\hat{\beta}^*$ لكل عينة.

	العينة	$\hat{\beta}^*$
1	(0.8, 20), (1.1, 25), (1.295, 30), (1.395, 35), (1.56, 40)	0.0408
2	(0.81, 20), (0.91, 25), (1.1, 30), (1.305, 35), (1.685, 40)	0.0390
3	(0.8, 20), (1.005, 25), (1.105, 30), (1.385, 35), (1.685, 40)	0.0400
4	(0.72, 20), (0.978, 25), (1.19, 30), (1.365, 35), (1.685, 40)	0.0399
5	(0.72, 20), (0.91, 25), (1.105, 30), (1..49, 35), (1.56, 40)	0.0389
6	(0.81, 20), (1.005, 25), (1.105, 30), (1.385, 35), (1.56, 40)	0.0390
7	(0.715, 20), (0.995, 25), (1.295, 30), (1.49, 35), (1.56, 40)	0.0405
8	(0.715, 20), (0.91, 25), (1.295, 30), (1.365, 35), (1.685, 40)	0.0409
9	(0.905, 20), (0.915, 25), (1.17, 30), (1.49, 35), (1.495, 40)	0.0396
10	(0.72, 20), (0.911, 25), (1.19, 30), (1.305, 35), (1.56, 40)	0.0381

المتوسط $\bar{\hat{\beta}}^* = 0.03967$ وتقدير التباين إلى $\hat{\beta}$ باستخدام البوتستراب هو

$$\frac{1}{B-1}\sum_{j=1}^{B}(\hat{\beta}_j^* - \bar{\hat{\beta}}^*)^2 = \frac{1}{9}(0.01574469 - 0.01573709) = 0.00000084$$

نلاحظ أن تقدير التباين باستخدام البوتستراب أقل بكثير من التقدير الحقيقي.

6.22 التقدير بفترة

إذا كنا نرغب في حساب 95% فترة ثقة إلى المعلمة θ يمكننا القيام بذلك على الوجه الآتي:

$$\hat{\theta} \pm 2\,\mathrm{SD}$$

ولكن إذا أردنا أن نكون أكثر دقة، يمكننا أن نحسب فترة ثقة بعد إيجاد قيم البوتستراب إلى $\hat{\theta}^*_1, ..., \hat{\theta}^*_B$، يمكننا إيجاد (1-α) فترة ثقة والتي هي

$$(\theta^*_{k_1}, \theta^*_{k_2})$$

حيث

$$k_1 = [n\alpha/2] \text{ و } k_2 = [n(1-\alpha/2]\,.$$

إذا استخدمنا البيانات الواردة في المثال أعلاه وبثقة 80%، فسنحصل على فترة ثقة إلى β وهي (0.0381, 0.0408).

References

1. Abramovitch, L. and Singh, K (1985). Edgeworth Corrected Pivotal Statistics and the Bootstrap, *Ann. Statist.*, 13, 116-132.
2. Arcones, M. A. and Gine, E. (1991). Some Bootstrap Tests of Symmetry for Univariate Continuous Distributions, *Ann. Statist.*, 19, 1496-1551.
3. Babu, G. J. and Bose, A. (1989). Bootstrap Confidence Intervals, *Statist. Prob. Letters*, **7**, 151-160.
4. Babu, G. J. and Singh K. (1983). Inference on Means using Bootstrap, *Ann. Statist.*, 11, 999-1003.
5. Bickel, P. J. and Freedman, D. A.(1984). Asymptotic Normality and the Bootstrap in Stratified Sampling, *Ann. Statist.*, 12, 470-482.
6. Boos, D. D. and Brownie, C. (1989). Bootstrap Methods foe Testing Homogeneity of Variances, *Technometrics*, 31, 69-82.
7. Booth, J. G. and Hall, P. (1994). Monte-Carlo Approximation and the Iterated Bootstrap, *Biometrika*, 81, 331-340.
8. Bose, A. (1990). Bootstrap in Moving Average Models, *Ann. Inst. Statist. Math.*, 42, 753-768.
9. Chao, M. T. and Lo, S. H. (1985). A Bootstrap Method for Finite Populations, *Sankhya A*, 47, 399-405.
10. Chen, Z. and Do, K. A. (1992). Important Resampling for the Smoothed Bootstrap, *J. Statist. Compu. Simul.* 40, 107-124.
11. Davison, A. C. and Hall, P. (1993). On Studentizing and Blocking Methods for Implementing the Bootstrap with Dependent Data, *Austral. J. Statist.*, 35, 215-224.
12. Diaconis, P. and Eforn, B. (1983). Computer-Intensive Methods in Statistics, *Scientific American*, May 1983, 116-130.
13. DiCiccio, T. J. and Tibshirani, R. J. (1987).Bootstrap Confidence Intervals and Bootstrap Approximations, *J. Amer. Statist. Assoc.*, 82, 163-170.
14. Do, K. A. and Hall, P. (1991). On Importance Resampling for the Bootstrap, *Biometrika*, 78, 161-167.
15. Eforn, B. (1981). The Jackknife Estimation of Variance, *Ann. Statist.*, 9, 586-596.
16. Eforn, B. (1982). The Jackknife, Bootstrap and other Resampling Plans, Siam Publication No. 38, Philadelphia, PA.
17. Eforn, B. (1992). Jackknife-after-Bootstrap Standard Errors and Influence Functions (with Discussions), *J. R. Statist. Soc.*, B54, 83-127.
18. Eforn, B. and Tibshirani, R. (1993). Introduction to the Bootstrap, Chapman and Hall, New York.
19. Falk, M. (1992). Bootstrap Optimal Bandwidth Selection for Kernel Density Estimates, *J. Statist. Plan. Inference*, 30, 13-32.
20. Freedman, D. A. and Peters, S. C. (1984). Bootstrapping a Regression Equation: Some Empirical Results, *J. Amer. Statist. Assoc.*, 79, 97-106.
21. Frangos, C. C. and Schucany, W. R. (1990). Jackknife Estimation of the Bootstrap Acceleration Constant, *Compu. Statist. Data Anal.*, 9, 271-282.

22. Franklin, L. A. and Wesserman, G. S. (1992). Bootstrap Lower Confidence Limits for Capability Indices, *Journal of Quality Technology*, 24, 196-210.
23. Ghosh, M. Parr, W. C., Singh, K. and Babu, G. J. (1984). A Note on Bootstrapping the Sample Median, *Ann. Statist.*, 12, 1130-1135.
24. Good, P. H. (2001). Resampling Methods: A Practical Guide to Data Analysis, 2end., Birkhauser, Boston.
25. Govindarajulu, Z. (1999). Elements of Sampling Theory and Methods, Prentice Hall.
26. Hall, P. (1986). On the Bootstrap and Confidence Intervals, *Ann. Statist.*, 14, 1431- 1452.
27. Hall, P. (1987). On the Bootstrap and Likelihood-Based Confidence Regions, *Biometrika*, 74, 481-493.
28. Hall, P. (1988). Theoretical Comparisons of Bootstrap and Confidence Intervals (with Discussions), *Ann. Statist.*, 16, 927-953.
29. Hall, P. (1989). On efficient Bootstrap Simulation, *Biometrika*, 76, 613-617.
30. Hall, P. (1990). Using the Bootstrap to Estimate Mean Square Error and Select Smoothing Parameters in Nonparametric Problems, *J. Multivariate Anal.*, 32, 177-203.
31. Hall, P. (1992). On Bootstrap Confidence Intervals in Nonparametric Regression, *Ann. Statist*, 20, 689-711.
32. Hall, P. and Martin, M. A. (1988). On the Bootstrap and Two Sample Problems, *Austral. J. Statist.*, 30, 179-182.
33. Hall, P. and Wilson, S. R. (1991). Two Guide Lines for Bootstrap Hypothesis Testing, *Biometrics*, 47, 757-762.
34. Hinkely, D. V. (1987). Bootstrap Significance Tests, *Proceeding of the 47th Session of International Statistical Institute*, 65-74, Paris.
35. Janas, D. (1993). Bootstrap Procedures for Time Series, Shaker, Aachen.
36. Knight, K. (1989). On the Bootstrap of the Sample Mean in the Infinite Variance Case, *Ann. Statist*, 17, 1168-1175.
37. Kuk, A. Y. C. (1987). Bootstrap Estimators of Variance under Sampling with Proportional to Aggregate Size, *J. Statist. Compu. Simul.* 28, 303-311.
38. Kuk, A. Y. C. (1989). Double Bootstrap Estimation of Variance under Systematic Sampling with Probability Proportional to Size, *J. Statist. Compu. Simul.* 31, 303-311
39. Kunsch, H. R. (1989). The Jackknife and Bootstrap for General Stationary Observations. *Ann. Statist*, 17, 1217-1241.
40. Lahiri, S. N. (1991). Second Order Optimality of Stationary Bootstrap, *Statist. Prob. Letters*, 14, 335-341.
41. Lee, K. W. (1990). Bootstrapping Logistic Regression Models with Random Regression, *Comm. Statist. A*, 19, 2527-2539.
42. Liu, R. Y. (1988). Bootstrap Procedures under some non-i.i.d. Models, *Ann. Statist*, 16, 1697-1708.
43. Lo, A. Y. (1987). A Large Sample Study of the Bayesian Bootstrap, *Ann. Statist*, 15, 360-375.
44. Loh, W. Y. (1991). Bootstrap Calibration for Confidence Construction and Selection, *Statist. Sinica*, 1, 479-495
45. Parr. W. C. (1983). A Note on the Jackknife, Bootstrap, and the Delta Method Estimates of Bias and Variance, *Biometrika* 70, 719-722.

46. Politis, D. N. and Romano, J. P. (1994). The Stationary Bootstrap, *J. Amer. Statist. Assoc.*, 89, 1303-1313.
47. Romano, J. P. (1988). Bootstrapping the Mode, *Ann. Inst. Statist. Math.*, 40, 565-586.
48. Shao, J. (1988). Bootstrap Variance and Bias Estimation in Linear Models, *Canadian, J. Statist.*, 16, 371-382.
49. Shao, J. (1992). Bootstrap Variance and Bias Estimators with Truncation, *Statist. Prob. Letters*, 15, 95-101.
50. Shao, J. (1994). Bootstrap Sample Size in Nonlinear Cases, *Proceeding of the Amer. Math. Soc,*. 122, 1251-1262.
51. Shao, J. and Tu, D. (1995). The Jackknife and the Bootstrap, Springer, New York.
52. Sitter, R. R. (1992). Comparing Three Bootstrap Methods for Survey Data, *Canadian, J. Statist.*, 20, 135-154.
53. Stute, W. (1990). Bootstrap of the Linear Correlation Model, *Statistics*, 21, 433-436.
54. Tibshirani, R. J. (1988). Variance Stabilization and the Bootstrap, *Biometrika*, 75, 433-444.
55. Tu, D. and Zhang, L. (1992). On the Estimation of Skewness of a Statistic using the Jackknife and the Bootstrap, *Statistical Papers*, 33, 39-56.
56. Wang, S. (1989). On the Bootstrap and Smoothed Bootstrap, *Comm. Statist. A*, 18, 3949-3962.
57. Wu, C. F. (1986). Jackknife, Bootstrap and other Resampling Methods in Regression Analysis (with Discussions). *Ann. Statist.*, 14, 1261-1350.
58. Yang, S. S. (1988). A Central Limit Theorem for the Bootstrap Mean, *Amer. Statist.*, 42, 202-203.
59. Zhang, J. Boos, D. D. (1992). Bootstrap Critical Values for Testing Homogeneity of Covariance Matrices, *J. Amer. Statist. Assoc.*, 87, 425-429.

الفصل الثالث والعشرون

أخطاء عدم الإجابة

Nonresponse Errors

1.23 مقدمة

بالإضافة إلى أخطاء المعاينة التي هي جزء من عملية المسح بالعينة والتي لا يمكن التخلص منها بشكل نهائي ولكن يمكن تقليلها، هناك أخطاء أخرى يمكن أن تحدث، منها:

1. **أخطاء التقارير.** دَخلُ الشخص الذي يخبر به دائرة الضرائب ربما يختلف عن دخله الحقيقي، كثيرٌ من الناس يعرفون وزنهم وطولهم ومقدار ما يصرفونه شهرياً على أنفسهم ولكن نادراً ما نحصل منهم على المعلومات الدقيقة حول هذه الأمور.
2. **أخطاء عدم الإجابة.** غالباً ما يقوم بعض الأشخاص الذين ترسل لهم الاستبانات برميها في سلة المهملات، ولا يجيبون عليها على الرغم من التذكيرات التي ترسل لهم لاحقاً، وفي بعض الأحيان الذين يجيبون ربما يعودن إلى طبقة معينة، والذين لا يستجيبون يعودن إلى طبقة أخرى، وهذا يؤدي إلى التحيز في العينات.
3. **أخطاء باختيار العينة.** إن اختيار العينة ربما يكون متحيزاً بسبب المسح أو عندما لا يتم اتباع المعاينة العشوائية بشكل صحيح وصارم. على سبيل المثال الفواكه المعروضة في نوافذ المحلات ليست كالتي هي موجودة داخل المحل.

4. **أخطاء عدم قياس بعض الوحدات.** أحياناً لا يقوم العادون أو الباحثون بقياس بعض الوحدات التي هي من ضمن وحدات العينة، وهذا يحدث عندما تكون عملية العد في الليل، كذلك يحدث في المجتمعات الإنسانية وذلك بسبب عدم الاستطاعة من تحديد أماكن بعض الوحدات.
5. **أخطاء في قياس الوحدة.** ربما تكون وحدة القياس متحيزة أو غير دقيقة. في المجتمعات الإنسانية ربما لا توجد لدى الأشخاص المعلومات الدقيقة، أو ربما يقومون بإعطاء أجوبة متحيزة.
6. **أخطاء فنية.** تحدث هذه الأخطاء في عملية معالجة البيانات مثل عملية الترميز أو التحرير أو جدولة البيانات أي وضعها أو تبويبها في جداول.

2.23 تأثير عدم الإجابة

لنمعن النظر في تأثير عدم الإجابة على تقدير العينة للمتوسط والنسبة. لنفترض أن N_1 يمثل عدد المستجيبين إلى استبانة معينة (لنرمز لها بالطبقة I). لنفترض أن $W_1 = N_1/N$ و $W_2 = N_2/N$ حيث إن $N = N_1 + N_2$. ذلك يعني أن W_2 تشير إلى نسبة عدم الإجابة في المجتمع. إذا سحبنا عينة عشوائية بحجم n من الطبقة I و $\overline{y}_1$ يمثل تقدير العينة، لذا فإن التحيز في متوسط العينة بسبب عدم الإجابة هو

$$E(\overline{y}_1) - \overline{Y} = \overline{Y}_1 - \overline{Y} = \overline{Y}_1 - (W_1 \overline{Y}_1 - W_2 \overline{Y}_2) = W_2(\overline{Y}_1 - \overline{Y}_2)$$

بما أنه لا يمكن مشاهدة الطبقة الثانيةII، لذا فإن حجم التحيز غير معروف فيما عدا إذا قمنا بتخمين $\overline{Y}_2$. ومع ذلك إذا كانت البيانات مستمرة فإن فترة الثقة للمتوسط $\overline{Y}_2$ تكون واسعة جداً لدرجة لا يمكن الاستفادة منها. ولكن إذا كنا نرغب في تقدير النسبة P_2 فإن هناك أمل أن تكون

نسبة عدم الإجابة P_2 محدودة من الأعلى بواحد عندما نقوم بحساب فترة ثقة إلى نسبة المجتمعP. لنفترض أن حجم العينة العشوائية الكلية هوn وهنالك n_1 استجابة من بينn وحدة. إذا كان حجم n_1 كبير يمكننا أن نجد فترة 95% ثقة إلى P_1 وهي

$$p_1 \pm 2\sqrt{p_1(1-p_1)/n_1}$$

عندما يمكن إهمال معامل التصحيح للمجتمعات المحدودة، حيث P_1 يمثل نسبة العينة. يمكننا أن نجد فترة ثقة متحفظة إلى نسبة المجتمعP وذلك بوضع قيمة $P_2 = 0$ عندما نجد الحد الأدنى لفترة الثقة $\hat{P}_L$ و$P_2=1$ عندما نجد الحد الأعلى لفترة الثقة $\hat{P}_U$ ، وعليه فإن فترة 95% ثقة إلىP ستكون

$$\hat{P}_L = W_1[p_1 - 2\sqrt{p_1(1-p_1)/n_1}\,] + 0(W_2)$$

$$\hat{P}_u = W_1[p_1 + 2\sqrt{p_1(1-p_1)/n_1}\,] + 1(W_2)$$

عندما تكون قيمة W_2 غير معلومة فإن Cochran (1977) اقترح الطريقة الآتية لإيجاد فترة ثقة إلى P. عند حساب $\hat{P}_L$ نفترض أن جميع عدم المستجيبين في العينة أعطوا جواباً سالباً، وهذا يعني أن $W_1=1$ و$P_2=0$. أما عندما نقوم بحساب $\hat{P}_U$ فنفترض أن جميع عدم المستجيبين أعطوا جواباً موجباً، وهذا يعني أن $W_1=1$ و$P_2=1$. لاحظ عندما تكون $W_1=1$ سنضع $n_1=n$.

مثال: لنفترض أن n=500 و$n_1=400$ و$p_1=15\%$ ، لذا فإن 75 عضواً من العينة أجابوا بالإيجاب، ومعدل عدم الإجابة هو 20% . أوجد 95% فترة ثقة إلىP.

الحل:

$$\hat{P}_L = 0.15 - 2\sqrt{0.15(0.85)/500} = 0.15 - 0.032 = 11.8\%$$

$$\hat{P}_U = 0.35 - 2\sqrt{0.35(0.65)/500} = 0.35 - 0.043 = 39.3\%$$

3.23 أنواع عدم الإجابة

يكننا أن نعطي تصنيفاً تقريبياً لأنواع عدم الإجابة وهي

1. **عدم التغطية**. وهو عبارة عن الفشل في تحديد بعض وحدات العينة أو الوصول إليها، ويبرز هذا غالباً عندما يفشل العاد بالوصول إلى بعض الوحدات بسبب رداءة الطقس، أو أن تكون القائمة والكشف الذي عنده غير كامل، أو بسبب عدم توافر المواصلات الجيدة في أثناء مدة المسح ليتسنى الوصول إلى جميع الوحدات.
2. **غير موجود في المنزل**. تتكون هذه الفئة من مجموعة من الأشخاص الذين يقيمون في المنزل ولكنهم غير موجودين في المنزل ساعة حضور العاد إليه. غالباً ما يصعب الحصول على شخص داخل المنزل إذا كان جميع أفراد العائلة يعملون، بعكس العوائل التي فيها أطفال أو عجزة حيث غالباً ما تجد شخصاً ما داخل المنزل.
3. **غير قادر على الإجابة**. قد لا يمتلك الشخص المسؤول الإجابة، التي تتطلب معلومات محددة أو ربما لا يرغب في إعطاء الإجابة، تتولى عملية حسن صياغة الأسئلة في الاستبانة معالجة مثل هذه الأمور.
4. **الفريق الصعب**. الأشخاص الذين يرفضون الإجابة باستمرار أو العاجزين عن إجراء المقابلة أو الذين غالباً ما يكونون غير موجودين في المنزل طيلة مدة المسح هم الذين يشكلون هذا القطاع. وهم يشكلون مصدراً للتحيز لا يتزحزح مهما بذلت الجهود.

إن معالجة مشكلة عدم التغطية غالباً ما تكون صعبة، ولكن هناك بعض الطرق التي يمكن أن نتبعها للتغلب أو التخفيف من هذه المشكلة منها: إعادة زيارة الوحدات التي لم يكن بالمستطاع الوصول إليها في المحاولة الأولى، القيام بوضع القوائم بعناية بحيث يمكن استخدامها للتحقق من أن المعلومات

كاملة. وأحياناً نقوم بالمقارنة بين أعداد الأشخاص أو المنازل التي أحصيناها وبين المعلومات المتوافرة حول الموضوع نفسه من مسوحات أخرى. وعندما يكون الدليل غير كامل وذلك بسبب إنشاء مبانٍ جديدة يتم تدعيم الكشف الموجود بعينة مساحية، الهدف منها معاينة أجزاء من المدينة للتعرف على المباني الجديدة وإضافتها إلى الكشف الموجود، لمزيد من المعلومات حول المسوحات التي تعاني من مشكلة عدم التغطية وكيفية معالجة هذه المشكلة يرجى الرجوع إلى Kish and Hess (1958) وWoolesy (1956).

فيما يتعلق بمشكلة الغائبين عن المنزل تكون المشكلة أسهل إذا كان أي بالغ في المنزل يمكنه الإجابة على الأسئلة، كما هو الحال في المسوحات التي نريد أن نقابل بها شخصاً واحداً محدداً جرى سحبه بطريقة عشوائية. ونفضل مقابلة شخصٍ بالغٍ بمفرده، إذا كان المسح من النوع الذي يتضمن أشخاصاً لا يستطيع أحدهم الإجابة بدقة نيابة عن الآخرين، أو إذا كان هناك ارتباطات عالية ما بين المعلومات ضمن المنزل الواحد، بحيث يصبح قياس أكثر من شخص واحد غير مُجدٍ اقتصادياً. لقد طور Kish (1949) طريقة جيدة لاختيار شخص واحد من أسرة واحدة.

4.23 الرجوع مرة أخرى

لتفادي الأعداد الكبيرة من عدم الإجابة نقوم بالرجوع للوحدة أكثر من مرة، ومن الأشياء المتعارف عليها هو الرجوع مرة أخرى خصوصاً إذا لم يكن هناك أحد في المنزل. لنمعن النظر في الجدول الآتي الذي يتضمن نسبة الإجابة مصحوبة بنسبة المنازل التي لديها أطفال بعمر أقل من سنتين موزعة على عدد مرات الرجوع، هذه البيانات تم الحصول عليها من Hilgard and Payne (1944)

التي رجع إليها P.S.R.S. Rao(1983) والتي تمثل مسحاً يتألف من3265 منزل، والتي نفذت من خلال مقابلة شخصية لأصحاب المنازل.

عدد مرات الرجوع

	1	2	>2	عدم الإجابة	المجموع
نسبة الإجابة	63.5	22.2	14.3	0	10.0
نسبة المنازل التي لديها أطفال صغار من بين المستجيبين	17.2	9.5	6.2	0	13.9

يوضح هذا الجدول أن الشخص البالغ الذي لديه أطفال صغار يكون احتمالية وجوده في المنزل للمقابلة أكبر عندما يقوم العاد بزيارتهم. وهذا يعني إذا توقف العاد بعد الزيارة الأولى فإن العينة تمثل العوائل التي لديها أطفال صغار بنسبة أكبر من العوائل الأخرى. فإذا كان المجتمع الذي نستهدفه هو جميع العوائل فإن ذلك سيؤدي إلى تحيز في تقدير جميع المتغيرات التي لها علاقة بالأطفال الصغار. نتوقف عن إعادة الزيارة أو الرجوع في الحياة العملية بعد محاولات قليلة. والسبب الرئيس للتوقف هو التكاليف، والتي تلعب الدور الكبير في كثيرٍ من المسوحات. ربما نحتاج إلى نموذج لتحديد مجموع تكاليف العمليات الميدانية للمسح بالعينة، يراجع Govindarajulu (1999) و Birnbaum and Sirken (1950) لمزيد من المعلومات.

5.23 أخطاء القياس

تحدث أخطاء القياس في الحياة العملية، ولقد دُرست هذه المسألة من قبل أكثر من باحث من بينهم Sukhatme et al. (1984). لنفترض أن Y_i تمثل القيمة الحقيقية للإجابة من قبل الوحدة i ونرمز لها بـ U_i. لنفترض أننا سحبنا عينة عشوائية بسيطة بحجم n وحدة من مجتمع بحجم N وحدة، وأيضًا

نفترض أن x_{ij} تمثل القيمة التي قدمها العاد j للوحدة U_i حيث إن i=1,...,h و j=1,...,m.

لندرس أبسط النماذج وهو

$$x_{ij} = y_i + \alpha_j + \varepsilon_{ij}$$

حيث

α_j : يمثل التحيز للعاد j للمشاهدات المعادة لجميع الوحدات

ε_{ij} : يمثل الخطأ الذي ارتكبه العاد j في الوحدة i

حيث إن

$$E(\varepsilon_{ij} | i, j) = 0$$

و

$$E(\varepsilon_{ij}^2) = S_e^2$$

و ε_{ij} غير مرتبطة لجميع قيم i و j.

نفترض أن عدد التكرارات n_{ij} تساوي 1 أو 0. لنفترض

$n_{i.} = \sum_{j=1}^{m} n_{ij}$: عدد المشاهدات للوحدة U_i

$n_{.j} = \sum_{i=1}^{h} n_{ij}$: عدد المشاهدات التي قام بها العاد j

$\overline{x}_{.j} = \frac{1}{n_{.j}} \sum_{i=1}^{h} n_{ij} x_{ij}$: المتوسط لجميع المشاهدات التي قام بها العادj والتي عددها $n_{.j}$

$\overline{x}_{..} = \frac{1}{n} \sum_{i}^{h} \sum_{j}^{m} n_{ij} x_{ij}$: المتوسط لجميع المشاهدات n والتي أجريت على h من الوحدات في العينة.

لذا

$$\bar{x}.j = \frac{1}{n_{.j}}\sum_{i=1}^{h}(n_{ij}y_i + n_{ij}\alpha_j + n_{ij}\varepsilon_{ij}) = \frac{1}{n_{.j}}\sum_{i}^{h} y_i n_{ij} + \alpha_j + \frac{1}{n_{.j}}\sum_{i}^{h} n_{ij}\varepsilon_{ij}$$

كذلك

$$\bar{x}_{..} = \frac{1}{n}\sum_{i=1}^{h} y_i n_{i.} + \frac{1}{n}\sum_{j=1}^{m}\alpha_j n_{.j} + \sum_{i}^{h}\sum_{j}^{m}\varepsilon_{ij} n_{ij}$$

لنفترض أننا وضعنا الفروض الآتية:

1. العادون m يمثلون عينة عشوائية بسيطة من مجتمع حجمه M من العادين.
2. الوحدات h في العينة جرى توزيعها بصورة عشوائية بين العادين.
3. $n_{.j} = n/m = \bar{n}$.
4. $n_{i.} = n/h = p$.

لذا فإن المعادلتين أعلاه تصبحان

$$\bar{x}.j = \frac{1}{n}\sum_{i=1}^{h} y_i n_{ij} + \alpha_j + \frac{1}{n}\sum_{i}^{h} n_{ij}\varepsilon_{ij}$$

و

$$\bar{x}_{..} = \frac{1}{h}\sum_{i=1}^{h} y_i + \frac{1}{m}\sum_{j=1}^{m}\alpha_j + \frac{1}{n}\sum_{i}^{h}\sum_{j}^{m}\varepsilon_{ij} n_{ij}$$

كذلك

$$E(\bar{x}.j) = \frac{1}{N}\sum_{i=1}^{N} y_i + \frac{1}{M}\sum_{i}^{M}\alpha_j = \mu + \bar{\alpha}$$

و

$$E(\bar{x}_{..}) = \mu + \bar{\alpha}$$

يمكننا أن نجد تباين $\bar{x}.j$

$$\mathrm{var}(\bar{x}_{.j}) = (\frac{1}{n} - \frac{1}{N})S_y^2 + (1 - \frac{1}{M})S_\alpha^2 + \frac{S_\varepsilon^2}{n}$$

إذا كانت M وN كبيرتين، فيمكن كتابة العلاقة الأخيرة كما يأتي

$$\text{var}(\bar{x}_{.j}) \quad (\frac{1}{n})(S_y^2+S_\varepsilon^2) + S_\alpha^2$$

و تباين $\bar{x}_{..}$

$$\text{var}(\bar{x}_{..}) = (\frac{1}{h}-\frac{1}{N})S_y^2+(\frac{1}{m}-\frac{1}{M})S_\alpha^2+\frac{S_\varepsilon^2}{n} \quad \frac{S_y^2}{h}+\frac{S_\alpha^2}{m}+\frac{S_\varepsilon^2}{n}$$

حيث

$$S_\alpha^2 = \frac{1}{M-1}\sum_{j=1}^{M}(\alpha_j-\bar{\alpha})^2 \text{ و } S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-m)^2$$

نلاحظ أنه بالرغم من كون α تختلف من عاد إلى آخر لكننا لا نستطيع إهمال $\bar{\alpha}$.

بما أن تباين المشاهدة الواحدة التي سحبت من مجتمع غير متناهٍ عندما تكون M كبيرة يمكننا أن نكتب

$$S_x^2 = S_y^2 + S_\alpha^2 + S_\varepsilon^2$$

يمكننا أن نكتب تباين $\bar{x}_{..}$

$$\text{var}(\bar{x}_{..}) = \frac{S_x^2}{h}+(\frac{1}{m}-\frac{1}{h})S_\alpha^2+(\frac{1}{n}-\frac{1}{h})S_\varepsilon^2$$

نلاحظ من المعادلة الأخيرة أن التباين للمقدر $\bar{x}_{..}$ ليس مصدره الاختلاف بين وحدات العينة، ولكن تم تضخيمه بسبب التغاير أو الاختلاف بين العادين. لذا فإن الصيغ التي مرت معنا في الفصول السابقة قد تُقدر التباين بأقل مما يجب عليه. وأخيراً نلاحظ أن

$$\text{var}(\bar{x}_{..}) = \frac{S_x^2}{h}$$

إذا كان m= h= n أو إذا كانت α_j ثابتة لجميع قيم j وn =h. لمزيد من المعلومات يراجع Sukhatme et al. (1984).

6.23 تأثير التحيز الثابت

لنفترض أن هناك تحيزاً ثابتاً مقداره β في مقاسات المتغير y_i لجميع الوحدات ومقداره غير معلوم. لذا فإن متوسط العينة العشوائية البسيطة $\overline{y}$ سيكون عرضة لخطأ مقداره β. ولكن التحيز يلغى عند تقدير خطأ التباين على اعتبار أنه يعتمد على مجموع المربعات $(y_i - \overline{y})^2$. وكنتيجة لذلك فإن حساب فترة الثقة لمتوسط المجتمع لن تخضع لهذا التحيز، وسنحصل على النتيجة نفسها إذا استخدمنا العينة الطبقية.

وتبقى الحالة نفسها بالنسبة لتقدير النسبة والانحدار. لنأخذ تقدير الانحدار

$$\overline{y}_{lr} = \overline{y} + b(\overline{X} - \overline{x})$$

إذا كان y_i و x_i يخضع كل منهما إلى تحيز ثابت مقداره β_y و β_x على التوالي. بما أن تقدير المربعات الصغرى إلى b لم يتغير بسبب التحيز والتحيز يلغى من الحد $(\overline{X} - \overline{x})$ و هذا يعني أن $\overline{y}_{lr}$ سيتأثر بتحيز β_y. ويمكن بسهولة إثبات أن تقدير تباين $\overline{y}_{lr}$ لا يحتوي على أي تأثير للتحيز.

أما فيما يخص تقدير النسبة

$$\overline{y}_R = \frac{\overline{y}}{\overline{x}}\,\overline{X}$$

فإن التحيز الثابت سيكون مقداره β_y ، بما أن $E(\overline{X}|\overline{x}) \approx 1$ إذا كان حجم العينة كبيراً حتى ولو كان x_i يتعرض إلى تحيز ثابت. في حالة العينات الكبيرة فإن تقدير التباين هو

$$s^2(\overline{y}_R) = \frac{(N-n)}{Nn}\,\frac{\sum(y_i - \hat{R}x_i)^2}{n-1}$$

سيكون حراً من التحيز باعتباره تقديراً إلى

$$E(\overline{y}_R - \overline{Y})^2$$

خلاصة القول أن التحيز الثابت سوف يمر من غير أن تكتشفه بيانات الإحصائية للعينة.

7.23 الإجابة العشوائية

عندما يطلب من الشخص الإجابة عن بعض الأسئلة الحساسة مثل: الإصابة بمرض فقد المناعة المكتسبة أو لديه علاقة غير شرعية، غالباً ما يتهرب الشخص من الإجابة عن مثل هذا النوع من الأسئلة أو لا يعطي الجواب الصحيح. في مثل هذه الحالات ربما نلجأ إلى طريقة الإجابة العشوائية، على سبيل المثال إذا كانت π_A تمثل النسبة الحقيقية من الناس الذين يحملون الصفة A، لقد أثبت Warner (1965) أنه يمكن تقدير π_A من دون أن يحتاج الشخص المجيب أن يكشف نفسه فيما يخص هذا السؤال الحساس أو المحرج.

نختار الإجابة العشوائية مثل رمي زهر النرد، صندوق يحتوي على كرات حمراء وبيضاء، اختيار عبارة من عبارتين كل منهما يتطلب الإجابة عليها بنعم أو لا، كي يجري تقديمه إلى المستجيب، ولا يعلم العاد أو المعاين أي سؤال أجاب عليه المستجيب. ولكن العاد يعلم الاحتمالات P و1-P التي قدمت بها العبارتان. الفكرة الأساسية هنا أن المستجيب متأكد أنه لم يفشِ سره حول السؤال الحساس.

لنفترض أن العبارتين أو الجملتين كما يأتي

" أنا أحمل الصفةA " (وتقدم باحتمالP).

" أنا لا أحمل الصفةA " (وتقدم باحتمال1-P).

في عينة عشوائية حجمها n، نفترض أن لدينا m، جواب إيجابي أي أنا أحمل الصفة A. لذا فإن

$$\hat{\phi} = m/n$$

حيث

$$\phi = P\pi_A + (1-P)(1-\pi_A) = (2P-1)\pi_A + 1 - P$$

عندما تكون P معلومة

$$\hat{\pi}_A = \frac{\hat{\phi} - (1-P)}{2P-1};\ P \neq 1/2$$

يمكننا أن نلاحظ أن $\hat{\pi}_A$ عبارة عن تقدير الاحتمالية العظمى (MLE) إلى π_A و $E(\hat{\pi}_A) = \pi_A$ وبما أن $\hat{\phi}$ هو تقدير الاحتمالية العظمى إلى ϕ و $E(\hat{\phi}) = \phi$ وتباين $\hat{\pi}_A$ هو

$$\text{var}(\hat{\pi}_A) = \frac{\phi(1-\phi)}{n(2P-1)^2}$$

باستخدام قيمة ϕ أعلاه يمكننا إعادة كتابة تباين $\hat{\pi}_A$ كما يأتي:

$$\text{var}(\hat{\pi}_A) = \frac{\pi_A(1-\pi_A)}{n} + \frac{P(1-P)}{n(2P-1)^2}$$

سيكون القسم الأول من تباين $\hat{\pi}_A$ أعلاه التباين الفعلي لو أن جميع الأشخاص أجابوا على السؤال الحساس بصورة صحيحة فيما إذا كان يحملون الصفة A. ممكن أن يكون القسم الثاني كبيراً جداً، وهذا يعتمد على قيمة P.

مثال: ترغب إحدى المدارس أن تقدر نسبة الطلاب الذين يغشون في الاختبارات من بين طلاب المدرسة. قام العاد بسحب عينة عشوائية حجمها

n=500 طالب من طلاب المدرسة التي عددها N=2500. كل طالب سئل أن يسحب ورقة من بين مجموعة أوراق، في هذه المجموعة 85% من الأوراق تحمل العبارة " أنا أغش في الاختبارات" و15% تحمل العبارة " أنا لا أغش في الاختبارات". ولقد أجاب 300 طالب بنعم للورقة التي رآها، علماً بأن العاد لا يعرف ما كتب على الورقة التي سحبها الطالب وأجاب عليها. أوجد فترة 95% ثقة لنسبة الطلاب الذين يغشون من بين طلاب المدرسة.

الحل:

$$\hat{\phi} = \frac{m}{n} = \frac{300}{500} = 0.6$$

$$\hat{\pi}_A = \frac{\phi - (1-P)}{2P-1} = \frac{0.6-(1-0.85)}{2(0.85)-1} = 0.643$$

$$var(\hat{\pi}_A) = \frac{\pi_A(1-\pi_A)}{n} + \frac{P(1-P)}{n(2P-1)^2} = \frac{0.64(1-0.64)}{500} + \frac{0.85(1-0.85)}{500(2(0.85)-1)^2}$$

$$= 0.0004591 + 0.0005204 = 0.0009795$$

لذا فإن فترة 95% ثقة لنسبة الغشاشين في المدرسة هي

$$0.643 \pm 2\,(0.0313) = 0.643 \pm 0.0626$$

لمزيد من المعلومات ولتفاصيل أكثر يراجع Warner (1965) و Cochran (1977) و Govindaraj ulu (1999).

8.23 اختيار إعادة المقابلة

لقد لاحظنا أعلاه أن هناك نسبة من عدم الاستجابة، سببها في الغالب عدم وجود الشخص في وقت المقابلة أو الذهاب إليه لجمع المعلومات منه في

وقت غير مناسب. إذا كان لدينا عينة حجمها n من الوحدات وحصلنا على n_1 من الإجابات حيث إن $n_1 < n$ ، سيكون لدينا مجموعة من عدم الإجابة، يمكن أن نتصور أن المجتمع مكون من طبقتين، عدد وحدات العينة التي اختيرت من الطبقة الأولى n_1 ومن الطبقة الثانية $n_2 = n-n_1$ ، ولكن في الحقيقة ليس لدينا طبقية بالمعنى الصحيح، لأن حجم العينة من الطبقة الثانية n_2 متغير عشوائي يمكن تحديده فقط بعد الانتهاء من جمع البيانات، ولكن إذا فكرنا فيها كعينة طبقية نستطيع أن نحصل على طريقة مثالية لتقسيم الموارد المالية بين المسح الأولي والإعادة.

لنفترض أننا قررنا أن نعيد مقابلة r من الذين لم نستطع مقابلتهم في المحاولة الأولى والبالغ عددهم n_2 حيث إن $r = n_2/k$ و $k > 1$ عدد ثابت.

إذا كان $\bar{y}_1$ يمثل معدل العينة في المحاولة الأولى و $\bar{y}_2$ معدل العينة إلى r من الوحدات في المحاولة الثانية، إذاً

$$\bar{y}^{*} = \frac{1}{n}(n_1 \bar{y}_1 + n_2 \bar{y}_2)$$

تقدير غير متحيز إلى الوسط الحسابي للمجتمع، لمزيد من المعلومات يراجع Scheaffer, Mendenhall and ott (1996).

References

1. Armitage, P. (1947). A Comparison of Stratified with Unrestricted Random Sampling from a finite Population. *Biometrika*, 34, 273-280.
2. Bartholomew, D. J. (1961). A Method for "not-at-home" Bias in Sample Surveys, *App. Stat.* 10, 52-59.
3. Birnbaum, Z., W. and Sirken, M. G. (1950). Basis Due to Nonavailability In Sampling Surveys, *J. Amer. Statist. Assoc.*, 45, 98-111.
4. Cochran, W. G. (1977). Sampling Technique, 3rd Ed. Wiley, New York.
5. Deming, W. E. (1953). On a Probability Mechanism to Attain an Economic Balance Between the Resultant Error of Non-Response and the Bias of Non-Response, *J. Amer. Statist. Assoc.*, 48, 743-772.
6. Durbin, J. (1954). Non-Response and Call-Backs in Surveys, *Bull. Int. Stat. Inst.*, 34, 72-86.
7. Durbin, J. and Stuart, A. (1954). Callbacks and Clustering in Sample Surveys: An Experimental Study, *J. Roy. Stat. Soc.* A117, 387-428.
8. Govindarajulu, Z. (1999). Elements of Sampling Theory and Methods, Prentice Hall.
9. Hansen, M. H. (1951). Response Errors in Surveys, *J. Amer. Statist. Assoc.*, 46, 147-190.
10. Hansen, M. H. and Hurwitz, W. N. (1946). The Problem of Nonresponse in Sample Surveys, *J. Amer. Statist. Assoc.*, 41, 517-529.
11. Hansen, M. H. and Waksberg, J. (1970). Research on Non-Response in Censuses and Surveys, *Rev. Int. Stat. Inst.*, 38, 318-332.
12. Hoewitz, D. G., Shah, B. V. and Simmons, W. R. (1967). The Unrelated Nonresponse Randomized Response Model, *Poroc. Soc. Statist. Section Amer. Statist. Assoc.*,663-685.
13. Kish, L (1949). A Procedure for Objective Respondent Selection within the Household, *J. Amer. Statist. Assoc.*, 44, 380-387.
14. Kish, L and Hess, I (1958). On Noncoverage of Sample Dwelling, *J. Amer. Statist. Assoc.*, 53, 509-524.
15. Kish, L. and Lansing, J, B. (1954). Response Errors in Estimating the Value of Homes, *J. Amer. Statist. Assoc.*, 49, 520-538.
16. Madow, G. A. (1965). On some Aspects of Response Error Measurement, *Poroc. Soc. Statist. Section Amer. Statist. Assoc.*, 182-192.
17. Oh, H. L. and Scheuren, F. J. (1983). Weighting Adjustment for Unit Nonresponse. In *Incomplete Data in Sample Surves,* Vol. 2, (ed. I. Olkin and D. B. Rubin), New York, Academic Press, 143-184.
18. Politz, A. N. and Simmons, W. R. (1949). An Attempt to Get the "Not at Homes" into the Sample without Callbacks, *J. Amer. Statist. Assoc.*, 44, 9-31.
19. Politz, A. N. and Simmons, W. R. (1950). An Attempt to Get the "Not at Homes" into the Sample without Callbacks, *J. Amer. Statist. Assoc.*, 45, 136-137.
20. Rao, P. S. R. S. (1983). Callbacks, Followups and Repeated Telephone Calls. In *Incomplete Data in Sample Surves,* Vol. 2, (ed. I. Olkin and D. B. Rubin), New York, Academic Press, 33-44.
21. Scheaffer, R.L., Mendenhall, W. and Ott L. (1996). Elementary Sampling Survey, 5th ed., Duxbury, New York.

22. Sukhatme, P. V. and Seth, G. R. (1952). Non-Sampling Errors in Surveys, *j. Ind. Soc. Agr. Stat.*, 4, 5-41.
23. Sukhatme, P. V. and Sukhatme, B. V. (1984). Sampling Theory of Surveys with Applications, (3rd ed.) Ames, Iowa State Univ. Press.
24. Warner, S. L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias, *J. Amer. Statist. Assoc.*, **60**, 63-69.
25. Woolsey, T. D. (1956). Sampling Methods for Small Household Survey, *Pub. Health Monograph,* No. 40.

الفصل الرابع والعشرون

طرق أخرى للمعاينة

Other Sampling Methods

1.24 مقدمة

إن الهدف من هذا الفصل إعطاء القارئ فكرة عن بعض طرق المعاينة الأخرى، التي قد يحتاجها الباحث في عمله أو في أثناء قيامه بالبحث العلمي. سوف نتناول باختصار شديد بعض طرق المعاينة التي قد تستعمل أكثر من غيرها في الحياة العملية، وسوف يعطى القارئ مراجع لكل طريقة سنتكلم عنها في هذا الفصل للرجوع إليها ودراستها بشكل مفصل.

2.24 المعاينة العشوائية مع الإرجاع

لقد تكلمنا بالتفصيل عن العينة العشوائية البسيطة بدون إرجاع. وكذلك غالبية طرق المعاينة التي تكلمنا عنها في هذا الكتاب هي بدون إرجاع بشكل أو آخر. وذلك لأهمية وانتشار استخدامها في الحياة العملية.

إن الفكرة الأساسية لعملية المعاينة مع الإرجاع كما أشرنا إلى ذلك في الفصل الأول تقوم على سحب وحدات العينة من المجتمع ، باستخدام إحدى الطرق العشوائية المعروفة ولكن نقوم بإرجاع الوحدة التي تم سحبها إلى المجتمع بعد مشاهدتها قبل أن نقوم بسحب وحدة جديدة وهكذا. تسمى هذه الطريقة (Random Sampling with Replacement) المعاينة العشوائية مع الإرجاع. قد يكون سبب عدم انتشار أو استعمال هذه الطريقة في الحياة العملية على الرغم من سهولة التعامل مع التقديرات الإحصائية باستخدامها،

أنَّ الوحدة إذا تم سحبها أكثر من مرة في العينة فإنها في المرات اللاحقة لن تضيف أي معلومات جديدة، لأنها مكررة، لمزيد من المعلومات يراجع Cochran (1977) أو ترجمة أنيس كنجو(1995).

3.24 التقدير في المجتمعات الجزئية

نرغب في بعض الأحيان التقدير لمجموعة جزئية من المجتمع أو لمجتمع جزئي، بالإضافة إلى المجتمع كاملاً، فمثلاً قد نرغب في تقدير دخل النساء في المجتمع بالإضافة إلى تقدير دخل الفرد في المجتمع، تشكل النساء هنا مجموعة جزئية من المجتمع الكلي الذي يحتوي على الرجال والنساء، أو قد نرغب في تقدير دخل العاملين في قطاع الزراعة بالإضافة إلى دخل جميع أفراد المجتمع، لمزيد من المعلومات يراجع(1977) Cochran أو ترجمة أنيس كنجو (1995) وكذلك Raj (1972).

4.24 نموذج الإجابة العشوائية

غالباً ما يرفض كثير من الأشخاص الذين نجمع المعلومات منهم الإجابة عن بعض الأسئلة التي قد تسبب لهم في بعض الأحيان أذى، فعلى سبيل المثال غالباً يرفض الأشخاص الإجابة عن الأسئلة السياسية في البلدان التي تكون فيها الديمقراطية معطلة، لأن الإجابة عن هذه الأسئلة قد تسبب لهم أذى.

سوف نحاول تقدير نسبة الأشخاص الذين يحملون صفة معينة دون الحاجة إلى مساءلتهم بطريقة مباشرة في المقابلة. إن أول من اقترح هذه الطريقة هو(1965) Warneer، فمثلاً إذا أردنا تقدير نسبة الطلاب الذين يحاولون الغش في الامتحانات لن نحصل على إجابات صحيحة إذا سألنا الطلاب بصورة مباشرة، لذلك نقوم بتقسيم المجتمع إلى مجموعتين A وB في مثالنا هذا قد تكون A يغش وB لا يغش، وأعضاء المجتمع هنا إما يقعون في A أو B.

لنفترض أن p تمثل نسبة الطلاب الذين يغشون في الامتحان. إن الهدف هو تقدير p دون أن نسأل الطلاب مباشرة في المقابلة.

نبدأ بأخذ مجموعة من الأوراق البيضاء التي كلها تحمل نفس المواصفات، ما عدا جزءاً ونسبة θ يحمل علامة A وجزءاً آخر ونسبة $(1-\theta)$ يحمل علامة B. نقوم بسحب عينة عشوائية بسيطة حجمها n طالب، ونطلب من كل واحد منهم سحب ورقة واحدة، ويجيب بنعم إذا كان الحرف المكتوب على الورقة يتفق مع مجموعته، ولا، إذا كان يختلف عن مجموعته. إن الشخص الذي يقوم بالمقابلة لا يرى ما هو مكتوب على الورقة، فقط يسجل الإجابات نعم أو لا وبعد الانتهاء من جميع أفراد العينة، نستخدم هذه البيانات لتقدير p، يراجع Scheaffer, Mendenhall and ott (1996) لمزيد من المعلومات.

5.24 المعاينة العنقودية بثلاث مراحل

إن طريقة العينة العنقودية بمرحلتين يمكن أن توسع إلى مرحلة ثالثة، وذلك بسحب عينة عشوائية من كل وحدة من وحدات المرحلة الثانية بدلاً من مشاهدة جميع عناصر الوحدة، على سبيل المثال في مسح الحقول لتقدير إنتاجية أحد المحاصيل كالقمح مثلاً، يمكن أن نَعُدّ المرحلة الأولى هي سحب عينة من القرى الموجودة في المنطقة التي يجري فيها تقدير المحصول، أما المرحلة الثانية فتكون اختيار مجموعة من الحقول المزروعة بالقمح الموجودة في القرى التي سحبت في المرحلة الأولى، أما المرحلة الثالثة فتكون اختيار عدد من القطع الصغيرة داخل كل حقل سحب في المرحلة الثانية. نلاحظ أننا حصلنا على العينة العنقودية في ثلاث مراحل وذلك من خلال توسيع المرحلة الثانية في العينة العنقودية بمرحلتين، مع الافتراض أن جميع الوحدات في المرحلة الثالثة نختارها باحتمالات متساوية. لمزيد من المعلومات يراجع Singh and Chaudhary (1986).

6.24 الطبقية مع المعاينة العنقودية

إن من أهم الطرق الشائعة الاستعمال على مستوى كبير في المسوحات ما يسمى (Cluster Sampling with Stratification) الطبقية مع العينة العنقودية. لا توجد هنالك مفاهيم جديدة مع هذه الطريقة. عندما نريد تقدير الوسط الحسابي للمجتمع على سبيل المثال نقوم بتقسيم المجتمع إلى k من الطبقات، وتكون المعاينة ضمن كل طبقة مستقلة عن الطبقة الأخرى، حيث نقوم بسحب عينة عنقودية داخل كل طبقة.

لنفترض أن الطبقة i تحتوي N_i من الوحدات (المرحلة الأولى) وكل وحدة تحتوي على M_i من الوحدات (المرحلة الثانية) لذلك فإن n_i هو حجم العينة المسحوبة في المرحلة الأولى و m_i هو حجم العينة المسحوبة في المرحلة الثانية إن تقدير الوسط الحسابي للمجتمع.

حيث إن $\bar{y}_i$ يمثل الوسط الحسابي للطبقة i و

$$W_i = \frac{N_i M_i}{\sum_{i=1}^{k} N_i M_i}$$

يمثل وزن الطبقة لمزيد من المعلومات يراجع Singh and Chaudhary (1986).

7.24 المعاينة بالحصة

تؤدي السرعة عملاً مهماً في مسوحات رأي الناس، لذلك أصبح شائعاً استخدام ما يسمى بالمعاينة بالحصة (Quota Sampling) لاختبار العينة من المجتمع. إن طريقة المعاينة بالحصة قائمة على مبدأ أن العينة توزع جيداً على المجتمع، ويجب أن تحتوي على نفس النسبة التي يمثلها أشخاص يحملون صفة مميزة في المجتمع، فمثلاً إذا كانت نسبة الرجال في المجتمع 60% فيجب أن تحتوي العينة على 60% من الرجال. إن أهم الصفات التي في العادة تؤخذ في

الحسبان في هذا النوع من المعاينة هي الجنس، والعمر، والوظيفة، والحالة الاقتصادية، وحجم السكان بالإضافة إلى التوزيع الجغرافي للمجتمع.

إن المعاينة بالحصة تستخدم المسوحات الشاملة أو عينات كبيرة سحبت من المجتمع في مُددٍ سابقة متقاربة. لتقسيم المجتمع إلى طبقات والحصول على أوزان هذه الطبقات، يقوم العادون بسحب أشخاص من المجتمع بأي طريقة تعجبهم على شرط أن يحافظوا على الحصة المحددة لكل طبقة، أي مثلاً يجب أن يكون عدد الرجال في العينة 60% والنساء 40% أو أي صفة تستخدم كالعمر أو الوظيفة ...إلخ.

إن الفرق الرئيس بين العينة بالحصة والعينة العشوائية البسيطة أو الطبقية هو أن العينة بالحصة لا تستخدم العشوائية في اختيار الوحدات من المجتمع، لذلك ربما يقرر العادُّ إهمال أجزاء من المجتمع إذا كان لا يستطيع الوصول إليها بسهولة. لذلك من الناحية النظرية البحتة يمكن أن نقول إن العينة بالحصة تفتقر إلى الأساس العلمي، لأنه لا يدخل فيها عنصر العشوائية في الاختيار، أي اختيار وحدات العينة من المجتمع، لذلك لا نستطيع استخدام نتائجها دون الخوف من أن هنالك تحيزاً في اختيار وحدات العينة، ولكنها تستخدم كثيراً في أبحاث السوق واستطلاعات الرأي، وأحياناً تعطي نتائج قريبة من العينة الاحتمالية.

إن من أهم ميزات المعاينة بالحصة أنها سريعة، وتقلل التكاليف والجهد والوقت. لمزيد من المعلومات يراجع Cochran (1977) أو ترجمة أنيس كنجو (1995) و Barnett (1991) و Raj (1968).

المراجع العربية

1. وليم كوكوران – ترجمة أنيس كنجو .(1995) تقنية المعاينة الإحصائية، جامعة الملك سعود، الرياض.

References

1. Barnett, V. (1991). Sample Survey Principles and Methods, Oxford Uni. Press, New York.
2. Chaudhuri, A. and Stenger, H. (1992). Survey Sampling: Theory and Methods, Marcel Dekker, New York.
3. Cochran, W. G. (1977). Sampling Technique, 3rd Ed. Wiley, New York.
4. Deming, W. E. (1960). Sampling Design in Business Research. Wiley, New York
5. Foreman, E. K. (1991). Survey Sampling Principles, Marcel Dekker, New York.
6. Govindarajulu, Z. (1999). Elements of Sampling Theory and Methods, Prentice Hall.
7. Hajek, J (1981). Sampling from Finite Population, Marcel Dekker, New York.
8. Hansen, M. H., and Hurwitz, W. N. (1943). On the Theory of Sampling from Finite Populations. *Ann. Math. Statist.* 14, 333-362.
9. Hedayat, A. S. and Sinha, B. K. (1991). Design and Inference in Finite Population Sampling, Wiley, New York.
10. Kish, L (1995). Survey Sampling, Wiley, New York.
11. Levy, P. S. and Lemeshow S. (1991). Sampling of Population: Methods and Applications, Wiley, New York.
12. Lohr, S. L. (1999). Sampling: Design and Analysis, Duxbury, New York.
13. Raj, D. (1968). Sampling Theory, McGraw Hill, New York.
14. Sampath, S. (2000). Sampling Theory and Methods, CRC Press.
15. Scheaffer, R.L., Mendenhall, W. and Ott L. (1996). Elementary Sampling Survey, 5th ed., Duxbury, New York.
16. Singh, D. and Chaudhary, F. S. (1986). Theory and Analysis of Sample Survey Designs, Wily Eastern, New Delhi.
17. Sturat A. (1984). The Ideas of Sampling, Revised Edition, Griffin, London.
18. Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S., and Ashok, C. (1984). Sampling Theory of Surveys with Applications, 3rd ed., Ames (Iowa): Iowa State University Press.
19. Thompson, S. K (2002). Sampling, 2end Wiley, New York.
20. Warner, S. L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias, *J. Amer. Statist. Assoc.*, 60, 63-69.