

KINH NGHIỆM THIẾT KẾ MẪU CHỦ cho các cuộc điều tra hộ gia đình của Bangladesh

Cục Thống kê Bangladesh (BBS) là cơ quan chính phủ được ủy quyền thực hiện thu thập dữ liệu phục vụ biên soạn số liệu thống kê chính thức cho Bangladesh. Các cuộc điều tra hộ gia đình phạm vi toàn quốc là những công cụ thu thập dữ liệu ban đầu của BBS. Trước khi tiến hành Khảo sát lực lượng lao động (LFS) năm 2009-2010 thì các cuộc điều tra như Điều tra lao động việc làm 2005, Điều tra thu nhập và chi tiêu hộ gia đình (HIES) năm 2005 là các cuộc điều tra hộ gia đình cuối cùng được tiến hành bởi BBS và sử dụng các thiết kế mẫu tích hợp nhiều mục đích (Integrated Multi-Purpose Sampling Design, viết tắt là IMPS). Tuy nhiên, các nghiên cứu trước đó của Maligalig và Barcenas (2008) đã xác định thiếu sót trong kỹ thuật của IMPS, đặc biệt giá trị “hiệu quả thiết kế (deff)”. Bằng chứng là nghiên cứu một tiêu thức quan trọng được quan tâm trong cuộc điều tra, chỉ tiêu “tỷ lệ thất nghiệp” thì thấy rằng giá trị hiệu quả thiết kế (deff) lớn ở Vùng thống kê đô thị (SMA) và cho các khu vực lớn như Dhaka, Chittagong và Rajshahi, nguyên nhân là do các biện pháp phân tầng không hiệu quả. Ngoài ra, quyền số điều tra được sử dụng trong IMPS không phản ánh xác suất lựa chọn các hộ gia đình lấy ra đưa vào mẫu đã được áp dụng tại thời điểm điều tra. Hơn nữa, Maligalig và Barcenas (2008) cũng lưu ý rằng số lượng các hộ gia đình được lấy mẫu của mỗi đơn vị chọn mẫu cấp một (PSU) vẫn có thể giảm và số lượng của PSU tăng để giảm thiểu những hiệu quả thiết kế lớn. Do những vấn đề này, BBS yêu cầu hỗ trợ kỹ thuật từ Ngân hàng Phát

triển châu Á trong năm 2008 để phát triển một thiết kế mẫu mới và dàn mẫu chủ có thể được sử dụng cho cuộc Khảo sát lực lượng lao động giai đoạn 2009-2010 sắp tới. Hơn nữa, mục tiêu chính của hoạt động này là chuẩn bị cho thiết kế chọn mẫu khi dữ liệu từ Tổng điều tra Dân số năm 2011 có sẵn. Ngoài ra, các phiên bản được cập nhật sẽ được sử dụng như thiết kế mẫu chủ cho các cuộc điều tra hộ gia đình mà BBS sẽ tiến hành. Bài viết giới thiệu các quy trình đã được thực hiện để xây dựng thiết kế mẫu cho LFS năm 2009-2010 của BBS, đó cũng sẽ là cơ sở cho một thiết kế mẫu chủ mới. Mong muốn các thông tin đưa ra sẽ cung cấp, hướng dẫn kinh nghiệm cho các nhà điều tra thống kê từ các nước đang phát triển trong việc xây dựng một thiết kế mẫu chủ cho các cuộc điều tra hộ gia đình.

Bài viết đề cập các nội dung sau: (1) Hướng dẫn xác định các đơn vị chọn mẫu ban đầu (còn gọi là đơn vị chọn mẫu cấp 1, viết tắt là PSU) cho các cuộc điều tra hộ gia đình; (2) Thảo luận về sự phân tầng của điều tra như một công cụ để cải thiện độ chính xác của ước lượng điều tra; (3) Thảo luận về kế hoạch lựa chọn mẫu để có thể kiểm soát hiệu quả thiết kế của các cuộc điều tra phức tạp.

1. Hướng dẫn xác định các đơn vị chọn mẫu ban đầu

Đơn vị chọn mẫu ban đầu hay còn gọi là đơn vị chọn mẫu cấp 1 và được viết tắt là PSU. Đối với các cuộc điều tra hộ gia đình phạm vi toàn quốc ở các nước phát triển, chọn mẫu nhiều tầng thường phù hợp nhất cả về chi phí

và thiết kế kỹ thuật. Các hộ gia đình (hoặc hộ dân cư) là đơn vị chọn mẫu cuối cùng trong khi các đơn vị chọn mẫu cấp 1 (PSU) thường là cụm các hộ liền kề. Mặc dù chọn mẫu ngẫu nhiên đơn giản phân tầng có lẽ là hiệu quả nhất trong số các thiết kế chọn mẫu thông thường, nhưng nó không thiết thực và khả thi cho hầu hết các cuộc điều tra hộ gia đình đối với những nước đang phát triển vì danh sách cập nhật tất cả các hộ gia đình của cả nước thường không có sẵn. Nói chung, cần thiết phải có một dàn mẫu tốt để đảm bảo rằng mỗi đơn vị chọn mẫu cuối cùng có cùng cơ hội được lựa chọn, đảm bảo việc suy rộng kết quả tổng thể từ giá trị tính toán được từ mẫu.

Xây dựng dàn đơn vị chọn mẫu cấp 1 là bước đầu tiên trong việc xây dựng một thiết kế chọn mẫu nhiều giai đoạn. Điều quan trọng là quyết định một cách cẩn thận về cái gì được chọn lựa là đơn vị mẫu cấp 1, cần có một số cân nhắc, đảm bảo rằng tất cả các đơn vị trong tổng thể mục tiêu thuộc một và chỉ một đơn vị chọn mẫu cấp 1. Để đạt được mục đích này, đơn vị mẫu cấp 1 phải có ranh giới rõ ràng. Ngoài ra, thông tin phụ về "kích cỡ" của một PSU sẽ được sử dụng để lựa chọn đơn vị sẽ được chọn vào mẫu. Nếu tổng số hộ gia đình được sử dụng như là cách tính kích cỡ của PSU, thì làm thế nào để PSU càng nhỏ càng tốt nhưng cũng phải đảm bảo đủ số lượng mẫu cuối cùng đặt ra. Dàn mẫu phải thiết kế sao cho có thể thuận lợi cho việc chọn mẫu quay vòng đối với các cuộc điều tra khác nhau là một yêu cầu quan trọng khi thiết kế dàn mẫu chủ. Hơn nữa, thông tin sẵn có cũng được sử dụng cho vấn đề phân tầng và phân bổ mẫu, hoặc đưa ra cân nhắc trong việc tạo ra các PSU.

Ở Bangladesh, các đơn vị được ngầm có khả năng để chọn làm PSU là: liên bang

(unions)², mauza³, thôn và địa bàn điều tra theo quy định tại Tổng điều tra dân số năm 2001 (enumeration areas - EA). Tuy nhiên, phân tích sơ bộ cho thấy rằng các "union" rất khác nhau về kích thước và nói chung, chúng quá lớn để quản lý phạm vi hoạt động. Đối với "Thôn", một mặt nó gần như giống với EA đối với khu vực nông thôn, nhưng ranh giới của nó không rõ ràng đối với ở khu vực đô thị. Với thông tin này, chỉ có "mauza" và EA được xem xét chọn làm đơn vị mẫu cấp 1 trong các thảo luận tiếp theo. Sử dụng dữ liệu từ Tổng điều tra Dân số năm 2001, Bảng 1 và Bảng 2 tóm tắt sự phân bố của số hộ theo mauza và các EA tương ứng.

Nhìn vào Bảng 1 và Bảng 2 bên dưới thấy rằng, phân theo "mauza" tổng số hộ gia đình tương đối chênh lệch nhau, nằm trong khoảng 1 đến 22.366 (khu đô thị khác, Dhaka). Nếu mauza được chỉ định là đơn vị mẫu cấp 1, sau đó một số mauza sẽ vẫn phải được phân nhỏ để đảm bảo rằng mỗi đơn vị mẫu cấp 1 sẽ không được lựa chọn nhiều hơn một lần. Đồng thời, một số mauza có thể cần phải được kết hợp với nhau để đảm bảo rằng có đủ số hộ gia đình có thể được rút ra từ mỗi đơn vị mẫu cấp 1. Ngược lại, khoảng cách về số hộ giữa EA nhỏ nhất và lớn nhất nhỏ hơn, nếu chọn đơn vị chọn mẫu cấp 1 là đơn vị địa bàn (EA), nhiều địa bàn vẫn cần phải được chia nhỏ hơn nữa vì tổng hộ gia đình của địa

² Union: Là đơn vị hành chính của chính phủ và các địa phương nông thôn nhỏ ở Bangladesh. Mỗi liên bang được tạo thành bởi 9 phường. Thông thường, một ngôi làng được chỉ định là một phường. Có 4.550 liên bang ở Bangladesh.

³ Mauza: Còn được gọi là mouza là một loại huyện hành chính, tương ứng với diện tích đất cụ thể trong đó có thể có một hoặc nhiều hơn các khu định cư. Trước thế kỷ 20, thuật ngữ gọi là đơn vị thu ngân trong một pargana hoặc huyện. Do dân số tăng và thôn đã trở thành phổ biến hơn và phát triển, khái niệm của mouza giảm tầm quan trọng. Ngày nay, nó tương đương với khái niệm "gam" hoặc làng.

bản lớn nhất là 497 (Bảng 2). Xem xét nhiều là một lựa chọn tốt hơn so với chọn mauzas yếu tố khác, đơn vị mẫu cấp 1 sử dụng địa bàn làm đơn vị mẫu cấp 1.

Bảng 1: Số liệu tóm tắt số hộ gia đình của Bangladesh phân theo làng

Vùng	Số lượng (làng)	Số lượng làng phân theo từng khu vực (làng)		Phân bố số lượng hộ gia đình trong làng (hộ)					
				Tổng số	Nhỏ nhất	Trung vị	Trung bình	Lớn nhất	Độ lệch chuẩn
Barisal	3,414	Nông thôn	2,896	1,411,766	1	321	487.49	5,126	481.19
		Khu đô thị	419	144,911	11	249	345.85	3,196	333.86
		Vùng đô thị lớn (SMA) ⁴	99	91,408	26	718	923.31	4,342	794.92
		Khu đô thị khác	-	-	-	-	-	-	-
Chittagong	8,879	Nông thôn	7,367	3,317,141	1	258	450.27	11,943	600.50
		Khu đô thị	1,175	743,076	1	284	632.41	9,831	1,130.94
		Vùng đô thị lớn	207	257,432	9	794	1,243.63	8,045	1,398.13
		Khu đô thị khác	130	154,899	10	823	1,191.53	5,328	1,208.63
Dhaka	18,295	Nông thôn	14,660	5,399,312	1	219	368.30	8,820	473.40
		Khu đô thị	2,616	1,824,745	1	303	697.53	9,218	1,175.43
		Vùng đô thị lớn	289	254,248	1	576	879.75	5,325	952.55
		Khu đô thị khác	730	758,382	1	316	1,038.88	22,366	2,283.58
Khulna	7,483	Nông thôn	6,300	2,472,098	1	264	392.04	5,119	422.51
		Khu đô thị	913	433,156	1	307	474.43	5,823	544.95
		Vùng đô thị lớn	166	131,468	51	583	791.98	4,101	705.05
		Khu đô thị khác	105	82,880	1	408	789.33	4,938	1,015.84
Rajshahi	18,887	Nông thôn	16,423	5,643,537	1	221	343.64	5,758	382.17
		Khu đô thị	1,951	645,620	1	232	330.92	2,597	304.51
		Quận	340	280,392	5	588	824.68	4,042	703.76
		Khu đô thị khác	173	58,248	1	278	336.69	2,026	301.33
Sylhet	5,708	Nông thôn	4,989	1,213,085	1	167	243.15	3,052	256.24
		Khu đô thị	608	110,982	1	136	182.54	1,328	166.86
		Vùng đô thị lớn	111	64,155	20	427	577.97	2,865	559.31
		Khu đô thị khác	-	-	-	-	-	-	-

Nguồn: Tính toán của Author sử dụng dữ liệu Tổng điều tra dân số năm 2001 được tiến hành bởi BBS

⁴ SMA - Statistical metropolitan areas: Là Vùng đô thị lớn, với mật độ dân số tương đối cao và có quan hệ kinh tế chặt chẽ trong toàn khu vực nhưng không được tách riêng thành một thành phố, nó bao gồm các quận và những đơn vị tương đương quận.

Bảng 2: Số liệu tóm tắt số hộ gia đình của Bangladesh phân theo đơn vị địa bàn

Vùng	Số lượng địa bàn (địa bàn)		Hộ gia đình phân theo đơn vị địa bàn (hộ)					
			Tổng	Nhỏ nhất	Trung vị	Trung bình	Lớn nhất	Độ lệch chuẩn
Barisal	Nông thôn	14,473	1,411,766	1	96	97.54	354	25.58
	Khu đô thị	1,573	144,911	1	88	92.12	233	29.94
	Vùng đô thị lớn	898	91,408	2	98	101.79	267	28.13
	Khu đô thị khác	-	-	-	-	-	-	-
Chittagong	Nông thôn	36,172	3,317,141	1	94	91.70	321	34.25
	Khu đô thị	7,943	743,076	1	92	93.55	339	31.73
	Vùng đô thị lớn	2,997	257,432	1	87	85.90	237	39.46
	Khu đô thị khác	1,428	154,899	2	107	108.47	310	35.52
Dhaka	Nông thôn	14,660	5,399,312	1	219	368.30	483	31.38
	Khu đô thị	18,819	1,824,745	1	93	96.96	471	37.97
	Vùng đô thị lớn	2,418	254,248	1	102	105.15	404	36.49
	Khu đô thị khác	7,030	758,382	1	100	107.88	478	43.49
Khulna	Nông thôn	23,530	2,472,098	1	104	105.06	320	30.28
	Khu đô thị	3,998	433,156	1	103	108.34	344	35.53
	Vùng đô thị lớn	1,187	131,468	8	108	110.76	239	31.07
	Khu đô thị khác	744	82,880	1	106	111.40	305	34.66
Rajshahi	Nông thôn	55,004	5,643,537	1	101	102.60	463	29.94
	Khu đô thị	6,707	645,620	1	93	96.26	497	35.07
	Vùng đô thị lớn	2,639	280,392	1	103	106.25	286	33.12
	Khu đô thị khác	546	58,248	1	104	106.68	295	34.59
Sylhet	Nông thôn	14,875	1,213,085	1	84	81.55	258	36.29
	Khu đô thị	1,302	110,982	1	84.5	85.24	276	39.11
	Vùng đô thị lớn	723	64,155	1	90	88.73	258	37.99
	Khu đô thị khác	-	-	-	-	-	-	-

Nguồn: Tính toán của Author sử dụng dữ liệu Tổng điều tra dân số năm 2001 được tiến hành bởi BBS

Như đã đề cập trước đó, mong muốn mọi PSU phải đủ lớn để có đủ số lượng đơn vị mẫu cuối cùng để đảm bảo tính khả thi của việc áp dụng một thiết kế mẫu luân phiên cho các cuộc điều tra khác nhau, sẽ áp dụng nhiều khi thiết kế mẫu chủ. Trong trường hợp của Bangladesh, nhóm chuyên gia thiết lập các

ngưỡng để có 40 hộ gia đình mỗi PSU. Trong số 259.828 địa bàn, 12.273 địa bàn có ít hơn 40 hộ gia đình. Những địa bàn nhỏ nên được xem xét sát nhập. Khi kết hợp địa bàn nhỏ để tạo thành PSU, các địa bàn điều tra kết hợp phải liền kề nhau. Tuy nhiên, bản đồ địa lý những địa bàn sự thiếu tin cậy, nhóm chuyên

gia quyết định kết hợp các địa bàn điều tra nhỏ dựa trên các tiêu chí giới thiệu ngay dưới đây. Ngoài ra, do các vấn đề về khái niệm và chuẩn bị trong việc phân loại các Vùng thống kê đô thị (SMA) và các khu đô thị khác, quyết định rằng hai lĩnh vực này sẽ được phân loại theo khu vực đô thị thay thế.

Tiêu chuẩn để sát nhập địa bàn điều tra (EA) để tạo thành một đơn vị chọn mẫu cấp 1

1. Một EA có hơn 40 hộ gia đình là một PSU.

2. EA nhỏ được gắn vào một EA liền kề thuộc về việc phân loại thành thị/nông thôn và mauza.

3. EA nhỏ duy nhất trong một mauza có thể được kết hợp với một EA của một mauza khác miễn là cả hai mauzas thuộc về cùng một "union" và hai EA được kết hợp đó phải thuộc cùng loại đô thị hoặc loại nông thôn.

Theo tiêu chuẩn này, có 248,904 PSU đã được xây dựng từ 259,828 EA ban đầu. Bảng 3 cung cấp sự phân bố số lượng hộ gia đình trong các PSU.

Bảng 3: Số lượng hộ gia đình ở Bangladesh theo đơn vị chọn mẫu cấp 1

Vùng	Số lượng đơn vị chọn mẫu cấp 1 phân theo Khu vực (đơn vị)		Hộ gia đình phân theo đơn vị chọn mẫu cấp 1 (hộ)					
			Tổng	Tối thiểu	Trung vị	Trung bình	Tối đa	Độ lệch chuẩn
Barisal	Nông thôn	14,280	1,411,766	41	97	98.86	354	24.30
	Thành thị	2,414	236,319	42	94	97.90	267	27.36
Chittagong	Nông thôn	33,721	3,317,141	41	97	98.37	321	28.17
	Thành thị*	11,810	1,155,407	23	95	97.84	339	30.98
Dhaka	Nông thôn	52,667	5,399,312	19	100	102.52	483	27.88
	Thành thị*	27,317	2,837,375	21	98	103.88	478	36.93
Khulna	Nông thôn	22,886	2,472,098	31	105	108.02	320	27.01
	Thành thị	5,823	647,504	41	105	111.20	344	32.69
Rajshahi	Nông thôn	53,554	5,643,537	41	102	105.38	463	27.28
	Thành thị	9,614	984,260	13	98	102.38	497	32.20
Sylhet	Nông thôn	12,992	1,213,085	41	92	93.37	266	29.79
	Thành thị	1,826	175,137	21	93	95.91	296	33.27

Ghi chú: * - Vùng có 3 đơn vị chọn mẫu cấp 1 (PSU) có rất ít số hộ gia đình, được xác định trên cơ sở dữ liệu điều tra dân số mới nhất. Đặc biệt có 1 PSU thuộc vùng Chittagong (thành thị) và hai PSU thuộc vùng Dhaka (đô thị) có ít hơn 10 hộ gia đình. Những đơn vị này không đưa vào Bảng trên.

Như được hiển thị trên bảng, còn có những PSU nhỏ hơn 40 hộ gia đình. Những trường hợp này ở các "union" là rất nhỏ về số lượng hộ gia đình. Có khoảng 11PSU như vậy, nhóm chuyên gia quyết định loại trừ các PSU

này khỏi dàn mẫu.

Như vậy, cần nghiên cứu lý luận phù hợp thực tiễn của từng quốc gia để lựa chọn đơn vị chọn mẫu cấp 1, số lượng đơn vị mẫu cấp 1

khá quan trọng sẽ quyết định số lượng mẫu của toàn cuộc điều tra.

2. Sự phân tầng điều tra, công cụ để cải thiện độ chính xác của ước lượng điều tra

Tầng trong một cuộc điều tra là một bộ phận của tổng thể mà ở đó việc chọn mẫu là riêng biệt, được lên kế hoạch, chỉ định và lựa chọn cụ thể (Kish, 1987). Lựa chọn tầng ở từng cuộc điều tra phụ thuộc vào nhiều yếu tố như: Yêu cầu báo cáo, thiết kế chọn mẫu và quan trọng hơn, ngân sách được cấp và khối lượng công việc sẽ được dự kiến làm (Kish, 1965; 1987). Cả những phát sinh về phương pháp luận thống kê và phát sinh trong thực tế phải được xem xét khi chỉ định các tầng. Nói chung, yêu cầu quan trọng nhất là cho các giá trị thống kê phải ở mức độ "mịn" (Elbers, Lanjouw và Lanjouw 2003). Muốn tăng độ mịn của tầng thì yêu cầu tăng số lượng tầng. Do cỡ mẫu thường được xác định ở cấp độ tầng, tăng số lượng tầng nhất thiết sẽ kéo theo việc tăng cỡ mẫu chung. Cỡ mẫu được điều chỉnh bởi giá trị hiệu quả thiết kế. Thông thường, khi thiết kế mẫu sử dụng phương pháp chọn mẫu chùm để xác định mẫu cuối cùng. Hiệu quả thiết kế trung bình cho các mẫu chùm kỳ vọng sẽ lớn hơn hoặc bằng 3, do đó, cỡ mẫu cuối cùng sẽ phải tăng lên bởi giá trị này. Tuy nhiên, những điều này cũng được xem xét trong bối cảnh ngân sách được cấp cho thu thập dữ liệu điều tra là bao nhiêu.

Một khi các tầng đã được quy định rõ ràng, cỡ mẫu cho mỗi tầng sẽ được xác định sao cho các ước lượng suy ra đủ tin cậy trong phạm vi tầng. Thông tin về sự biến động của các đơn vị lấy mẫu trong mỗi tầng, sai số chấp nhận được và các chi phí liên quan là những yếu tố cần thiết để xác định cỡ mẫu. Ví dụ, giả sử các tiêu thức quan tâm là tiêu thức tỷ lệ.

Chọn mẫu ngẫu nhiên đơn giản (SRS), cỡ mẫu dự kiến cho một tầng nào đó được tính dựa vào công thức.

$$n_{SRS} = \frac{t_{(\alpha, N-1)}^2 P(1-P)}{1 + \frac{1}{N} \left(\frac{t_{(\alpha, N-1)}^2 P(1-P)}{d^2} - 1 \right)} \quad (1)$$

Trong đó $t_{(\alpha, N-1)}$ là giá trị tới hạn của phân phối t với độ tin cậy $1 - \alpha$, và bậc tự do N (cỡ tổng thể); P là tỷ lệ thực tế của tiêu thức quan tâm và phạm vi sai số d (Cochran, 1977). Nếu P chưa biết, ta có thể coi như $P = 0,5$ hoặc sử dụng bất kỳ thông tin đã có về giá trị của P dựa vào các nghiên cứu trước đó. Lưu ý rằng xác định $P = 0,5$ cho ta cỡ mẫu bảo toàn nhất hay là lớn nhất. Các kết quả cỡ mẫu thu được từ việc áp dụng (1) sau đó được tăng lên bởi giá trị hiệu quả thiết kế ($deff$), giả định biết trước về giá trị $deff$.

$$n_{complex} = Deff * n_{SRS} \quad (2)$$

Trong trường hợp của Bangladesh, các phân khu địa lý⁵ đã được chỉ định là các tầng cụ thể. Nếu 64 zilas⁶ (tỉnh) được quy định là 64 tầng, cỡ mẫu sau khi điều chỉnh $deff$ sẽ vượt quá ngân sách của BBS. Hơn nữa, nhóm chuyên gia sử dụng số liệu tỷ lệ thất nghiệp ước tính qua điều tra LFS năm 2005 để cho giá trị P cụ thể. Bảng 4 cho thấy cỡ mẫu dự kiến đã được tính toán với độ tin cậy $(1 - \alpha) = 0.95$ và phạm vi sai số d . Các hiệu quả thiết kế tương ứng với tỷ lệ thất nghiệp từ LFS năm 2005 cũng được thể hiện trong Bảng 4. Lưu ý rằng với $d = 0.01$, cỡ mẫu là 115.277. Trong khi đó, cỡ mẫu là 100.000 hộ gia đình thì đã nhiều hơn số hộ gia đình của điều tra LFS

⁵ Phân khu địa lý: Hay còn gọi là Vùng, ở Bangladesh có 6 phân khu là: Barisal, Chittagong, Dhaka, Khuha, Rajshahi, Sylhet

⁶ Zilas: Tên tiếng Bangladesh, tương đương tỉnh.

2009-2010 được BBS phân bổ ngân sách để thu thập. Nếu chọn $d = 0,03$ (có thể không phù hợp nếu tỷ lệ thất nghiệp là khá nhỏ) thì cỡ mẫu là khoảng 12.814 hộ gia đình, nằm trong vòng ngân sách. Đối với khu vực Dhaka và Khulna, cỡ mẫu trở nên rất lớn do ảnh hưởng của hiệu quả thiết kế. Các biến đổi lớn được quan sát thấy trong những khu vực có thể không thực sự phản ánh khác nhau lớn về số hộ gia đình nhưng các biến đổi lớn về quyền sở giả đã được quy cho các hộ gia đình đó. Với thực trạng này, cỡ mẫu được tính toán trong Bảng 4 chỉ được sử dụng như hướng dẫn để xác định cỡ mẫu. Đặc biệt, nhóm chuyên gia đề xuất để lấy mẫu trong mỗi PSU là 10 hộ gia đình theo Maligalig và Barcenias (2008) thay cho 40 hộ gia đình mỗi PSU trong IMPS. Điều này cho phép chúng ta tăng số lượng PSU lấy mẫu từ 1000 (trong IMPS) lên đến 1500 khi thiết kế mẫu mới. Xét thấy rằng những tương quan chặt chẽ giữa các hộ gia đình ở trong nội bộ các PSU, tăng số lượng PSU và giảm số

lượng hộ gia đình trong mỗi PSU được coi là hợp lý.

Nếu quyền số điều tra được sử dụng để tính toán kích thước mẫu trong Bảng 4 đã chính xác, các ước tính ở mức độ phân khu địa lý có sai số khoảng 0.03. Mức độ sai số là khá lớn, không chấp nhận được, trong khi tỷ lệ thất nghiệp theo vùng chỉ thay đổi từ 0.01 (Sylhet) đến 0.06 (Barisal). Mặt khác, do quyền số điều tra trong LFS năm 2005 có lỗi kỹ thuật và sử dụng giải pháp phân tầng không hiệu quả trong việc kiểm soát các hiệu quả thiết kế, kết quả ước lượng từ sử dụng thiết kế mẫu chủ của LFS 2009-2010 có thể làm cho hiệu quả thiết kế được chấp nhận thậm chí chỉ với tổng cỡ mẫu là 15.000 hộ gia đình. Kết quả thuận lợi này phụ thuộc vào chất lượng của việc xây dựng một thiết kế mẫu chủ, tiêu chuẩn định rõ các quyền số điều tra chính xác và cách phân tầng tốt hơn.

Bảng 4: Kích thước mẫu dự kiến

Vùng	Tỷ lệ thất nghiệp (%)	Số hộ gia đình (hộ)	Giá trị Deff	Cỡ mẫu theo chọn mẫu ngẫu nhiên giản đơn			Cỡ mẫu sau hiệu chỉnh		
				d = 0.05	d = 0.03	d = 0.01	d = 0.05	d = 0.03	d = 0.01
Barisal	0.0622	1,648,085	5.12	89.57	248.77	2236.24	460.31	1278.51	11492.75
Chittagong	0.0461	4,472,548	8.38	67.51	187.53	1687.17	567.31	1575.81	14177.52
Dhaka	0.0474	8,236,687	27.00	69.37	192.70	1733.99	1878.49	5217.95	46952.74
Khuha	0.0545	3,119,602	18.58	79.18	219.92	1978.19	1475.64	4098.83	36868.64
Rajshahi	0.0311	6,627,797	3.41	46.26	128.51	1156.41	158.07	439.08	3951.07
Sylhet	0.0182	1,388,222	2.66	27.53	76.47	687.90	73.40	203.88	1834.07
Tổng							4613.22	12814.05	115276.79

Nguồn: Nghiên cứu của Authors sử dụng dữ liệu từ LFS năm 2005 tiến hành bởi BBS.

Một số chiến lược phân bổ đã được kiểm tra để phân bổ 15.000 hộ mẫu theo địa bàn như: phân bổ đều, phân bổ theo tỷ lệ, phân bổ tỷ lệ với căn bậc và phân bổ Kish.

Phân bổ đều:

$$n_d = \frac{n}{D} = \frac{n}{6}$$

Phân bổ theo tỷ lệ:

$$n_d = n \frac{N_d}{N} = nW_d$$

Phân bố tỷ lệ căn bậc hai:

$$n_d = \frac{\sqrt{N_d}}{\sum_m \sqrt{N_m}}$$

Phân bố Kish:

$$n_d = n \frac{\sqrt{D^{-2} + IW_m^2}}{\sum_m \sqrt{D^{-2} + IW_m^2}} = n \frac{\sqrt{\frac{1}{36} + IW_m^2}}{\sum_m \sqrt{\frac{1}{36} + IW_m^2}}$$

Trong đó: n_d là cỡ mẫu trong tầng, n là tổng số mẫu, D là số lượng tầng, N_d là tổng số hộ gia đình trong tầng d , N là tổng số hộ gia đình ở Bangladesh. Tổng điều tra dân số năm 2001, W_d là tỷ lệ các hộ gia đình trong tầng d , I là chỉ số phân bố Kish biểu thị sự liên quan giữa các ước lượng tại một quốc gia hoặc các phân nhóm được cắt bởi tầng (loại (i)), liên quan giữa các ước lượng ở cấp tầng (loại (ii)). Để minh họa, chúng ta có thể liên hệ loại (i) với tiêu thức quan tâm như số lượng nông dân trồng trọt và người lao động nữ không được trả lương, tỷ lệ người nghèo ở Bangladesh, số người trong lực lượng lao động, người đang thất nghiệp, tỷ lệ hộ có điện, và ước lượng về sự khác nhau giữa các phân nhóm. Khi tính toán ở cấp độ tầng,

chúng trở thành các tham số loại (ii). Nếu mỗi quan tâm chính là để lấy được ước lượng cho tiêu thức quan tâm của loại (ii), thì phương pháp tốt nhất để phân bố tổng số mẫu là sử dụng phân bố mẫu tỷ lệ với quy mô dân số của mỗi tầng. Tuy nhiên, cách tiếp cận lý tưởng cho loại (ii) là phân chia tổng số mẫu đều giữa các tầng (Kish, 1987). Hơn nữa, cần phải nhấn mạnh rằng hai phương pháp này có thể mang lại phân bố mẫu rất đa dạng đặc biệt là khi các vùng khác nhau về quy mô. Bên cạnh đó, có một cách tiếp cận hiệu quả hơn khi muốn ước lượng một loại nhất định của đặc trưng quan tâm nhưng không nhất thiết phải cho các loại khác đó là sử dụng phân bố Kish, về bản chất là sự kết hợp giữa phân bố đều và phân bố tỷ lệ. Với $I = 0$, nó làm giảm các phân bố đều trong khi có xu hướng tiếp cận phân bố tỷ lệ với $I \rightarrow \infty$. Bảng 5 cung cấp các ước lượng cỡ mẫu cho mỗi tầng bằng cách sử dụng các phân bố khác nhau. Phân bố Kish tại $I = 1$ đã được lựa chọn để đảm bảo rằng độ chính xác của cả hai loại (i) và loại (ii) và tiêu thức quan tâm sẽ là xấp xỉ như nhau.

Bảng 5: Phân bố mẫu số lượng hộ gia đình cho mỗi tầng

Vùng	Tổng số hộ gia đình (hộ)	W_d	Phân bố đều (hộ)	Phân bố tỷ lệ (hộ)	Phân bố căn bậc 2 (hộ)	Phân bố Kish (hộ)
Barisal	1,648,085	0.064649	2,500	969.7	1,633.65	1,817.68
Chittagong	4,472,548	0.175443	2,500	2,631.6	2,691.21	2,460.51
Dhaka	8,236,687	0.323097	2,500	4,846.4	3,652.13	3,696.56
Khulna	3,119,602	0.122371	2,500	1,835.5	2,247.60	2,102.39
Rajshahi	6,627,797	0.259986	2,500	3,899.7	3,276.08	3,140.06
Sylhet	1,388,222	0.054455	2,500	816.8	1,499.34	1,782.81
Bangladesh	25,492,941	1.000000	15,000	15,000.0	15,000.00	15,000.00

Nguồn: Tính toán của Authors sử dụng thủ tục phân bố mẫu không đồng đều

(Xem tiếp trang 7)

(Tiếp theo trang 27)

3. Kinh nghiệm lựa chọn mẫu để có thể kiểm soát hiệu quả thiết kế của các cuộc điều tra phức tạp

3.1 Phân lớp ngầm trong các PSU

Sự phân lớp (ngầm) trong PSU là rất quan trọng để đảm bảo rằng (giới hạn) cỡ mẫu được cấp bởi BBS vẫn sẽ làm cho các ước lượng đáng tin cậy ở cấp PSU và cấp được phân chia bởi PSU. Thông tin để tạo ra các lớp trong PSU gọi là đơn vị phân lớp ngầm. Lý tưởng nhất là đơn vị phân lớp ngầm có sẵn và được đo lường thống nhất cho tất cả các PSU trong tầng. Ví dụ, đơn vị phân lớp là thông tin địa lý như zila (tỉnh) và các khu vực thành thị/nông thôn vì mỗi PSU mang mã tỉnh hoặc

được phân loại khu vực đô thị. Phân lớp khác có thể được áp dụng để đảm bảo rằng các lớp cuối cùng của PSU là đồng nhất hơn. Các ứng cử cho đơn vị phân lớp sẵn có để sử dụng cho tất cả các PSU là những biến có trong Tổng điều tra dân số năm 2001. Ngoài ra, đơn vị phân lớp xem hiệu quả là đơn vị có quan hệ tương quan cao với những chỉ tiêu chủ yếu được quan tâm trong cuộc điều tra đó. Ví dụ, những chỉ tiêu được coi là có tương quan với thu nhập và việc làm là những chỉ tiêu chính trong LFS bao gồm: Tỷ lệ hộ có nhà ở vật liệu tốt (PStrong), tỷ lệ hộ gia đình có sản xuất nông nghiệp là nguồn thu nhập chính (PAgri); và tỷ lệ các hộ gia đình sở hữu đất nông nghiệp (POal).

(Còn nữa)

Vũ Văn Anh, Trần Thị Thu Hằng (lược dịch)

Nguồn: Developing a Master Sample Design for Household Surveys in Developing Countries: A Case Study in Bangladesh; Dalisay S. Maligalig and Arturo Martinez Jr; 12.07.2013 (<http://surveyinsights.org/?p=2151>)