

Chọn mẫu trong các nghiên cứu khảo sát dựa trên nền tảng web trong lĩnh vực truyền thông

TS BÙI THU HƯƠNG

Học viện Báo chí và Tuyên truyền; Email: dangmyhanhajc@gmail.com

Nhận ngày 15 tháng 2 năm 2022; chấp nhận đăng tháng 7 năm 2022.

Tóm tắt: Các nghiên cứu khảo sát dựa trên nền tảng web càng ngày càng trở nên phổ biến hơn trong các ngành khoa học xã hội và nghiên cứu truyền thông. Sự chấp nhận nhanh chóng của một chiến lược nghiên cứu quen thuộc trên một nền tảng mới này đã thách thức những hiểu biết vốn có về phương pháp luận và đặc biệt là cách thức chọn mẫu. Trong khuôn khổ nhận thức này, bài viết tập trung phân tích về một số vấn đề liên quan đến cách thức chọn mẫu trong các nghiên cứu khảo sát dựa trên nền tảng web trong lĩnh vực truyền thông.

Từ khóa: nghiên cứu khảo sát dựa trên web; nghiên cứu truyền thông; chọn mẫu; chọn mẫu xác suất; chọn mẫu phi xác suất.

Abstract: Web-based surveys are becoming more and more popular in the social sciences and media studies. The rapid adoption of this familiar research method on a new platform has challenged common knowledge in methodology and sampling. Within this cognitive framework, the article focuses on analyzing some issues related to sampling in web-based surveys in the field of media.

Keywords: web-based survey research; media studies; sampling; probability sampling; non-probability sampling.

Nghiên cứu khảo sát hay nghiên cứu mô tả là một dạng nghiên cứu khá được ưa chuộng trong nhiều ngành khoa học bao gồm cả nghiên cứu truyền thông. Đây là một hình thức nghiên cứu thu thập thông tin sử dụng các bảng hỏi tiêu chuẩn hoá (sử dụng các câu hỏi giống nhau, thường là đóng, cho tất cả những người tham gia) và người tham gia có thể là một phần hoặc toàn bộ một nhóm dân cư mục tiêu (khách thể nghiên cứu). Chính sự xuất hiện và áp dụng chiến lược chọn mẫu xác suất giúp chọn ra một phần của một nhóm dân cư mục tiêu vào những năm 30 của thế kỷ XX đã khiến cho nghiên cứu khảo sát càng trở nên phổ biến. Đặc biệt, cùng với sự phát triển và phổ biến của Internet, công nghệ di động, các ứng dụng và các phần mềm hỗ trợ thu thập và lưu trữ thông tin, nghiên cứu khảo sát và đặc biệt là nghiên cứu khảo sát dựa trên nền tảng web càng ngày càng chứng tỏ tầm quan trọng của nó.

Tuy nhiên, sự phát triển nhanh chóng của dạng nghiên cứu này cũng đặt ra không ít vấn đề về phương pháp luận. Bên cạnh những quan ngại về thiết kế bảng hỏi, chi phí, quản lý, tỷ lệ trả lời, điều chỉnh và xử lý số liệu, nhà nghiên cứu cũng gặp không ít khó khăn khi đối diện với các câu hỏi liên quan đến việc chọn mẫu như thế nào để tránh sai sót, bảo đảm độ ổn định và độ chuẩn xác của số liệu. Trong phần tiếp theo, bài viết giới thiệu khái niệm, phân loại và một số vấn đề liên quan đến chọn mẫu trong các nghiên cứu khảo sát dựa trên nền tảng web trong lĩnh vực truyền thông.

1. Khái niệm và phân loại chọn mẫu trong nghiên cứu khảo sát dựa trên nền tảng web

Khảo sát dựa trên nền tảng web (khảo sát trên web) sử dụng một bảng câu hỏi được truy cập qua trình duyệt Internet (chẳng hạn như Internet Explorer của Microsoft hoặc Firefox của Mozilla).

Các khảo sát này thường được tạo và vận hành thông qua phần mềm khảo sát trực tuyến chuyên dụng (chẳng hạn như các sản phẩm của các công ty như Conconfirm, Nebo hoặc Voxco)⁽¹⁾. Cũng giống như các nghiên cứu khảo sát truyền thống, nghiên cứu khảo sát dựa trên nền tảng web (gọi tắt là khảo sát dựa trên web) được thực hiện nhằm thu thập thông tin về một nhóm dân số nào đó. Một khảo sát dựa trên web đồng nghĩa với khả năng thực hiện với toàn bộ một nhóm dân số cao hơn so với các khảo sát truyền thống, tuy nhiên việc này không phải lúc nào cũng khả thi nhất là khi ngân sách nghiên cứu hạn hẹp và không phải lúc nào nhóm nghiên cứu cũng có thể xác định được tất cả các thành viên của nhóm dân số đó.

Chính vì vậy, trên thực tế, các khảo sát dựa trên web cũng thường chỉ làm việc với một mẫu hay một bộ phận được lựa chọn từ nhóm dân số cần nghiên cứu, và từ đó có thể suy rộng ra cho toàn bộ nhóm dân số đó. Nhóm dân số cần nghiên cứu còn được gọi là tổng thể/quần thể nghiên cứu. Việc chọn một bộ phận của nhóm dân số tham gia khảo sát như vậy gọi là chọn mẫu. Để có thể suy rộng ra tổng thể, mẫu được lựa chọn cần phải mang tính đại diện cho nhóm dân số đó để từ đó nhà nghiên cứu có thể tính toán và dự tính được các sai số tiêu chuẩn của các số liệu thống kê mẫu phù hợp. Tính đại diện ở đây không nên chỉ hiểu đơn thuần là mẫu lựa chọn thoả mãn các đặc điểm có thể quan sát được của quần thể nghiên cứu. Hơn thế, đó là kết quả số liệu mà nhà nghiên cứu thu thập được từ mẫu này trùng khớp với kết quả mà nhà nghiên cứu có thể thu thập được nếu thực hiện nghiên cứu với toàn bộ nhóm dân số.

Khi thiết kế các khảo sát trên web, nhà nghiên cứu có thể lựa chọn sử dụng hai hình thức chọn mẫu điển hình là chọn mẫu theo xác suất và phi xác suất. *Thứ nhất*, mẫu theo xác suất là mẫu trong đó những người trả lời được chọn bằng cách sử dụng một cơ chế xác suất nào đó giúp xác định trước xác suất được lựa chọn vào mẫu của các thành viên/phần tử trong quần thể nghiên cứu. Cơ hội được chọn vào mẫu của mỗi phần tử trong tổng thể là như nhau. Quá trình chọn mẫu theo xác suất tuân theo một quy luật và không thể tự ý thay đổi. Các tham số của mẫu có thể được dùng để ước lượng các thông số của

tổng thể. Các chiến lược chọn mẫu theo xác suất hay chọn mẫu ngẫu nhiên bao gồm: 1) Chọn mẫu ngẫu nhiên đơn giản, 2) Chọn mẫu ngẫu nhiên hệ thống, 3) Chọn mẫu phân tầng và 4) Chọn mẫu phân cụm⁽²⁾⁽³⁾. Thông thường khi có được danh sách các đơn vị lấy mẫu có sẵn hay còn gọi là khung lấy mẫu thì việc chọn mẫu phục vụ cho khảo sát trực tuyến hay trực tiếp truyền thống hoàn toàn như nhau. Tuy nhiên, với các khảo sát trực tuyến, nếu không có thông tin gì thêm về nhóm khách thể nghiên cứu ngoài địa chỉ email, các nhà nghiên cứu thường sử dụng các dạng chọn mẫu ngẫu nhiên đơn giản. Các phương pháp chọn mẫu phức tạp hơn như phân tầng và phân cụm đều cần thêm nhiều thông tin hơn về nhóm khách thể.

Các khảo sát trên web áp dụng các chiến lược chọn mẫu xác suất này bao gồm: 1) Các khảo sát chặn hay khảo sát xuất hiện bất ngờ dành cho những người truy cập một/nhiều trang web nào đó có số thứ tự cách nhau một khoảng cách hằng định (bước nhảy k) do nhà nghiên cứu mặc định trước; 2) Các khảo sát tuyển dụng người tham gia là những người có phản hồi đối với thư mời tham gia nghiên cứu được gửi qua email; 3) Các khảo sát sử dụng kết hợp nhiều hình thức, trong đó người tham gia có thể lựa chọn tham gia khảo sát trực tiếp (giấy bút) hoặc trực tuyến và 4) Các khảo sát với các nhóm dân số thoả thuận trước. Về bản chất, nhà nghiên cứu cũng sử dụng một/nhiều danh sách sẵn có, lựa chọn ngẫu nhiên một mẫu, liên lạc trước với họ qua điện thoại hay bưu điện, giới thiệu, thuyết phục và nhận được sự cam kết tham gia của họ trước khi gửi thư mời và đường dẫn tới khảo sát⁽⁴⁾.

Thứ hai và ngược lại, khi không thể xác định được xác suất được lựa chọn của các cá thể hoặc người trả lời vào mẫu hay không biết xác suất xuất hiện của các phần tử, nhà nghiên cứu sử dụng các chiến lược chọn mẫu phi xác suất hay còn gọi là chọn mẫu thuận tiện. Sử dụng các chiến lược chọn mẫu này, ngoài việc mẫu được chọn tùy thuộc vào nhà nghiên cứu, cơ cấu mẫu còn tùy thuộc vào quyết định tham gia khảo sát của mỗi cá nhân. Ví dụ, như nhà nghiên cứu có thể đăng một lời mời và đường dẫn đến một khảo sát trên một trang web nào đó, và sau đó chờ đợi những người viếng thăm trang web

này sẽ thể hiện một chút thiện chí nào đó và quyết định tham gia khảo sát. Trong trường hợp chọn các phần tử vào mẫu không theo quy luật ngẫu nhiên như thế này, không thể dùng các tham số của mẫu để ước lượng hay kiểm nghiệm các thông số của tổng thể. Các chiến lược chọn mẫu phi xác suất bao gồm một số dạng cơ bản như: 1) Chọn mẫu phán đoán, 2) Chọn mẫu theo chỉ tiêu và 3) Chọn mẫu tích lũy nhanh/quả bóng tuyết⁽⁵⁾. Theo đó, các dạng khảo sát trên web sử dụng các chiến lược chọn mẫu phi xác suất thường là các khảo sát mang mục đích giải trí, các khảo sát ngỏ, tự lựa chọn hay các khảo sát dựa vào các nhóm tình nguyện viên là những người sử dụng Internet. Các dạng khảo sát này không thỏa mãn các yêu cầu về tính đại diện trong chọn mẫu và cũng không nhằm mục đích suy rộng ra tổng thể nào hết⁽⁶⁾.

Cần lưu ý rằng ngay cả trong các khảo sát trên web sử dụng các chiến lược chọn mẫu xác suất, người trả lời cũng hoàn toàn có thể quyết định không tham gia và nhiệm vụ của các nhà nghiên cứu là phải tìm mọi cách giảm tối đa số lượng này. Cả hai chiến lược chọn mẫu này đều có những sai lệch. Tuy nhiên, chọn mẫu phi xác suất có nguy cơ sai lệch nhiều hơn vì trong nhiều trường hợp, những người quyết định tham gia khảo sát lại không đại diện cho quần thể nghiên cứu và, bên cạnh đó, nhà nghiên cứu thường không có thông tin về những người lựa chọn không tham gia vào khảo sát. Do vậy, dù tiết kiệm được thời gian và công sức, nhưng nhìn chung các chiến lược chọn mẫu phi xác suất không phù hợp với các mục đích suy luận thống kê chính thức. Tuy vậy, các chiến lược chọn mẫu phi xác suất khá hữu ích trong một số giai đoạn của nghiên cứu như giai đoạn xây dựng giả thuyết, xác định vấn đề, xác định giải pháp hay thu thập các dữ liệu không phục vụ cho mục đích suy luận thống kê.

2. Một số vấn đề đặt ra liên quan đến chọn mẫu trong khảo sát trực tuyến

Thứ nhất là, khung chọn mẫu và phạm vi nghiên cứu. Cũng giống như các nghiên cứu quy mô lớn truyền thống, các khảo sát trên nền tảng Internet luôn gặp khó khăn trong việc xác định khung lấy mẫu. Hay nói cách khác, cho đến nay vẫn chưa có một danh mục đầy đủ các địa chỉ thư điện tử đã được

tạo và đang được sử dụng. Trên thực tế, các khung lấy mẫu thường được xây dựng dựa trên các danh sách địa chỉ thư điện tử có sẵn của một vài tổ chức, đoàn thể, danh sách những người đăng ký sử dụng một dịch vụ nào đó hay các danh sách thư điện tử được tổng hợp và mua bán trên mạng. Tuy nhiên, sự tập trung vào một/nhiều nhóm dân số cụ thể như vậy chính là hạn chế về phạm vi nghiên cứu của đề tài. Ngoài ra, ngay cả khi càng ngày số lượng người sử dụng Internet càng tăng và có một danh bạ thư điện tử lớn hơn nữa thì vẫn cần phải lưu ý rằng danh sách tiềm năng này cũng khó có thể đầy đủ vì rõ ràng nó chỉ bao gồm những người có sử dụng Internet, thư điện tử và có khai báo. Chưa kể, không ít người trong số này có nhiều hơn một địa chỉ thư điện tử. Những người này không thể đại diện cho toàn bộ những người sử dụng Internet và càng không đại diện cho nhóm dân số không sử dụng Internet. Do vậy, các khảo sát trên web sử dụng các mẫu từ quần thể những người sử dụng Internet thì cũng không thể suy rộng cho dân số nói chung. Do vậy, nếu mục tiêu của nghiên cứu là suy rộng cho tổng thể dân số nói chung thì nhà nghiên cứu luôn cần phải ý thức được sự hạn chế và sai số về phạm vi bao phủ của nghiên cứu dù mẫu nghiên cứu có được lựa chọn một cách thận trọng như thế nào từ quần thể những người sử dụng Internet⁽⁷⁾.

Thậm chí nếu nhà nghiên cứu giảm thiểu được sai số về phạm vi bao phủ của khảo sát trên web của mình, khi chỉ sử dụng phương thức khảo sát trên web, nhóm dân số mục tiêu cũng cần phải có kiến thức và kỹ năng nhất định về máy tính và có điều kiện truy cập dễ dàng và thường xuyên vào mạng Internet để có thể tham gia khảo sát. Điều này có nghĩa là ngay cả khi nhà nghiên cứu có được danh sách đầy đủ các tài khoản thư điện tử của các thành viên của một cơ quan, tổ chức thì điều này cũng không có nghĩa là các thành viên trong danh sách đó đều có các cơ hội tiếp cận như nhau. Thiếu đi cơ hội tiếp cận công bằng có thể ảnh hưởng đáng kể đến kết quả lựa chọn khảo sát và những thiên lệch liên quan đến tỷ lệ không phản hồi⁽⁸⁾.

Thứ hai, cỡ mẫu lớn hơn không phải lúc nào cũng tốt hơn. Thông thường, với các chiến lược chọn mẫu xác suất, cỡ mẫu lớn hay nhỏ phụ thuộc vào một số

tiêu chí và yêu cầu như: (1) Mức độ biến đổi trong dân cư- Biến đổi càng cao cỡ mẫu càng lớn; (2) Cách thức chọn mẫu - Mẫu chùm cần cỡ mẫu lớn; (3) Tỷ lệ dân số có những đặc tính mà nhà nghiên cứu quan tâm hay quy mô của hiện tượng mà nhà nghiên cứu muốn quát sát tìm hiểu - Càng hiếm thì cỡ mẫu càng cần nhiều; (3) Mục đích phân tích (đơn biến, đa biến) -Phân tích đa biến cần cỡ mẫu lớn, và (4) Phạm vi sai số và hệ số tin cậy cho phép. Cả phạm vi sai số và hệ số tin cậy đều được tính bằng tỷ lệ phần trăm và do nhà nghiên cứu và người đặt hàng nghiên cứu quyết định ở mức chấp nhận được và phù hợp. Phạm vi sai số (có thể là $\pm 3\%$, $\pm 5\%$, $\pm 7\%$, $\pm 10\%$) tỷ lệ nghịch với cỡ mẫu và hệ số tin cậy (thường là $\pm 95\%$) tỷ lệ thuận với cỡ mẫu ⁽⁹⁾. Khi xác định được phạm vi sai số và hệ số tin cậy, nhà nghiên cứu có thể sử dụng các công thức hoặc các trang web giúp ước tính cỡ mẫu để xác định cỡ mẫu cần thiết. Ngược lại, việc xác định cỡ mẫu trong các chiến lược phi xác suất là không thể và không cần thiết ⁽¹⁰⁾.

Cần lưu ý, trong các khảo sát trên mạng, dù chọn mẫu theo xác suất hay phi xác suất, cỡ mẫu lớn không phải lúc nào cũng đồng nghĩa với việc mẫu có tính đại diện cao hơn cho tổng thể nghiên cứu. Mẫu dù to hay nhỏ vẫn luôn có sự thiên lệch. Trong thực tế, nếu có được các khung lấy mẫu có sẵn, nhà nghiên cứu thường gửi đường link khảo sát tới toàn bộ các đơn vị trong khung lấy mẫu. Mặc dù, có nhiều người cho rằng bằng cách đó nhà nghiên cứu đang cố gắng thực hiện một dạng tổng điều tra và như vậy đã thiên theo chiều hướng chọn mẫu thuận tiện, tạo cơ hội cho tất cả các thành viên trong khung chọn mẫu lựa chọn tham gia vào khảo sát. Tuy nhiên, việc này chỉ có ý nghĩa làm giảm đi sai số liên quan đến chọn mẫu. Các sai số khác liên quan đến phạm vi bao phủ, công cụ đo lường và tỷ lệ không phản hồi cũng rất quan trọng và ảnh hưởng không nhỏ đến mục đích suy rộng một cách chính xác ra tổng thể từ dữ liệu mẫu. Chính vì những điểm tồn tại như vậy và trong khi không thể phủ định tầm quan trọng và sự phổ biến của các khảo sát trên web, một số nhà nghiên cứu cho rằng không nhất thiết đề cao tính đại diện và mục đích suy rộng ra tổng thể của các khảo sát dựa trên web. Thay vào đó, các nghiên cứu khảo sát trên web có thể cung cấp các ví

dụ trường hợp, các số liệu có tính chỉ dẫn và so sánh hữu ích. Theo đó, hoàn toàn có thể sử dụng các khung lấy mẫu sẵn có (dù hạn hẹp và thiên lệch) và các chiến lược chọn mẫu xác suất phù hợp ⁽¹¹⁾.

Tóm lại, sự sẵn có và phát triển của Internet và các công nghệ phần mềm đặc dụng đã khiến cho các khảo sát dựa trên web hiện nay trở nên dễ dàng và thịnh hành hơn bao giờ hết. Bên cạnh những lợi thế như tiết kiệm thời gian và kinh phí thực địa, các câu trả lời có thể được lưu giữ trực tiếp và dễ dàng chiết xuất sang các dạng phù hợp với các phần mềm xử lý và quản lý số liệu sẵn có, các khảo sát dựa trên web đang bị thách thức bởi các vấn đề liên quan đến chọn mẫu và phạm vi nghiên cứu. Cũng giống như các khảo sát truyền thống, việc chọn mẫu trong các khảo sát trên web có thể được thực hiện dựa theo các chiến lược xác suất và phi xác suất tùy thuộc vào khả năng có thể xác lập được khung lấy mẫu hay không. Tuy nhiên, việc chọn mẫu cho các nghiên cứu dựa trên nền tảng web dù có được khung lấy mẫu và thực hiện theo các chiến lược xác suất thì cũng không nên quá đề cao tính đại diện và khả năng suy rộng ra tổng thể từ dữ liệu mẫu. Đơn giản là vì cho đến nay vẫn chưa có sáng kiến nào giúp xác lập được khung lấy mẫu trên toàn Internet có thể phản ánh đầy đủ sự đa dạng của tổng thể cần nghiên cứu./.

(1) Poynter, R. (2010), *The Handbook of Online and Social Media Research- Tools and Techniques for Market Researchers*, John Wiley & Sons, Inc., UK, tr. 4-5.

(2) (9) Bùi Thu Hương (2017), *Thiết kế nghiên cứu- Một số vấn đề cơ bản*, Nxb. Thế giới, Hà Nội;

(3) Sekaran, U. (2003), "Sampling", *Research Methods for Business- A Skill-building Approach*, John Wiley & Sons, Inc., USA, tr. 263-299.

(4) (6) Couper, M.P. (2008), *Designing Effective Web Surveys*, Cambridge University Press, Cambridge.

(5) Patton, M.Q. (2002), *Qualitative Evaluation and Research Methods*, Sage Publications, London.

(7) (8) Fielding, N.G, R.M. Lee & G. Blank (2017), *The Sage Handbook of Online Research Methods*, Sage, London.

(10) Sue, M. V & L.A. Ritter (2007), "Chapter 3: Sampling", *Conducting Online Surveys*, Sage, London, tr. 25-37.

(11) Andrews, D., B. Nonnecke & J. Preece (2003), "Electronic survey methodology: A case study in reaching hard to involve Internet Users", *International Journal of Human-Computer Interaction*. 16 (2), tr. 185-210.