Những kết quả cần lưu ý trong tổng hợp và phân tích thống kê

Trường hợp 1

Trường hợp được giới thiệu dưới đây thường được biết đến như là một "nghịch lý".

Để hiểu nghịch lý này, hãy bắt đầu bằng việc phân tích kết quả điều tra về tình trạng việc làm tại một tỉnh theo trình độ người lao động

	Lực lượng lao động (1000 người)			Tỷ lệ thất nghiệp (%)
	Tổng số	Chia theo tình trạng việc làm		
		Có việc làm	Thất nghiệp	
Chung	610	580	30	4,92
Chia theo trình độ				
- Có trình độ	500	475	25	**5,00**
- Không có trình độ	110	105	5	**4,55**

Kết quả trên đây đã cho thấy: *tính chung toàn tỉnh, lao động có trình độ có tỷ lệ thất nghiệp **cao** hơn so với lao động không có trình độ*. Tình trạng này đôi khi được giải thích bằng nhiều lý do khác nhau như: lao động có trình độ thường kén chọn công việc hơn lao động không có trình độ; đối với lao động không qua đào tạo, cơ hội tìm được việc làm dễ hơn khi nhu cầu lao động của nền kinh tế phần lớn là lao động phổ thông ...

Để tìm hiểu nguyên nhân tỷ lệ thất nghiệp của lao động có trình độ lại cao hơn lao động không có trình độ, cần có những phân tích sâu hơn về tình trạng thất nghiệp theo một số tiêu thức.

Thông thường, bước tiếp theo cần tìm hiểu tỷ lệ thất nghiệp với các trình độ khác nhau của lao động sẽ như thế nào đối với khu vực thành thị và khu vực nông thôn. Tỷ lệ thất nghiệp cho mỗi trình độ người lao động được tính cho từng khu vực thành thị, nông thôn.

Kết quả như sau:

	Lực lượng lao động (1000 người)			Tỷ lệ thất nghiệp (%)
	Tổng số	Chia theo tình trạng việc làm		
		Có việc làm	Thất nghiệp	
Thành thị				
- Có trình độ	300	282,5	17,5	**5,83**
- Không có trình độ	15	14	1	**6,67**
Nông thôn				
- Có trình độ	200	192,5	7,5	**3,75**
- Không có trình độ	95	91	4	**4,21**

Khi phân tích theo khu vực thành thị, nông thôn thì kết quả là một xu hướng hoàn toàn ngược lại với phân tích chung toàn tỉnh*: trong từng khu vực, lao động có trình độ có tỷ lệ thất nghiệp **thấp** hơn lao động không có trình độ*. Cụ thể là trong khi tính chung toàn tỉnh tỷ lệ thất nghiệp của lao động có trình độ và lao động không có trình độ lần lượt là 5% và 4,55%; nhưng khi tính riêng cho từng khu vực thì ở khu vực thành thị tỷ lệ này lần lượt là 5,83% và 6,67%, còn khu vực nông thôn là 3,75% và 4,21%.

Hiện tượng này cũng xuất hiện trong một số trường hợp khác và chỉ xảy ra khi tính toán tỷ lệ của các biểu hiện khác nhau theo các phân tổ kết hợp của một tổng thể nghiên cứu.

Trong ví dụ trên, tổng thể nghiên cứu là những người lao động. Trong quá trình phân tích đã sử dụng hình thức phân tổ kết hợp giữa phân tổ theo khu vực, phân tổ theo tình trạng việc làm, phân tổ theo trình độ người lao động. Tỷ lệ thất nghiệp (hoặc ngược lại là tỷ lệ có việc làm) được tính cho từng trạng thái của phân tổ kết hợp

Ở đây:

Tỷ lệ thất nghiệp được tính bằng số lao động thất nghiệp chia cho tổng số lao động.

Để mô tả bằng toán học, chúng ta sử dụng 2 ký hiệu A và S:

- A là số người thất nghiệp
- S là số lao động

Như vậy, tỷ lệ thất nghiệp là $\frac{A}{S}$

Để nhận diện tỷ lệ thất nghiệp theo các phân tổ, có thể bổ sung 2 chỉ số vào các ký hiệu:

- Chỉ số thứ 1 dùng để nhận diện trình độ: 1 là lao động có trình độ, 2 là lao động không có trình độ
- Chỉ số thứ 2 dùng để nhận diện khu vực: 1 là khu vực thành thị, 2 là khu vực nông thôn

Như vậy:

Tại khu vực thành thị, những người có trình độ có tỷ lệ thất nghiệp thấp hơn so những người không có trình độ:

$$\frac{A_{11}}{S_{11}} < \frac{A_{21}}{S_{21}} \quad (1)$$

Tại khu vực nông thôn, những người có trình độ có tỷ lệ thất nghiệp thấp hơn so những người không có trình độ:

$$\frac{A_{12}}{S_{12}} < \frac{A_{22}}{S_{22}} \quad (2)$$

Tính chung toàn tỉnh thì những người có trình độ có tỷ lệ thất nghiệp cao hơn so những người không có trình độ:

$$\frac{A_{11}+A_{12}}{S_{11}+S_{12}} > \frac{A_{21}+A_{22}}{S_{21}+S_{22}} \quad (3)$$

Nghịch lý xuất hiện nghĩa là cả 3 bất đẳng thức trên đều đúng.

Trong thực tế, không phải lúc nào tình trạng trên của xảy ra. Chỉ trong một số trường hợp không nhiều mới xuất hiện tình trạng cả 3 bất đẳng thức cùng đúng. Ở đây không phải là một nghịch lý hoặc sai sót và hoàn toàn đúng thực tế

Hiện tượng này đã được phát hiện từ lâu và được gọi là nghịch lý Simpson (Simpson Paradox). Mặc dầu nhà thống kê người Anh Udny Yule lần đầu tiên đề cập đến hiện tượng này vào năm 1903, nhưng phần lớn tài liệu đều cho rằng nghịch lý này do nhà toán học Edward Simpson phát hiện vào năm 1951. Tại website http://en.wikipedia.org/wiki/Simpson's_paradox có thể tìm hiểu kỹ hơn về nghịch lý này.

Ngoài ra hiện tượng trên cũng xảy ra trong trường hợp việc phân tổ theo những tiêu thức có nhiều hơn 2 biểu hiện. Ví dụ cho những trường hợp này có thể được tìm thấy tại wesite nói trên hoặc tại website http://vudlab.com/simpsons/

Theo nội dung của nghịch lý, những thay đổi về cấu trúc số liệu có thể dẫn đến kết luận sai cho một tổng thể.

Như trong trường hợp trên, nếu chỉ phiến diện kết luận ngay từ tính toán đầu tiên thì chúng ta đã kết luận sai. Bản chất của hiện tượng là ngược lại nếu khi chúng ta đi sâu vào chi tiết của từng khu vực. Nếu xem xét kỹ thì có thể thấy nguyên nhân gây ra hiện tượng đảo ngược xu hướng khi tổng hợp chung có thể do các nguyên nhân sau:

- Do kết cấu lao động theo trình độ rất khác biệt nhau ở từng khu vực: tỷ lệ lao động không có trình độ ở khu vực nông thôn là 32,2%, nhưng tỷ lệ này tại khu vực thành thị là 5%.
- Do kết cấu lao dộng theo khu vực: mặc dầu số liệu lao động theo khu vực trong ví dụ trên không thay đổi; nhưng nếu số lao động theo khu vực thành thị hoặc nông thôn thay đổi đáng kể cũng dẫn đến kết quả khác với ví dụ đã nêu.

Từ đây có thể thấy bản thân các con số (tỷ lệ thất nghiệp theo trình độ) không còn quan trọng bằng cách thức thu thập dữ liệu để tính toán con số đó. Nếu số liệu nói trên dược thu thập đúng phương pháp và cỡ mẫu phản ánh đúng thực trạng kết cấu lao động theo trình độ và theo khu vực thì có thể nói bản chất của sự việc đã được phản ánh đúng. Còn nếu vì lý do nào đó mà số liệu thu thập không đại diện cho thực tế của mỗi trình độ, mỗi khu vực thì có thể nói kết quả khi tính cho tổng thể chung không chính xác.

Trường hợp 2

Trường hợp lưu ý thứ hai được giới thiệu ở đây xuất phát từ một quan niệm sau:

Đối với một chỉ tiêu thống kê là số tương đối, giá trị của chỉ tiêu khi tính cho một tổng thể chung sẽ nằm trong khoảng giữa các giá trị của chỉ tiêu đó khi tính theo từng phân tổ của tổng thể đó. Điều này dường như đã trở thành qui luật và thường được sử dụng để kiểm tra kết quả tính toán.

Ví dụ: giá trị của tỷ lệ thất nghiệp tính chung cho một tỉnh sẽ nằm trong khoảng tỷ lệ thất nghiệp khi tính cho từng khu vực thành thị, nông thôn, hoặc tính cho từng quận, huyện; CBR (tỷ suất sinh thô) của tỉnh phải nằm trong khoảng giữa CBR của từng khu vực hoặc quận, huyện; hệ số Gini của cả nước phải nằm trong khoảng giữa hệ số Gini của từng tỉnh...

Tuy nhiên trong thực tế có một vài chỉ tiêu thống kê không tuân theo qui luật trên. Trong trường hợp này sẽ nêu ra 2 ví dụ: ví dụ về chỉ tiêu *Hệ số chênh lệch giàu nghèo* và ví dụ về *Tỷ lệ thu nhập của 40% dân số có thu nhập thấp*.

Hệ số chênh lệch giàu nghèo

Hệ số chênh lệch giàu nghèo được tính bằng cách lấy thu nhập bình quân (TNBQ) của 20% dân số giàu nhất chia cho TNBQ của 20% dân số nghèo nhất.

Để minh họa, có thể lấy ví dụ đơn giản nhất cho một địa phương giả định với 10 người dân. Có 5 người dân thuộc khu vực thành thị và 5 người dân thuộc khu vực nông thôn. Số liệu về TNBQ của 10 người này được xếp theo thứ tự từ cao xuống thấp trong bảng sau:

Thứ tự	1	2	3	4	5	6	7	8	9	10
Thu nhập bình quân	100	90	80	70	60	50	40	35	30	25
Mã số khu vực 1=thành thị; 2=nông thôn	1	1	2	2	1	1	1	2	2	2

Hệ số chênh lệch giàu nghèo tính chung cho toàn tỉnh:

TNBQ của 20% dân số có thu nhập cao nhất = (100 + 90)/2 = 95

TNBQ của 20% dân số có thu nhập thấp nhất = (30 + 25)/2 = 27,5

Hệ số chênh lệch giàu nghèo = 95/27,5 = **3,45**

Hệ số chênh lệch giàu nghèo tính cho khu vực thành thị:

TNBQ của 20% dân số có thu nhập cao nhất = 100

TNBQ của 20% dân số có thu nhập thấp nhất = 40

Hệ số chênh lệch giàu nghèo = 100/40 = **2,5**

Hệ số chênh lệch giàu nghèo tính cho khu vực nông thôn:

TNBQ của 20% dân số có thu nhập cao nhất = 80

TNBQ của 20% dân số có thu nhập thấp nhất = 25

Hệ số chênh lệch giàu nghèo = 80/25 = **3,2**

Như vậy trong ví dụ này hệ số chênh lệch giàu nghèo khi tính chung cho toàn tỉnh cao hơn khi tính cho từng khu vực. Trong khi theo suy nghĩ thông thường là giá trị tính cho toàn tỉnh phải nằm trong khoảng giá trị tính cho từng khu vực (ở đây nghĩa là phải nằm giữa 2,5 - giá trị thấp nhất và 3,2 - giá trị cao nhất).

Hiện tượng này xuất hiện khá thường xuyên. Có thể nhìn vào giá trị của hệ số này tính chung cho cả nước và chia theo khu vực thành thị, nông thôn qua các cuộc điều tra Khảo sát mức sống hộ gia đình tiến hành 2 năm một lần. Từ năm 1999 đến 2012 có 7 kỳ điều tra thì tình trạng này đã xuất hiện 6 lần.

	1999	2002	2004	2006	2008	2010	2012
Chung	**8,9**	**8,1**	**8,3**	**8,4**	**8,9**	**9,2**	**9,4**
Thành thị	9,8	8,0	8,1	8,2	8,3	7,9	7,1
Nông thôn	6,3	6,0	6,4	6,5	6,9	7,5	8,0

Nguồn: Kết quả Khảo sát mức sống dân cư Việt Nam 2012

Đây là một thực tế, không phải là một nghịch lý. Nguyên nhân ở đây là do phương pháp tính của hệ số và TNBQ của dân số.

Hiện tượng nói trên thường xảy ra khi có sự phân hóa quá mạnh mẻ với dân số của một khu vực nào đó quá tập trung vào đầu danh sách hoặc cuối danh sách.

Hiện tượng này sẽ còn có thể tìm thấy trong quá trình phân tích một số chỉ tiêu khác theo phương pháp ngũ phân vị. Ví dụ như khi tính mức độ chênh lệch về chi tiêu cho y tế, giáo dục,... giữa nhóm 5 và nhóm 1 theo khu vực trong phân tích đời sống.

Tỷ lệ thu nhập của 40% dân số có thu nhập thấp

Tỷ lệ thu nhập của 40% dân số có thu nhập thấp được tính bằng cách lấy tổng thu nhập của 40% dân số có thu nhập thấp chia cho tổng thu nhập của toàn bộ dân số.

Để minh họa, có thể lấy ví dụ đơn giản nhất cho một địa phương giả định với 10 người dân. Có 5 người dân thuộc khu vực thành thị và 5 người dân thuộc khu vực nông thôn. Số liệu về thu nhập của 10 người này được xếp theo thứ tự từ cao xuống thấp trong bảng sau:

Thứ tự	1	2	3	4	5	6	7	8	9	10
Thu nhập	180	150	100	80	60	50	40	30	20	15
Mã số khu vực 1=thành thị; 2=nông thôn	1	1	1	2	1	1	2	2	2	2

Tỷ lệ thu nhập của 40% dân số có thu nhập thấp tính chung cho toàn tỉnh:

Thu nhập của 40% dân số có thu nhập thấp = 15 + 20 + 30 + 40 = 105

Tổng thu nhập của dân số toàn tỉnh = 725

Tỷ lệ 40% thu nhập của dân số có thu nhập thấp = 105/725 = **14,5**%

Tỷ lệ thu nhập của 40% dân số có thu nhập thấp tính cho khu vực thành thị:

Thu nhập của 40% dân số có thu nhập thấp = 50 + 60 = 110

Tổng thu nhập của dân số khu vực thành thị = 540

Tỷ lệ 40% thu nhập của dân số có thu nhập thấp = 110/540 = **20,4**%

Tỷ lệ thu nhập của 40% dân số có thu nhập thấp tính cho khu vực nông thôn:

Thu nhập của 40% dân số có thu nhập thấp = 15 + 20 = 35

Tổng thu nhập của dân số khu vực nông thôn = 185

Tỷ lệ 40% thu nhập của dân số có thu nhập thấp = 35/185 = **18,9**%

Như vậy tỷ lệ thu nhập của 40% dân số có thu nhập thấp trong ví dụ này khi tính chung cho toàn tỉnh thấp hơn khi tính cho từng khu vực. Trong khi theo suy nghĩ thông thường là giá trị tính cho toàn tỉnh phải nằm trong khoảng giá trị tính cho từng khu vực (ở đây nghĩa là phải nằm giữa 18,9% - giá trị thấp nhất và 20,4% - giá trị cao nhất).

Nhìn chung, tỷ lệ thu nhập của 40% dân số có thu nhập thấp rất ít khi rơi vào trường hợp như trên nếu cỡ mẫu đưa vào phân tích đủ lớn.

Cả 2 trường hợp với hiện tượng đã nêu đều xuất hiện trong thực tế. Hiện tượng này chỉ có thể xảy ra khi tính toán các chỉ tiêu mà phương pháp tính không thực hiện trên toàn bộ tổng thể mà chỉ một phần. Chúng ta sẽ không thấy hiện tượng này xuất hiện khi tính những chỉ tiêu mà tất cả đơn vị tổng thể đều tham gia vào tính toán ở cấp toàn bộ cũng như từng phân tổ.

Ví dụ, tỷ lệ thất nghiệp tính chung toàn tỉnh luôn luôn phải nằm trong giới hạn giữa các tỷ lệ thất nghiệp chia theo khu vực hoặc quận, huyện. Bởi vì trong quá trình tính toán, mỗi lao động thất nghiệp sẽ được kể đến khi tính cho toàn tỉnh và khi tính theo vùng hoặc quận, huyện mà lao động thất nghiệp đó đang có mặt.

Trái lại, trong cách tính hệ số chênh lệch giàu nghèo. Khi tính toán hệ số chung cho toàn tỉnh thì có 40% dân

số (20% dân số nghèo nhất và 20% dân số giàu nhất) tham gia vào quá trình tính toán. Khi tính hệ số này theo khu vực thì trong 40% dân số được tính của khu vực đó sẽ có những người không thuộc về 40% dân số khi tính chung cho toàn tỉnh.

Tương tự, trong cách tính tỷ lệ thu nhập của 40% dân số có thu nhập thấp. Khi tính chung cho toàn tỉnh thì có 40% dân số nghèo nhất được tính. Nhưng không phải những người thuộc 40% dân số này lại thuộc về 40% dân số nghèo nhất của khu vực mà họ đang sống; do đó khi tính tỷ lệ cho khu vực thì họ bị bỏ ra.

Và tất cả các trường hợp đều không phải là nghịch lý. Đó là thực tế muôn màu mà phương pháp phân tích thống kê đã thể hiện. Là người phân tích phải hiểu đúng bản chất hiện tượng để kết luận đúng.

Trần Triết Tâm

Phòng Dân số-Văn xã, Cục Thống kê Đà Nẵng

Tài liệu tham khảo:

http://en.wikipedia.org/wiki/Simpson's_paradox

http://vudlab.com/simpsons/