Sử dụng kiểm định chi bình phương kiểm tra tính đại biểu của mẫu

Thạc sĩ HÀ VĂN SƠN

Số liệu của mẫu sau khi điều tra sẽ được tính toán để suy rộng cho tổng thể, để số liệu suy rộng có thể sử dụng được, ta phải kiểm tra tính đại biểu của mẫu. Trong thực tế ta có thể dùng các cách như: kiểm định chi bình phương- χ^2, kiểm định bằng biểu đồ, dùng tỷ lệ sai số chọn mẫu...

Ta biết, muốn cho tổng thể mẫu có tính đại biểu cao, nghĩa là các kết quả ước lượng hay kiểm định giả thuyết đảm bảo độ tin cậy mong muốn thì mẫu phải được chọn theo đúng phương pháp khoa học, đồng thời sau khi chọn được mẫu cụ thể cần kiểm tra lại tính đại biểu của nó. Tính đại biểu của mẫu thể hiện trước tiên ở sự giống nhau về luật phân phối và độ biến thiên của tiêu thức. Riêng trong trường hợp tổng thể chung được phân phối theo quy luật chuẩn (trường hợp này rất hay xảy ra), thì nhiệm vụ của chúng ta lại là kiểm định tính chuẩn của phân phối mẫu. Để kiểm định tính chuẩn của phân phối mẫu có rất nhiều cách, một trong các cách đó là sử dụng một kiểm định phi tham số phù hợp, kiểm định chi bình phương.

Trong điều tra chăn nuôi, số liệu sau khi thu thập từ mẫu sẽ được suy rộng cho tổng thể. Để việc suy rộng đảm bảo độ tin cậy mong muốn ta phải kiểm tra tính chuẩn của mẫu và người ta thường sử dụng kiểm định . Ví dụ ở đây ta có danh sách chăn nuôi heo của 700 hộ được chọn từ 32 ấp mẫu của TP.HCM vào ngày 1.8.2004.

Để thực hiện kiểm định ta sử dụng phần mềm thống kê SPSS nhập số liệu của 700 hộ vào. Ta có bảng trình bày các tham số của mẫu.

Bảng 1: Các tham số của mẫu từ SPSS

Frequencies

Statistics

Số heo các hộ

N	Valid	700
	Missing	0
Mean		15.80
Std. Error of Mean		.532
Std. Deviation		14.079
Minimum		1
Maximum		150

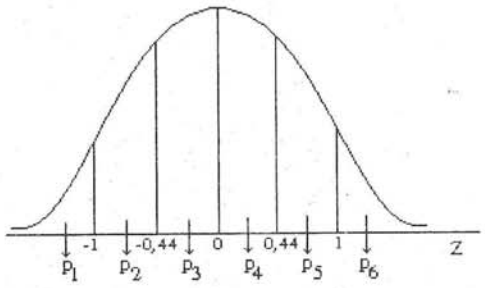
Để kiểm định xem số lượng heo chăn nuôi trong các hộ gia đình mẫu có tuân theo phân phối chuẩn hay không ta đặt giả thuyết H_0 và H_1 như sau:

Giả thuyết H_0 : Số heo nuôi trong các hộ gia đình có phân phối chuẩn.

Giả thuyết H_1 : Số heo nuôi trong các hộ gia đình không có phân phối chuẩn.

Trước tiên, chúng ta xác định các xác suất để một đại lượng phân phối chuẩn có trị số rơi vào các khoảng nhất định. Từ bảng phân phối chuẩn, ta xác định được các xác suất của đại lượng phân phối chuẩn Z. Chẳng hạn, tra bảng phân phối chuẩn ta có xác suất để đại lượng phân phối chuẩn Z rơi vào khoảng từ 0 đến 1 là 0,3413 và gần phân nửa của xác suất này là 0,1700 ứng với trị số giới hạn z = 0,44. Vậy xác suất Z có trị số rơi vào khoảng từ 0,44 đến 1 bằng 0,1713 và xác xuất Z rơi vào khoảng từ 1 $\rightarrow \infty$ sẽ bằng 0,1587 (0,5 - 0,3413).

Hình 1: Các xác suất để Z nằm giữa các khoảng giá trị



Tương tự chúng ta xác định được các trị số giới hạn của biến Z và các xác xuất để Z nhận các trị số nằm giữa các trị số giới hạn này đối xứng qua 0:

z=-1 z=-0,44 z=0 z=0,44 z=1

$p_1 = 0,1587$, $p_2 = 0,1713$, $p_3 = 0,17$, $p_4 = 0,17$, $p_5 = 0,1713$, $p_6 = 0,1587$

Từ công thức $E_i = n.p_i$, (n = 700), các trị số lý thuyết E_i có kết quả tính toán như sau:

$E_1 = 111,09$, $E_2 = 119,91$, $E_3 = 119$, $E_4 = 119$, $E_5 = 119,91$, $E_6 = 111,09$

Dựa vào công thức $Y = \mu + \sigma Z$ (suy ra từ công thức chuẩn hóa các trị số quan sát), chuyển các trị số giới hạn của đại lượng ngẫu nhiên có phân phối chuẩn Z thành trị số của yếu tố đang nghiên cứu là số heo nuôi trong một hộ gia đình. Chúng ta có thể dùng $\bar{y}$ và s (tham số mẫu) thay cho μ và σ (tham số tổng thể).

Dựa vào Bảng 1, Statistics ta có ngay: trung bình Mean = 15,80 và độ lệch tiêu chuẩn Std. Deviation = 14,079.

Trị số giới hạn của các nhóm được xác định như sau:

$y_1 = 15,7986 + (-1)(14,0787) = 1,7199$

$y_2 = 15,7986 + (-0,44)(14,0787) = 9,604$

$y_3 = 15,7986 + (0)(14,0787) = 15,7986$

$y_4 = 15,7986 + (0,44)(14,0787) = 21,9932$

$y_5 = 15,7986 + (1)(14,0787) = 29,8773$

Như vậy, ta đã xác định được các tổ, ở đây ta có 6 tổ, xác xuất để một quan sát rơi vào tổ thứ i (p_i), và số lượng quan sát theo lý thuyết (E_i).

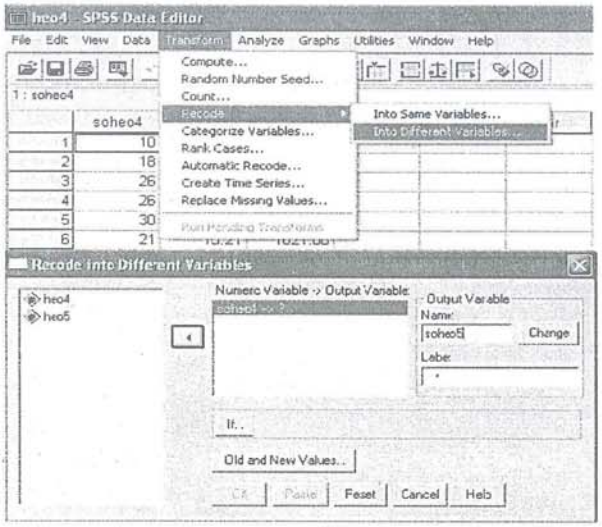
Bảng 2: Bảng phân phối tần số lý thuyết

Y_i(con)	P_i	Tần số lý thuyết $E_i = (n.p_i)$
< 1,72	0,1587	111,09
1,72 - 9,60	0,1713	119,91

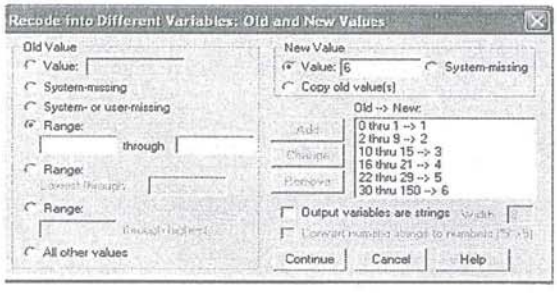
9,60 – 15,80	0,1700	119
15,80 – 21,99	0,1700	119
21,99 – 29,88	0,1713	119,91
≥ 29,88	0,1587	111,09
Tổng cộng	1	700

Tiếp theo ta phải xác định các tần số thực tế 0i , tức là phân phối 700 hộ vào 6 tổ trên. Rõ ràng là ta không thể làm bằng tay. Ta có thể làm nhanh bằng cách thực hiện lệnh Recode trong SPSS để tiến hành phân tổ lại.

Vào menu Transform Recode Into Different Variables, hộp thoại sau xuất hiện:



Trong hộp thoại Recode này, ta chọn biến cần mã hóa lại, nhấn vào nút Old and New Values để xác định các giá trị cũ và chỉ định mã mới tương ứng, hộp thoại sau xuất hiện:



Sau khi đã tạo được biến mới, ta lập bảng phân phối tần số theo các tổ định sẵn, ta có kết quả sau:

Frequencies

Statistics

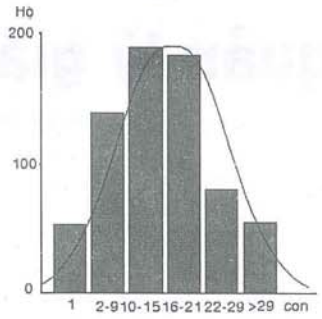
N	Valid	700
	Missing	0

Bảng 3: Bảng phân tổ lại

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	< 1,72	53	7.6	7.6	7.6
	1,72 – 9,60	139	19.9	19.9	27.4
	9,60 – 15,80	189	27.0	27.0	54.4
	15,80 – 21,99	184	26.3	26.3	80.7
	21,99 – 29,88	80	11.4	11.4	92.1
	≥ 29,88	55	7.9	7.9	100.0
	Total	700	100.0	100.0	

Hình 2: Biểu đồ phân phối tần số theo bảng phân tổ lại

Như vậy ta đã có tần số thực tế của các tổ, ta lập bảng tính toán đại lượng kiểm định. Vì số lượng heo là số nguyên dương, nên ta có thể viết lại các tổ trong bảng 3 thành bảng 4.



Bảng 4: Bảng tính toán đại lượng kiểm định χ^2

Y_i (con)	P_i	Tần số lý thuyết ($E_i = nP_i$)	Tần số thực tế (0_i)	$(0_i - E_i)^2/E_i$
1	0,1587	111,09	53	30.3758
2 – 9	0,1713	119,91	139	3.0392
10 – 15	0,1700	119	189	41.1765
16 – 21	0,1700	119	184	35.5042
22 – 29	0,1713	119,91	80	13.2834
> 29	0,1587	111,09	55	28.3202
Tổng	1	700	700	151.6992

Tính giá trị kiểm định χ^2 theo công thức $\chi^2 = \sum_{i=1}^{k} \frac{(0_i - E_i)^2}{E_i}$

Từ bảng 4 ta có $\chi^2 = 151,7$ và trong 6 tổ có hai tham số được ước lượng ($\bar{y}$ cho μ và s cho σ) nên số bậc tự do là (k -1 - số tham số) = 6 -1 -2 = 3.

Tra bảng phân phối χ^2, ta có: $\chi^2_{3,\,0,05} = 7,815 < \chi^2 = 151,7$. Do vậy ta bác bỏ giả thuyết H_0 ở mức ý nghĩa 5%, tức là số heo nuôi trong các hộ gia đình mẫu không có phân phối chuẩn, do đó số liệu của điều tra mẫu không dùng để suy rộng cho đàn heo của thành phố được.

Người ta cũng có thể kiểm định tính chuẩn của mẫu bằng cách dùng biểu đồ. Nhìn vào hình 2 ta thấy hình dáng của biểu đồ có vẻ giống phân phối chuẩn. Tuy nhiên nếu nhìn kỹ hai nhánh của biểu đồ ta thấy không đối xứng. Kiểm định bằng biểu đồ thường dùng trực quan để kiểm tra nên thường không chính xác lắm, ta thường phải kết hợp với các phương pháp khác.

Qua trình bày ở trên, ta thấy kiểm định chi bình phương tính toán tương đối phức tạp, tuy nhiên cho kết quả chính xác. Hiện nay với các phần mềm thống kê thông dụng việc tính toán trở nên đơn giản và vì vậy phương pháp kiểm định chi bình phương trở nên khá phổ biến trong nghiên cứu kinh tế ■