KIỂM TRA TÍNH CHUẨN BẰNG ĐỒ THỊ PHÂN SỐ VỊ CHUẨN

Phân phối chuẩn là một trong những dạng phân phối hay gặp nhất trong nhiều bài toán thống kê. Thực tế có rất nhiều tệp số liệu có phân phối (tương đối) chuẩn, chẳng hạn như điểm của các môn thi do nhiều người tiến hành; các đo đạc cẩn thận được lặp đi lặp lại trên một đại lượng; các thuộc tính của các tổng thể sinh học đồng nhất... Rất nhiều phép phân tích số liệu được thực hiện có dựa vào giả thiết về tính chuẩn của phân phối của số liệu. Điều này rất quan trọng vì các bước thống kê suy luận tiếp theo đều phụ thuộc vào sự đúng đắn của giả thiết này. Vậy bằng cách nào mà chúng ta có thể kiểm tra rằng số liệu quan sát được có phân phối (tương đối) chuẩn? Khoa học thống kê cung cấp cho chúng ta một số phương pháp để có thể khảo cứu sơ bộ về tính phân phối chuẩn của một tệp số liệu - đó

là ***đồ thị phân số vị chuẩn*** {Normal Quantile Plot} - thường được gọi là đồ thị xác xuất chuẩn {normal probability plot}, còn trong một số phần mềm vi tính, ví dụ như SPSS, lại gọi nó là Q-Q Plot.

Trước tiên cần lưu ý khái niệm về các ***phân số vị*** {quantile}. Chúng chính là các phân vị {percentile} nhưng được biểu diễn dưới dạng phân số chứ không phải là phần trăm. Ví dụ như số trung vị là phân vị thứ 50 và là phân số vị 0,50 của một phân phối.

Ý tưởng về một đồ thị phân số vị chuẩn là vẽ từng trị số quan sát x với phân số vị z tương ứng của một phân phối chuẩn hoá {standard normal distribution}. Chúng ta có thể hiểu rõ ý tưởng của đồ thị phân số vị chuẩn từ một ví dụ đơn giản dưới đây. Giả sử chúng ta có 20 quan sát như sau:

11,5	5,1	12,1	7,8	15,9	8,2	10,7	6,8	10,7	12,9
11,7	12,4	8,1	9,4	12,9	8,2	3,8	11,4	10,3	6,1

Quan sát nhỏ nhất là x = 3,8. Nếu xếp theo độ lớn dần nó là con số đầu tiên của 20 quan sát, do vậy x = 3,8 là phân số vị 1/20 hoặc 0,05 của dãy số. Phân số vị 0,05 của một phân phối chuẩn hoá là trị số z mà tại đó diện tích nằm ở bên trái của nó dưới đường cong phân phối chuẩn hoá $N \sim (0, 1)$ là 0,05. Tìm kiếm con số gần 0,05 nhất trong Bảng Z[2] ta có được z xấp xỉ -1,65. Vậy điểm đầu tiên trên đồ thị phân số vị chuẩn là (3,8, -1,65). Tương tự như vậy điểm thứ hai trên đồ thị này là (5,1, -1,28) trong đó 5,1 là trị số x nhỏ thứ hai trong dãy số trên, còn -1,28 là giá trị của z có phân số vị chuẩn là 0,1 trong phân phối chuẩn hoá, v.v...

Tính chất của một đồ thị phân số vị chuẩn có thể được phát biểu như sau: Nếu các điểm trên một đồ thị phân số vị chuẩn nằm gần một đường thẳng thì đồ thị đó chỉ ra rằng số liệu quan sát được có phân phối chuẩn. Các sự sai lệch có hệ thống so với đường thẳng chỉ ra rằng phân phối của dữ liệu là không chuẩn. Những số ngoại lai được bộc lộ bằng các điểm nằm tách xa hẳn dạng mẫu chung của phần đông các điểm. Nói cách khác, nếu các quan sát của biến X mà có phân phối chuẩn $N(\mu, \sigma)$ thì biến được chuẩn hoá $Z=(X-\mu)/\sigma$ cũng có phân phối chuẩn hoá $N(0, 1)$. Vì thế sẽ có một mối quan hệ tuyến tính $X = \sigma Z + \mu$ giữa biến nghiên cứu và một biến chuẩn hoá, và do vậy các điểm trên một đồ thị phân số vị chuẩn sẽ nằm gần một đường thẳng $z=x$.

Dưới đây là 4 đồ thị được lập cho bốn tệp số liệu khác nhau, có sử dụng phần mềm SPSS 9.0. Chúng ta sẽ lần lượt khảo cứu từng đồ thị một.

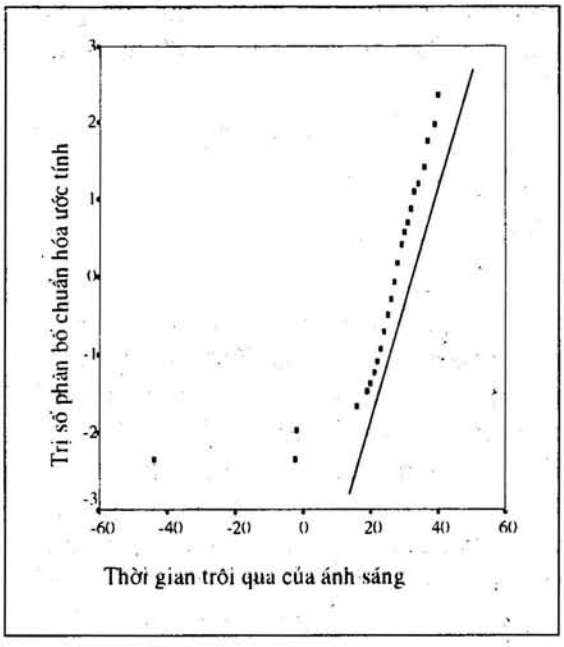
Số liệu của Newcomb (sử dụng cho Hình 1.a và 1.d)

28 26 33 24 34 -44 27 16 40 -2 29 22 24 21 25 30 23 29 31 19 24 20 36 32 36 28 25 21 28 29 37 25 28 26 30 32 36 26 30 22 36 23 27 27 28 27 31 27 26 33 26 32 32 24 39 28 24 25 32 25 29 27 28 29 16 23

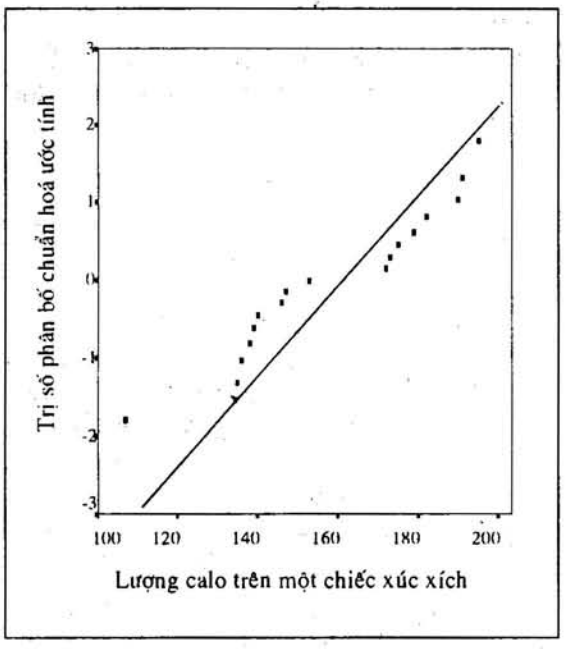
[2] Là bảng tính sẵn mà các trị số của bảng là các diện tích (xác xuất) nằm dưới Z=z, trong đó Z có phân phối chuẩn hoá với trung bình 0 và độ lệch chuẩn 1.

Số liệu chi tiêu của 50 khách hàng (sử dụng cho Hình 1.c)

2,32	6,61	6,90	8,04	9,45	10,26	11,34	11,63	12,66	12,95
13,67	13,72	11,35	14,52	14,55	15,01	15,33	16,55	17,15	18,22
18,30	18,71	19,54	19,55	20,58	20,89	20,91	21,13	23,85	26,04
27,07	28,76	29,15	30,54	31,99	32,82	33,26	33,80	34,76	36,22
37,52	39,28	40,80	43,97	45,58	52,36	61,57	63,85	64,30	69,49



Hình 1,(a) Đồ thị các phân số vị chuẩn về số liệu về thời gian trôi qua của ánh sáng của nhà thiên văn học Newcomb.

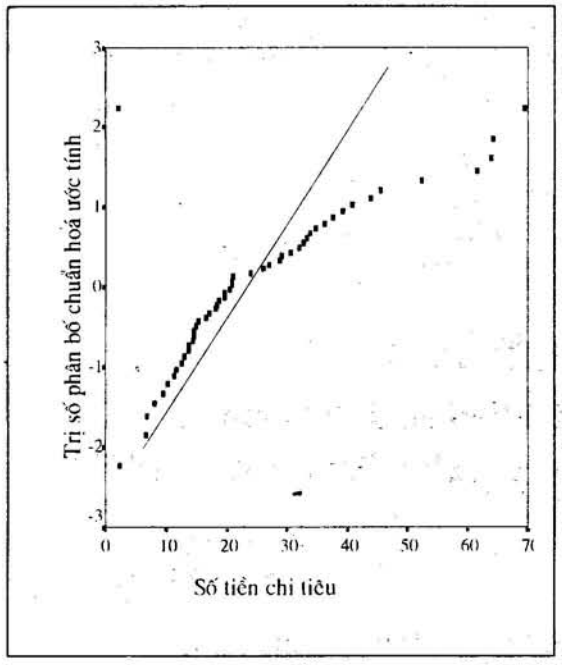


Hình 1,(b) Đồ thị các phân số vị chuẩn về số liệu về lượng calo trong các loại xúc xích thịt hỗn hợp.

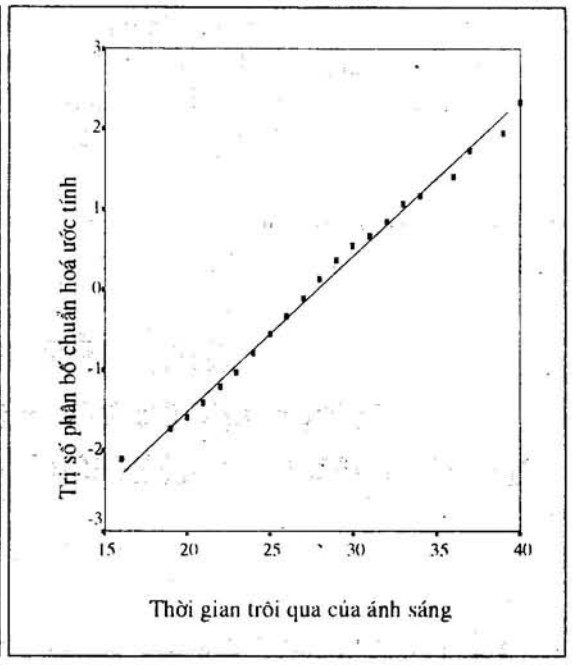
Hình 1,(a) là đồ thị phân số vị chuẩn cho số liệu thời gian trôi qua của ánh sáng do nhà thiên văn học Newcomb đo đạc. Hầu hết các điểm đều nằm gần một đường thẳng, chỉ ra rằng một mô hình phân phối chuẩn là rất phù hợp. Hai số ngoại lai/cực trị nằm xa lệch hẳn so với đường thẳng là một minh hoạ cho cách mà đồ thị đã bộ lộ các số ngoại lai cực bé ra sao. Những số cực bé sẽ nằm ở bên trái (cạnh thấp) của dạng tuyến tính chung của phần đông các điểm. Ngược lại các số cực lớn sẽ nằm ở bên phải (cạnh cao) của dạng mẫu chung đó.

Đồ thị của số liệu về lượng calo chứa trong các xúc xích thịt hỗn hợp[2] ở Hình 1,(b) thể hiện rất rõ ràng sự tồn tại của hai cụm số liệu riêng biệt và một số ngoại lai bé.

[2] Riêng số liệu cho ví dụ này, do khuôn khổ bài báo có hạn, không được viện dẫn ra đây. Bạn đọc nào quan tâm xin hãy vui lòng tìm đọc bài *Đồ thị cành và lá* của Việt Linh trong **Thông tin Khoa học Thống kê** số 2/1998, trang 24.



Hình 1.(c) Đồ thị các phân số vị chuẩn về số liệu về lượng tiền chi tiêu của 50 khách hàng tại một cửa hàng tạp phẩm.



Hình 1.(d) Đồ thị các phân số vị chuẩn về số liệu về thời gian trôi qua của ánh sáng của nhà thiên văn học Newcomb sau khi loại bỏ 2 quan sát ngoại lai.

Hình 1.(c) là đồ thị phân số vị chuẩn cho lượng tiền chi tiêu của 50 khách hàng tại một cửa hiệu tạp hoá. Một phân phối lệch phải thấy rất rõ nếu bạn kẻ một đường thẳng dọc theo các điểm nằm ở bên trái nhất (là những điểm tương ứng với các quan sát nhỏ nhất). Rõ ràng là các quan sát lớn hơn nằm một cách có hệ thống (không phải là ngẫu nhiên) ở bên phải đường thẳng. Chúng nằm xa về phía bên phải hơn chứ không giống như ở trong một phân phối chuẩn. Trong một phân phối lệch phải, các quan sát lớn nhất nằm lệch một cách rõ ràng sang bên phải của một đường thẳng được vẽ qua phần chính của các điểm. Tương tự như vậy các quan sát bé nhất nằm lệch sang bên trái của đường thẳng sẽ cho ta một bằng chứng về sự lệch trải của phân phối. Không giống như Hình 1.(a), ở Hình 1.(c) này không có các quan sát ngoại lai riêng rẽ nào.

Hình 1.(d) là một đồ thị phân số vị chuẩn khác cho các quan sát của Newcomb về thời gian trôi quan của ánh sáng nhưng lần này đã loại bỏ hai số ngoại lai -44 và -2. Mục đích của việc loại trừ hai số ngoại lai là khuếch đại tính chuẩn của các quan sát còn lại. Như có thể nhìn thấy trong Hình 1(d), một mô hình chuẩn là hoàn toàn phù hợp. Như Hình 1(d) đã minh hoạ, các số liệu thực hầu như luôn thể hiện một vài sự sai lệch nhỏ so với các mô hình phân phối chuẩn lý thuyết (đó là các điểm không nằm hoàn toàn trên một đường thẳng). Thật quan trọng khi tập trung việc kiểm tra của bạn

vào tìm kiếm các dạng mẫu bộc lộ các sai lệch rõ ràng so với phân phối chuẩn trong một đồ thị phân số vị chuẩn, và đừng khăng khăng quá mức đối với các sai lệch nhỏ trong đồ thị.

Tóm lại, các đồ thị các phân số vị chuẩn là một công cụ hữu ích được khuyến nghị nên dùng để khảo xét tính phân phối chuẩn của một tệp số liệu.

Việt Linh

Tổng thuật từ Davis S, Moore và George P. McCabe, Introduction to the Practice of Statistics, 2nd Edition. pp 63-86, W.H. Freeman & Company, New York, 1993.