XÁC ĐỊNH CÁC ĐẶC TRƯNG THỐNG KÊ CỦA CHUỖI SỐ LIỆU TRONG LÂM HỌC

1. Một số thuật ngữ

(a) Các quan sát. Thuật ngữ này biểu thị các trường hợp[1] hay chủ thể. Một trường hợp bao gồm nhiều thông tin của một đại lượng nghiên cứu. Ví dụ: Đại lượng quan sát là đường kính ngang ngực (D, cm), chiều cao thân cây (H, m), tuổi của rừng; (A, năm), tiết diện ngang của quần thụ (G, m^2/ha), trữ lượng hay sản lượng gỗ của quần thụ (M, m^3/ha)…Đó là những đại lượng quan sát trong khi mô tả đối tượng nghiên cứu.

(b) Các biến[2] hay tiêu thức. Thuật ngữ này biểu thị các thông tin cần được thu thập, xác định hay dự đoán. Ví dụ: D, H, G, M…

(c) Tên biến. Đó là tên gọi của biến quan sát. Mỗi biến được đặt một tên biến[3] cụ thể và ngắn gọn. Ví dụ: Kí hiệu D là tên ngắn gọn của biến đường kính ngang ngực. Sở dĩ tên biến phải được viết ngắn gọn là vì bảng số liệu có thể chứa nhiều biến. Để mô tả tên đầy đủ của một biến, người ta dùng nhãn biến[4]. Ví dụ: Nhãn biến D là "Đường kính ngang ngực".

2. Phương pháp xử lý số liệu

Số liệu thống kê thường là rất nhiều và chúng được ghi lại trong một bảng ma trận. Nếu quan sát số liệu trong bảng ma trận này, chúng ta khó nhận ra những thông tin mong muốn. Để phát hiện tính quy luật của chuỗi số liệu, chúng ta cần phải vận dụng công cụ toán học. Thống kê mô tả (Descriptive Statistics) giúp chúng ta xác định những đặc trưng thống kê của chuỗi số liệu. Đó là dung lượng mẫu (N); giá trị trung bình (Mean, Average = X_{bq}); trung vị (Me = Median); giá trị xuất hiện nhiều nhất (M_0 = Mode); giá trị lớn nhất (X_{Max}); giá trị

[1] Case
[2] Variables
[3] Variable name
[4] Variable label

lớn nhỏ nhất (X_{Min}); phạm vi biến động (Range = R = X_{Max} - X_{Min}); phương sai (Variance = S^2); sai tiêu chuẩn (SEE = Standard Error of the Estimate hoặc S = Standard Deviation); độ lệch (Skewness = Sk), độ nhọn (Kurtosis = Ku), khoảng tin cậy của trung bình mẫu (Confidence interval of the sample mean)…

2.1. Những số đo mức độ tập trung của chuỗi số liệu

Mức độ tập trung của chuỗi số liệu được đo lường bằng ba thống kê là trung bình (Mean), trung vị (Median) và số xuất hiện nhiều nhất (Mode).

(a) Trung bình mẫu (Sample Mean). Trong thống kê toán học, người ta sử dụng nhiều khái niệm khác nhau về số trung bình như trung bình cộng (trung bình số học = Arithmetic Average), trung bình nhân (Multiplictional Average), trung bình điều hòa (Harmonic Average), trung bình hình học (Geometric Average), trung bình bình phương (Squared Average)…Tuy nhiên, trong lĩnh vực lâm học, hai trung bình được sử dụng phổ biến là trung bình số học (trung bình cộng) và trung bình nhân. Ý nghĩa cơ bản của giá trị trung bình cộng là nó được sử dụng để phân tích độ lớn của các thành phần của một biến số và dung hòa các dao động cao thấp. Giá trị trung bình của một mẫu quan sát được gọi là trung bình mẫu. Giá trị trung bình của một biến số (X_{Bq}) được tính theo trung bình cộng đơn giản và trung bình cộng gia quyền. Trong trường hợp mẫu lớn (n > 30), giá trị X_{Bq} đơn giản được xác định theo công thức 1. Khi mẫu nhỏ (n < 30), giá trị X_{Bq} đơn giản được xác định theo công thức 2.

$$X_{Bq} = \frac{(X_1 + X_2 + \ldots + X_n)}{n} \qquad (1)$$

$$X_{Bq} = \frac{(X_1 + X_2 + \ldots + X_n)}{n - 1} \qquad (2)$$

Khi nhiều biến quan sát nhận những giá trị như nhau, người ta xây dựng bảng phân phối tần số theo nhóm (tổ, cấp) quan sát. Khi số liệu được ghép thành nhóm, thì giá trị trung bình của chuỗi số liệu được tính theo trung bình cộng gia quyền (Trung bình có trọng số = The Weighted Mean). Trong trường hợp mẫu

lớn (n > 30), giá trị X_{Bq} gia quyền được xác định theo công thức 3. Khi mẫu nhỏ (n < 30), giá trị X_{Bq} gia quyền được xác định theo công thức 4. Ở công thức 3 và 4, n = tổng số quan sát; f_i (i = 1 – n) là tần số quan sát của các biến X_i, n = f_1+ f_2 +…+ f_n.

$$X_{Bq} = \frac{(f_1X_1 + f_2X_2 + \ldots + f_nX_n)}{n} \qquad (3)$$

$$X_{Bq} = \frac{(f_1X_1 + f_2X_2 + \ldots + f_nX_n)}{n - 1} \qquad (4)$$

Ngoài cách xác định trung bình cộng đơn giản, người ta còn xác định trung bình nhân. Trung bình nhân là căn bậc n của tích các giá trị của chuỗi số liệu (Công thức 5). Khi chuỗi số liệu được ghép thành nhóm, thì trung bình nhân được xác định theo Công thức 6.

$$X_{Bq} = \sqrt[n]{(Xi * Xi + 1 * \ldots * Xi + k)} \qquad (5)$$

$$X_{Bq} = \sqrt[n]{X1^{f1} * X2^{f2} * \ldots * Xk^{fk}} \qquad (6)$$

(b) Trung vị mẫu. Trung vị mẫu (Kí hiệu = Me = Sample Median) là số đứng ở vị trí trung tâm của chuỗi số liệu theo thứ tự tăng dần. Nếu số lượng của chuỗi số liệu là số lẻ, thì Me là số nằm ở trung tâm của chuỗi số liệu. Nếu số lượng của chuỗi số liệu là số chẵn, thì Me là số trung bình của cặp số nằm chính giữa của chuỗi số liệu được sắp xếp theo thứ tự từ nhỏ đến lớn.

Khi chuỗi số liệu bao gồm n phần tử và không được phân chia thành các tổ. Nếu n là số lẻ, thì Me được xác định theo công thức 7. Nếu n là số chẵn, thì Me được xác định theo công thức 8.

Khi chuỗi số liệu bao gồm n phần tử được phân chia thành các tổ; trong đó cự ly tổ là K. Trong trường hợp này, Me là giá trị thuộc tổ có tần số tích lũy lớn nhất. Gọi f_M là tần số của tổ M; F_M là tần số tích lũy từ tổ đầu tiên đến tổ M. Để xác định Me, trước hết xác định tổ có tần số lớn nhất (f_{Max}). Kế đến xác định tần số tích lũy từ tổ đầu tiên đến tổ có tần số lớn nhất (F_M) và tổ M - 1 (F_{M-1}). Sau đó xác định Me theo công thức 9; trong đó X_M là giới hạn dưới của tổ thứ M

có tần số lớn nhất (f_{Max}); K là khoảng cách của mỗi tổ; $X_{(n/2)}$ là giá trị ở vị trí n/2; F_M là tần số tích lũy ở tổ có tần số lớn nhất, $F_{(M-1)}$ là tần số tích lũy ở tổ M - 1.

$$Me = X_{(n+1)/2} \quad (7)$$

$$Me = [\frac{(X_{(n/2)} + X_{(n/2)+1})}{2}] \quad (8)$$

$$Me = X_M + K\times[\frac{(n/2 - F_{(M-1)})}{f_M}] \quad (9)$$

Trung vị là một đặc trưng thống kê của chuỗi số liệu. Ưu điểm của Me là nó không phụ thuộc vào kiểu phân bố của chuỗi số liệu. Khi chuỗi số liệu có phân bố lệch chuẩn, thì trung vị là đại diện tốt hơn so với giá trị trung bình.

(c) Giá trị xuất hiện nhiều nhất. Trong một chuỗi số liệu, những số xuất hiện nhiều nhất được gọi là Mốt (Kí hiệu = Mo = Mode). Nói cách khác, Mo là số có tần số cao nhất trong bảng số liệu đã được ghép thành tổ hay nhóm. Tùy theo cách thức tập hợp số liệu, giá trị Mo được xác định khác nhau.

Khi chuỗi số liệu được phân tổ với khoảng cách đều nhau, Mo là giá trị thuộc tổ có tần số lớn nhất và được xác định theo công thức 10. Ở công thức 10, X_M là giá trị ở cận dưới của tổ chứa Mo; K là khoảng cách của mỗi tổ; f_{Mo} là tần số của tổ chứa Mo; $f_{(Mo-1)}$ là tần số của tổ đứng trước tổ chứa Mo; $f_{(Mo+1)}$ là tần số của tổ đứng sau tổ chứa Mo.

$$Mo = X_M + K\times[\frac{(f_{Mo} - f_{(Mo-1)})}{(f_{Mo} - f_{(Mo-1)}) + (f_{Mo} - f_{(Mo+1)})}] \quad (10)$$

Ưu điểm của Mo là nó không chịu ảnh hưởng của các giá trị cực trị (giá trị lớn nhất và giá trị nhỏ nhất). Nhược điểm của Mo là nó kém nhạy bén với sự biến thiên của chuỗi số liệu. Trong thực tế, giá trị Mo được sử dụng ít hơn so với Me và số trung bình.

2.2. Những số đo mức độ phân tán của chuỗi số liệu

Các giá trị trung bình, Me và Mo chỉ cho biết những giá tri trung tâm của chuỗi số liệu. Chúng phản ánh không đầy đủ các tính chất của chuỗi số liệu. Vì thế, để đo lường mức độ phân tán của chuỗi số liệu, người ta sử dụng một số

thống kê như khoảng biến thiên hay phạm vi biến động (Range = R = Max-Min), độ lệch (Bias), phương sai (Variance = S^2), sai tiêu chuẩn hay độ lệch chuẩn (Standard Deviation = S), hệ số biến động (Coefficient of Variation = CV%), độ lệch (Skewness = Sk) và độ nhọn (Kurtosis = Ku).

(1) Khoảng biến thiên (R). Đó là thống kê phản ánh sự chênh lệch giữa giá trị lớn nhất (X_{max}) và giá tri nhỏ nhất (X_{min}) của chuỗi số liệu. Giá trị R được xác định theo công thức 11. Khoảng biến thiên càng nhỏ thì tổng thể càng đồng đều và số trung bình có tính đại diện càng cao và ngược lại.

$$R = X_{Max} - X_{Min} \qquad (11)$$

(2) Độ lệch (Bias). Đó là chênh lệch ay sai số hệ thống giữa các giá trị quan sát (X_i) so với giá trị trung bình mẫu (X_{Bq}) của chuỗi số liệu. Thống kê Bias được xác định theo công thức 12.

$$Bias = X_i - X_{Bq} \qquad (12)$$

Trong thống kê, người ta sử dụng Bias để đo sự phân tán của các quan sát so với trung bình mẫu. Các giá trị Bias có thể nhận giá trị âm hoặc giá trị dương. Nếu tổng Bias dương, thì những giá trị lớn hơn giá trị trung bình nhiều hơn so với những giá trị nhỏ hơn giá trị trung bình. Ngược lại, nếu tổng Bias âm, thì những giá trị lớn hơn giá trị trung bình nhỏ hơn so với những giá trị nhỏ hơn giá trị trung bình. Bởi vì các giá trị Bias có thể nhận giá trị âm hoặc giá trị dương, nên khi dung lượng mẫu lớn (n > 30), thì tổng Bias của chuỗi số liệu bằng Zero. Điều đó gây ra khó khăn trong việc phân tích và so sánh giữa các tập số liệu. Để loại bỏ hiện tượng Bias = 0, người ta xác định sai lệch tuyệt đối trung bình (MAE = Mean Absolute Error) (Công thức 13) và sai lệch tuyệt đối trung bình theo phần trăm (MAPE = Mean Absolute Percent Error) (Công thức 14).

$$MAE = \left| \frac{(X_i - X_{Bq})}{n} \right| \qquad (13)$$

$$MAPE = \left| \frac{(X_i - X_{Bq})}{X_{Bq}} \right| 100 \qquad (14)$$

(3) Phương sai (S^2). Đó là số bình quân cộng của bình phương các độ lệch (Bias) giữa các giá trị quan sát (X_i) so với số trung bình của chuỗi số liệu. Khi chuỗi số liệu không được ghép thành nhóm và n > 30, thì phương sai mẫu được xác định theo công thức 15. Khi chuỗi số liệu không được ghép thành nhóm và n < 30, thì phương sai mẫu được xác định theo công thức 16. Khi chuỗi số liệu được ghép thành nhóm và n > 30, thì phương sai mẫu được xác định theo công thức 17; trong đó X^*_i và f_i tương ứng là giá trị giữa và tần số của mỗi nhóm, còn tổng $f_i = n$. Khi chuỗi số liệu không được ghép thành nhóm và n < 30, thì phương sai mẫu được xác định theo công thức 18.

$$S^2 = \frac{(X_i - X_{Bq})^2}{n} \qquad (15)$$

$$S^2 = \frac{(X_i - X_{Bq})^2}{n - 1} \qquad (16)$$

$$S^2 = \frac{f_i(X_i^* - X_{Bq})^2}{n} \qquad (17)$$

$$S^2 = \frac{(X_i - X_{Bq})^2}{n - 1} \qquad (18)$$

(4) Sai tiêu chuẩn (SEE). Đó là căn bậc 2 của phương sai. Khi chuỗi số liệu không được ghép thành nhóm và n > 30, thì sai tiêu chuẩn mẫu được xác định theo công thức 19. Khi chuỗi số liệu không được ghép thành nhóm và n < 30, thì sai tiêu chuẩn mẫu được xác định theo công thức 20. Khi chuỗi số liệu được ghép thành nhóm và n > 30, thì sai tiêu chuẩn mẫu được xác định theo công thức 21. Khi chuỗi số liệu được ghép thành nhóm và n < 30, thì sai tiêu chuẩn mẫu được xác định theo công thức 22.

$$SEE = \sqrt[2]{\frac{(Xi - XBq)\text{^}2}{n}} \qquad (19)$$

$$SEE = \sqrt[2]{\frac{(Xi - XBq)^2}{n - 1}} \qquad (20)$$

$$SEE = \sqrt[2]{\frac{fi(Xi - XBq)^2}{n}} \qquad (21)$$

$$SEE = \sqrt[2]{\frac{fi(Xi - XBq)^2}{n - 1}} \qquad (22)$$

(5) Chuẩn hóa số liệu. Chuẩn hóa số liệu là loại bỏ các đơn vị đo của các biến số trong chuỗi số liệu. Trong thống kê mô tả, người ta gọi chuẩn hóa số liệu là Z – Score (*Giá trị chuẩn hóa*). Giá trị Z - Score là tỉ lệ giữa các sai lệch của các giá trị quan sát so với giá trị trung bình và sai tiêu chuẩn mẫu. Giá trị Z – Score được xác định theo công thức 23. Trong phân tích hồi quy tuyến tính đa biến, chuẩn hóa số liệu giúp ích cho việc xác định vai trò của các biến dự đoán trong hàm phản hồi.

$$Z - Score = [\frac{(X_i - X_{Bq})}{S}] \qquad (23)$$

(6) Hệ số biến động (CV%). Hệ số biến động biểu thị biến động theo phần trăm giữa các giá trị quan sát so với giá trị trung bình mẫu. Hệ số CV% của mẫu được xác định theo công thức 24.

$$CV\% = \frac{SEE}{X_{Bq}} 100 \qquad (24)$$

(7) Độ lệch (Sk). Độ lệch (Sk) đo tính đối xứng hoặc tính bất đối xứng của chuỗi số liệu. Phân bố của tập dữ liệu là đối xứng nếu chuỗi số liệu ở bên trái và bên phải của giá trị trung tâm (trung bình, Me, Mo) là như nhau. Vì thế, giá trị Sk được sử dụng để biểu thị độ lệch (độ xiên) của chuỗi số liệu so với giá trị trung tâm.

Khi chuỗi số liệu không được ghép thành nhóm, thì hệ số Sk được xác định theo công thức 25. Khi chuỗi số liệu được ghép thành nhóm, thì hệ số Sk được xác định theo công thức 26.

$$Sk = \frac{(X_i - X_{Bq})^3}{n \times S^3} \qquad (25)$$

$$Sk = \frac{f_i(X_i - X_{Bq})^3}{n \times S^3} \qquad (26)$$

Hệ số Sk có thể nhận giá trị bằng zero, giá trị âm hoặc giá trị dương. Khi chuỗi số liệu phân bố chuẩn, thì hệ số Sk = 0. Khi hệ số Sk > 0, thì chuỗi số liệu phân bố lệch trá . Nguyên nhân là vì tổng tam thừa của những chênh lệch dương lớn hơn so với tổng tam thừa của những chênh lệch âm so với giá trị trung bình mẫu. Trái lại, khi hệ số Sk < 0, thì chuỗi số liệu phân bố lệch phải. Nguyên nhân là vì tổng tam thừa của những chênh lệch dương nhỏ hơn so với tổng tam thừa của những chênh lệch âm so với giá trị trung bình mẫu.

(8) Độ nhọn (Ku). Độ nhọn (Ku) biểu thị mức độ tập trung của các giá trị quan sát xung quanh giá trị trung tâm (trung bình, Me, Mo). Khi chuỗi số liệu không được ghép thành nhóm, thì hệ số Ku được xác định theo công thức 3.27. Khi chuỗi số liệu được ghép thành nhóm, thì hệ số Ku được xác định theo công thức 3.28.

$$Ku = \frac{(X_i - X_{Bq})^4}{n \times S^4} - 3 \qquad (27)$$

$$Ku = \frac{f_i(X_i - X_{Bq})^4}{n \times S^4} - 3 \qquad (28)$$

Hệ số Ku có thể nhận giá trị bằng Zero, giá trị âm hoặc giá trị dương. Khi chuỗi số liệu phân bố chuẩn, thì hệ số Ku = 3. Khi hệ số Ku > 0, thì đỉnh đường cong của chuỗi số liệu có dạng nhọn. Điều đó có nghĩa là chuỗi số liệu có nhiều giá trị phân bố xung quanh giá trị trung bình. Trái lại, khi hệ số Ku < 0, thì đỉnh đường cong của chuỗi số liệu có dạng bẹt. Điều đó có nghĩa là chuỗi số liệu có ít giá trị phân bố xung quanh giá trị trung bình.

2.3. Ước lượng khoảng của trung bình mẫu

Một câu hỏi đặt ra là trung bình mẫu được ước lượng với độ tin cậy như thế nào? Để chỉ rõ mức độ tin cậy của trung bình mẫu, người ta sử dụng thống kê khoảng ước lượng của trung bình mẫu (hay khoảng tin cậy của trung bình mẫu = Confidence Interval).

Khi dung lượng mẫu nhỏ (n < 30), khoảng ước lượng của trung bình mẫu được xác định theo công thức 29. Khi dung lượng mẫu lớn (n > 30), khoảng ước lượng của trung bình mẫu được xác định theo công thức 30. Ở công thức (29) và (30), X_{Bq} là trung bình mẫu; SEE là sai tiêu chuẩn mẫu; n là dung lượng mẫu quan sát; $t_{\alpha/2}$ là giá trị t tính toán (t – value) của phân bố t – Strudent ở mức ý nghĩa α/2; $z_{\alpha/2}$ là giá trị Z tính toán (Z – value) của phân bố chuẩn tiêu chuẩn ở mức ý nghĩa α/2. Giá trị $Z_{\alpha/2}(SEE/\sqrt{n})$ = E là giới hạn sai số của trung bình mẫu.

$$X_{Bq} \pm t_{\alpha/2} \times \left(\frac{SEE}{\sqrt{n}}\right) \qquad (29)$$

$$X_{Bq} \pm Z_{\alpha/2} \times \left(\frac{SEE}{\sqrt{n}}\right) \qquad (30)$$

Trong lâm học, trung bình mẫu được ước lượng với khoảng tin cậy 95% hay mức ý nghĩa α = 0,05 hay 5%. Giả sử quần thể có phân bố chuẩn, trung bình mẫu X_{Bq} = 100; SEE = 10 và n = 25. Bởi vì n < 30, nên khoảng tin cậy của trung bình mẫu được xác định theo công thức 29. Tra bảng phân bố t với n – 1 = 25 – 1 = 24 độ tự do và mức ý nghĩa α/2 = 0,05/2 = 0,025, chúng ta nhận được t – value = 2,064. Như vậy, khoảng tin cậy của trung bình mẫu là 100 ± 2,064*(10/Sqrt(25)) = 100 ± 4,1 hay [95,9; 104,1]. Trong phần thuyết minh kết quả, chúng ta tuyên bố giá trị trung bình mẫu bằng 100 với khoảng tin cậy 95% từ 95,9 đến 104,1. Khi tuyên bố khoảng tin cậy 95%, chúng ta tin đến 95% rằng giá trị trung bình mẫu nằm trong khoảng từ 95,9 đến 104,1. Nếu sai, thì sai số là 5%.

Quy tắc kinh nghiệm

Nếu chuỗi số liệu có phân bố chuẩn, thì ba khoảng tin cậy của trung bình mẫu được xác định theo công thức 31÷33. Ký hiệu "P" chỉ xác suất mà giá trị X_{bq} nằm trong khoảng từ X_i đến X_j.

$$P(X_{Bq} - SEE < X_{Bq} < X_{Bq} + SEE) \geq 68\% \quad (31)$$

$$P(X_{Bq} - 2\times SEE < X_{Bq} < X_{Bq} + 2\times SEE) \geq 95\% \quad (32)$$

$$P(X_{Bq} - 3\times SEE < X_{Bq} < X_{Bq} + 3\times SEE) \geq 99{,}7\% \quad (33)$$

Theo định lý Chebyshev, nếu chuỗi số liệu có phân bố tiệm cận chuẩn, thì ba khoảng tin cậy của trung bình mẫu được xác định theo công thức 34÷36.

$$P(X_{Bq} - SEE < X_{Bq} < X_{Bq} + SEE) \geq 68\% \quad (34)$$

$$P(X_{Bq} - 2\times SEE < X_{Bq} < X_{Bq} + 2\times SEE) \geq 95\% \quad (35)$$

$$P(X_{Bq} - 3\times SEE < X_{Bq} < X_{Bq} + 3\times SEE) \geq 99{,}7\% \quad (36)$$

Hiện nay tất cả những thống kê mô tả (trung bình, phương sai, giá trị lớn nhất và giá trị lớn nhỏ nhất, phạm vi biến động, độ lệch, độ nhọn, khoảng tin cậy của trung bình mẫu...) có thể được xác định dễ dàng bằng các phần mềm thống kê như Excel, SPSS (Statistical Products For The Special Sciences), SAS (Statistical Analysis System), STAT (Statgraphics...)...

Xác định dung lượng mẫu thích hợp

Độ tin cậy của các thống kê mô tả phụ thuộc vào dung lượng mẫu. Dung lượng mẫu nhỏ dẫn đến kết quả kém tin cậy. Dung lượng mẫu lớn đảm bảo nhận được kết quả với độ tin cậy cao nhưng lại làm tăng chi chí nghiên cứu (thời gian, nhân lực, kinh phí…). Để xác định dung lượng mẫu thích hợp, chúng ta cần biết trung bình và sai tiêu chuẩn của tổng thể. Tuy vậy, hai giá trị này thường không biết trước khi nghiên cứu. Trong thực hành, trước hết sử dụng một mẫu với kích thước nhất định. Kế đến, từ mẫu này, xác định X_{Bq} và SEE. Sau đó, từ $E = Z_{\alpha/2}(SEE/\sqrt{n})$, chúng ta xác định dung lượng mẫu thích hợp (n) theo công thức 37. Thông thường, khoảng tin cậy của trung bình mẫu là 95%, còn $\alpha = 0{,}05$ và $\alpha/2 = 0{,}025$, $Z_{\alpha/2} = 1{,}96$.

Khi biết hệ số biến động (CV%) và yêu cầu đối với sai số của số liệu điều tra (p%), thì dung lượng mẫu thích hợp được xác định gần đúng theo công thức 38. Trong lâm học và điều tra rừng, yêu cầu đối với sai số của số liệu điều tra là p = 5%.

$$n = (Z_{\alpha/2})^2 S^2 / E^2 \qquad (37)$$

$$n = \frac{CV^2}{p^2} \qquad (38)$$