Chuyển đổi dữ liệu trong nghiên cứu thực nghiệm

■ **TS. NGUYỄN THỊ THU NGÀ; ThS. KIỀU LAN HƯƠNG**
Trường Đại học Công nghệ Giao thông vận tải
■ **ThS. NGUYỄN TUẤN TÚ**
Trường Đại học Công nghiệp Hà Nội

TÓM TẮT: Trong nghiên cứu thực nghiệm, một trong những giả định phổ biến nhất đối với phương pháp phân tích thống kê là giả định dữ liệu phân phối chuẩn, với gần như tất cả các phân tích tham số đều yêu cầu giả định này hoặc theo cách này hay cách khác. Mặc dù không phải tất cả các giả định chuẩn tắc đều liên quan trực tiếp đến phân phối của một biến riêng lẻ, nhưng việc đáp ứng giả định này thường dễ dàng hơn nếu mỗi biến trong phân tích tuân thủ theo phân phối chuẩn. Song, không phải lúc nào dữ liệu thực nghiệm cũng có phân phối chuẩn. Để khắc phục điều này, có thể áp dụng phương pháp chuyển đổi dữ liệu bằng cách áp dụng một hàm toán học cho giá trị dữ liệu ban đầu để tạo ra bộ dữ liệu mới gần hơn với dạng phân phối chuẩn, điều này sẽ thuận tiện cho việc phân tích mẫu dữ liệu thực nghiệm. Trong phạm vi bài báo, một số cách đánh giá phân phối của dữ liệu và phép chuyển đổi được đề cập, giúp người nghiên cứu có cái nhìn tổng quan hơn trong việc đánh giá dữ liệu của mình.

TỪ KHÓA: Phân phối chuẩn, dữ liệu, chuyển đổi.

ABSTRACT: In empirical research, one of the most common assumptions for statistical analysis methods is the assumption that the data are normally distributed, with nearly all parametric analyzes requiring this assumption or in some way. one way or another. Although not all normal assumptions are directly related to the distribution of an individual variable, it is often easier to satisfy this assumption if each variable in the analysis conforms to a normal distribution. But, the experimental data is not always normally distributed. To overcome this, it is possible to apply a data transformation method by applying a mathematical function to the original data value to create a new dataset that is closer to the normal distribution, which will be convenient. convenient for sample analysis of experimental data. Within the scope of the article, a number of ways to evaluate the distribution of the data and the transformation are mentioned, helping researchers have a better overview in evaluating their data.

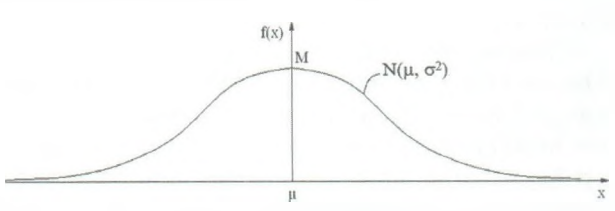
KEYWORDS: Normal distribution, data, transformation.

1. ĐẶT VẤN ĐỀ

Phân phối chuẩn còn được gọi là phân phối Gauss (Gaussian distribution), một phân phối xác suất liên tục rất phổ biến. Trong cuốn *Theorie Analytique des Probabilites,* nhà toán học danh tiếng Carl F.Gauss phát triển các đặc điểm của luật phân phối chuẩn và chỉ ra rằng luật phân phối này phù hợp với các hiện tượng tự nhiên. Hầu hết các hiện tượng sinh học tự nhiên (như chiều cao, trọng lượng cơ thể, huyết áp, mật độ xương, chỉ số IQ...) đều có thể mô tả bằng luật phân phối chuẩn một cách chính xác. Phân phối chuẩn của đại lượng ngẫu nhiên X, có thể nhận mọi giá trị trên trục số thực thỏa mãn phân phối chuẩn nếu hàm mật độ cho bởi:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

Trong đó: μ - Giá trị trung bình của phân phối; σ^2 - Phương sai của phân phối; σ - Độ lệch chuẩn của phân phối; $-\infty \le x \le \infty$ [1]. Người ta kí hiệu đơn giản phân phối chuẩn là $X \in N(\mu, \sigma^2)$. Có nhiều cách để thể hiện các đặc tính của một phân phối xác suất, như thông qua hàm mật độ xác suất (có dạng như hình chuông, đối xứng qua giá trị trung bình x = μ), nó cho biết khả năng xảy ra của mỗi giá trị của biến ngẫu nhiên.



Hình 1.1: Đồ thị hàm phân phối chuẩn

Phân phối chuẩn là phân phối xác suất liên tục đối xứng xung quanh giá trị trung bình của nó, hầu hết các quan sát tập hợp xung quanh đỉnh trung tâm và xác suất đối với các

iá trị xa trung bình giảm dần theo cả hai hướng. Các giá trị ực trị ở cả hai phía của phân phối là khó xảy ra tương tự. rong khi phân phối chuẩn là đối xứng, nhưng không phải ất cả các phân phối đối xứng đều chuẩn như phân phối tudent's t, Cauchy và logistic.

Nếu Xi ~ N(μ, σ²), 1≤ i ≤ n là các biến ngẫu nhiên độc ập, có giá trị trung bình $\overline{X}$ thì $\overline{X} \sim N(\mu,\sigma^2)$ => định lý giới ạn trung tâm được xem là một trong những định lý quan rọng nhất của lý thuyết xác suất. Nó có thể giải thích tại sao ất nhiều hiện tượng tự nhiên đã quan sát được là tương tự hư phân phối chuẩn bởi chúng có thể coi như được hợp hành từ rất nhiều các biến cố nhỏ hơn. Trong ngành kỹ huật nói chung, thuyết này cho phép thực hiện tham chiếu hống kê các đặc tính của toàn bộ đám đông dựa trên cơ ở của một tập mẫu có kích thước n, đặc biệt nó cho phép ánh giá mức thống kê khả năng sai số đi liền với việc ước ượng kỳ vọng và phương sai của đám đông từ một số liệu. Chính vì vậy, luật phân phối chuẩn được ứng dụng cực kì ộng rãi trong khoa học thực nghiệm, thị trường chứng hoán và trong những loại phân tích đo lường khác. Có thể ói rằng, phân phối chuẩn là nền tảng, là trụ cột của tất cả ác phân tích thống kê. Không có luật phân phối này cũng ó nghĩa là không có khoa học thống kê hiện đại. Khi có dãy ố liệu quan sát được, dựa vào cơ sở lý thuyết của phân phối huẩn để đưa ra các kết luận quan trọng làm tiền đề cho hững vấn đề giải quyết tiếp sau.

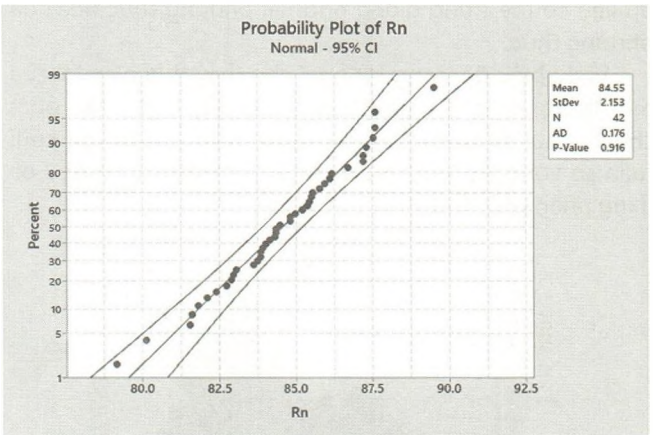
Các phương pháp thống kê tham số được phổ biến ộng rãi và được sử dụng trong các nghiên cứu thực nghiệm ói chung. Trong quá trình phân tích cần phải hoàn thành ý thuyết các giả định, chẳng hạn như tính chuẩn, tính đồng hất của phương sai và không có mối tương quan giữa các ai số. Việc thiếu tính chuẩn của các lỗi là ít quan trọng trong hử nghiệm của Fisher. Tuy nhiên, nó có thể ảnh hưởng đến ính đồng nhất của phương sai chủ yếu khi có sự khác biệt ớn về số lượng quan sát trong các nhóm hoặc phương pháp. Sự không đồng nhất này thường đi kèm với các biến không bình thường, vì vậy, khuyến nghị rằng các phép biến đổi được áp dụng để ổn định phương sai và chuẩn hóa các phản hồi [2, 3, 4].

Trong thực tế, nhiều dữ liệu nghiên cứu nói chung không hoàn thành các giả định cơ bản để áp dụng thống kê tham số bởi vì chúng về cơ bản là rời rạc. Điều này giải thích tại sao các biến thường không được điều chỉnh đến phân phối chuẩn như phân phối nhị thức hay phân phối xác suất rời rạc Poisson [5]. Các phương pháp thống kê phi tham số là cần thiết vì chúng không phụ thuộc vào dạng phân phối dữ liệu, có thể được sử dụng cho các mẫu nhỏ và thường nhanh hơn, đơn giản hơn để áp dụng.

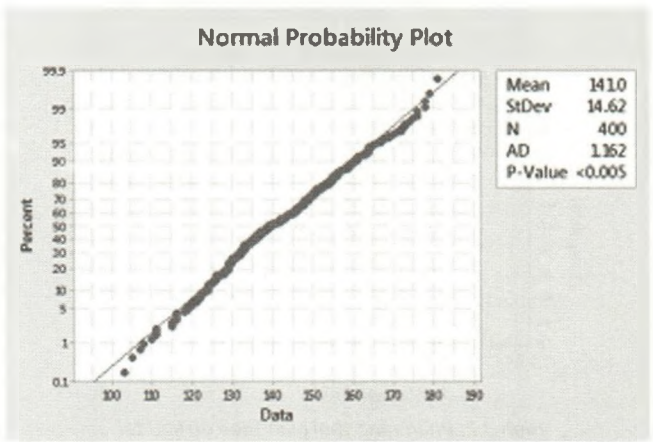
2. PHƯƠNG PHÁP ĐÁNH GIÁ DỮ LIỆU

Một số công cụ đồ họa và thống kê có thể được sử dụng để đánh giá xem dữ liệu của bạn có tuân theo phân phối chuẩn hay không, bao gồm:

- Biểu đồ Histogram cho biết tần suất xuất hiện các giá trị khác nhau của một biến trong dữ liệu. Biểu đồ thường được mô tả bằng một loạt các thanh được sắp xếp dọc theo trục x (đại diện cho các giá trị của biến) với độ dài của các thanh được hiển thị dọc theo trục y (đại diện cho tần suất của các giá trị), xem đường cong phân phối dữ liệu có giống hình chuông hay không;
- Kiểm tra phân phối chuẩn xem giá trị P có lớn hơn mức ý nghĩa α = 0,05;
- Xem biểu đồ phân phối chuẩn các điểm dữ liệu có bám sát đường thẳng.



Hình 2.1: Phân phối chuẩn của Rn



Hình 2.2: Phân phối không chuẩn

Ví dụ tập dữ liệu ở *Hình 2.1* cho thấy các điểm phân phối nằm dọc theo đường chuẩn và nằm trong khoảng giới hạn, có thể coi các dữ liệu tuân theo quy luật phân phối chuẩn. Thêm vào đó, giá trị p của bộ dữ liệu là 0,916 lớn hơn nhiều mức ý nghĩa α là 0,05. Trong khi đó ở *Hình 2.2*, với kích thước mẫu lớn (n = 400 > 200), phép kiểm định Anderson-Darling có thể phát hiện ra những biến bất thường nhỏ, chỉ ra tính không bình thường (p < 0,005) mặc dù dữ liệu phân phối bám sát đường chuẩn.

3. CÁC NGUYÊN NHÂN DẪN ĐẾN DỮ LIỆU PHÂN PHỐI KHÔNG CHUẨN

** Giá trị cực trị, ngoại lai:*

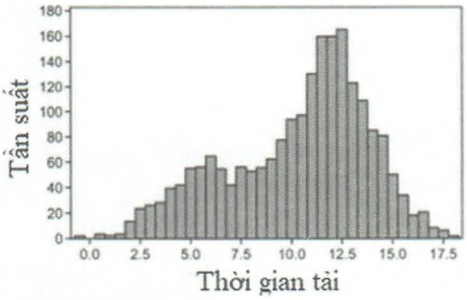
Quá nhiều giá trị cực trị, ngoại lai trong một tập dữ liệu sẽ dẫn đến phân phối sai lệch. Tính chuẩn của dữ liệu có thể đạt được bằng cách làm sạch dữ liệu. Điều này liên quan đến việc xác định lỗi đo lường, lỗi nhập dữ liệu và các lỗi ngoại lệ, đồng thời xóa chúng khỏi dữ liệu vì những lý do hợp lệ. Bản chất của dữ liệu được phân phối chuẩn là có thể mong đợi một tỷ lệ nhỏ các giá trị cực trị. Các giá trị cực trị

chỉ nên được giải thích và loại bỏ khỏi dữ liệu nếu có nhiều giá trị hơn mong đợi trong điều kiện bình thường.

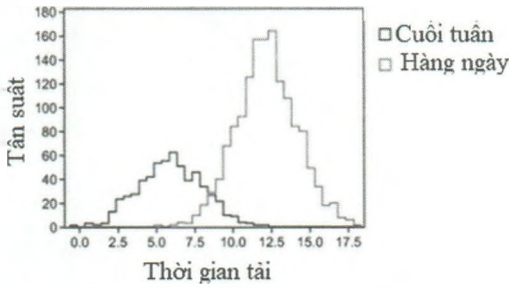
** Chồng chéo của hai hoặc nhiều quy trình:*

Dữ liệu có thể không được phân phối bình thường vì nó thật sự đến từ nhiều quá trình hoặc từ một quá trình thường xuyên thay đổi. Nếu hai hoặc nhiều tập dữ liệu thường được phân phối riêng của chúng bị chồng chéo, dữ liệu có thể trông giống như hai phương thức hoặc đa phương thức.

Ví dụ biểu đồ trong *Hình 3.1* cho thấy thời gian tải của một trang website không tuân theo quy luật phân phối chuẩn. Sau khi phân tầng thời gian tải theo dữ liệu cuối tuần so với ngày làm việc *(Hình 3.2)*, cả hai nhóm đều có dạng phân phối chuẩn.



Hình 3.1: Dữ liệu thời gian tải trang website



Hình 3.2: Phân tầng thời gian theo dữ liệu tải

** Phân biệt dữ liệu không đủ:*

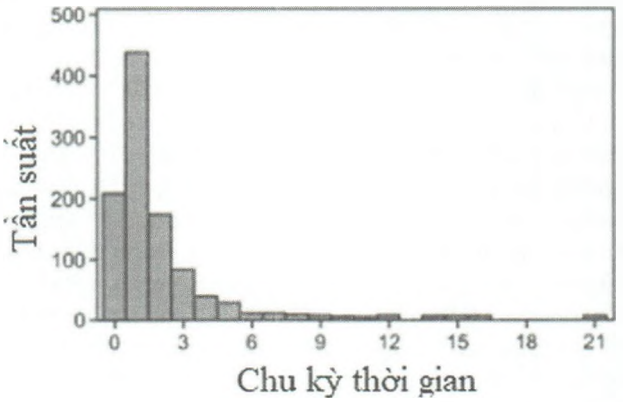
Lỗi vòng lặp hoặc thiết bị đo lường có độ phân giải kém có thể làm cho dữ liệu thật sự liên tục và được phân phối chuẩn trông rời rạc và không bình thường. Sự phân biệt dữ liệu không đủ - không đủ số lượng các giá trị khác nhau - có thể được khắc phục bằng cách sử dụng các hệ thống đo lường chính xác hơn hoặc bằng cách thu thập nhiều dữ liệu hơn.

** Dữ liệu được sắp xếp:*

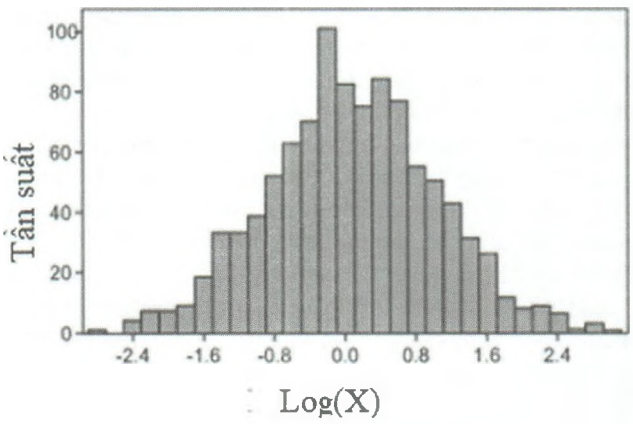
Dữ liệu đã thu thập có thể không được phân phối bình thường nếu nó chỉ đại diện cho một tập hợp con của tổng sản lượng mà một quá trình tạo ra, điều này có thể xảy ra nếu dữ liệu được thu thập và phân tích sau khi phân loại.

** Giá trị gần bằng 0 hoặc giới hạn tự nhiên:*

Nếu một quy trình có nhiều giá trị gần bằng 0 hoặc giới hạn tự nhiên, thì phân phối dữ liệu sẽ lệch sang phải hoặc trái. Trong trường hợp này, một phép biến đổi, chẳng hạn như phép biến đổi Box-Cox, có thể giúp dữ liệu tuân theo quy luật phân phối chuẩn. Khi so sánh dữ liệu được chuyển đổi, mọi thứ được so sánh phải được chuyển đổi theo cùng một cách. Ví dụ trong *Hình 3.3* cho thấy một tập hợp dữ liệu chu kỳ-thời gian (X); *Hình 3.4* cho thấy cùng một dữ liệu được biến đổi với logarit tự nhiên.



Hình 3.3: Tập dữ liệu chu kỳ thời gian



Hình 3.4: Log dữ liệu chu kỳ thời gian

** Dữ liệu theo một phân phối khác:*

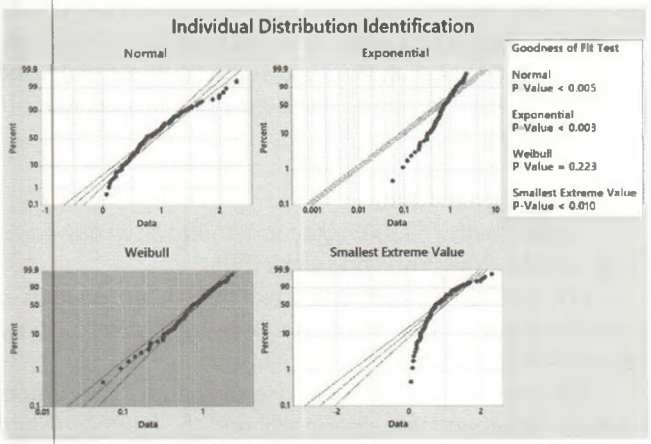
Khi dữ liệu không chuẩn cần xem xét các cách sau:

- Kiểm tra đánh giá các điểm ngoại lai (outliers), nên loại bỏ chúng và kiểm tra lại dữ liệu có phân phối chuẩn không. Đây là một phần thiết yếu của việc phân tích dữ liệu ban đầu, tuy nhiên cần lưu ý tác động của việc loại bỏ chúng.
- Sử dụng các thủ tục thống kê phi tham số (non-parametric statistics);
- Cân nhắc việc chuyển đổi dữ liệu, rất hữu ích trong những trường hợp nếu dữ liệu bị lệch, việc biến đổi dữ liệu sẽ giảm thiểu ảnh hưởng của các giá trị ngoại lai. Song, các phép biến đổi không nên được sử dụng thường xuyên bởi các thống kê như F-test, T-test nói chung là rất mạnh nên việc giải thích các giá trị sau khi biến đổi có thể có vấn đề.

Đối với một số phân tích nhất định có độ nhạy cao với giả định về tính phân phối chuẩn, chẳng hạn như độ tin cậy và khả năng tồn tại, xác suất, tìm một phân phối phù hợp với dữ liệu là rất quan trọng. Đối với các ứng dụng đáng tin cậy, việc có dữ liệu phân phối không chuẩn là điều khá bình thường, thì dữ liệu đó phải được xử lý bằng các công cụ tương tự như với dữ liệu không thể được "tạo ra" chuẩn như phân phối nhị thức, phân phối Poisson, phân phối Weibull...

Ví dụ phân phối Weibull thường gặp khi lập mô hình dữ liệu thời gian-lỗi (time - to - failure), phân phối này có thể lệch trái, lệch phải hoặc thậm chí gần đúng đối xứng

Khi không chắc cách phân phối nào phù hợp nhất với dữ liệu, sau đó bạn có thể sử dụng các công cụ như Minitab's Individual nhận dạng phân phối để tìm hiểu. Khi dữ liệu tuân theo Weibull, theo cấp số mũ hoặc một số phân phối không chuẩn khác thì không nhất thiết phải sử dụng phân phối chuẩn để phân tích dữ liệu, thay vào đó nên sử dụng bản phân phối phù hợp nhất với dữ liệu để phân tích. Ví dụ *Hình 3.5* cho thấy, các đồ thị xác suất và giá trị p thì phân phối Weibull phù hợp nhất cho tập dữ liệu này.



Hình 3.5: Một số dạng phân phối

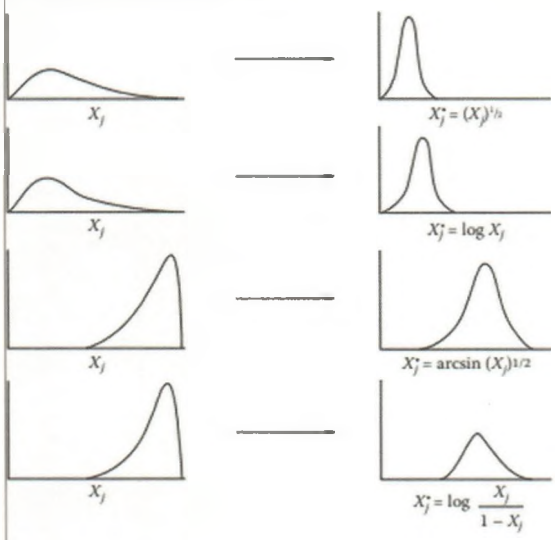
4. CHUYỂN ĐỔI DỮ LIỆU SANG PHÂN PHỐI CHUẨN

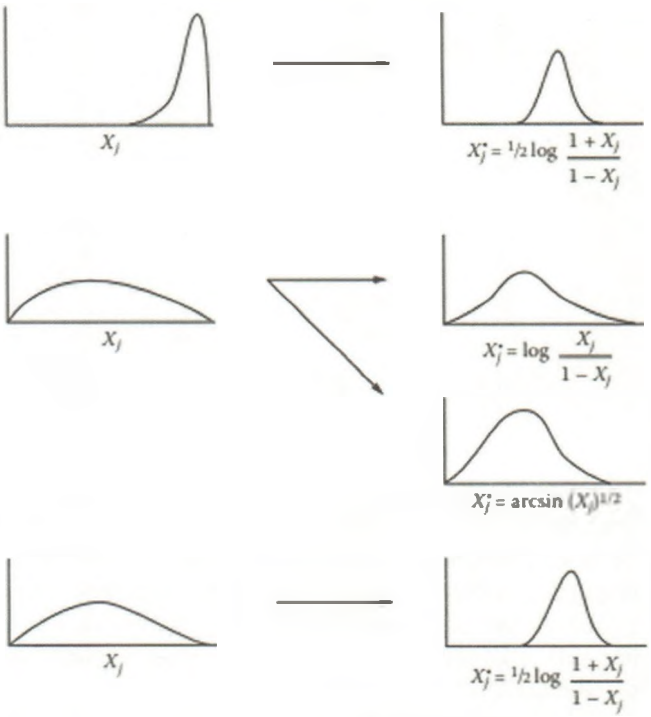
Các phương pháp chuyển đổi dữ liệu thường được thực hiện với mục đích:

Làm cho các phân phối lệch đối xứng và gần với phân phối chuẩn hơn;

- Để có sự đồng nhất phương sai;
- Để đạt được một thang đo có ý nghĩa hơn của sự đo lường.

Thông thường, phép biến đổi có thể xảy ra đối với dữ liệu có độ lệch dương dùng phép biến đổi căn bậc hai cho độ lệch vừa phải và phép biến đổi logarit đối với dữ liệu có độ lệch dương nghiêm trọng. *Hình 4.1* mô tả một số phép biến đổi thường được sử dụng để kéo vào phần đuôi bên phải, độ xiên bị ảnh hưởng bởi các yếu tố ngoại lai nên cần kiểm tra các yếu tố này trước.





X_j là phân phối dữ liệu thô ban đầu; X_j^* là phân phối dữ liệu chuyển đổi

Hình 4.1: Một số mô hình chuyển đổi dữ liệu phân phối [6]

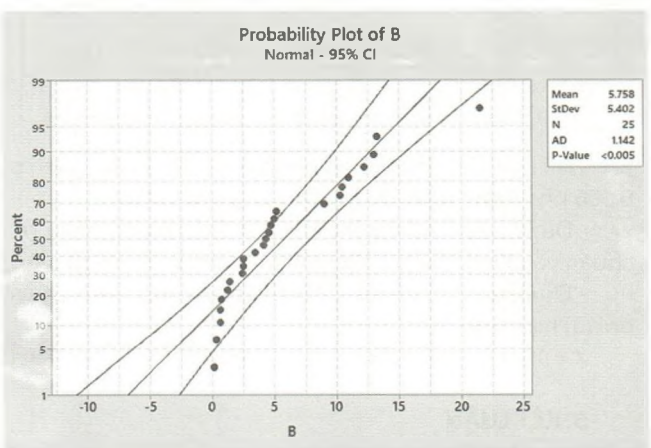
** Phép chuyển đổi Box-Cox:*

Được sử dụng khi tập dữ liệu phân bố rất lệch về một phía. Điều kiện để sử dụng phép chuyển đổi này là các dữ liệu phải có giá trị lớn hơn 0 và chia thành các nhóm nhỏ. Chẳng hạn, tập dữ liệu được thu thập định kỳ, mỗi lần thu được n số liệu, ta nói cỡ nhóm nhỏ là n. Phép chuyển đổi Box- Cox thực hiện phép đổi biến có dạng (2):

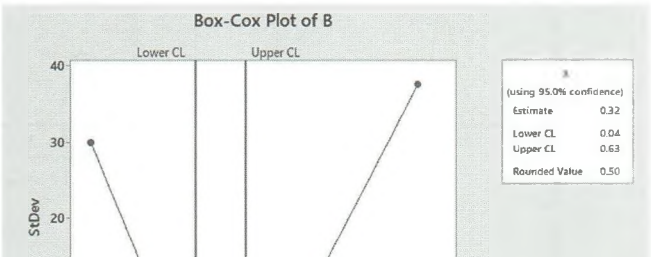
$$x(\lambda) = \frac{x^{\lambda} - 1}{\lambda} \ (\lambda \neq 0) \qquad (2)$$

Hoặc: $x(\lambda) = \ln(x) \ (\lambda = 0)$

Trong đó: x(λ) - Giá trị dữ liệu mới; x - Giá trị dữ liệu cũ; λ - Số mũ chuyển đổi. Giá trị λ được xác định bằng cách dò tìm, sao cho độ lệch chuẩn của tập dữ liệu đã được chuyển đổi là nhỏ nhất. Phương pháp Box-Cox thực hiện việc dò tìm λ trong khoảng từ -5 đến 5.



Hình 4.2: Đồ thị phân phối dữ liệu

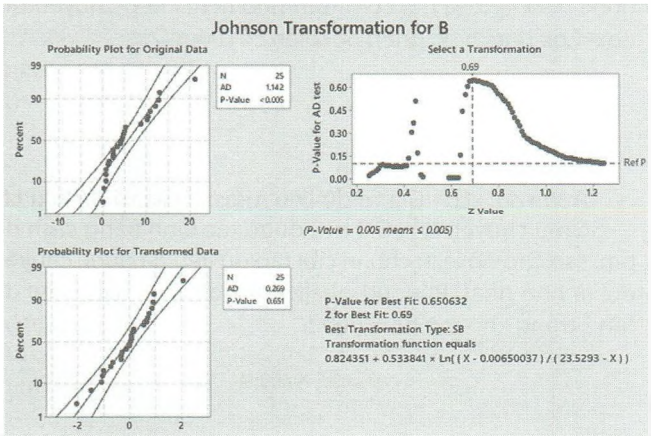


Hình 4.3: Đồ thị Box-Cox

Hình 4.2 cho thấy tập dữ liệu ban đầu của B không có phân phối chuẩn (p < 0,005). Bằng phương pháp Box-Cox *(Hình 4.3)*, khoảng tin cậy 95%CI (confidence interval), dữ liệu chuyển đổi mới tuân theo quy luật phân phối chuẩn. Giá trị λ (0,04 0,63) không chứa 1 nên việc chuyển đổi dữ liệu là hợp lý, còn nếu chứa 1 thì sự chuyển đổi là không cần thiết.

** Chuyển đổi Johnson:*

Chuyển đổi Johnson, còn gọi là phép đổi biến Johnson, được sử dụng để chuyển đổi một tập dữ liệu không theo quy luật phân phối chuẩn thành dạng phân phối chuẩn. Phương pháp chuyển đổi Johnson thực hiện một giải thuật phức tạp hơn so với phương pháp Box-Cox và thường được áp dụng cho các trường hợp phương pháp Box-Cox không hữu hiệu hoặc không áp dụng được.



Hình 4.4: Chuyển đổi dữ liệu bằng phương pháp Johnson

- Dữ liệu ban đầu không phân phối chuẩn do giá trị p < 0,006 nhỏ hơn rất nhiều so với mức ý nghĩa 0,05.
- Dữ liệu sau khi đã chuyển đổi đã có dạng phân phối chuẩn với p = 0,866 lớn hơn nhiều so với mức ý nghĩa 0,05.
- Dữ liệu được chuyển đổi theo phương trình chuyển đổi (3) như sau:

$$X' = 0.824351 + 0.533841 \times Ln((X - 0.00650037)/(23.5293 - X)) \quad (3)$$

5. KẾT LUẬN

Trong nghiên cứu thực nghiệm, rất nhiều mẫu dữ liệu không thuộc phân phối chuẩn, có thể xảy ra như vậy vì nhiều lý do như mẫu được lấy từ các quần thể khác nhau (vị trí, giới tính, mùa), chứa các yếu tố ngoại lai hay quá ít thông số... Dù thuộc trường hợp nào, bước đầu tiên trong phân tích dữ liệu là tìm hiểu, đánh giá các nguyên nhân và sao mẫu dữ liệu không thuộc dạng phân phối chuẩn. Thực tế thì hầu hết các đại lượng trong kỹ thuật đều có giá trị thay đổi ngẫu nhiên, khi bộ dữ liệu thu được không tuân theo phân phối chuẩn thì với phương pháp chuyển đổi sẽ đưa dữ liệu về dạng gần chuẩn nếu cần thiết. Tuy nhiên, không có phương pháp nào đảm bảo dữ liệu sau khi chuyển đổi sẽ có phân phối chuẩn, nên luôn cần kiểm tra lại bằng biểu đồ xác suất để xác định mẫu dữ liệu đó đã thuộc phân phối chuẩn hay chưa, rồi mới có các bước phân tích dữ liệu thực nghiệm tiếp theo.

Tài liệu tham khảo

[1]. Bùi Minh Trí (2005), *Xác suất thống kê và quy hoạch thực nghiệm*, NXB. Khoa học và Kỹ thuật.

[2]. Box, G.E.P., and D.R. Cox. (1964), *An analysis o transformations*, Journal of the Royal Statistical Society Series B 26:211-252.

[3]. Font, H., V. Torres, M. Herrera and R. Rodrígue (2007), *Fulfillment of the normality and the homogeneity o the variance in frequencies of accumulated measurement o the egg production variable in White Leghorn hens*, Cuban Journal of Agricultural Science 41:207-211.

[4]. Steel, R.G., e I.H. Torrie. (1992), *Bioestadística principios y procedimientos*, 740 p. McGraw-Hil Interamericana, México.

[5]. Sokal, R.R., and F.J. Rohlf. (1995), *Biometry*, 776 p. 3r ed. Freeman, New York, USA.

[6]. Rummel, R.J. (1970), *Applied factor analysis* Evanston, IL: Northwestern University Press.

Ngày nhận bài: 15/4/2022
Ngày chấp nhận đăng: 26/6/2022
Người phản biện: TS. Nguyễn Anh Tuấn
TS. Trần Việt Hưng