Thống kê mô tả chỉ phản ánh dữ liệu mà chúng được áp dụng. Thống kê mô tả có thể là:

* Một thước đo xu hướng trung tâm, chẳng hạn như trung bình, trung vị hoặc mốt: Những thước đo này được sử dụng để xác định điểm trung bình hoặc điểm trung tâm trong một tập dữ liệu.
* Một thước đo độ phân tán hoặc độ biến thiên, chẳng hạn như phương sai, độ lệch chuẩn, độ lệch chuẩn hoặc khoảng: Những thước đo này phản ánh sự phân tán của các điểm dữ liệu.
* Một thước đo phân phối, chẳng hạn như số lượng hoặc tỷ lệ phần trăm của một kết quả cụ thể: Những thước đo này thể hiện tần suất của kết quả đó trong một tập dữ liệu.
* Hầu như bất kỳ dữ liệu định lượng nào cũng có thể được phân tích bằng thống kê mô tả, chẳng hạn như kết quả từ một thử nghiệm lâm sàng liên quan đến tác dụng phụ của một loại thuốc cụ thể.
* Thống kê mô tả thể hiện thước đo xu hướng trung tâm có thể cho thấy độ tuổi trung bình của những người đã thử thuốc là 37. Ngoài ra, như một thước đo phân phối, thống kê mô tả có thể cho thấy 25% nhóm gặp phải tác dụng phụ nhẹ, trong khi 2% gặp tác dụng phụ từ trung bình đến nặng và 73% không gặp tác dụng phụ nào.
* Dữ liệu thô có thể được biểu diễn dưới dạng số liệu thống kê và đồ thị, sử dụng các hình ảnh trực quan như biểu đồ hình tròn, biểu đồ đường, bảng và các cách biểu diễn khác tóm tắt dữ liệu thu thập được về một nhóm dân số nhất định. [1]

Thống kê mô tả là loại thống kê có thể xuất hiện trong tâm trí của hầu hết mọi người khi họ nghe từ "thống kê". Trong nhánh thống kê này, mục tiêu là để mô tả. Các thước đo số được sử dụng để cho biết về các tính năng của một tập hợp dữ liệu. Có một số mục thuộc phần thống kê này, chẳng hạn như:

* Giá trị trung bình, hoặc thước đo của tâm của một tập dữ liệu, bao gồm giá trị trung bình, trung bình, chế độ hoặc tầm trung
* Sự lan truyền của một tập dữ liệu, có thể được đo bằng phạm vi hoặc độ lệch chuẩn
* Mô tả tổng thể về dữ liệu như tóm tắt năm số
* Các phép đo như độ lệch và kurtosis
* Khám phá các mối quan hệ và mối tương quan giữa dữ liệu được ghép nối
* Trình bày kết quả thống kê dưới dạng đồ họa [2]

Thống kê mô tả (Descriptive Statistics)

* Định nghĩa: Là các phương pháp dùng để tóm tắt, mô tả, trình bày dữ liệu đã thu thập được.
* Mục đích: Giúp ta hiểu dữ liệu “trông như thế nào” mà không đưa ra dự đoán hay kết luận cho quần thể lớn hơn.
* Ví dụ:
* Tính giá trị trung bình, trung vị, mode của điểm thi.
* Tính độ lệch chuẩn, phương sai để xem mức độ phân tán.
* Vẽ biểu đồ cột, histogram, boxplot.
* Chỉ nói lên đặc điểm của dữ liệu hiện có, không suy rộng ra ngoài mẫu. [3]

Thống kê mô tả (Descriptive Statistics) là phương pháp trong thống kê nhằm tóm tắt, sắp xếp, đơn giản hóa, mô tả và trình bày các đặc điểm chính của một tập dữ liệu đã thu thập, có thể là toàn bộ tổng thể hoặc một mẫu nghiên cứu.

Mục đích chính là giúp người nghiên cứu hiểu rõ hơn về tính chất và cấu trúc cơ bản của dữ liệu thông qua các số liệu hoặc hình ảnh trực quan.

Các chỉ số thường dùng trong thống kê mô tả bao gồm:

* Đo lường xu hướng tập trung: Giá trị trung bình (Mean), Trung vị (Median), Yếu vị (Mode).
* Đo lường biến động/phân tán: Độ lệch chuẩn (Standard Deviation), Phương sai (Variance), Khoảng giá trị (Range), Giá trị nhỏ nhất (Min), Giá trị lớn nhất (Max).
* Trình bày trực quan: Bảng tần số, Biểu đồ (Histogram, Bar chart, Pie chart,...). [4]

**1.1.1 Ôn tập lý thuyết**

* ***Thống kê mô tả (Descriptive Statistics) là gì?***

**Thống kê mô tả** là một nhánh của thống kê, bao gồm các phương pháp và kỹ thuật được sử dụng để tóm tắt, sắp xếp, mô tả và trình bày các đặc điểm chính của một tập dữ liệu đã được thu thập (có thể là một mẫu hoặc toàn bộ tổng thể).

**Mục tiêu cốt lõi** là giúp người nghiên cứu hiểu rõ dữ liệu "trông như thế nào", bao gồm cấu trúc, xu hướng và sự phân tán của nó, mà không đưa ra bất kỳ kết luận hay suy luận nào về một tổng thể rộng lớn hơn.

* ***Thống kê Suy luận (Inferential Statistics) là gì?***

**Thống kê suy luận** là tập hợp các phương pháp sử dụng lý thuyết xác suất để phân tích dữ liệu mẫu và đưa ra các dự đoán, kết luận hoặc tổng quát hóa có ý nghĩa về các đặc điểm (tham số) của toàn bộ tổng thể mà mẫu đó được rút ra.

**Nó giải quyết** **câu hỏi**: "Dữ liệu mẫu này có ý nghĩa gì đối với toàn bộ nhóm lớn hơn?"

* ***Bảng so sánh****:*
* Khác biệt cốt lõi

|  |  |  |
| --- | --- | --- |
| **Loại Thống Kê** | **Thống Kê MÔ TẢ** | **Thống Kê SUY LUẬN** |
| Mục tiêu Chính | **Tóm tắt và Trình bày** dữ liệu. | **Tổng quát hóa và Dự đoán** dữ liệu. |
| Phạm vi | Dữ liệu **đã có sẵn** (Mẫu → Mẫu). | Dữ liệu **mẫu** được dùng cho **Tổng thể** (Mẫu → Tổng thể). |

+ Các tiêu chí khác:

|  |  |  |
| --- | --- | --- |
| **Đặc điểm** | **Thống Kê Mô Tả** | **Thống Kê Suy Luận** |
| Mục tiêu Chính | Mô tả, tóm tắt, sắp xếp các đặc điểm của dữ liệu đã thu thập. | Dự đoán, tổng quát hóa, và rút ra kết luận về Tổng thể dựa trên dữ liệu từ mẫu. |
| Phạm vi | Chỉ giới hạn trong tập dữ liệu hiện có (Mẫu hoặc Tổng thể). | Mở rộng ra ngoài mẫu, hướng tới toàn bộ Tổng thể lớn hơn. |
| Các Công cụ/Chỉ số | 1. Xu hướng tập trung: Mean, Median, Mode.  2. Độ phân tán: Standard Deviation, Variance, Range.  3. Trình bày: Bảng tần số, Biểu đồ (Histogram, Pie chart). | 1. Ước lượng: Khoảng tin cậy (Confidence Intervals).  2. Kiểm định: T-test, ANOVA, Chi-square, P-value.  3. Phân tích quan hệ: Hồi quy (Regression). |
| Tính Chất Kết quả | Phản ánh sự thật của dữ liệu. Kết quả là các con số chắc chắn (ví dụ: điểm trung bình là 7.5). | Mang tính xác suất và rủi ro (biên độ lỗi). Kết quả là các dự đoán đi kèm với mức độ tin cậy. |
| Câu hỏi Trả lời | Dữ liệu này là gì? Đặc điểm của nó như thế nào? | Dữ liệu này có ý nghĩa gì đối với Tổng thể? Liệu có sự khác biệt có ý nghĩa nào không? |
| Đầu ra | Các Thống kê (Statistics) và đồ thị. | Các Tham số Parameters) của Tổng thể, Giá trị P, Khoảng tin cậy. |
| Ví dụ | "45% sinh viên trong mẫu khảo sát này là nữ." | "Chúng tôi ước tính rằng tỷ lệ sinh viên nữ trong toàn trường là 45% ±5%, với độ tin cậy 95%." |

**Giải thích Điểm Khác Biệt Cốt Lõi**

**1. Sự Khác Biệt về Hành động: Mô tả vs. Dự đoán**

* **Thống kê Mô tả:** Thực hiện hành động "Kể lại" hay "Tóm tắt". Nó giống như việc bạn đọc một cuốn nhật ký và ghi lại tóm tắt các sự kiện đã xảy ra.
* Ví dụ: Bạn đếm được 500 người trong phòng và tính ra tuổi trung bình của 500 người đó là 35. Đây là một sự thật không cần suy đoán.
* **Thống kê Suy luận:** Thực hiện hành động "Suy đoán" hay "Mở rộng". Nó giống như việc bạn đọc một vài trang nhật ký (mẫu) và cố gắng dự đoán toàn bộ cuộc đời của người viết (tổng thể).
* Ví dụ: Bạn khảo sát tuổi của 100 người (mẫu) trong phòng và dùng con số đó để ước tính tuổi trung bình của tất cả 5.000 người (tổng thể) trong tòa nhà. Kết luận này luôn đi kèm với rủi ro và xác suất (ví dụ: Khoảng tin cậy 95%).

**2. Sự Khác Biệt về Kết quả: Chắc chắn vs. Xác suất**

* **Thống kê mô tả** đưa ra các kết quả là sự thật về dữ liệu được quan sát.
* **Thống kê suy luận** đưa ra các kết luận có thể đúng về Tổng thể và luôn gắn liền với xác suất (p-value) và biên độ lỗi (margin of error).

Nói tóm lại, nếu bạn chỉ muốn biết điều gì đang xảy ra trong tập dữ liệu bạn có, bạn dùng **Mô tả**. Nếu bạn muốn biết điều gì xảy ra ngoài tập dữ liệu bạn có, bạn dùng **Suy luận**. [5]

[1] [What's the difference between descriptive and inferential statistics? | Bradley University Online](https://onlinedegrees.bradley.edu/blog/whats-the-difference-between-descriptive-and-inferential-statistics)

[2] [Descriptive vs. Inferential Statistics](https://www.thoughtco.com/differences-in-descriptive-and-inferential-statistics-3126224)

[3] ChatGPT

[4] Gemini AI

[[5] Thống kê mô tả và suy luận: Khác biệt](https://g.co/gemini/share/11303c161d07)

**Vai trò và Ứng dụng của Thống kê Mô tả**

**Thống kê mô tả** (Descriptive Statistics) đóng vai trò quan trọng trong phân tích dữ liệu, giúp nhà nghiên cứu tóm tắt và làm rõ các đặc điểm cơ bản của một tập dữ liệu trước khi tiến hành các phân tích sâu hơn. Các thước đo chính như trung bình, trung vị, phương sai và độ lệch chuẩn được sử dụng để phản ánh xu hướng trung tâm và mức độ phân tán của dữ liệu, qua đó mang lại cái nhìn khái quát về cấu trúc thông tin thu thập được ([1]; [2]).

**Trung bình** (Mean) là thước đo phổ biến nhất, thể hiện giá trị trung tâm thông qua trung bình cộng. Nó phù hợp khi dữ liệu phân phối đối xứng, không có giá trị ngoại lai. Tuy nhiên, trong thực tế, dữ liệu thường tồn tại các giá trị cực đoan làm méo mó kết quả. Trong những trường hợp này, **trung vị** (Median) là lựa chọn tốt hơn vì phản ánh giá trị ở giữa của tập dữ liệu và ít bị ảnh hưởng bởi ngoại lai, đặc biệt hữu ích với dữ liệu lệch hoặc dữ liệu dạng thứ bậc như thang đo mức độ hài lòng [3].

Bên cạnh đó, **phương sa**i (Variance) và **độ lệch chuẩn** (Standard Deviation) là các chỉ số quan trọng để đo lường độ biến thiên. Phương sai cho biết mức độ dữ liệu phân tán quanh trung bình, nhưng đơn vị tính bị bình phương nên khó trực quan. Ngược lại, độ lệch chuẩn là căn bậc hai của phương sai, giữ nguyên đơn vị ban đầu của dữ liệu nên thường được dùng để diễn giải mức độ dao động trong thực tế [4].

**Tóm lại**, thống kê mô tả không chỉ giúp đơn giản hóa dữ liệu phức tạp mà còn hỗ trợ nhà nghiên cứu xác định xu hướng và mức độ phân tán, từ đó tạo nền tảng vững chắc cho các phân tích thống kê suy luận tiếp theo.

**Ứng dụng: Khi nào Dùng Trung vị Thay vì Trung bình?**

Mặc dù Trung bình là thước đo phổ biến, **Trung vị** được coi là thước đo đáng tin cậy hơn để đại diện cho "trung tâm" của dữ liệu trong các trường hợp sau:

1. **Khi dữ liệu bị lệch (*Skewed*)**: Đặc biệt là trong các phân phối lệch phải (ví dụ: thu nhập, giá nhà) [1]. Trung bình sẽ bị kéo về phía đuôi dài (giá trị cao), khiến nó cao hơn Trung vị và không còn phản ánh đúng mức giá/thu nhập "điển hình" của đa số.
2. **Khi có Giá trị Ngoại lai (*Outliers*)**: Một hoặc vài giá trị cực đoan sẽ làm Trung bình bị sai lệch lớn, trong khi Trung vị hầu như không bị ảnh hưởng và giữ được tính đại diện [2].
3. **Khi dữ liệu là Thứ bậc (*Ordinal Data*)**: Dữ liệu chỉ mang tính chất thứ tự (ví dụ: thang đo mức độ hài lòng 1-5). Trung vị phản ánh tốt nhất giá trị trung tâm vì tính Trung bình cho loại dữ liệu này có thể kém ý nghĩa [3].

**Tóm tắt:** Trung bình phù hợp cho dữ liệu đối xứng; **Trung vị là lựa chọn mặc định tốt nhất** khi phân phối dữ liệu không đối xứng, có ngoại lai, hoặc mang tính thứ bậc.

**Tài liệu tham khảo:**

[[1] Bradley University Online. What’s the difference between descriptive and inferential statistics? (2024).](https://onlinedegrees.bradley.edu/blog/whats-the-difference-between-descriptive-and-inferential-statistics?utm_source=chatgpt.com)

[[2] Laerd Statistics. Descriptive and inferential statistics (2018).](https://online225.psych.wisc.edu/wp-content/uploads/225-Master/225-UnitPages/Unit-07/Laerd_Statistics_2013.pdf?utm_source=chatgpt.com)

[[3] UCLA Institute for Digital Research and Education (IDRE). Introduction to statistics concepts (2023).](https://arxiv.org/abs/1302.2525?utm_source=chatgpt.com)

[4] Moore, D.S., McCabe, G.P., & Craig, B.A. Introduction to the Practice of Statistics. (2017).

**Làm thế nào để xác định phân bố của một tập dữ liệu?**

Để xác định phân bố của một tập dữ liệu, có thể dùng hai cách chính: trực quan và định lượng.

**Phương pháp trực quan** thường bao gồm việc vẽ các loại đồ thị khác nhau. Histogram (biểu đồ tần suất) là công cụ phổ biến nhất, cho phép ta quan sát trực tiếp hình dạng phân bố dữ liệu: đối xứng, lệch trái, hay lệch phải (Moore, McCabe & Craig, 2017) [1]. Boxplot (biểu đồ hộp) lại hữu ích trong việc tóm tắt dữ liệu, xác định trung vị, phạm vi tứ phân vị và đặc biệt là phát hiện các giá trị ngoại lai (Seltman, 2018) [4]. Ngoài ra, Q-Q plot (Quantile-Quantile plot) được dùng để kiểm tra tính phù hợp với một phân bố lý thuyết, chẳng hạn như phân bố chuẩn. Nếu các điểm dữ liệu nằm gần đường chéo 45°, ta có cơ sở để cho rằng dữ liệu phân bố chuẩn (UCLA IDRE, 2023) [2].

**Phương pháp định lượng** dựa vào các số liệu thống kê mô tả. Việc so sánh trung bình và trung vị giúp nhận diện dạng phân bố: nếu trung bình gần bằng trung vị, dữ liệu có xu hướng đối xứng; nếu trung bình lớn hơn trung vị, dữ liệu thường lệch phải; còn nếu nhỏ hơn, dữ liệu có thể lệch trái (Laerd Statistics, 2018) [3]. Thêm vào đó, hệ số độ lệch (skewness) cung cấp một thước đo chính xác hơn: giá trị dương biểu thị lệch phải, âm biểu thị lệch trái, còn gần bằng 0 gợi ý tính đối xứng (Seltman, 2018) [4]. Bên cạnh đó, độ nhọn (kurtosis) đo mức độ “nhọn” hay độ dày của đuôi phân bố; phân bố chuẩn có kurtosis xấp xỉ 3 (Moore, McCabe & Craig, 2017) [1].

**Các loại phân bố phổ biến là gì ?**

Một số dạng phân bố cơ bản thường gặp trong phân tích dữ liệu gồm:

1. **Phân bố đối xứng (Symmetrical Distribution).** Đây là dạng phân bố mà đồ thị có dạng đối xứng qua trung tâm. Trường hợp điển hình nhất là **phân bố chuẩn (Normal/Gaussian Distribution)**, có hình chuông, đối xứng hoàn hảo và chỉ có một đỉnh. Trong phân bố này, trung bình = trung vị = mốt (Moore, McCabe & Craig, 2017) [1]; nhiều hiện tượng tự nhiên như chiều cao, huyết áp hay điểm số thi thường tuân theo dạng này (UCLA IDRE, 2023) [2].
2. **Phân bố lệch (Skewed Distributions).** Nếu dữ liệu không đối xứng, ta gặp hai trường hợp chính. **Phân bố lệch phải (positively skewed)** có đuôi kéo dài về phía các giá trị lớn, khiến trung bình lớn hơn trung vị; ví dụ thường gặp là thu nhập, giá nhà hoặc thời gian chờ đợi (Laerd Statistics, 2018) [3]. Ngược lại, **phân bố lệch trái (negatively skewed)** có đuôi kéo dài về phía giá trị nhỏ, với trung bình nhỏ hơn trung vị; ví dụ là điểm thi đại học (nhiều học sinh đạt điểm cao) hoặc tuổi thọ (Seltman, 2018) [4].
3. **Phân bố đa đỉnh (Multimodal Distribution).** Khi dữ liệu có nhiều hơn một đỉnh, ta gọi là phân bố đa đỉnh. Dạng này thường phản ánh sự pha trộn của nhiều nhóm khác nhau trong cùng một tập dữ liệu. Ví dụ: chiều cao của một mẫu gồm cả nam và nữ thường tạo ra hai đỉnh riêng biệt (Moore, McCabe & Craig, 2017) [1].

**Tài liệu tham khảo**

[1] Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). *Introduction to the Practice of Statistics* (9th ed.). W.H. Freeman and Company.  
[2] UCLA Institute for Digital Research and Education (IDRE). (2023). *What is a normal distribution?* Retrieved from <https://stats.oarc.ucla.edu/>  
[3] Laerd Statistics. (2018). *Descriptive and inferential statistics.* Retrieved from <https://statistics.laerd.com/>  
[4] Seltman, H. J. (2018). *Experimental Design and Analysis.* Carnegie Mellon University. Retrieved from <http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf>

**Độ lệch chuẩn và phạm vi (range) có ý nghĩa gì trong việc đánh giá sự phân tán của dữ liệu?**

Độ lệch chuẩn (Standard Deviation) và Phạm vi (Range) là hai thước đo quan trọng trong thống kê mô tả để đánh giá sự phân tán của dữ liệu.

Trước hết, **phạm vi (range)** là khoảng cách giữa giá trị lớn nhất và nhỏ nhất trong tập dữ liệu. Đây là thước đo đơn giản nhất về mức độ trải rộng, cho ta biết dữ liệu “trải” rộng bao nhiêu (Moore, McCabe & Craig, 2017) [1]. Tuy nhiên, phạm vi chỉ dựa trên hai điểm cực trị nên dễ bị ảnh hưởng bởi các giá trị ngoại lai, vì vậy nó thường được dùng để có một cái nhìn sơ bộ, nhanh chóng về độ phân tán (Seltman, 2018) [2].

Trong khi đó, **độ lệch chuẩn (standard deviation)** phản ánh mức độ mà các điểm dữ liệu dao động quanh giá trị trung bình. Nếu độ lệch chuẩn nhỏ, dữ liệu tập trung gần trung bình; nếu lớn, dữ liệu phân tán rộng hơn (UCLA IDRE, 2023) [3]. Khác với phạm vi, độ lệch chuẩn tận dụng toàn bộ dữ liệu trong mẫu, nên là một thước đo ổn định và đáng tin cậy hơn để so sánh sự biến thiên giữa các tập dữ liệu (Laerd Statistics, 2018) [4].

Nói cách khác, phạm vi cho thấy sự chênh lệch tối đa, còn độ lệch chuẩn cho thấy “mức độ dao động trung bình” của dữ liệu. Vì vậy, cả hai thước đo thường được dùng bổ sung cho nhau: phạm vi để tóm tắt nhanh sự trải rộng, và độ lệch chuẩn để phân tích sâu hơn về mức độ phân tán tổng thể.

**Tài liệu tham khảo**

[1] Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). *Introduction to the Practice of Statistics* (9th ed.). W.H. Freeman and Company.  
[2] Seltman, H. J. (2018). *Experimental Design and Analysis.* Carnegie Mellon University.  
[3] UCLA Institute for Digital Research and Education (IDRE). (2023). *Introduction to statistics concepts.*  
[4] Laerd Statistics. (2018). *Descriptive and inferential statistics.*

**Sự khác biệt giữa các thước đo như Q1, Q2, Q3 trong biểu đồ hộp (boxplot) là gì?**

Trong biểu đồ hộp (boxplot), ba thước đo Q1, Q2, Q3 chính là các tứ phân vị (quartiles), có vai trò chia dữ liệu thành bốn phần bằng nhau và cho thấy sự phân tán của dữ liệu.

* **Q1 (Quartile 1 hay Lower Quartile):** Là giá trị phân vị thứ nhất, tức điểm nằm tại vị trí 25% dữ liệu nhỏ nhất. Q1 thể hiện ranh giới giữa nhóm dữ liệu thấp và phần còn lại. Nó thường được dùng để xác định “đáy” của hộp trong boxplot (Moore, McCabe & Craig, 2017) [1].
* **Q2 (Quartile 2 hay Median):** Chính là **trung vị (median)** của tập dữ liệu, chia dữ liệu thành hai nửa bằng nhau. Trong boxplot, Q2 được vẽ dưới dạng đường ngang bên trong hộp, thể hiện giá trị trung tâm của dữ liệu (Laerd Statistics, 2018) [2].
* **Q3 (Quartile 3 hay Upper Quartile):** Là giá trị phân vị thứ ba, tức điểm nằm tại vị trí 75% dữ liệu nhỏ nhất (hoặc 25% dữ liệu lớn nhất). Q3 biểu thị ranh giới của nhóm dữ liệu cao. Nó tạo thành “đỉnh” của hộp trong boxplot (UCLA IDRE, 2023) [3].

Sự khác biệt chính giữa Q1, Q2 và Q3 nằm ở **vị trí trong phân phối dữ liệu**: Q1 phản ánh ngưỡng thấp, Q2 phản ánh điểm giữa, còn Q3 phản ánh ngưỡng cao. Khoảng cách giữa Q3 và Q1 gọi là **khoảng tứ phân vị (Interquartile Range – IQR = Q3 – Q1)**, được dùng để đo mức độ phân tán chính của dữ liệu và phát hiện giá trị ngoại lai (Seltman, 2018) [4].

**Tài liệu tham khảo**

[1] Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). *Introduction to the Practice of Statistics* (9th ed.). W.H. Freeman and Company.  
[2] Laerd Statistics. (2018). *Descriptive and inferential statistics.* Retrieved from <https://statistics.laerd.com>  
[3] UCLA Institute for Digital Research and Education (IDRE). (2023). *Introduction to statistics concepts.* Retrieved from <https://stats.oarc.ucla.edu>  
[4] Seltman, H. J. (2018). *Experimental Design and Analysis.* Carnegie Mellon University.

**Làm thế nào để xử lý giá trị thiếu (missing values) trước khi tính toán các chỉ số thống kê mô tả?**

**Xử lý giá trị thiếu (missing values) là bước quan trọng trong tiền xử lý dữ liệu, giúp đảm bảo tính chính xác của các chỉ số thống kê mô tả như trung bình, trung vị, phương sai, hay độ lệch chuẩn.** Có một số cách tiếp cận phổ biến:

1. **Loại bỏ giá trị thiếu (Deletion).** Nếu số lượng giá trị thiếu nhỏ và phân bố ngẫu nhiên, ta có thể loại bỏ các quan sát chứa missing values. Phương pháp này đơn giản nhưng có nguy cơ làm giảm kích thước mẫu và tính đại diện của dữ liệu (Moore, McCabe & Craig, 2017) [1].
2. **Thay thế bằng thống kê tóm tắt (Simple Imputation).** Một cách phổ biến là thay thế giá trị thiếu bằng **trung bình (mean), trung vị (median) hoặc mốt (mode)** của biến. Trung bình thường dùng với dữ liệu đối xứng, còn trung vị được ưu tiên khi dữ liệu có ngoại lai hoặc bị lệch (Laerd Statistics, 2018) [2].
3. **Sử dụng phương pháp nội suy hoặc ước lượng nâng cao.** Trong các tập dữ liệu phức tạp, có thể dùng **nội suy (interpolation), hồi quy (regression imputation), hoặc kỹ thuật Multiple Imputation** để dự đoán giá trị bị thiếu dựa trên mối quan hệ với các biến khác (Seltman, 2018) [3].
4. **Gán giá trị đặc biệt hoặc giữ nguyên.** Trong một số trường hợp (ví dụ: khảo sát xã hội học), giá trị thiếu có thể mang ý nghĩa riêng. Khi đó, nhà nghiên cứu có thể gán nhãn “không trả lời” hoặc giữ nguyên để phân tích riêng nhóm này (UCLA IDRE, 2023) [4].

Nói chung, lựa chọn phương pháp xử lý phụ thuộc vào **tỷ lệ dữ liệu thiếu, cơ chế thiếu (ngẫu nhiên hay có hệ thống), và mục tiêu phân tích**. Nếu dữ liệu thiếu ít và ngẫu nhiên, loại bỏ có thể chấp nhận được; nhưng với dữ liệu quan trọng hoặc thiếu nhiều, cần dùng đến kỹ thuật bù đắp hợp lý để tránh sai lệch trong các chỉ số thống kê mô tả.

**Tài liệu tham khảo**

[1] Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). *Introduction to the Practice of Statistics* (9th ed.). W.H. Freeman and Company.  
[2] Laerd Statistics. (2018). *Handling missing data.* Retrieved from <https://statistics.laerd.com>  
[3] Seltman, H. J. (2018). *Experimental Design and Analysis.* Carnegie Mellon University.  
[4] UCLA Institute for Digital Research and Education (IDRE). (2023). *Missing data: overview.* Retrieved from <https://stats.oarc.ucla.edu>

**Bạn có thể giải thích cách đọc và diễn giải một biểu đồ histogram hoặc boxplot từ dữ liệu thực tế không?**

Khi phân tích dữ liệu, **histogram** là công cụ trực quan giúp nhận diện hình dạng phân bố. Nếu histogram có dạng chuông đối xứng, dữ liệu có xu hướng gần với phân bố chuẩn. Nếu biểu đồ có đuôi kéo dài về bên phải, phân bố bị lệch phải – thường gặp ở các biến như thu nhập hoặc giá nhà. Ngược lại, nếu đuôi kéo dài về bên trái, phân bố bị lệch trái – ví dụ như điểm thi khi phần lớn học sinh đạt điểm cao. Histogram cũng có thể cho thấy dữ liệu có một đỉnh hay nhiều đỉnh; trường hợp có nhiều đỉnh thường gợi ý rằng dữ liệu gồm nhiều nhóm khác biệt [1].

Trong khi đó, **boxplot** tập trung mô tả dữ liệu thông qua năm số thống kê: giá trị nhỏ nhất, Q1, Q2 (trung vị), Q3 và giá trị lớn nhất. Hộp trong boxplot biểu diễn khoảng tứ phân vị (IQR), phản ánh sự phân tán của 50% dữ liệu trung tâm. Nếu hộp dài, dữ liệu biến động mạnh; nếu ngắn, dữ liệu ổn định hơn. Vị trí trung vị trong hộp cũng cung cấp thông tin về xu hướng phân bố: trung vị ở giữa thể hiện phân bố gần đối xứng, lệch xuống dưới gợi ý phân bố lệch phải, và lệch lên trên phản ánh phân bố lệch trái. Các “râu” cho biết phạm vi dữ liệu chính, còn những điểm ngoài râu được coi là ngoại lai cần được chú ý [2].

Như vậy, histogram chủ yếu cho thấy **hình dạng phân bố**, trong khi boxplot nhấn mạnh **độ phân tán, xu hướng trung tâm và ngoại lai**. Kết hợp cả hai sẽ giúp ta có cái nhìn toàn diện và chính xác hơn về tập dữ liệu.

**Tài liệu tham khảo**  
[1] Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). *Introduction to the Practice of Statistics* (9th ed.). W.H. Freeman and Company.  
[2] UCLA Institute for Digital Research and Education (IDRE). *Chapter 1: Introduction to statistics concepts* (2023). <https://stats.oarc.ucla.edu>

**Khi gặp một tập dữ liệu có giá trị ngoại lai (outliers), bạn sẽ xử lý chúng như thế nào trước khi thực hiện thống kê mô tả?**

Khi gặp một tập dữ liệu có giá trị ngoại lai (outliers), bước đầu tiên là cần **xác định** chúng bằng các công cụ trực quan như boxplot hoặc bằng các quy tắc thống kê, chẳng hạn như quy tắc 1.5\*IQR (giá trị nhỏ hơn Q1 – 1.5×IQR hoặc lớn hơn Q3 + 1.5×IQR được xem là ngoại lai) hoặc kiểm tra điểm z-score (|z| > 3). Sau khi phát hiện, việc xử lý phụ thuộc vào mục tiêu phân tích và bản chất của dữ liệu.

Trong một số trường hợp, ngoại lai có thể là **lỗi nhập liệu** hoặc lỗi đo lường, khi đó cần kiểm tra và loại bỏ hoặc sửa chữa giá trị đó để dữ liệu phản ánh chính xác hơn thực tế [1]. Nếu ngoại lai là các quan sát hợp lệ nhưng hiếm gặp (ví dụ: một người có thu nhập cực cao trong khảo sát thu nhập), chúng không nên bị loại bỏ ngay. Thay vào đó, có thể sử dụng các thước đo **ít nhạy cảm với ngoại lai** như trung vị thay cho trung bình, hoặc dùng IQR thay cho độ lệch chuẩn để mô tả sự phân tán [2]. Một số kỹ thuật khác bao gồm **biến đổi dữ liệu** (log, căn bậc hai) để giảm tác động của ngoại lai, hoặc phân tích song song hai trường hợp: giữ nguyên và loại bỏ ngoại lai để so sánh kết quả.

Tóm lại, xử lý ngoại lai đòi hỏi sự cân nhắc cẩn thận. Việc loại bỏ, giữ lại hay biến đổi cần dựa trên mục tiêu nghiên cứu, bản chất của dữ liệu, và tác động mà ngoại lai gây ra cho các chỉ số thống kê mô tả.

**Tài liệu tham khảo:**  
[1] Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). *Introduction to the Practice of Statistics* (9th ed.). W.H. Freeman and Company.  
[2] UCLA Institute for Digital Research and Education (IDRE). *Chapter 1: Introduction to statistics concepts* (2023). <https://stats.oarc.ucla.edu>