Ôn tập lý thuyết.

1. Thống kê mô tả là gì? Nó khác gì với thống kê suy luận (inferential statistics)?

**Thống kê mô tả (Descriptive Statistics)** là phương pháp thu thập, tổ chức, tóm tắt và trình bày dữ liệu nhằm cung cấp cái nhìn tổng quan về đặc điểm của tập dữ liệu. Nó tập trung vào việc mô tả các đặc trưng cơ bản của dữ liệu như:

* **Các chỉ số trung tâm**: trung bình (mean), trung vị (median), mode (giá trị xuất hiện nhiều nhất).
* **Các chỉ số phân tán**: độ lệch chuẩn (standard deviation), phương sai (variance), khoảng giá trị (range).
* **Hình dạng phân phối**: độ lệch (skewness), độ nhọn (kurtosis).
* **Biểu đồ**: biểu đồ cột, biểu đồ tròn, histogram, boxplot.

Ví dụ: Nếu bạn có điểm thi của một lớp học, thống kê mô tả sẽ cho biết điểm trung bình, điểm cao nhất, thấp nhất, và cách điểm số phân bố.

**Thống kê suy luận (Inferential Statistics)** sử dụng dữ liệu từ một mẫu để đưa ra kết luận hoặc dự đoán về tổng thể (population). Nó dựa trên xác suất và thường được dùng để kiểm định giả thuyết, ước lượng tham số, hoặc dự báo. Các phương pháp bao gồm:

* **Kiểm định giả thuyết**: Kiểm tra xem một giả thuyết (như "trung bình điểm thi của nam và nữ không khác nhau") có đúng không.
* **Ước lượng**: Ước tính giá trị trung bình hoặc tỷ lệ của tổng thể dựa trên mẫu (ví dụ: ước tính tỷ lệ cử tri ủng hộ một ứng viên).
* **Mô hình hóa**: Dự đoán xu hướng hoặc quan hệ giữa các biến (như hồi quy tuyến tính).

Ví dụ: Dựa trên điểm thi của một lớp (mẫu), bạn suy ra điểm trung bình của toàn bộ học sinh trong trường (tổng thể).

Sự khác biệt:

| **Tiêu chí** | **Thống kê mô tả** | **Thống kê suy luận** |
| --- | --- | --- |
| **Mục đích** | Tóm tắt và mô tả dữ liệu hiện có | Đưa ra kết luận hoặc dự đoán về tổng thể |
| **Phạm vi** | Chỉ tập trung vào mẫu dữ liệu | Suy rộng từ mẫu ra tổng thể |
| **Ví dụ phương pháp** | Trung bình, phương sai, biểu đồ | Kiểm định t, ANOVA, hồi quy |
| **Kết quả** | Cung cấp số liệu cụ thể của mẫu | Đưa ra xác suất hoặc khoảng tin cậy |
| **Tính phức tạp** | Đơn giản, trực quan | Phức tạp hơn, cần giả định và xác suất |

2. Các thước đo thống kê mô tả chính (ví dụ: trung bình, trung vị, phương sai, độ lệch chuẩn) được sử dụng để làm gì? Trong trường hợp nào thì nên dùng trung vị thay vì trung bình?

**a. Các thước đo thống kê mô tả chính và mục đích**

* **Trung bình (Mean)**: Là giá trị trung bình cộng của tất cả các giá trị trong tập dữ liệu.
  + **Mục đích**: Cung cấp một giá trị đại diện cho mức trung bình của dữ liệu.
  + **Ví dụ**: Tính điểm trung bình của một lớp học (tổng điểm chia cho số học sinh).
  + **Ứng dụng**: Phù hợp khi dữ liệu phân bố đối xứng và không có giá trị ngoại lai lớn.
* **Trung vị (Median)**: Là giá trị nằm ở giữa khi dữ liệu được sắp xếp theo thứ tự tăng dần hoặc giảm dần.
  + **Mục đích**: Đại diện cho giá trị trung tâm của tập dữ liệu, ít bị ảnh hưởng bởi các giá trị ngoại lai.
  + **Ví dụ**: Nếu thu nhập của một nhóm người có sự chênh lệch lớn, trung vị sẽ cho biết mức thu nhập "điển hình" hơn trung bình.
  + **Ứng dụng**: Dùng khi dữ liệu có phân bố lệch (skewed) hoặc có giá trị ngoại lai.
* **Phương sai (Variance)**: Đo lường mức độ phân tán của dữ liệu so với giá trị trung bình.
  + **Mục đích**: Cho biết dữ liệu dao động nhiều hay ít quanh giá trị trung bình.
  + **Ví dụ**: Phương sai lớn trong điểm thi cho thấy sự chênh lệch lớn giữa các học sinh.
  + **Ứng dụng**: Đánh giá sự biến thiên trong dữ liệu, thường dùng trong phân tích tài chính hoặc nghiên cứu khoa học.
* **Độ lệch chuẩn (Standard Deviation)**: Là căn bậc hai của phương sai, biểu thị mức độ phân tán của dữ liệu trong cùng đơn vị với dữ liệu gốc.
  + **Mục đích**: Cung cấp một thước đo dễ hiểu hơn về sự phân tán so với phương sai.
  + **Ví dụ**: Độ lệch chuẩn của chiều cao học sinh giúp đánh giá mức độ đồng đều của chiều cao trong lớp.
  + **Ứng dụng**: Được dùng rộng rãi trong kiểm định thống kê, phân tích rủi ro, và so sánh các tập dữ liệu.

**b. Khi nào nên dùng trung vị thay vì trung bình?**

Trung vị được ưu tiên hơn trung bình trong các trường hợp sau:

* **Dữ liệu có phân bố lệch (Skewed Distribution)**: Khi dữ liệu không đối xứng (ví dụ: phân bố lệch phải hoặc lệch trái), trung bình có thể bị kéo lệch bởi các giá trị cực đại hoặc cực tiểu, không phản ánh đúng giá trị trung tâm. Trung vị sẽ đại diện tốt hơn.
  + **Ví dụ**: Thu nhập của một nhóm người thường lệch phải do một vài cá nhân có thu nhập rất cao. Trung vị sẽ cho biết mức thu nhập điển hình của nhóm.
* **Có giá trị ngoại lai (Outliers)**: Trung bình rất nhạy cảm với các giá trị ngoại lai, trong khi trung vị không bị ảnh hưởng nhiều.
  + **Ví dụ**: Nếu điểm thi của một lớp là [50, 60, 55, 58, 100], trung bình là 64.6, nhưng trung vị là 58, phản ánh tốt hơn mức điểm của đa số học sinh.
* **Dữ liệu thứ tự (Ordinal Data)**: Với dữ liệu thứ tự (ví dụ: xếp hạng 1, 2, 3), trung bình không có ý nghĩa thống kê, nhưng trung vị có thể được sử dụng để tìm giá trị trung tâm.
  + **Ví dụ**: Trong khảo sát mức độ hài lòng (1 = không hài lòng, 5 = rất hài lòng), trung vị sẽ phù hợp hơn để biểu thị mức độ hài lòng trung tâm.

3. Làm thế nào để xác định phân bố của một tập dữ liệu? Các loại phân bố phổ biến là gì (ví dụ: phân bố chuẩn, lệch trái, lệch phải)?

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

**Bước 1: Sử dụng các công cụ trực quan**

* **Biểu đồ histogram**: Hiển thị tần suất của các giá trị trong các khoảng (bin). Hình dạng của histogram giúp nhận biết phân bố (đối xứng, lệch trái, lệch phải, v.v.).
* **Biểu đồ boxplot (hộp râu)**: Cho thấy trung vị, tứ phân vị, và các giá trị ngoại lai, giúp nhận diện sự lệch và phân tán của dữ liệu.
* **Biểu đồ Q-Q (Quantile-Quantile)**: So sánh phân bố của dữ liệu với phân bố chuẩn. Nếu các điểm nằm gần đường thẳng, dữ liệu có khả năng phân bố chuẩn.
* **Biểu đồ mật độ (Density Plot)**: Tương tự histogram nhưng mượt mà hơn, giúp hình dung phân bố liên tục.

**Bước 2: Tính các thước đo thống kê mô tả**

* **Trung bình, trung vị, mode**: So sánh các giá trị này để đánh giá sự lệch.
  + Nếu trung bình ≈ trung vị ≈ mode → Phân bố đối xứng (có thể là phân bố chuẩn).
  + Nếu trung bình > trung vị > mode → Lệch phải (positively skewed).
  + Nếu trung bình < trung vị < mode → Lệch trái (negatively skewed).
* **Độ lệch (Skewness)**: Đo lường mức độ lệch của phân bố.
  + Skewness ≈ 0: Phân bố đối xứng.
  + Skewness > 0: Lệch phải.
  + Skewness < 0: Lệch trái.
* **Độ nhọn (Kurtosis)**: Đo lường độ "nhọn" của phân bố so với phân bố chuẩn.
  + Kurtosis = 3 (hoặc 0 khi chuẩn hóa): Phân bố chuẩn.
  + Kurtosis > 3: Phân bố nhọn hơn (leptokurtic).
  + Kurtosis < 3: Phân bố phẳng hơn (platykurtic).

**Bước 3: Kiểm định thống kê**

* **Kiểm định Shapiro-Wilk, Kolmogorov-Smirnov, hoặc Anderson-Darling**: Dùng để kiểm tra xem dữ liệu có tuân theo phân bố chuẩn hay không.
* **Kiểm định khác**: Nếu nghi ngờ dữ liệu theo các phân bố khác (như Poisson, exponential), có thể dùng các kiểm định chuyên biệt.

**Bước 4: Phân tích ngữ cảnh dữ liệu**

* Hiểu nguồn gốc và bản chất của dữ liệu. Ví dụ, thu nhập thường lệch phải, thời gian chờ thường theo phân bố exponential, còn chiều cao con người thường gần với phân bố chuẩn.

**b. Các loại phân bố phổ biến**

* **Phân bố chuẩn (Normal Distribution)**:
  + **Đặc điểm**: Đối xứng, hình chuông, trung bình = trung vị = mode. Độ lệch chuẩn quyết định độ rộng của phân bố.
  + **Ví dụ**: Chiều cao, cân nặng, điểm thi (trong điều kiện lý tưởng).
  + **Ứng dụng**: Rất phổ biến trong thống kê suy luận vì nhiều phương pháp giả định dữ liệu phân bố chuẩn.
* **Phân bố lệch phải (Positively Skewed)**:
  + **Đặc điểm**: Đuôi bên phải dài hơn, trung bình > trung vị > mode. Có nhiều giá trị nhỏ, một vài giá trị lớn bất thường.
  + **Ví dụ**: Thu nhập, thời gian thất bại của thiết bị, số lượng khách hàng trong hàng đợi.
  + **Ứng dụng**: Thường gặp trong dữ liệu tài chính, thời gian chờ.
* **Phân bố lệch trái (Negatively Skewed)**:
  + **Đặc điểm**: Đuôi bên trái dài hơn, trung bình < trung vị < mode. Có nhiều giá trị lớn, một vài giá trị nhỏ bất thường.
  + **Ví dụ**: Tuổi nghỉ hưu, điểm kiểm tra trong bài thi rất dễ.
  + **Ứng dụng**: Ít phổ biến hơn lệch phải, nhưng xuất hiện trong một số dữ liệu xã hội học.
* **Phân bố đều (Uniform Distribution)**:
  + **Đặc điểm**: Tất cả các giá trị trong một khoảng có xác suất như nhau.
  + **Ví dụ**: Kết quả của một lần tung xúc xắc, số ngẫu nhiên từ máy tính.
  + **Ứng dụng**: Mô phỏng dữ liệu ngẫu nhiên hoặc thử nghiệm.
* **Phân bố nhị thức (Binomial Distribution)**:
  + **Đặc điểm**: Mô tả số lần thành công trong một số lần thử cố định, mỗi lần thử có hai kết quả (thành công/thất bại).
  + **Ví dụ**: Số lần tung đồng xu được mặt sấp trong 10 lần tung.
  + **Ứng dụng**: Phân tích dữ liệu sự kiện nhị phân.
* **Phân bố Poisson**:
  + **Đặc điểm**: Mô tả số lần xảy ra của một sự kiện trong một khoảng thời gian hoặc không gian cố định.
  + **Ví dụ**: Số cuộc gọi đến tổng đài trong một giờ.
  + **Ứng dụng**: Phân tích dữ liệu đếm (count data).
* **Phân bố mũ (Exponential Distribution)**:
  + **Đặc điểm**: Mô tả thời gian giữa các sự kiện trong một quá trình Poisson.
  + **Ví dụ**: Thời gian chờ giữa các cuộc gọi đến tổng đài.
  + **Ứng dụng**: Phân tích thời gian chờ hoặc độ bền

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

**1. Độ lệch chuẩn (Standard Deviation)**

* **Định nghĩa**: Độ lệch chuẩn là căn bậc hai của phương sai, đo lường mức độ trung bình mà các giá trị trong tập dữ liệu lệch khỏi giá trị trung bình. Nó được tính bằng công thức: A math equation with white lines

  AI-generated content may be incorrect.

 **Đo lường sự phân tán**: Độ lệch chuẩn cho biết dữ liệu dao động nhiều hay ít quanh giá trị trung bình. Giá trị độ lệch chuẩn lớn hơn có nghĩa là dữ liệu phân tán rộng hơn (ít đồng đều), trong khi giá trị nhỏ hơn cho thấy dữ liệu tập trung gần trung bình (đồng đều hơn).

* **Đơn vị tương thích**: Độ lệch chuẩn có cùng đơn vị với dữ liệu gốc, giúp dễ dàng diễn giải. Ví dụ, nếu dữ liệu là chiều cao (cm), độ lệch chuẩn cũng được tính bằng cm.
* **Ứng dụng trong phân bố chuẩn**: Trong phân bố chuẩn, khoảng 68% dữ liệu nằm trong khoảng ±1 độ lệch chuẩn từ trung bình, 95% trong ±2 độ lệch chuẩn, và 99,7% trong ±3 độ lệch chuẩn.

 **Ví dụ**: Nếu điểm thi trung bình của một lớp là 70 với độ lệch chuẩn là 5, hầu hết học sinh có điểm từ 65 đến 75 (70 ± 5). Nếu độ lệch chuẩn là 15, điểm số sẽ phân tán rộng hơn (55 đến 85).

 **Ưu điểm**:

* Xem xét tất cả các giá trị trong tập dữ liệu, không chỉ các giá trị cực đại/cực tiểu.
* Phù hợp cho các phân tích thống kê suy luận (như kiểm định giả thuyết, khoảng tin cậy).

 **Nhược điểm**: Nhạy cảm với các giá trị ngoại lai, vì các giá trị này có thể làm tăng đáng kể độ lệch chuẩn.

**2. Phạm vi (Range)**

* **Định nghĩa**: Phạm vi là hiệu số giữa giá trị lớn nhất và giá trị nhỏ nhất trong tập dữ liệu: RANGE=GTRIMAX-GTRIMIN
* **Ý nghĩa**:
  + **Đo lường sự phân tán cơ bản**: Phạm vi cho biết khoảng cách giữa giá trị cao nhất và thấp nhất, cung cấp cái nhìn nhanh về mức độ lan rộng của dữ liệu.
  + **Dễ tính toán và hiểu**: Đây là thước đo đơn giản nhất để đánh giá sự phân tán.
  + **Ví dụ**: Nếu điểm thi của một lớp nằm trong khoảng từ 40 đến 90, phạm vi là 90 - 40 = 50, cho thấy dữ liệu trải dài trên 50 đơn vị điểm.
* **Ưu điểm**:
  + Đơn giản, dễ tính toán và trực quan.
  + Hữu ích khi cần một thước đo nhanh về sự phân tán.
* **Nhược điểm**:
  + Chỉ dựa trên hai giá trị (tối đa và tối thiểu), bỏ qua cách các giá trị khác phân bố.
  + Rất nhạy cảm với giá trị ngoại lai. Ví dụ, một giá trị ngoại lai (như 100 trong tập dữ liệu [40, 50, 60, 100]) sẽ làm tăng phạm vi đáng kể mà không phản ánh đúng sự phân tán của phần lớn dữ liệu.

**3. So sánh và ứng dụng trong đánh giá sự phân tán**

| **Thước đo** | **Ưu điểm** | **Nhược điểm** | **Khi nào sử dụng?** |
| --- | --- | --- | --- |
| **Độ lệch chuẩn** | Xem xét toàn bộ dữ liệu, chính xác hơn | Nhạy cảm với ngoại lai, phức tạp hơn | Khi cần đánh giá phân tán chi tiết, đặc biệt trong phân bố chuẩn hoặc phân tích thống kê nâng cao. |
| **Phạm vi** | Đơn giản, dễ tính toán, trực quan | Chỉ dựa trên hai giá trị, dễ bị lệch bởi ngoại lai | Khi cần đánh giá nhanh hoặc dữ liệu không có ngoại lai rõ rệt. |

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

**1. Định nghĩa các tứ phân vị**

* **Q1 (Tứ phân vị thứ nhất)**: Là giá trị mà 25% dữ liệu nằm dưới nó (hoặc 75% dữ liệu nằm trên nó). Đây là giá trị trung tâm của 25% dữ liệu thấp nhất.
* **Q2 (Tứ phân vị thứ hai)**: Là **trung vị** (median), giá trị chia tập dữ liệu thành hai phần bằng nhau, với 50% dữ liệu nằm dưới và 50% nằm trên.
* **Q3 (Tứ phân vị thứ ba)**: Là giá trị mà 75% dữ liệu nằm dưới nó (hoặc 25% dữ liệu nằm trên nó). Đây là giá trị trung tâm của 25% dữ liệu cao nhất.

**2. Cách tính các tứ phân vị**

1. Sắp xếp dữ liệu theo thứ tự tăng dần.
2. **Q2 (trung vị)**: Tìm giá trị ở giữa.
   * Nếu số lượng dữ liệu n n n là lẻ, Q2 là giá trị ở vị trí (n+1)/2 (n+1)/2 (n+1)/2.
   * Nếu n n n là chẵn, Q2 là trung bình cộng của hai giá trị ở vị trí n/2 n/2 n/2 và (n/2)+1 (n/2)+1 (n/2)+1.
3. **Q1**: Tìm trung vị của nửa dưới của tập dữ liệu (từ giá trị nhỏ nhất đến Q2).
4. **Q3**: Tìm trung vị của nửa trên của tập dữ liệu (từ Q2 đến giá trị lớn nhất).

Ví dụ: Với tập dữ liệu [3, 5, 7, 8, 10, 12, 14]:

* n=7 n = 7 n=7, Q2 = giá trị ở vị trí (7+1)/2=4 (7+1)/2 = 4 (7+1)/2=4, tức là 8.
* Nửa dưới: [3, 5, 7] → Q1 = 5 (trung vị của nửa dưới).
* Nửa trên: [10, 12, 14] → Q3 = 12 (trung vị của nửa trên).

**3. Vai trò trong biểu đồ hộp (Boxplot)**

Biểu đồ hộp sử dụng Q1, Q2, Q3 để hiển thị phân bố của dữ liệu một cách trực quan. Các thành phần chính của boxplot bao gồm:

* **Hộp (Box)**: Được giới hạn bởi Q1 (cạnh dưới) và Q3 (cạnh trên), thể hiện **khoảng tứ phân vị (IQR = Q3 - Q1)**, chứa 50% dữ liệu ở giữa.
* **Đường giữa hộp**: Là Q2 (trung vị), biểu thị giá trị trung tâm của tập dữ liệu.
* **Râu (Whiskers)**: Mở rộng từ Q1 và Q3 đến các giá trị không được coi là ngoại lai, thường được tính là:
  + Râu dưới: min(dữliệu,Q1−1.5×IQR)
  + Râu trên: max(dữliệu,Q3+1.5×IQR)
* **Điểm ngoại lai (Outliers)**: Các giá trị nằm ngoài khoảng [Q1−1.5×IQR,Q3+1.5×IQR]

**4. Ý nghĩa của Q1, Q2, Q3**

* **Q1**: Cho biết mức giá trị mà 25% dữ liệu thấp nhất nằm dưới, giúp xác định phần thấp của phân bố.
* **Q2 (trung vị)**: Đại diện cho giá trị trung tâm, ít bị ảnh hưởng bởi giá trị ngoại lai so với trung bình, hữu ích khi dữ liệu lệch hoặc có ngoại lai.
* **Q3**: Cho biết mức giá trị mà 75% dữ liệu thấp hơn, giúp xác định phần cao của phân bố.
* **IQR (Q3 - Q1)**: Đo lường mức độ phân tán của 50% dữ liệu ở giữa, thường dùng để phát hiện giá trị ngoại lai.

**5. Ứng dụng của tứ phân vị và biểu đồ hộp**

* **Mô tả phân bố**: Boxplot giúp nhận diện nhanh hình dạng phân bố (đối xứng, lệch trái, lệch phải) và sự hiện diện của giá trị ngoại lai.
* **So sánh các tập dữ liệu**: Có thể so sánh nhiều tập dữ liệu bằng cách vẽ các boxplot cạnh nhau để thấy sự khác biệt về trung vị, phân tán và ngoại lai.
* **Phát hiện ngoại lai**: Các giá trị nằm ngoài khoảng râu được coi là ngoại lai, giúp xác định các điểm bất thường trong dữ liệu.

**6. Ví dụ minh họa**

Giả sử bạn có tập dữ liệu điểm thi: [40, 50, 60, 65, 70, 75, 80, 85, 90, 100].

* Sắp xếp: [40, 50, 60, 65, 70, 75, 80, 85, 90, 100].
* Q2 (trung vị): Trung bình của giá trị thứ 5 (70) và thứ 6 (75) = (70 + 75) / 2 = 72.5.
* Q1: Trung vị của nửa dưới [40, 50, 60, 65, 70] = 60.
* Q3: Trung vị của nửa trên [75, 80, 85, 90, 100] = 85.
* IQR = Q3 - Q1 = 85 - 60 = 25.
* Râu dưới: Q1−1.5×IQR=60−1.5×25=22.5 Q1 - 1. 5 = 22.5 Q1−1.5×IQR=60−1.5×25=22.5.
* Râu trên: Q3+1.5×IQR=85+1.5×25=122.5 Q3 + 1.5 = 122.5 Q3+1.5×IQR=85+1.5×25=122.5.
* Ngoại lai: Không có giá trị nào nằm ngoài khoảng [22.5, 122.5].

Trong boxplot, hộp sẽ kéo dài từ 60 (Q1) đến 85 (Q3), đường giữa ở 72.5 (Q2), và râu kéo dài từ 40 (giá trị nhỏ nhất) đến 100 (giá trị lớn nhất).

**7. Sự khác biệt giữa Q1, Q2, Q3**

| **Thước đo** | **Vị trí trong dữ liệu** | **Ý nghĩa** | **Vai trò trong boxplot** |
| --- | --- | --- | --- |
| **Q1** | 25% dữ liệu thấp nhất | Đánh dấu ranh giới của 25% dữ liệu thấp nhất | Cạnh dưới của hộp |
| **Q2** | 50% dữ liệu (trung vị) | Giá trị trung tâm, chia dữ liệu thành hai phần bằng nhau | Đường giữa hộp |
| **Q3** | 75% dữ liệu thấp hơn | Đánh dấu ranh giới của 75% dữ liệu thấp nhất | Cạnh trên của hộp |

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

**1. Xác định và hiểu giá trị thiếu**

Trước khi xử lý, cần:

* **Xác định giá trị thiếu**: Kiểm tra dữ liệu để tìm các giá trị bị thiếu (có thể là NaN, NULL, giá trị trống, hoặc ký hiệu đặc biệt như -).
  + Công cụ: Trong Python, dùng pandas.DataFrame.isna() hoặc isnull(); trong R, dùng is.na().
* **Hiểu nguyên nhân thiếu dữ liệu**:
  + **MCAR (Missing Completely at Random)**: Giá trị thiếu không liên quan đến bất kỳ biến nào (ví dụ: lỗi nhập liệu ngẫu nhiên).
  + **MAR (Missing at Random)**: Giá trị thiếu liên quan đến các biến khác nhưng không liên quan đến chính giá trị bị thiếu (ví dụ: người lớn tuổi ít trả lời khảo sát).
  + **MNAR (Missing Not at Random)**: Giá trị thiếu liên quan đến chính giá trị đó (ví dụ: người có thu nhập cao không tiết lộ thu nhập).
* **Đánh giá mức độ thiếu**: Tính tỷ lệ giá trị thiếu (ví dụ: 5% hay 50% dữ liệu bị thiếu) để quyết định phương pháp xử lý.

**2. Các phương pháp xử lý giá trị thiếu**

Dưới đây là các phương pháp phổ biến để xử lý giá trị thiếu:

**a. Loại bỏ giá trị thiếu**

* **Mô tả**: Xóa các hàng hoặc cột chứa giá trị thiếu.
  + **Xóa hàng**: Loại bỏ toàn bộ quan sát (record) có giá trị thiếu (dùng dropna() trong pandas).
  + **Xóa cột**: Loại bỏ biến (column) nếu tỷ lệ giá trị thiếu quá cao (thường >50%).
* **Khi nào sử dụng**:
  + Tỷ lệ giá trị thiếu thấp (ví dụ: <5%).
  + Dữ liệu MCAR, vì loại bỏ không làm sai lệch phân bố.
* **Ưu điểm**:
  + Đơn giản, dễ thực hiện.
  + Không cần giả định về dữ liệu.
* **Nhược điểm**:
  + Làm giảm kích thước mẫu, có thể gây mất thông tin quan trọng.
  + Không phù hợp nếu tỷ lệ thiếu cao hoặc dữ liệu MAR/MNAR.
* **Ví dụ**: Nếu tập dữ liệu điểm thi có 100 học sinh và chỉ 2 học sinh thiếu điểm, có thể xóa 2 hàng này.

**b. Thay thế bằng giá trị thống kê (Imputation)**

* **Thay bằng trung bình (Mean Imputation)**:
  + Dùng giá trị trung bình của biến để thay thế giá trị thiếu.
  + **Khi nào sử dụng**: Dữ liệu phân bố đối xứng, không có ngoại lai, và MCAR.
  + **Ưu điểm**: Đơn giản, giữ nguyên kích thước mẫu.
  + **Nhược điểm**: Làm giảm phương sai, không phù hợp với dữ liệu lệch hoặc có ngoại lai.
* **Thay bằng trung vị (Median Imputation)**:
  + Dùng trung vị để thay thế, phù hợp với dữ liệu lệch hoặc có ngoại lai.
  + **Khi nào sử dụng**: Dữ liệu MAR hoặc có phân bố lệch.
  + **Ưu điểm**: Ít bị ảnh hưởng bởi ngoại lai hơn trung bình.
  + **Nhược điểm**: Có thể không phản ánh đúng sự biến thiên của dữ liệu.
* **Thay bằng mode (Mode Imputation)**:
  + Dùng giá trị xuất hiện nhiều nhất, thường dùng cho dữ liệu danh mục (categorical data).
  + **Khi nào sử dụng**: Biến danh mục, dữ liệu MCAR.
  + **Ưu điểm**: Phù hợp với dữ liệu không phải số.
  + **Nhược điểm**: Có thể làm sai lệch nếu mode không đại diện.
* **Ví dụ**: Nếu điểm thi của một học sinh bị thiếu, bạn có thể thay bằng trung bình (70) hoặc trung vị (72) của lớp.

**c. Thay thế bằng giá trị dự đoán**

* **Sử dụng mô hình dự đoán**:
  + Dùng các mô hình như hồi quy tuyến tính, KNN (K-nearest neighbors), hoặc mô hình máy học để dự đoán giá trị thiếu dựa trên các biến khác.
  + **Khi nào sử dụng**: Dữ liệu MAR, có mối quan hệ rõ ràng giữa các biến.
  + **Ưu điểm**: Chính xác hơn nếu mô hình tốt, giữ được cấu trúc dữ liệu.
  + **Nhược điểm**: Phức tạp, cần nhiều dữ liệu và giả định về mối quan hệ.
* **Thay thế bằng giá trị gần nhất**:
  + Sử dụng giá trị của quan sát gần nhất (ví dụ: theo thời gian hoặc thứ tự).
  + **Khi nào sử dụng**: Dữ liệu chuỗi thời gian hoặc dữ liệu có tính liên tục.
* **Ví dụ**: Nếu thu nhập của một người bị thiếu, dùng mô hình hồi quy dựa trên tuổi, nghề nghiệp để dự đoán.

**d. Thay thế bằng giá trị cố định**

* Thay giá trị thiếu bằng một hằng số (ví dụ: 0, “Unknown”, hoặc giá trị trung bình toàn cầu).
* **Khi nào sử dụng**: Biến danh mục hoặc khi cần giữ nguyên dữ liệu mà không muốn xóa.
* **Ưu điểm**: Đơn giản, giữ kích thước mẫu.
* **Nhược điểm**: Có thể làm sai lệch phân bố, không phản ánh thực tế.
* **Ví dụ**: Trong khảo sát, thay các giá trị “không trả lời” bằng “Unknown” cho biến giới tính.

**e. Sử dụng phương pháp nội suy (Interpolation)**

* **Mô tả**: Ước lượng giá trị thiếu dựa trên các giá trị lân cận, thường dùng cho dữ liệu chuỗi thời gian.
  + Các phương pháp: Nội suy tuyến tính, nội suy đa thức, hoặc spline.
* **Khi nào sử dụng**: Dữ liệu có tính liên tục, như nhiệt độ hoặc giá cổ phiếu.
* **Ưu điểm**: Phù hợp với dữ liệu có xu hướng rõ ràng.
* **Nhược điểm**: Có thể không chính xác nếu dữ liệu dao động mạnh.
* **Ví dụ**: Nếu nhiệt độ lúc 10h và 12h được ghi nhận, nội suy tuyến tính có thể ước lượng nhiệt độ lúc 11h.

**f. Không xử lý (giữ nguyên giá trị thiếu)**

* Một số thuật toán (như cây quyết định) có thể xử lý giá trị thiếu mà không cần thay thế.
* **Khi nào sử dụng**: Khi phân tích không yêu cầu dữ liệu đầy đủ hoặc khi thuật toán tự động xử lý giá trị thiếu.
* **Ưu điểm**: Tránh làm sai lệch dữ liệu.
* **Nhược điểm**: Không thể tính toán một số chỉ số thống kê mô tả (như trung bình) nếu có giá trị thiếu.

**3. Quy trình xử lý giá trị thiếu**

1. **Kiểm tra dữ liệu**:
   * Xác định tỷ lệ và mẫu hình giá trị thiếu (MCAR, MAR, MNAR).
   * Ví dụ: Trong Python, dùng df.isna().sum() để đếm giá trị thiếu.
2. **Chọn phương pháp phù hợp**:
   * Nếu tỷ lệ thiếu thấp (<5%) và MCAR, cân nhắc loại bỏ.
   * Nếu dữ liệu lệch hoặc có ngoại lai, dùng trung vị hoặc mô hình dự đoán.
   * Với dữ liệu danh mục, dùng mode hoặc giá trị cố định.
   * Với chuỗi thời gian, cân nhắc nội suy.
3. **Áp dụng và kiểm tra**:
   * Thực hiện xử lý (dùng pandas, R, hoặc phần mềm như SPSS).
   * Kiểm tra xem phương pháp có làm sai lệch phân bố dữ liệu không (so sánh histogram trước và sau).
4. **Ghi nhận phương pháp**: Lưu ý phương pháp xử lý để đảm bảo tính minh bạch trong phân tích.

**4. Lưu ý khi tính toán chỉ số thống kê mô tả**

* **Trung bình**: Nhạy cảm với giá trị thiếu và cách thay thế (trung bình hoặc trung vị có thể làm thay đổi kết quả).
* **Trung vị**: Ít bị ảnh hưởng bởi giá trị thiếu, nhưng cần đảm bảo đủ dữ liệu để tính chính xác.
* **Độ lệch chuẩn/Phương sai**: Loại bỏ hoặc thay thế không đúng cách có thể làm sai lệch thước đo phân tán.
* **Boxplot**: Giá trị thiếu có thể ảnh hưởng đến Q1, Q2, Q3, vì vậy cần xử lý trước khi vẽ.

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

**1. Biểu đồ Histogram**

**Cách vẽ histogram**

* Histogram chia dữ liệu thành các **khoảng (bins)** và đếm số lượng giá trị rơi vào mỗi khoảng.
* Trục ngang (x): Giá trị của dữ liệu (điểm thi, ví dụ: 40–50, 50–60, v.v.).
* Trục dọc (y): Tần suất (số học sinh có điểm trong mỗi khoảng).
* Trong ví dụ này, giả sử chia dữ liệu thành các khoảng có độ rộng 10 (40–50, 50–60, ..., 90–100).

**Cách đọc histogram**

1. **Hình dạng phân bố**:
   * Xem hình dạng tổng thể của histogram để xác định phân bố:
     + **Đối xứng**: Nếu các cột đối xứng qua giá trị trung tâm, dữ liệu có thể gần với phân bố chuẩn.
     + **Lệch phải**: Nếu đuôi bên phải dài hơn (nhiều giá trị nhỏ, ít giá trị lớn).
     + **Lệch trái**: Nếu đuôi bên trái dài hơn (nhiều giá trị lớn, ít giá trị nhỏ).
   * Với dữ liệu trên, histogram có thể cho thấy **lệch trái** vì có nhiều giá trị cao (90–100) và một vài giá trị thấp (40–50).
2. **Tần suất**:
   * Các cột cao hơn cho biết khoảng điểm nào phổ biến nhất. Ví dụ, nếu cột 70–80 cao, nhiều học sinh có điểm trong khoảng này.
   * Trong dữ liệu, khoảng 90–100 có 6 giá trị (90, 90, 92, 95, 95, 100, 100), nên cột này sẽ cao.
3. **Phạm vi và ngoại lai**:
   * Phạm vi dữ liệu là từ giá trị nhỏ nhất (40) đến lớn nhất (100).
   * Giá trị 40 có thể xuất hiện như một cột thấp riêng lẻ, gợi ý khả năng ngoại lai.

**Diễn giải histogram**

* **Phân bố**: Dữ liệu lệch trái nhẹ, với nhiều học sinh đạt điểm cao (80–100) và một vài học sinh đạt điểm thấp (40–50). Điều này có thể cho thấy bài thi không quá khó, nhưng có một số học sinh gặp khó khăn.
* **Tập trung**: Khoảng 70–80 và 90–100 có tần suất cao, cho thấy đa số học sinh đạt điểm khá tốt.
* **Ngoại lai**: Giá trị 40 nằm xa phần còn lại, có thể là ngoại lai, cần kiểm tra thêm (lỗi nhập liệu hoặc học sinh có kết quả kém bất thường).

**Ví dụ minh họa histogram**

Nếu vẽ histogram với dữ liệu trên (bins rộng 10), ta có thể thấy:

* Khoảng 40–50: 2 học sinh (40, 45).
* Khoảng 50–60: 3 học sinh (50, 55, 58).
* Khoảng 60–70: 6 học sinh (60, 62, 65, 65, 68, 70).
* Khoảng 70–80: 5 học sinh (70, 72, 75, 75, 78).
* Khoảng 80–90: 5 học sinh (80, 82, 85, 85, 88).
* Khoảng 90–100: 9 học sinh (90, 90, 92, 95, 95, 98, 100, 100, 100).

Histogram sẽ có đuôi trái dài hơn (do giá trị 40), với đỉnh ở khoảng 90–100.

**2. Biểu đồ hộp (Boxplot)**

**Cách vẽ boxplot**

* Boxplot sử dụng các **tứ phân vị** (Q1, Q2, Q3) và khoảng tứ phân vị (IQR) để mô tả phân bố:
  + **Q1**: Giá trị mà 25% dữ liệu thấp hơn (tứ phân vị thứ nhất).
  + **Q2**: Trung vị (50% dữ liệu thấp hơn).
  + **Q3**: Giá trị mà 75% dữ liệu thấp hơn (tứ phân vị thứ ba).
  + **IQR**: Q3 - Q1.
  + **Râu**: Mở rộng từ Q1 - 1.5 × IQR (râu dưới) đến Q3 + 1.5 × IQR (râu trên).
  + **Ngoại lai**: Các giá trị nằm ngoài khoảng râu.

**Tính toán cho dữ liệu**

Sắp xếp dữ liệu: [40, 45, 50, 55, 58, 60, 62, 65, 65, 68, 70, 70, 72, 75, 75, 78, 80, 82, 85, 85, 88, 90, 90, 92, 95, 95, 98, 100, 100, 100]

* **Q2 (trung vị)**: Với n=30 n = 30 n=30, trung vị là trung bình của giá trị thứ 15 và 16: (75 + 78) / 2 = 76.5.
* **Q1**: Trung vị của nửa dưới (15 giá trị đầu: 40, 45, ..., 75) = 62.
* **Q3**: Trung vị của nửa trên (15 giá trị cuối: 78, ..., 100) = 90.
* **IQR**: Q3 - Q1 = 90 - 62 = 28.
* **Râu dưới**: Q1 - 1.5 × IQR = 62 - 1.5 × 28 = 20.
* **Râu trên**: Q3 + 1.5 × IQR = 90 + 1.5 × 28 = 132 (nhưng giá trị tối đa là 100, nên râu trên dừng ở 100).
* **Ngoại lai**: Giá trị < 20 hoặc > 132. Ở đây, 40 là ngoại lai vì 40 < 20.

**Cách đọc boxplot**

1. **Hộp (Box)**: Kéo dài từ Q1 (62) đến Q3 (90), chứa 50% dữ liệu ở giữa (điểm từ 62 đến 90).
2. **Đường giữa (trung vị)**: Ở 76.5, cho thấy giá trị trung tâm của điểm thi.
3. **Râu (Whiskers)**:
   * Râu dưới: Từ 62 đến giá trị nhỏ nhất không phải ngoại lai (45).
   * Râu trên: Từ 90 đến giá trị lớn nhất (100).
4. **Ngoại lai**: Giá trị 40 được vẽ dưới dạng điểm riêng lẻ, vì nó nằm dưới ngưỡng 20.

**Diễn giải boxplot**

* **Trung vị cao (76.5)**: Cho thấy đa số học sinh đạt điểm khá tốt, gần với mức trung bình cao.
* **Hộp bất đối xứng**: Khoảng từ Q2 đến Q3 (76.5–90) ngắn hơn khoảng từ Q1 đến Q2 (62–76.5), gợi ý dữ liệu **lệch trái**, với nhiều điểm số cao hơn trung vị.
* **IQR (28)**: Cho thấy 50% học sinh có điểm từ 62 đến 90, mức phân tán vừa phải.
* **Ngoại lai (40)**: Một học sinh có điểm rất thấp so với phần còn lại, có thể do lỗi nhập liệu hoặc gặp vấn đề trong kỳ thi.
* **Râu trên ngắn**: Do điểm tối đa là 100, các giá trị cao bị giới hạn, tạo phân bố lệch trái.

8. 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?Thống kê mô tả là gì? Nó khác gì với thống kê suy luận (inferential statistics)?