Chương 1: Dữ liệu và thu thập dữ liệu

Khoa Toán - Cơ - Tin học
Trường Đại học Khoa học Tự nhiên
Đại học Quốc gia Hà Nội

Ngày 5 tháng 9 năm 2023

1. Giới thiệu về thống kê và phân tích dữ liệu
2. Các phương pháp thống kê
3. Biến và dữ liệu
4. Thang đo
5. Phương pháp thu thập dữ liệu

Nội dung

1. Giới thiệu về thống kê và phân tích dữ liệu
2. Các phương pháp thống kê
3. Biến và dữ liệu
4. Thang đo
5. Phương pháp thu thập dữ liệu

Thống kê là gì?

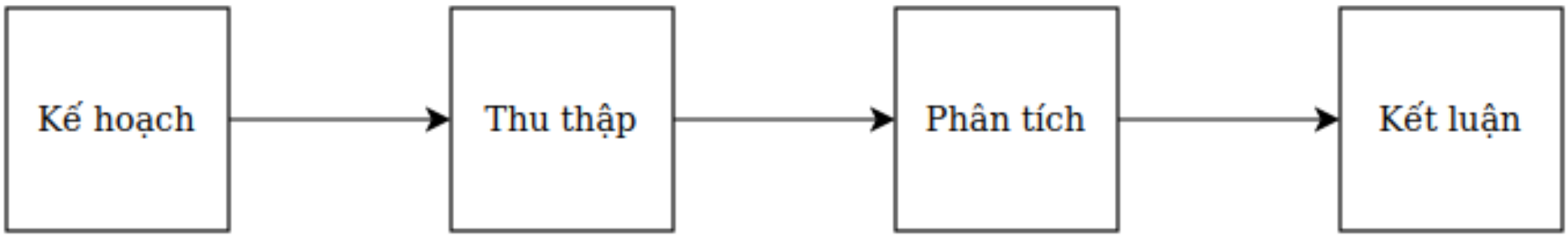
Định nghĩa

Thống kê là một khoa học, đồng thời là một công nghệ cung cấp cho ta những phương pháp, công cụ để thu thập và tạo dữ liệu, trình bày và phân tích dữ liệu để hiểu nội dung ẩn chứa trong dữ liệu. Từ đó rút ra những thông tin, tri thức hữu ích và đưa ra những quyết định, chính sách thích hợp.

Thống kê là gì?

Định nghĩa

Thống kê là một khoa học, đồng thời là một công nghệ cung cấp cho ta những phương pháp, công cụ để thu thập và tạo dữ liệu, trình bày và phân tích dữ liệu để hiểu nội dung ẩn chứa trong dữ liệu. Từ đó rút ra những thông tin, tri thức hữu ích và đưa ra những quyết định, chính sách thích hợp.



Vai trò của thống kê trong KHXH và KHSS

Khoa học xã hội

Trong Khoa học xã hội (Xã hội học, Tâm lý học, Kinh tế học, Giáo dục học, . . .), ngoài những yếu tố mang tính chất khách quan, tất yếu và phổ biến, các quy luật xã hội còn thể hiện những mỗi liên hệ và sự tác động lẫn nhau vô cùng phức tạp giữa người và người.
Do đó, nó đòi hỏi những phương pháp và công cụ nghiên cứu trong thống kê.

Vai trò của thống kê trong KHXH và KHSS

Khoa học xã hội

Trong Khoa học xã hội (Xã hội học, Tâm lý học, Kinh tế học, Giáo dục học, . . .), ngoài những yếu tố mang tính chất khách quan, tất yếu và phổ biến, các quy luật xã hội còn thể hiện những mỗi liên hệ và sự tác động lẫn nhau vô cùng phức tạp giữa người và người.
Do đó, nó đòi hỏi những phương pháp và công cụ nghiên cứu trong thống kê.

Khoa học sự sống

Trong Khoa học sự sống (sinh vật, y tế, . . .), các phương pháp thống kê giúp các nhà nghiên cứu thu thập dữ liệu, thiết kế thí nghiệm đúng đắn và đánh giá kết quả thí nghiệm.

Một số khái niệm thường dùng trong thống kê

Tổng thể

Tổng thể là tập hợp tất cả các đối tượng có chung đặc điểm mà ta cần điều tra. Các phần tử tạo thành tổng thể gọi là **đơn vị tổng thể**. Mỗi phần tử của tổng thể được gọi là một **cá thể** của tổng thể đó. Số lượng cá thể của tổng thể được gọi là **quy mô của tổng thể**.

Một số khái niệm thường dùng trong thống kê

Tổng thể

Tổng thể là tập hợp tất cả các đối tượng có chung đặc điểm mà ta cần điều tra. Các phần tử tạo thành tổng thể gọi là **đơn vị tổng thể**. Mỗi phần tử của tổng thể được gọi là một **cá thể** của tổng thể đó. Số lượng cá thể của tổng thể được gọi là **quy mô của tổng thể**.

Mẫu

Mẫu là một tập hợp con được rút ra từ tổng thể theo một cách nào đó. Số cá thể trong mẫu được gọi là **kích thước mẫu**. Nếu ta tiến hành lấy số đo mọi đơn vị tổng thể thì ta gọi là **tổng điều tra**.

Một số khái niệm thường dùng trong thống kê

Tiêu thức thống kê

Tiêu thức thống kê là đặc điểm của đơn vị tổng thể được chọn ra để nghiên cứu tùy theo mục đích nghiên cứu khác nhau (Ví dụ: Điều tra tình hình sức khỏe của trẻ sơ sinh, tổng điều tra dân số, . . .).

Ví dụ

Ta cần đo chiều cao của các cây gỗ "Sưa" tại rừng Quốc Gia Cúc Phương.

Một số khái niệm thường dùng trong thống kê

Tiêu thức thống kê

Tiêu thức thống kê là đặc điểm của đơn vị tổng thể được chọn ra để nghiên cứu tùy theo mục đích nghiên cứu khác nhau (Ví dụ: Điều tra tình hình sức khỏe của trẻ sơ sinh, tổng điều tra dân số, . . .).

Ví dụ

Ta cần đo chiều cao của các cây gỗ "Sưa" tại rừng Quốc Gia Cúc Phương.

- Tổng thể là tất cả các cây gỗ "Sưa" trong rừng Quốc Gia.
- Mẫu là một số lượng nhỏ các cây gỗ "Sưa" trong rừng được quan sát.
- Tiêu thức thống kê là chiều cao của cây gỗ.

Giới thiệu về phân tích dữ liệu

Dữ liệu

Dữ liệu là các con số, từ ngữ hay hình ảnh phản ánh thực tế của đối tượng nghiên cứu.

Phân tích dữ liệu

Phân tích dữ liệu là quá trình phát hiện, giải thích và truyền đạt các mô hình có ý nghĩa trong dữ liệu; đặc biệt có giá trị trong các lĩnh vực có nhiều thông tin được ghi lại, phân tích dựa vào sự ứng dụng đồng thời của số liệu thống kê, lập trình máy tính và nghiên cứu hoạt động để định lượng hiệu suất.

Giới thiệu về phân tích dữ liệu

Các bước phân tích dữ liệu

- Phân tích mô tả: Miêu tả những gì đã xảy ra trong một khoảng thời gian nhất định. VD: Doanh số tháng này có lớn hơn tháng trước không?
- Phân tích chẩn đoán: Tập trung nhiều hơn vào lí do tại sao một hiện tượng nào đó xảy ra. Điều này yêu cầu dữ liệu đầu vào đa dạng hơn và cần một vài giả thuyết. VD: Thời tiết có tác động đến doanh số bán hàng không?
- Phân tích dự đoán: Cho biết những gì có thể sẽ xảy ra trong thời gian tới. VD:Trong mùa hè nóng lần trước doanh số của chúng ta là bao nhiêu? Có bao nhiêu mô hình thời tiết dự đoán mùa hè năm nay sẽ nóng?
- Phân tích đề xuất: Đề xuất những hành động nên thực hiện. VD: nếu xác suất rằng mùa hè năm nay là nóng được đo theo các mô hình thời tiết mà công ty sử dụng là trên 58%, công ty nên tăng sản lượng bia

Nội dung

1. Giới thiệu về thống kê và phân tích dữ liệu
2. Các phương pháp thống kê
3. Biến và dữ liệu
4. Thang đo
5. Phương pháp thu thập dữ liệu

Các phương pháp thống kê

Thống kê mô tả

Thống kê mô tả có nhiệm vụ cung cấp các phương pháp để tổ chức, mô tả và trình bày các dữ liệu thu thập được sao cho người đọc sẽ hiểu được dữ liệu một cách tốt nhất. Các phương pháp đó là:

- Thu thập dữ liệu
- Trình bày dữ liệu
- Đưa ra các thông tin đặc trưng của dữ liệu

Các phương pháp thống kê

Thống kê mô tả

Thống kê mô tả có nhiệm vụ cung cấp các phương pháp để tổ chức, mô tả và trình bày các dữ liệu thu thập được sao cho người đọc sẽ hiểu được dữ liệu một cách tốt nhất. Các phương pháp đó là:

- Thu thập dữ liệu
- Trình bày dữ liệu
- Đưa ra các thông tin đặc trưng của dữ liệu

Thống kê suy diễn

Thống kê suy diễn có nhiệm vụ xây dựng các phương pháp để ta suy diễn ra các kết luận, lập các dự báo (với một độ chính xác nào đó) về toàn bộ tổng thể căn cứ trên một mẫu dữ liệu thu thập. Các phương pháp đó là:

- Ước lượng cho một tham số
- Kiểm định một giả thiết thống kê

Các phương pháp chọn mẫu

Chọn mẫu phi xác suất

Phương pháp chọn mẫu mà các đơn vị trong tổng thể chung không có khả năng ngang nhau để được chọn vào mẫu nghiên cứu.

Các phương pháp chọn mẫu

Chọn mẫu phi xác suất

Phương pháp chọn mẫu mà các đơn vị trong tổng thể chung không có khả năng ngang nhau để được chọn vào mẫu nghiên cứu.

Chọn mẫu xác suất

Phương pháp chọn mẫu mà khả năng được chọn vào tổng thể của tất cả các đơn vị của tổng thể đều như nhau.

Các phương pháp chọn mẫu xác suất

- Mẫu ngẫu nhiên đơn giản
- Mẫu theo khối
- Mẫu phân tầng

Các phương pháp chọn mẫu phi xác suất

- Judgement
- Chunk
- Quota

Chọn mẫu ngẫu nhiên đơn giản

Phép chọn mẫu ngẫu nhiên đơn giản là phép chọn mẫu mà trong đó mỗi cá thể của tổng thể được chọn một cách độc lập với xác suất như nhau.

Ví dụ

Viết tên của mỗi cá thể vào một phiếu và bỏ vào và bỏ vào một cái thùng trộn đều lên và lần lượt bốc thăm, Cá thể nào có tên trong phiếu được bốc ra thì được chọn là phần tử của mẫu. Tuy nhiên khi quy mô của tổng thể lớn thì người ra phải nhờ đến các phần mềm máy tính như: Excel, R, ... để thực hiện việc lấy mẫu ngẫu nhiên.

Chọn mẫu theo khối

1. Chia tổng thể chung gồm N khối, mỗi khối xem là một tổng thể con.
2. Từ mỗi khối, chọn ngẫu nhiên ra m khối trong N khối đó.
3. Tập hợp các cá thể trong m khối đó được chọn thành một mẫu để khảo sát.

Phương pháp chọn mẫu này được áp dụng khi ta không liệt kê được danh sách tất cả các cá thể trong tổng thể.

Ví dụ

Một nhà nghiên cứu muốn điều tra các hộ gia đình trên toàn quận Thanh Xuân. Giả sử ta không có danh sách tất cả các hộ gia đình trong quận mà chỉ có danh sách các tổ dân phố. Trong quận Thanh Xuân có 700 tổ dân phố, mỗi tổ dân phố xem là một tổng thể con. Chọn ngẫu nhiên 7 tổ dân phố. Tất cả các hộ gia đình của 7 tổ dân phố được chọn thành một mẫu để khảo sát.

Chọn mẫu phân tầng

1. Giả sử quy mô tổng thể là N, ta cần chọn mẫu cỡ n.
2. Chia tổng thể thành k tầng, tầng thứ i có N_i cá thể, $N = \sum_{i=1}^{k} N_i$.
3. Trong tầng thứ i chọn ngẫu nhiên ra $n_i = nf_i$ trong đó $f_i = \frac{N_i}{N}$ là tỷ lệ của tầng thứ i trong toàn bộ tổng thể.

Ví dụ

Tại một trường đại học có 20000 sinh viên có 5 hệ đào tạo khác nhau: Hệ chính quy: 10000 sinh viên, chiếm tỷ lệ 50%; Hệ hoàn thiện đại học: 2000 sinh viên, chiếm tỉ lệ 10%; Hệ văn bằng hai 2000 sinh viên, chiếm tỷ lệ 10%; Hệ tại chức: 5000 sinh viên, chiếm tỷ lệ 25% và Hệ cao học: 1000 sinh viên chiếm tỷ lệ 5%. Mỗi hệ đào tạo được xem như một tầng. Bộ phận đảm bảo chất lượng tiến hành cuộc khảo sát về chất lượng và mức độ hài lòng của sinh viên. Số sinh viên chọn ra để khảo sát là 1000. Như vạy ta lấy ở hệ chính quy $(1000)(0{,}5) = 500$ sinh viên, Hệ hoàn thiện đại học $(1000)(0{,}1) = 100$ sinh viên, Hệ văn bằng hai $(1000)(0{,}1) = 100$ sinh viên, Hệ tại chức $(1000)(0{,}25) = 250$ sinh viên và Hệ cao học $(1000)(0{,}05) = 50$ sinh viên.

Nội dung

1. Giới thiệu về thống kê và phân tích dữ liệu
2. Các phương pháp thống kê
3. Biến và dữ liệu
4. Thang đo
5. Phương pháp thu thập dữ liệu

Biến và dữ liệu

Biến là một dấu hiệu ta quan tâm nghiên cứu trên tổng thể. Ta gọi nó là biến vì nó thay đổi từ cá thể này sang cá thể khác. Biến có hai loại: biến định lượng và biến định tính.

Biến và dữ liệu

Biến là một dấu hiệu ta quan tâm nghiên cứu trên tổng thể. Ta gọi nó là biến vì nó thay đổi từ cá thể này sang cá thể khác. Biến có hai loại: biến định lượng và biến định tính.

Biến định lượng và dữ liệu định lượng

- Một biến được gọi là biến định lượng nếu nó có thể đo được trên mỗi cá thể và có giá trị là một số. Ta gọi đó là giá trị của biến.

Biến và dữ liệu

Biến là một dấu hiệu ta quan tâm nghiên cứu trên tổng thể. Ta gọi nó là biến vì nó thay đổi từ cá thể này sang cá thể khác. Biến có hai loại: biến định lượng và biến định tính.

Biến định lượng và dữ liệu định lượng

- Một biến được gọi là biến định lượng nếu nó có thể đo được trên mỗi cá thể và có giá trị là một số. Ta gọi đó là giá trị của biến.
- Tập hợp các giá trị của biến định lượng trên toàn bộ tổng thể cho ta dữ liệu định lượng.

Biến định tính và dữ liệu định tính

- Một biến được gọi là biến định tính nếu giá trị của biến đó trên mỗi cá thể là việc gán cho cá thể đó một thuộc tính hay gán nó vào một phàm trù, mức độ nào đó.

Biến định tính và dữ liệu định tính

- Một biến được gọi là biến định tính nếu giá trị của biến đó trên mỗi cá thể là việc gán cho cá thể đó một thuộc tính hay gán nó vào một phàm trù, mức độ nào đó.
- Giá trị của biến định tính là một trong các phạm trù (thuộc tính) mà nhà nghiên cứu đưa ra. Biến định tính là loại biến rất thường gặp trong khoa học xã hội.

Biến định tính và dữ liệu định tính

- Một biến được gọi là biến định tính nếu giá trị của biến đó trên mỗi cá thể là việc gán cho cá thể đó một thuộc tính hay gán nó vào một phàm trù, mức độ nào đó.
- Giá trị của biến định tính là một trong các phạm trù (thuộc tính) mà nhà nghiên cứu đưa ra. Biến định tính là loại biến rất thường gặp trong khoa học xã hội.
- Tập hợp các giá trị của biến định tính trên toàn bộ tổng thể cho ta dữ liệu định tính.

Biến và dữ liệu

Ví dụ

Các bác sĩ đã thực hiện việc đánh giá sức khoẻ (thông qua cân nặng) và nghiên cứu giới tính của trẻ sơ sinh ở Hà Nội.

Biến và dữ liệu

Ví dụ

Các bác sĩ đã thực hiện việc đánh giá sức khoẻ (thông qua cân nặng) và nghiên cứu giới tính của trẻ sơ sinh ở Hà Nội.

- Tổng thể là tập hợp tất cả các trẻ sơ sinh ở Hà Nội. Mỗi trẻ sơ sinh là một cá thể.

Biến và dữ liệu

Ví dụ

Các bác sĩ đã thực hiện việc đánh giá sức khoẻ (thông qua cân nặng) và nghiên cứu giới tính của trẻ sơ sinh ở Hà Nội.

- Tổng thể là tập hợp tất cả các trẻ sơ sinh ở Hà Nội. Mỗi trẻ sơ sinh là một cá thể.
- Biến định lượng là cân nặng của trẻ sơ sinh. Giá trị của biến là một số thực dương.
- Dữ liệu định lượng (thô) là tập hợp các giá trị về cân nặng của trẻ sơ sinh.

Biến và dữ liệu

Ví dụ

Các bác sĩ đã thực hiện việc đánh giá sức khoẻ (thông qua cân nặng) và nghiên cứu giới tính của trẻ sơ sinh ở Hà Nội.

- Tổng thể là tập hợp tất cả các trẻ sơ sinh ở Hà Nội. Mỗi trẻ sơ sinh là một cá thể.
- Biến định lượng là cân nặng của trẻ sơ sinh. Giá trị của biến là một số thực dương.
- Dữ liệu định lượng (thô) là tập hợp các giá trị về cân nặng của trẻ sơ sinh.
- Biến định tính là giới tính của trẻ sơ sinh. Mỗi trẻ sơ sinh được gán cho một trong hai thuộc tính: "Nam" hay "Nữ".
- Dữ liệu định tính là tập hợp các giá trị về giới tính của trẻ sơ sinh.

Các nguồn dữ liệu

Các nguồn dữ liệu

Có hai nguồn dữ liệu chính:

Các nguồn dữ liệu

Có hai nguồn dữ liệu chính:

Nguồn dữ liệu thứ cấp

Là dữ liệu từ một nguồn có sẵn đã công bố hoặc chưa công bố. Hiện nay khá nhiều dữ liệu thứ cấp (đã hoặc chưa qua xửa lý, tổng hợp) được đưa lên mạng dưới dạng nguồn dữ liệu mở. Nhà nghiên cứu dùng các công cụ tìm kiếm như Google để khai thác và tìm kiếm dữ liệu.

Các nguồn dữ liệu

Có hai nguồn dữ liệu chính:

Nguồn dữ liệu thứ cấp

Là dữ liệu từ một nguồn có sẵn đã công bố hoặc chưa công bố. Hiện nay khá nhiều dữ liệu thứ cấp (đã hoặc chưa qua xửa lý, tổng hợp) được đưa lên mạng dưới dạng nguồn dữ liệu mở. Nhà nghiên cứu dùng các công cụ tìm kiếm như Google để khai thác và tìm kiếm dữ liệu.

Nguồn dữ liệu sơ cấp

Là dữ liệu mà nhà nghiên cứu tự thu thập theo một quy trình bài bản phục vụ một mục tiêu và nội dung nghiên cứu đã xác định. Những phương pháp để thu thập dữ liệu sơ cấp thường dùng là:

Các nguồn dữ liệu

Có hai nguồn dữ liệu chính:

Nguồn dữ liệu thứ cấp

Là dữ liệu từ một nguồn có sẵn đã công bố hoặc chưa công bố. Hiện nay khá nhiều dữ liệu thứ cấp (đã hoặc chưa qua xửa lý, tổng hợp) được đưa lên mạng dưới dạng nguồn dữ liệu mở. Nhà nghiên cứu dùng các công cụ tìm kiếm như Google để khai thác và tìm kiếm dữ liệu.

Nguồn dữ liệu sơ cấp

Là dữ liệu mà nhà nghiên cứu tự thu thập theo một quy trình bài bản phục vụ một mục tiêu và nội dung nghiên cứu đã xác định. Những phương pháp để thu thập dữ liệu sơ cấp thường dùng là:

- Tiến hành thí nghiệm.
- Tiến hành quan sát, điều tra, khảo sát.

Nội dung

1. Giới thiệu về thống kê và phân tích dữ liệu
2. Các phương pháp thống kê
3. Biến và dữ liệu
4. Thang đo
5. Phương pháp thu thập dữ liệu

Thang đo

Mỗi một số đo của biên đều nằm trên một "thang đo" nào đó. Tùy mức độ tốt của thang đo, ta đề cập đến bốn thang đo sau:

- Thang đo định danh (Norminal).
- Thang đo thứ bậc (Ordinal).
- Thang đo khoảng (Interval).
- Thang đo tỉ lệ (Ratio).

Thang đo định danh

Định nghĩa

Thang đo định danh dùng cho các biến định tính. Số đo của các biến này là các mã số để phân loại đối tượng. Giữa các mã số ở đây không có quan hệ hơn kém, chỉ dùng để đếm tần số xuất hiện của của các biểu hiện.

Thang đo định danh

Định nghĩa

Thang đo định danh dùng cho các biến định tính. Số đo của các biến này là các mã số để phân loại đối tượng. Giữa các mã số ở đây không có quan hệ hơn kém, chỉ dùng để đếm tần số xuất hiện của của các biểu hiện.

Ví dụ

Số đo của biến giới tính (nam, nữ), biến màu sắc (xanh, đỏ, tím, ...), biến khu vực sống,.. thuộc thang đo định danh.

Thang đo thứ bậc

Thang đo thứ bậc thường dùng cho các biến định tính, đôi khi dùng cho cả biến định lượng. Trong thang đo này giữa các số đo của các biến có quan hệ bậc hơn kém. Tuy nhiên, sự chệnh lệch giữa các số đo không nhất thiết bằng nhau.

Thang đo thứ bậc

Ví dụ

Kết quả của các câu trả lời sau thuộc thang đo thứ bậc: 1. Bạn đánh giá thế nào về việc tiếp thu của mình đối với việc học các môn tự nhiên:

1. Tốt 2. Bình thường 3. Kém

2. Ý kiến của bạn về việc dọn dẹp vỉa hè của thành phố:
   - Rất không ủng hộ
   - Không ủng hộ
   - Không ý kiến
   - Ủng hộ
   - Rất ủng hộ

Thang đo khoảng

Định nghĩa

Thang đo khoảng là thang đo thứ bậc có khoảng cách đều nhau. Các phép tính cộng trừ đều cí nghĩa nhưng không có giá trị không chính xác và không lấy tỉ lệ giữa các số đo.

Ví dụ

Số đo nhiệt độ, chỉ số IQ,... thuộc thang đo khoảng.

Thang đo tỉ lệ

Định nghĩa

Thang đo tỉ lệ là thang đo khoảng, hơn nữa thang đo này có giá trị 0 xác định một cách chính xác và có thể lấy tỉ lệ giữa các số đo.

Ví dụ

Đơn vị đo tiền tệ (VND, dollar, pound, yen, ...); đơn vị đo chiều dài (cm, m, km,...); đơn vị đo khối lượng (kg, tấn, tạ, yến,...);... thuộc thang đo tỉ lệ.

Nội dung

1. Giới thiệu về thống kê và phân tích dữ liệu
2. Các phương pháp thống kê
3. Biến và dữ liệu
4. Thang đo
5. Phương pháp thu thập dữ liệu

Các phương pháp điều tra

Các phương pháp điều tra

- Điều tra trực tiếp: Phỏng vấn, quan sát, thí nghiệm, thực nghiệm.
- Sử dụng cơ sử dữ liệu có sẵn.

Các phương pháp điều tra - Phương án điều tra

Các phương pháp điều tra

- Điều tra trực tiếp: Phỏng vấn, quan sát, thí nghiệm, thực nghiệm.
- Sử dụng cơ sở dữ liệu có sẵn.

Phương án điều tra

- Xác định mục đích nghiên cứu.
- Xác định phạm vi, đối tượng và đơn vị điều tra.
- Xác định nội dung điều tra.
- Chọn phương pháp thu thập thông tin.
- Soạn thảo bảng hỏi.
- Chọn mẫu điều tra.
- Lập kế hoạch tổ chức và tiến hành điều tra.

Một số sai số khi điều tra

Dưới đây là một vài sai số chủ yếu khi khi điều tra, thu thập số liệu:

Một số sai số khi điều tra

Dưới đây là một vài sai số chủ yếu khi khi điều tra, thu thập số liệu:

- Sai số phi chọn mẫu: Là sai số xảy ra ở các cuộc điều tra do nhân viên cân đong, đo đếm sai, ghi chép sai, đơn vị điều tra cung cấp sai sự thật, . . . Sai số này không do việc chọn mẫu gây ra.
- Sai số chọn mẫu: Là sai số xảy ra trong điều tra chọn mẫu, do điều tra một số ít đơn vị nhưng kết quả lại ước lượng cho cả tổng thể.
- Sai số do không có câu trả lời.
- Sai số do thước đo: Chọn câu hỏi không tốt, chọn sai thang đo.