**1.2 Dữ liệu** (trang 6)

**Phần tử** là các thực thể mà từ đó dữ liệu được thu thập.

**Biến** là các một đặc tính quan tâm của phần tử.

**Quan sát** là tập hợp các số đo thu được của một phần tử.

**Thang đo**

Các loại thang đo: **danh nghĩa**, **thứ bậc**, **khoảng** hoặc **tỉ lệ**.

* Khi dữ liệu của một biến bao gồm các nhãn hoặc tên được sử dụng để phân biệt một thuộc tính của phần tử, thang đo này được coi là **thang đo danh nghĩa**.
* Nếu dữ liệu thể hiện tính chất của dữ liệu danh nghĩa và thứ tự hoặc xếp hạng của dữ liệu này có ý nghĩa thì sẽ được coi là **thang đo thứ bậc**.
* Thang đo đối với một biến là **thang đo khoảng** nếu dữ liệu có tất cả các thuộc tính của dữ liệu thứ tự và khoảng cách giữa các giá trị được thể hiện dưới dạng đơn vị đo lường cố định. Dữ liệu đo khoảng luôn là *dữ liệu số*.
* Thang đo lường đối với một biến là **thang đo tỷ lệ** nếu dữ liệu có tất cả các đặc tính của dữ liệu khoảng và tỷ lệ của hai giá trị có ý nghĩa. Các biến như khoảng cách, chiều cao, trọng lượng và thời gian đều sử dụng thang đo tỷ lệ. Thang đo này đòi hỏi một giá trị không (số 0) để chỉ ra rằng không có gì tồn tại trong biến tại điểm không này.

**Dữ liệu định tính và dữ liệu định lượng**

Dữ liệu định tính bao gồm các nhãn hay tên dùng để phân biệt các phần tử. **Dữ liệu định tính (****categorical)** sử dụng một trong 2 thang đo danh nghĩa hoặc thứ bậc.

**Dữ liệu định lượng (****quantitative)** sử dụng các giá trị số để biểu diễn mức độ bao nhiêu. Dữ liệu định lượng thu được bằng thang đo khoảng hoặc thang đo tỷ lệ.

**Biến định tính** là biến với các dữ liệu định tính, và **biến định lượng** là biến với dữ liệu định lượng.

**Dữ liệu thời điểm và dữ liệu chuỗi thời gian**

**Dữ liệu thời điểm (cross-sectional)** là dữ liệu được thu thập ở cùng hoặc xấp xỉ và cùng một thời điểm.

**Dữ liệu chuỗi thời gian (time series)** là dữ liệu thu thập được thông qua nhiều giai đoạn thời gian.

**Ghi chú**

1. Một quan sát là tập hợp các số đo thu được từ mỗi phần tử trong một tập dữ liêu. Do đó, số lượng các quan sát luôn luôn bằng số lượng các phần tử. Số lượng các số đo thu được cho mỗi phần tử bằng số lượng các biến. Do đó, tổng số các dữ liệu có thể được xác định bằng cách nhân số quan sát với số lượng các biến.

2. Dữ liệu định lượng có thể rời rạc hay liên tục. Dữ liệu định lượng để đo lường bao nhiêu (ví dụ, số lượng cuộc gọi nhận được trong 5 phút) là rời rạc. Dữ liệu định lượng để đo lường bao nhiêu (ví dụ, trọng lượng hoặc thời gian) là liên tục vì không có khoảng cách giữa các giá trị dữ liệu có thể thu được.

**Suy diễn thống kê**

**Tổng thể (population)** là tập hợp tất cả các phần tử được quan tâm trong một nghiên cứu cụ thể.

**Mẫu (sample)** là một tập con của tổng thể.

**Điều tra tổng thể (census)** là quá trình khảo sát để thu thập dữ liệu trên toàn bộ tổng thể.

**Điều tra mẫu (sample survey)** là quá trình thu thập dữ liệu trên một mẫu.

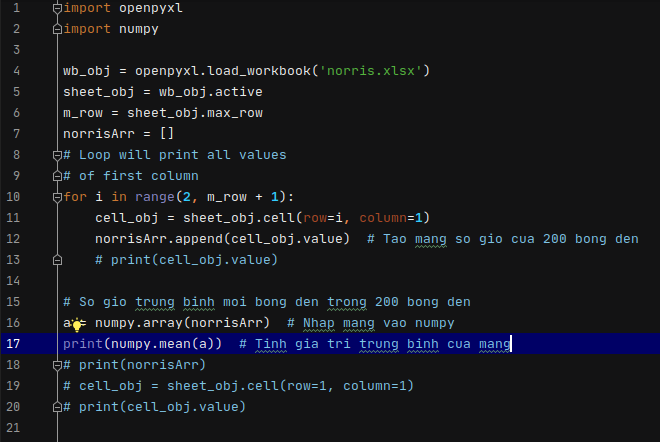
**Suy diễn thống kê (Statistical Inference)** là sử dụng dữ liệu từ một mẫu ước lượng và kiểm định giả thuyết về các đặc tính của một tổng thể.

Ví dụ nghiên cứu công ty sản xuất bóng đèn Norris Electronics.

Tổng thể là tất cả các bóng đèn có thể được sản xuất với sợi đốt mới. Để đánh giá những ưu điểm của sợi đốt mới, 200 bóng đèn với sợi đốt mới được sản xuất và kiểm tra. Dữ liệu thu thập ở bảng 1.5.

Để suy diễn tuổi thọ trung bình của tổng thể tất cả các bóng đèn được sản xuất với sợi đốt mới. Cộng 200 giá trị trong bảng **1.5** và chia tổng cho 200 sẽ tuổi thọ trung bình của bóng đèn: 76 giờ.

Cách tính trong Python



**Bài tập:**

2.

a) Số phần tử: 9 (Tên nơi lưu trú)

b) Biến: 4

c)

Biến định tính: Quốc gia, Giá phòng

Biến định lượng: Số lượng phòng, Điểm đánh giá chung

d)

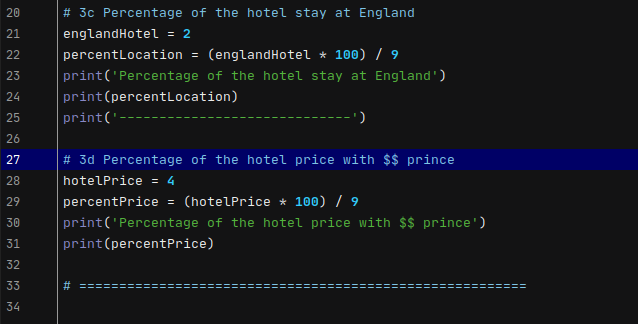
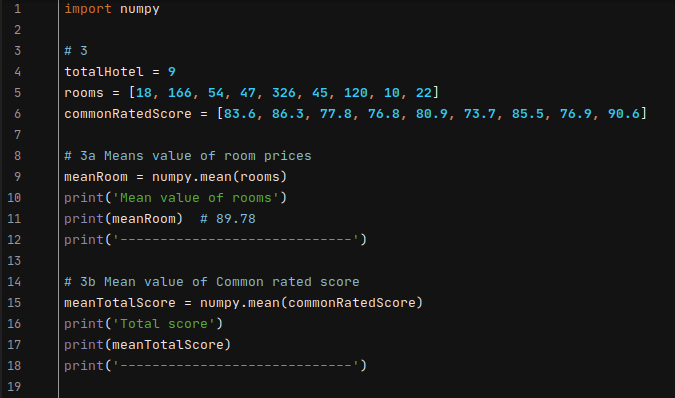
Quốc gia: Thang đo danh nghĩa

Giá phòng: Thang đo thứ bậc

Số lượng phòng: Thang đo tỷ lệ

Điểm đánh giá chung: Thang đo khoảng

3. Cách tính trong Python

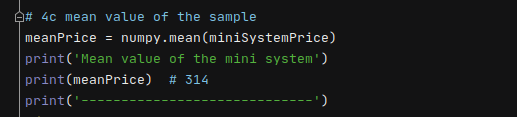


4.

a. Có 10 phần tử.

b. Tổng thể: Nhãn hiệu và kiểu

c.



d. Mức giá trung bình của 10 mẫu minisystem trong khảo sát là 314$.

5.

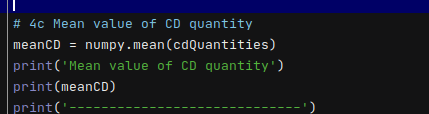
a. Có 5 biến trong tập dữ liệu.

b.

Biến định tính: Chất lượng âm thanh, Đài FM.

Biến định lượng: Giá (USD), Khả năng chứa đĩa CD, Số hộc băng.

c. Số CD chứa trung bình của đầu CD trong mẫu là: 3.



d. Phần trăm minisystem có đài FM được đánh giá rất tốt hay xuất sắc là 70%.

e. Phần trăm minisystem có 2 hộc băng cassette là 40%.

6.

a. Định lượng.

b. Định tính.

c. Định lượng.

d. Định lượng.

e. Định tính.

7.

a. Định tính

b. Thang đo thứ bậc.

8.

a. Cỡ mẫu khảo sát là 1005.

b. Định tính.

c. Tỷ lệ %.

d. ~291 người.

9.

a. Loại doanh nghiệp là biến định tính.

b. Tỷ lệ doanh nghiệp nhỏ: ~23%.

10.

a. Dữ liệu định lượng, thang đo tỉ lệ

b. Dữ liệu định tính, thang đo danh nghĩa

c. Dữ liệu định tính, thang đo thứ bậc

d. Dữ liệu định lượng, thang đo tỷ lệ

e. Dữ liệu định tính, thang đo danh nghĩa

11.

a. Dữ liệu định lượng, thang đo tỷ lệ

b. Dữ liệu định tính, thang đo thứ bậc

c. Dữ liệu định tính, thang đo thứ bậc

d. Dữ liệu định lượng, thang đo tỷ lệ.

e. Dữ liệu định tính, thang đo danh nghĩa

12.

a. Tổng thể thể nghiên cứu là số khác đến Hawaii.

b. Đúng

c. Định lượng, định tính, định tính, định lượng.

13.

a. Lợi nhuận tính bằng tỷ USD là dữ liệu định lương.

b. Dữ liệu chuỗi thời gian từ 1997 đến 2005.

c. Lợi nhuận của Volkswagen.

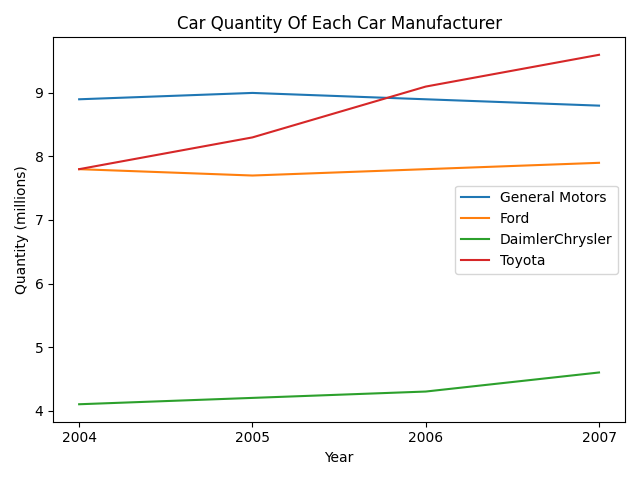
d. Lợi nhuận tương đối thấp trong các năm 1997 đến 1999, tăng mạnh trong 2001 và 2001 và suy giảm từ 2003 đến 2005, khuynh hướng suy giảm lợi nhuận dự phóng 600 triệu USD trong năm 2006 là khá hợp lý.

e. Vào tháng 7 năm 2001, xu hướng lợi nhuận đang dương, Volkswagen hứa hẹn là một đầu tư tốt trong năm 2001.

d. Hãy cẩn thận khi dữ phóng dữ liệu chuỗi thời gian trong tương lai vì xu hướng dữ liệu trong quá khứ có thể hoặc không thể tiếp tục.

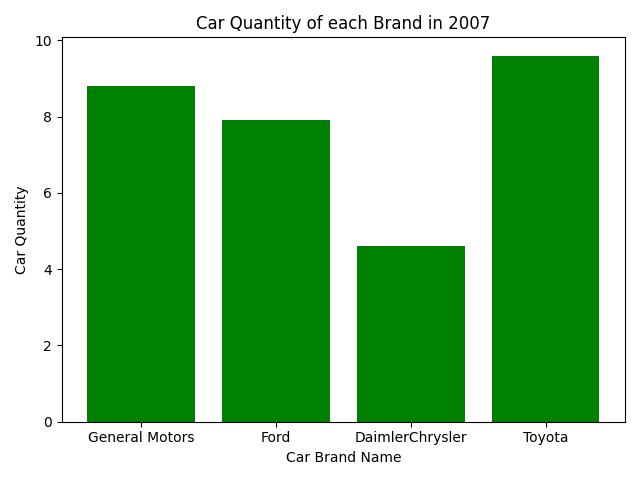
14.

a)



b) Hãng xe hơi lớn nhất thế giới: Toyota

c)



15)

a) Dữ liệu định lượng.

b) Dữ liệu là chuỗi thời gian.

c) Có khoảng 20 thuốc mới được phê duyệt trong năm 2003.

d) Năm có số lượng thuốc mới ít được kiểm duyệt là năm 2002. Có khoảng 26 loại thuốc được phê duyệt.

e) Số lượng thuốc mới được phê duyệt giảm dần theo thời gian. Ở các năm 1999 và năm 2003 có sự tăng nhẹ so với năm trước đó.

16)

a) Thử nghiệm vị sản phẩm và thử nghiệp thị trường

b) Nghiên cứu thống kê được thiết kế đặc biệt.

18)

a) 36%

b) 189 người

c) Định tính

19)

a) Tổng quan trong nghiên cứu là số người đăng ký Bussiness Week ở Bắc Mỹ.

b) Thu nhập hằng năm là biến định lượng.

c) Sỡ hữu thẻ American Express là biến định tính.

d) Nghiên cứu này liên quan đến dữ liệu thời điểm.

e)

20)

a) 43% nhà quản lý tự phân loại mình là lạc quan hay rất lạc quan trên thị trường chứng khoán, 21% lựa chọn chăm sóc sức khỏe như lĩnh vực có nhiều khả năng dẫn dắt thị trường trong 12 tháng kế tiếp.

b) Lợi nhuận trên vốn trung bình dự kiến trong 12 tháng tới được ước lượng là 11.2% cho tổng thể các nhà quản lý đầu tư.

c) Trung bình mẫu 2.5 năm là ước lượng thời gian mà tổng thể các nhà quản lý nghĩ cần để trở lại tăng trưởng bền vững.

21)

a) Hai tổng thể là phụ nữ có mẹ tuống thuốc DES trong suốt thai kỳ và phụ nữ có mẹ không uống thuốc DES.

b) Dữ liệu khảo sát này là một thực nghiệm.

c) Thống kê mô tả trong số 1000 người trong tổng thể này có xấp xỉ 16 người có mô bất thường.

d) Đối với phụ nữ có mẹ không dùng thuốc DES, ước lượng phụ nữ trong số 1000 người có mô bất thường là 984 người.

e)

22)

a) Tổng thể là các cử tri đã đăng ký.

b) Các cử tri.

c) Tốn quá nhiều thời gian và chi phí để tiếp xúc với toàn bộ tổng thể.

23)

a) Điểm đánh và thị phần.

b) Tổng thể là hộ gia đình có TV có xem chương trình

c) d)

24)

a) Đúng

b) Sai

c) Đúng

d) Sai

e) Sai

25) Cổ phiếu song hành là cổ phiếu phổ thông của các công ty nhỏ hơn không được các nhà phân tích phố Wall theo dõi chặt chẽ.

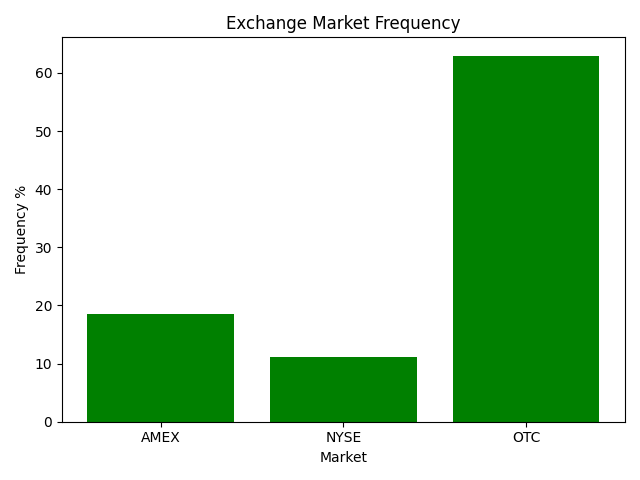
a) Có 5 biến trong tập dữ liệu

b) Biến định tính: Thị trường, Mã cổ phiếu. Biến định lượng: Giá trị vốn hóa thị trường (USD), Tỷ số giá/thu nhập, Lợi nhuận biên (%)

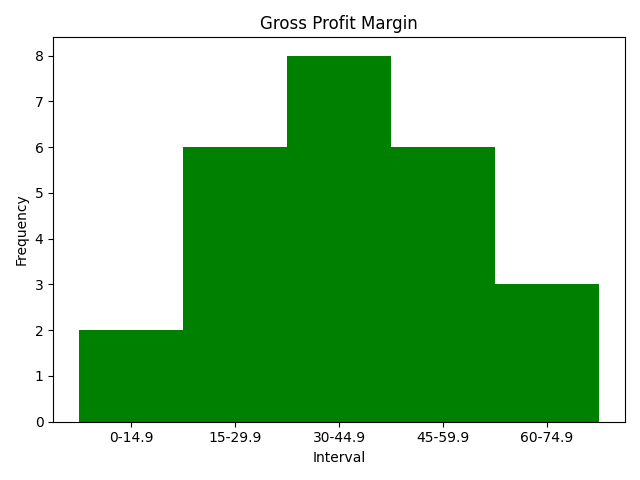
c)

Tần số: AMEX: 5, NYSE: 3, OTC: 17

Phần trăm : AMEX: ~19%, NYSE: ~11%, OTC: ~63%



d)



e) Tỷ số giá/thu nhập trung bình: ~20.22

