Thống kê cho Khoa học dữ liệu

Tiến sĩ Nguyễn Phúc Sơn

Trường Đại học Kinh tế - Luật
Đại học Quốc gia Thành phố Hồ Chí Minh

Ngày 3 tháng 10 năm 2018

Table of Contents

Thống kê là gì ?

- Thống kê làm việc với dữ liệu, cung cấp các insights để trợ giúp quá trình ra quyết định.
- Trong Khoa học dữ liệu (KHDL),

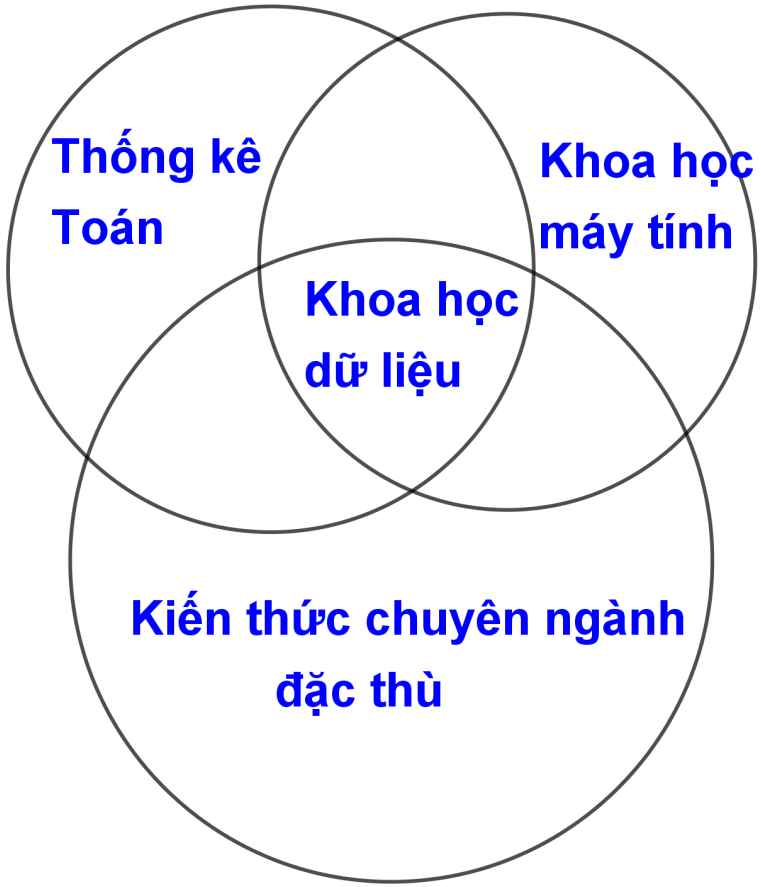


Table of Contents

Biểu đồ

- Hình vẽ dễ hiểu hơn lời nói. Chúng ta cần khi:
  1. Hiểu sơ bộ data đang có trong tay, bao gồm một số chẩn đoán ban đầu (diagnosis).
     - Xem xét outliers/anomaly's.
     - Phân cụm data (nếu cần)
  2. Trình bày kết quả cho khách hàng.
- Load dataset *customers.txt* và thử vài chart, diagrams

customers.txt

1. Đây là data về wholesale customers ở 1 nước Châu Âu (Nguồn: Kaggle).
2. Có 3 Regions. Ví dụ ta muốn xem nhanh mỗi vùng có bao nhiêu khách hàng trong data này.

```
df = pd.read_csv("customer.txt")
"""
Descriptive Statistics
"""
dnc = df["Region"].value_counts()
sns.barplot(dnc.index, dnc)

dfc = df["Channel"].value_counts()
sns.barplot(dfc.index, dfc)
```

customers.txt

- Tương tự cho 2 kênh mua sắm (Channels).
- Ngoài ra, ta cần nhìn qua xem các numeric columns giá trị phân bố thế nào

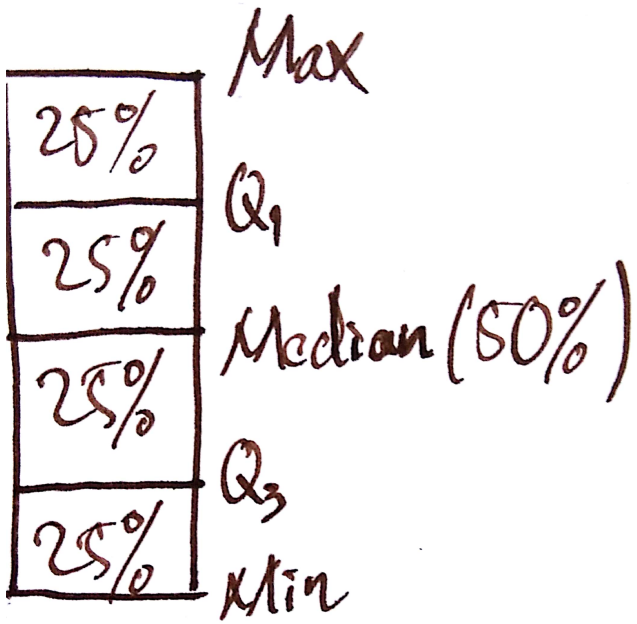
```
num = df.iloc[:,range(3,8)]
sns.boxplot(data=num)
```

- Nếu chúng ta xét nhiều nhóm ngành sản phẩm cùng lúc (nhiều biến) thì có thể phải tách riêng các giá trị cực hạn ra 1 nhóm đặc biệt

Mô tả bằng số

```
num.describe()
```

1. mean: trung bình
2. five-number summary:



Central tendency

- Mean và median(trung vị) cho biết vị trí trung tâm của data.
- Trong đời thường: dùng mean. Tuy nhiên, mean không phải lúc nào cũng thích hợp.
- Các bạn so sánh 2 data sau:
  1. điểm (nhóm 1): 89, 87, 91, 93, 94
  2. điểm (nhóm 2): 89, 87, 91, 93, 94, 4
  3. Tính mean và median của 2 datasets trên và nhận xét
  4. Theo các bạn mean hay median sẽ thích hợp hơn cho giá trị trung tâm ?
- mean1 = 90.8, mean2 = 76.33
- median1 = 91, median2 = 90.

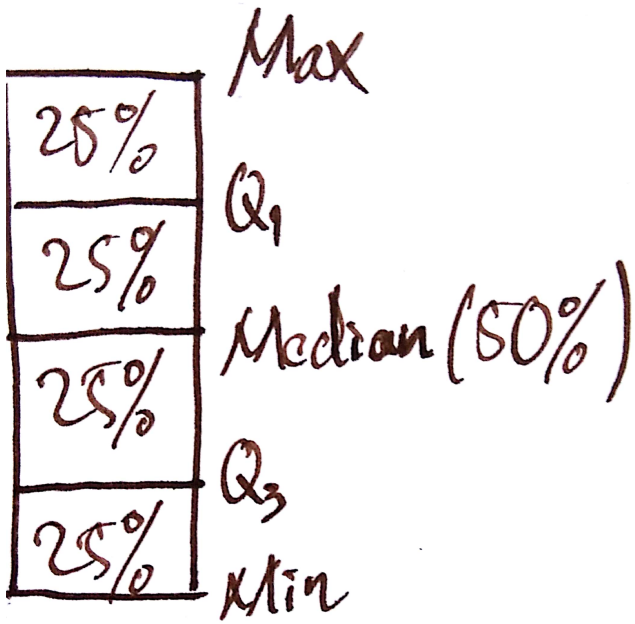
Central tendency

- Mean và median(trung vị) cho biết vị trí trung tâm của data.
- Trong đời thường: dùng mean. Tuy nhiên, mean không phải lúc nào cũng thích hợp.
- Các bạn so sánh 2 data sau:
  1. điểm (nhóm 1): 89, 87, 91, 93, 94
  2. điểm (nhóm 2): 89, 87, 91, 93, 94, 4
  3. Tính mean và median của 2 datasets trên và nhận xét
  4. Theo các bạn mean hay median sẽ thích hợp hơn cho giá trị trung tâm ?
- mean1 $= 90.8$, mean2 $= 76.33$
- median1 $= 91$, median2 $= 90$.
- 2 datasets khác nhau chỗ nào ?

Central tendency

- Mean và median(trung vị) cho biết vị trí trung tâm của data.
- Trong đời thường: dùng mean. Tuy nhiên, mean không phải lúc nào cũng thích hợp.
- Các bạn so sánh 2 data sau:
  1. điểm (nhóm 1): 89, 87, 91, 93, 94
  2. điểm (nhóm 2): 89, 87, 91, 93, 94, 4
  3. Tính mean và median của 2 datasets trên và nhận xét
  4. Theo các bạn mean hay median sẽ thích hợp hơn cho giá trị trung tâm ?
- mean1 $= 90.8$, mean2 $= 76.33$
- median1 $= 91$, median2 $= 90$.
- 2 datasets khác nhau chỗ nào ?

Central tendency



Central tendency

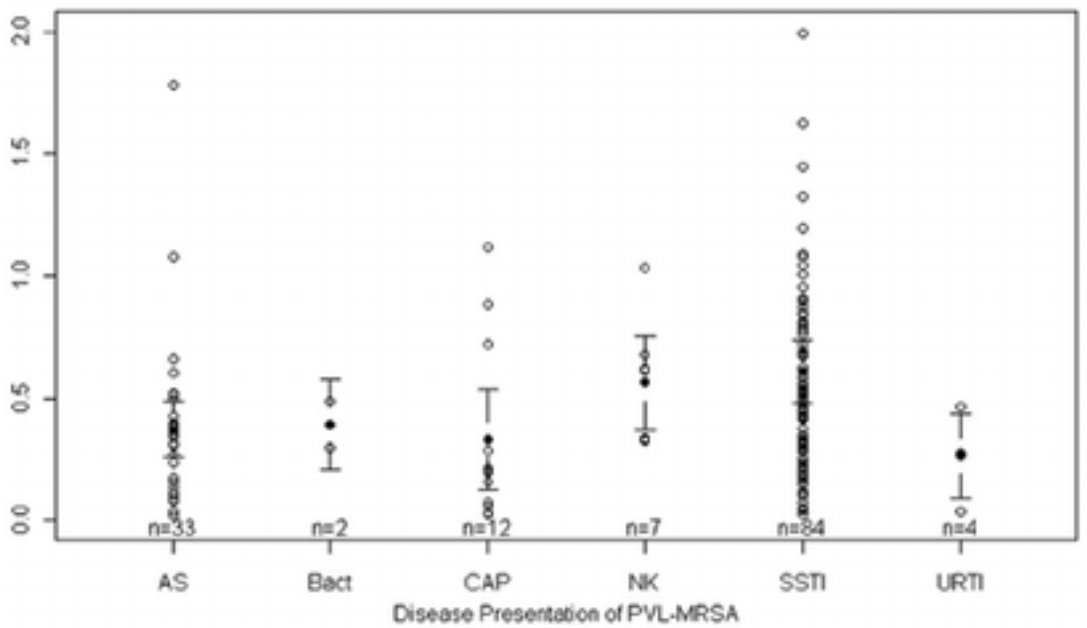
- Thử với Income datasets của Mỹ (nguồn: Kaggle)

```
income = pd.read_csv("Family Income and
    Expenditure.csv")
income = income["Total Household Income"]
sns.distplot(income, kde=False, rug=False)
sns.distplot(income)
sns.boxplot(income)
income.mean()
income.median()
```

- Theo bạn, nếu chúng ta báo cáo income ở trung tâm của data dân số thì nên dùng mean hay median ?

Dispersion

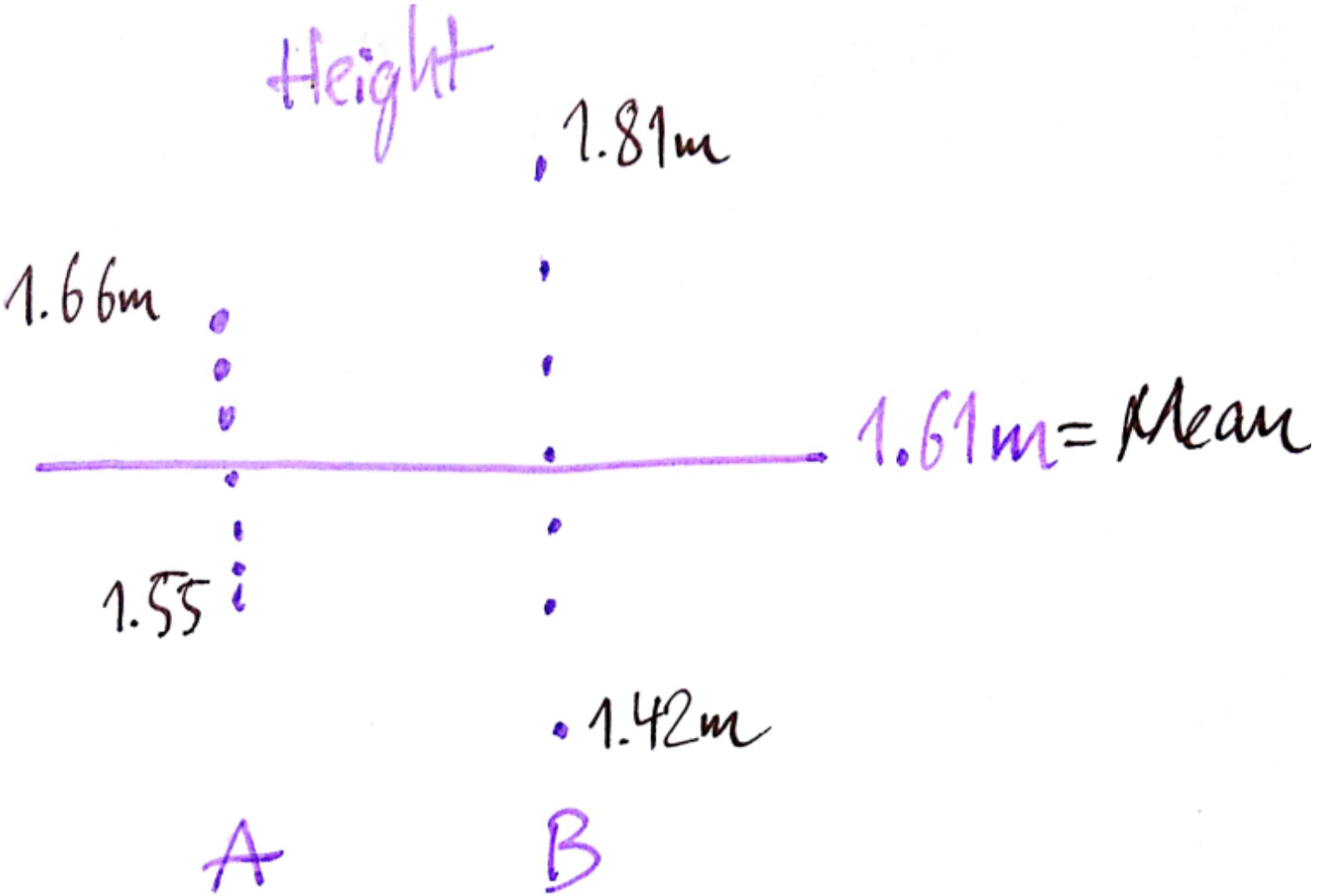
- Ngoài trung tâm, chúng ta cần biết data phân bổ trên diện rộng hay bó hẹp quanh trung tâm



Dispersion

- Vì sao như vậy ?
- Ví dụ:
  1. Chúng ta mang được đúng 100 áo thun đến bán ở hai trường A và B. Áo có 3 sizes: L, M, S. Để quyết định đem bao nhiêu áo theo các sizes đến từng trường, ta quan sát chiều cao của các bạn sinh viên theo hình dưới đây

Dispersion



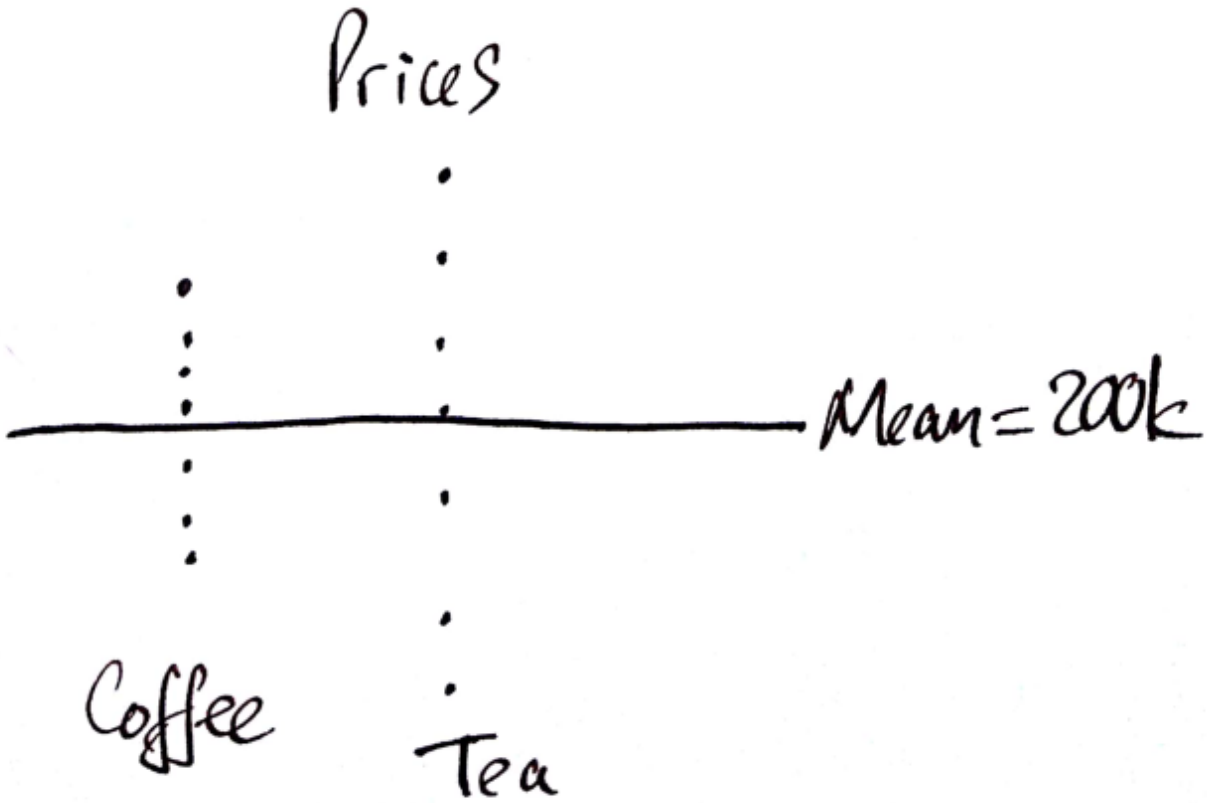
Dispersion

- Theo các bạn thì thành phần L, M, S đem đến A và B có giống nhau không ?
- 2 datasets này khác nhau chỗ nào ?

Dispersion

- Ví dụ 2: Giả sử chúng ta cân nhắc đầu tư hoặc vào trà, hoặc vào cà phê. Ta thu thập giá lịch sử của 2 mặt hàng này thì thấy như hình dưới.

Dispersion



Dispersion

- Bạn quyết định đầu tư vào trà hay cà phê ?
- Giải thích quyết định của mình ?

- Rủi ro.
- Phân loại nhà đầu tư
  1. risk-averse
  2. risk-seeking

Dispersion

- Bạn quyết định đầu tư vào trà hay cà phê ?
- Giải thích quyết định của mình ?

- Rủi ro.
- Phân loại nhà đầu tư
  1. risk-averse
  2. risk-seeking

Table of Contents