Bài 1: Tóm tắt về thống kê (Statistics)

Ôn tập khái niệm

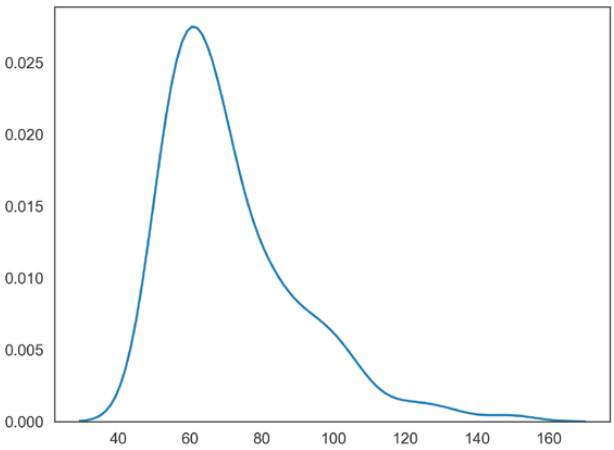
Thống kê (statistics) là công việc tổng hợp gồm nhiều việc nhỏ hơn như: **đặt câu hỏi**, **thu thập dữ liệu**, **trình bày dữ liệu, phân tích dữ liệu**, **diễn giải dữ liệu** và **suy diễn** (ra thông tin mới).

Xác suất (**probability**) là một cách đo khả năng xảy ra của một biến cố và được ước lượng bằng một con số từ 0 đến 1 (tương ứng từ 0% đến 100%).

Phân phối xác suất (**probability distribution**) là cách mô tả tất cả các khả năng xảy ra của biến cố.

Phân phối xác suất rời rạc (**discrete probability distribution**) thể hiện tất cả giá trị mà một biến ngẫu nhiên có thể có cùng với xác suất của nó.

Phân phối xác suất liên tục (**continuous probability distribution**): biểu diễn xác suất của mỗi giá trị có thể có của một biến ngẫu nhiên liên tục. Ví dụ hình bên dưới minh họa phân phối thời gian di chuyển từ chỗ làm về nhà. Trong đa số trường hợp thì mất khoảng 60 phút, nhưng thỉnh thoảng nhanh hơn vì không có kẹt xe, và thỉnh thoảng mất nhiều thời gian hơn nếu có kẹt xe.



Đánh giá dữ liệu

Một trong các cách để cảm nhận được dữ liệu là đánh giá chúng. Bạn nên làm quen với các khái niệm để đánh giá hoặc đo đạt (measure) dữ liệu như:

① Đo **sự tập trung của dữ liệu** (hoặc sự cô đặc của dữ liệu)

② Ngược lại với sự tập trung là sự phân tán của dữ liệu. Vì vậy ta cũng cần biết các khái niệm để đo **sự phân tán của dữ liệu**

Đo sự tập trung dữ liệu (Measure of Central Tendency)

Sự tập trung dữ liệu thường được đo bằng giá trị trung bình (average). Có 3 loại giá trị trung bình thường được sử dụng:

Mean: Giá trị trung bình được tính bằng tổng của các giá trị chia cho số lượng các quan sát.

Median

Median gọi là trung vị. Đây chính là giá trị của phần tử ở chính giữa một dãy giá trị có xếp theo thứ tự. Trong trường hợp dãy có số phần tử là chẵn thì trung vị được tính là trung bình của 2 phần tử ở giữa của dãy có thứ tự.

Mở rộng một chút: thay vì quan tâm đến phần tử chính giữa một tập dữ liệu, tức là phần tử mà tại đó chia đôi tập dữ liệu (có thứ tự) thì nếu bạn quan tâm đến phần tử mà tại đó chia ¼, ¾ hoặc 1 tỉ lệ bất kỳ mà bàn muốn thì xem khái niệm Quantile.

Mode

Mode là giá trị được lặp lại nhiều nhất.

Ví dụ theo dõi giá trị một cổ phiếu được giao dịch theo lô trong một ngày gồm có các mức giá tại mười thời điểm như sau: 127, 128, 128, 126, 127, 128, 129, 128, 127, 126.

Giá trị mean được tính bằng:

$$(127 + 128 + 127 + 126 + 128 + 128 + 129 + 128 + 127 + 126) / 10 = 127.4$$

Để tìm giá trị median thì ta cần sắp lại thứ tự của mười mức giá:

126, 126, 127, 127, **127, 128**, 128, 128, 128, 129

Nếu số phần tử là lẻ thì sau khi sắp thứ tự thì median sẽ là giá trị của phần tử chính giữa dãy. Tuy nhiên trong ví dụ này có 10 phần tử, nên median được tính bằng trung bình của 2 phần tử thứ 5 và 6 trong dãy đã xếp thứ tự: (127 + 128) / 2 = 127.5

Mode là giá trị 128 (được lặp lại 4 lần)

Đo sự phân tán (Dispersion)

Sự phân tán còn được gọi là tính dao động, hoặc mức độ dao động (varibility) của các giá trị.

Phương sai

Phương sai (variance) dùng để đo độ lệch, hoặc là mức độ cách biệt của các giá trị so với giá trị mean. Quay lại mười mức giá của cổ phiếu ở trên thì câu hỏi đặt ra là các giá trị dao động như thế nào? Cụ thể trong bảng bên dưới chúng ta tính độ lệch bằng cách đo khoảng cách của Giá cổ phiếu và Giá trị trung bình ở dòng 3. Vì giá trị này có thể là số âm nên nếu tính tổng các độ lệch thì sẽ không phản ánh được tổng các độ lệch của tất cả giá trị so với giá trung bình. Vì thế phải lấy bình phương các độ lệch sau đó chia cho 10. Kết quả phương sai là 0.84.

Các mức giá cổ phiếu	126	126	127	127	127	128	128	128	128	129
Giá trị mean	127.4	127.4	127.4	127.4	127.4	127.4	127.4	127.4	127.4	127.4

Khoảng cách của Giá và Mean	-1.4	-1.4	-0.4	-0.4	-0.4	0.6	0.6	0.6	0.6	1.6
Bình phương của khoảng cách	1.96	1.96	0.16	0.16	0.16	0.36	0.36	0.36	0.36	2.56
	0.84									

Tính phương sai: có 2 trường hợp

1) Trường hợp tổng thể (Population variance)

$$\text{Var(X)} = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$$

2) Trường hợp mẫu (Sample variance)

$$\text{Var(X)} = \frac{1}{N-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

Trong đó:

μ: Giá trị trung bình của tập dữ liệu.

Khi nào dùng cái nào?

- Khi bạn có toàn bộ dữ liệu (popluation) thì dùng Population variance. Ví dụ khi bạn có đầy đủ toàn bộ điểm của lớp học.
- Khi bạn chỉ có một phần (mẫu, sample) thì dùng Sample variance. Tình huống này rất phổ biến trong Machine learning, Data science, Thống kê thực tế.

Độ lệch chuẩn (standard deviation)

Độ lệch chuẩn (**standard deviation**): được tính bằng căn bậc hai của phương sai. Như vậy Độ lệch chuẩn là một giá trị mà nó có ý nghĩa cũng tương tự phương sai. Nó phản ánh mức độ chênh lệch của các giá trị quan sát so với giá trị trung bình. Độ lệch chuẩn càng nhỏ thì cho thấy dữ liệu tập trung càng dày đặc xung quanh giá trị trung bình.

Cách tính độ lêch chuẩn: Căn bậc 2 của Phương sai.

Z score

Một chỉ số liên quan đến độ lệch chuẩn là điểm chuẩn Z (Z score): để đo mức độ của một giá trị so với giá trị trung bình của tập dữ liệu

Công thức tính Z score: $Z = (X - \mu) / \sigma$

Trong đó:

X: Giá trị cần tính Z Score.

μ: Giá trị trung bình của tập dữ liệu.

σ: Độ lệch chuẩn của tập dữ liệu.

Range

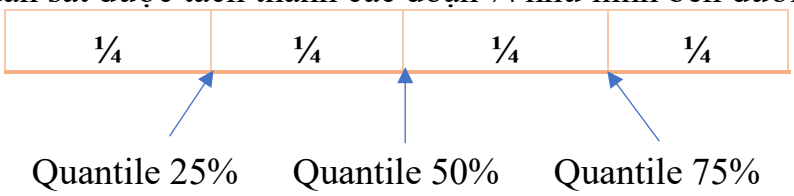
Range là giá trị khác biệt giữa phần tử lớn nhất và phần tử nhỏ nhất.

Quantile

Quantile là dạng tổng quát của Median. Tức là những giá trị (hay điểm cắt (cut points)) chia tập dữ liệu có thứ tự thành p phần có số phần tử bằng nhau. Khi đó ta có thể gọi các điểm này là p-quantiles. Median là 2-quantiles.

Một Quantile phổ biến khác thường được dùng là Tứ phân vị (4-quantiles). Tức là các điểm mà tại đó tập dữ liệu được chia làm 4 phần.

Quantile giúp chúng ta hình dung được sự phân bố, hoặc phân tán dữ liệu. Quantile có thể được xem là mở rộng khái niệm Trung vị (median). Cụ thể tìm Trung vị là tìm giá trị của phần tử để phân tách tập giá trị quan sát thành 2 nửa 50-50; 4-Quantiles sẽ cho biết các giá trị mà tại đó tập giá trị quan sát được tách thành các đoạn ¼ như hình bên dưới.



- Bách phân vị 50% chính là median (Trung vị)

Interquartile range

Interquartile range là khoảng giữa hai Bách phân vị 25% và 75%.

Correlation

Mối tương quan (Correlation): Các khái niệm đã thảo luận ở phần trước dùng để đánh giá các biến đơn lẻ (single variable). Để đánh giá mối quan hệ xác suất giữa hai hay nhiều biến thì người ta dùng khái niệm **correlation**.

Bài tập

Nếu tra Internet có thể bạn sẽ thấy thông báo tuyển dụng lập trình viên của các công ty đưa ra nhiều mức lương tháng khác nhau như: $300, $400, $1000, $1200 và $700.

Bạn có thể tính các giá trị sau:

Giá trị trung bình (Mean) = $\frac{\$300+\$400+\$1000+\$1200+\$700}{5} = \$720$

Trung vị (Median) = $700

Độ lệch chuẩn (SD) = $\sqrt[2]{\frac{(300-720)^2+(400-720)^2+(1000-720)^2+(1200-720)^2+(700-720)^2}{5}} = 343$

Range: $1200 - $300 = $900

Kiểu dữ liệu (Data Types)

Tùy theo đối tượng và thông tin chúng ta cần đo đạt, quan sát và phân tích thì có nhiều dạng thông tin khác nhau gọi Data Types.

Có thể chia Data Types thành các nhóm như:

① Các thông tin mô tả về đặt tính, đặc trưng của đối tượng như màu sắc, giới tính, v.v…gọi là kiểu dữ liệu Danh mục (Categorical data) hay còn gọi là dữ liệu **Định tính** (Qualitative data).

- Các dữ liệu dạng Danh mục không có ý nghĩa về thứ tự (vd: màu sắc, giới tính) gọi là `Nominal data.`
- Các dữ liệu về Danh mục nhưng có thêm ý nghĩa thứ tự như: các bậc học (Tiểu học, Phổ thông, Trung học, Đại học, Sau đại học) thì gọi là `Ordinal data.`

② Các thông tin mô tả về đối tượng dưới dạng con số như chiều cao, cân nặng, giá trị cổ phiếu, v.v…thì gọi là Numerical data hay còn gọi là dữ liệu **Định lượng** (Quantitative data).

- Giá trị định lượng có thể là liên tục (Continuous data) như chiều cao, cân năng.
- Giá trị định lượng có thể là rời rạc (Discrete data) như số chân của con vật.

Ví dụ mô tả con mèo hàng xóm có các thông tin sau:

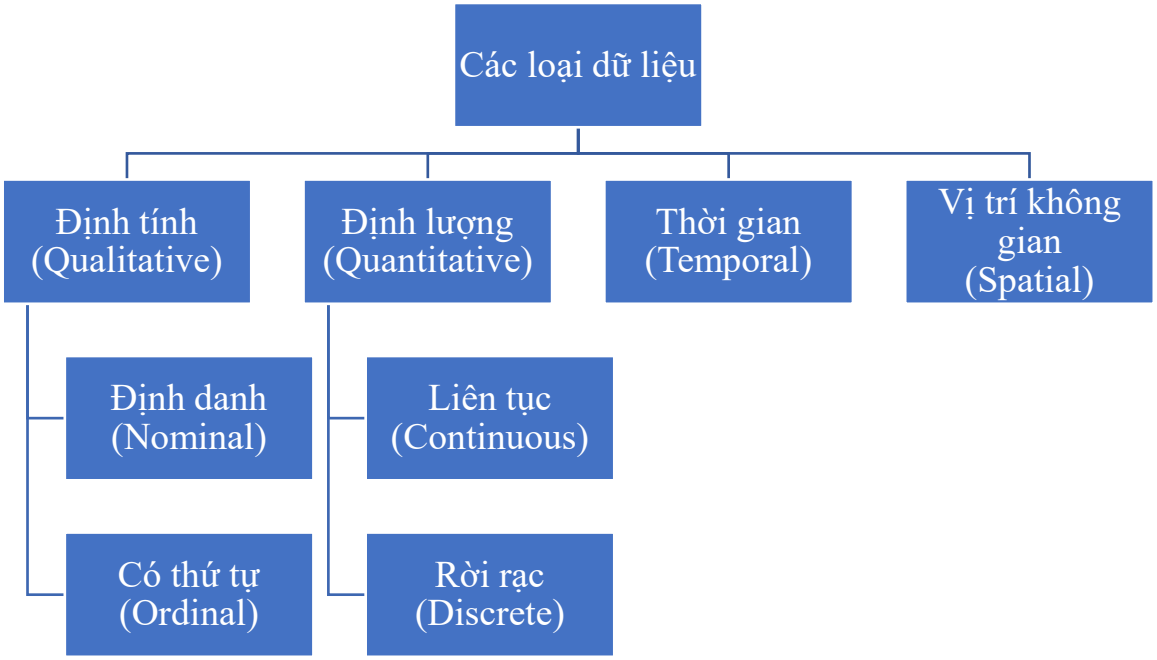
- Màu sắc: đen (Nominal data)
- Giống: cái (Nominal data)
- Nặng: 1.3 kg (Quantitative data - Continous data)
- Số chân: 4 (Quantitative data – Discrete data)

Ngoài hai dạng dữ liệu Định tính và Định lượng mà bạn thường gặp ở trên thì còn có hai loại khác như:

③ Dạng dữ liệu có đi kèm thêm yếu tố thời gian (`Temporal data`). Ví dụ giá cổ phiếu. Khi nói đến giá cổ phiếu thì phải nói thêm giá vào ngày nào. Ví dụ giá cổ phiếu của VNM ngày 10/10/2019 là 127 nghìn đồng.

④ Dạng dữ liệu liên quan đến vị trí địa lý (Spatial data). Ví dụ vị trí vật lý trên bản đồ (gồm có kinh độ và vĩ độ), hoặc đơn giản hơn là ví trí gồm x và y trong một hệ trục hai chiều.

Sơ đồ bên dưới tổng hợp các loại dữ liệu:



Hình 2: Sơ đồ các loại dữ liệu

Trên đây là các khái niệm phân loại dữ liệu ở mức trừu tượng.

Để biểu diễn dữ liệu trong máy vi tính và để cho các phần mềm có thể xử lý được dễ dàng thì bạn cần nắm các loại dữ liệu cơ bản sau:

- Số nguyên (Integer)
- Số thực (Real / Double number)
- Kí tự (Character)
- Luận lý (Logical)
- Chuỗi (String)
- Thời gian (Date, Time)
- Mảng (Array)

Dùng khái niệm / thước đo nào để quan sát dữ liệu?

Một trong các cách để có cái cảm nhận nhanh về dữ liệu là đo sự tập trung của dữ liệu (central tendency). Như phần trên đã trình bày thì có nhiều thước đo như Mean, Median, Mode.

Cần ôn lại một chút là:

- Mean để thấy giá trị trung bình
- Median để thấy trung vị
- Mode để thấy sự lặp lại của dữ liệu

Câu hỏi đặt ra là với loại dữ liệu nào thì cần dùng thước đo nào?

Bảng bên dưới sẽ gợi ý cho bạn nên dùng thước đo nào cho các kiểu dữ liệu khác nhau.

Kiểu dữ liệu	Đo sự tập trung của dữ liệu	Ghi chú
Nominal	Mode	Nominal là dữ liệu định danh không có thứ tự. Vì vậy cần biết là có bao nhiêu dữ liệu được lặp lại. Ví dụ hôm nay ra đường bạn thấy xe hơi màu nào nhiều nhất. Tức ra đường bạn sẽ thấy rất nhiều xe với nhiều màu khác nhau. Nhưng tựu trung lại ngày hôm nay bạn thấy màu nào nhiều nhất! Nên dùng Mode để tính. *Biết đâu màu xe mà bạn gặp nhiều nhất có tác động đến kết quả làm việc của ngày hôm đó*?
Ordinal	Median	Ordinal là dữ liệu danh mục có tính thứ tự. Ví dụ trong một doanh nghiệp thì có 5 cấp độ nhân viên lập trình (Dev 1, Dev 2, Dev 3, Dev 4, Dev 5) thì Median của Cấp độ lập trình viên là Dev 3.
Numerical	Mean/Median	Đối với dữ liệu số thì dễ dàng tính giá trị trung bình và trung vị. Ví dụ trong nhóm bạn học của mình thì trung bình chiều cao là bao nhiêu? Nếu đứng xếp hàng theo thứ tự chiều cao thì bạn nào sẽ bạn đứng giữa cao bao nhiêu? (nếu số người là chẵn thì lấy chiều cao trung bình của 2 bạn đứng giữa).

Biến phụ thuộc và biến tiên lượng

Phần lớn các nghiên cứu, mô hình phân tích dữ liệu phân biệt hai loại biến số:

- Biến độc lập (independent variable). Đôi khi gọi là biến tiên lượng (predictor variable), hoặc đặc trưng (feature)
- Biến phụ thuộc (dependent variable). Đôi khi gọi là outcome.

Liên lạc | Tài trợ

Lê Ngọc Thạch	
(+084) 0908 550 642	facebook.com/ThachLN
Website: https://xmyworkspace.com/learn/sponsor