**ĐẠI HỌC MỎ - ĐỊA CHẤT**

# BÁO CÁO HỌC THUẬT

## NGHIÊN CỨU VAI TRÒ CỦA THỐNG KÊ TRONG KHOA HỌC DỮ LIỆU

**Cán bộ thực hiện**
**ThS. Trương Xuân Bình**

**Hà Nội - 2023**

# MỤC LỤC

# MỞ ĐẦU

Khoa học dữ liệu (KHDL) là khoa học về việc quản trị và phân tích dữ liệu để tìm ra các hiểu biết, các tri thức hành động, các quyết định dẫn dắt hành động. KHDL gồm ba phần chính: Tạo ra và quản trị dữ liệu, phân tích dữ liệu, và chuyển kết quả phân tích thành giá trị của hành động. Nôm na bước thứ nhất là về số hóa và bước thứ hai là về dùng dữ liệu. Việc phân tích và dùng dữ liệu lại dựa vào ba nguồn tri thức: toán học (thống kê toán học), công nghệ thông tin (máy học) và tri thức của lĩnh vực ứng dụng cụ thể.

ỨNG DỤNG CỦA KHOA HỌC DỮ LIỆU

Nếu phân tích dữ liệu về nhu cầu thị trường ta có thể quyết định cần nuôi bao nhiêu lợn mỗi nơi mỗi lúc. Nếu có và phân tích được dữ liệu mô phỏng các phương án xả lũ vào mùa mưa ta có thể chọn được cách xả lũ ít thiệt hại nhất. Nếu có và phân tích được các bệnh án điện tử của người bệnh ta có thể tìm ra được phác đồ thích hợp hơn cả cho người bệnh. Amazon đã phân tích các lần mua hàng trước của bạn để dự đoán những món đồ bạn có thể sẽ thích mua và gửi quảng cáo tới, v.v. Khi nghe nói về các thành tựu đột phá gần đây của Trí tuệ nhân tạo người nghe có thể cũng chưa biết rằng phần lớn chúng đều dựa vào các phương pháp và đột phá của KHDL.

Mạng xã hội và dữ liệu người dùng


CEO của Facebook, Mark Zuckerberg thuyết trình tại hội nghị phát triển F8

Facebook, mạng xã hội lớn nhất hành tinh, một trong những cái tên được nhắc tới nhiều nhất trong giới trẻ hiện nay, là một trong những ứng dụng nổi tiếng của khoa học dữ liệu.

Tại hội nghị các nhà phát triển F8 đầu năm 2016, CEO Mark Zurkerberg cũng đã thông báo về một lộ trình mười năm tới. Trong đó, sẽ tạo ra một hệ sinh thái với những sản phẩm và công nghệ tiên tiến như trí tuệ nhân tạo (Artificial Intelligence). Tất cả đều dựa trên toàn bộ nguồn dữ liệu từ người dùng và các thuật toán máy học (Machine Learning Algorithms).

"Mỗi cú click chuột, mỗi cái like, mỗi bình luận và tất cả các kết nối đều được sử dụng để xây dựng một hồ sơ hoàn chỉnh cho mỗi người dùng."

Đằng sau những trải nghiệm kết nối và tương tác giữa bạn bè và người thân, đó là sự vận hành của các thuật toán đánh giá người dùng được xây dựng bởi những kỹ sư hàng đầu thế giới.

Tính tới tháng 8 năm 2016, tổng số lượng người dùng trên trang này cán đến mốc 1,750,000 người, gấp 5 lần dân số nước Mỹ, tương đương với 1/3 dân số thế giới và lớn hơn tổng số dân của châu Âu, châu Úc và Nam Mỹ cộng lại.

Và những nhà khoa học phân tích dữ liệu …

Theo thống kê của Glassdoor, một trong những trang web việc làm lớn nhất thế giới, ngành khoa học dữ liệu đứng đầu trong số 25 nghề nghiệp tốt nhất, đứng thứ 16 về mức lương với trung bình hơn $116,000 và có nhiều vị trí được tìm kiếm tuyển dụng nhất trong năm 2015 ở Hoa Kỳ

Trong một nghiên cứu của O'Reilly, một trong những nhà phát hành chuyên về mảng công nghệ và khoa học máy tính, có 4 dạng nhà khoa học dữ liệu tiêu biểu.

1. Doanh nhân (Data Businesspeople)

Quan tâm vào sản phẩm và phát triển lợi nhuận, họ là các nhà lãnh đạo, nhà quản lý và doanh nhân có sự am hiểu về mặt kỹ thuật. Đa phần đều có nền tảng giáo dục xuất phát bằng kỹ sư kết hợp với một MBA.

2. Nhà sáng tạo (Data Creatives)

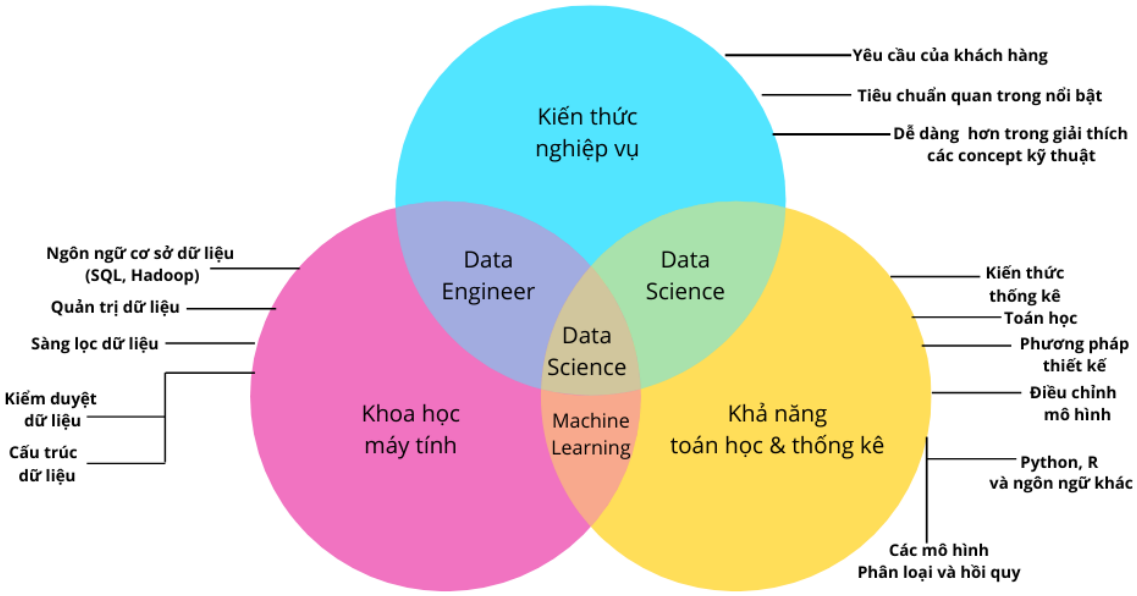
Có nhiều biệt tài và kinh nghiệm với nhiều dạng dữ liệu và công cụ, những nhà sáng tạo thường ví von mình như là một nghệ sĩ hoặc tin tặc. Điểm nhấn thường thấy là sự xuất sắc sử dụng các công nghệ minh họa (Visualization Techonology) và mã nguồn mở.

3. Nhà phát triển (Data Developers)

Nhà phát triển dữ liệu thường tập trung vào việc viết phần mềm để làm phân tích, thống kê, và nhiệm vụ học máy, thường xuyên trong môi trường sản xuất. Họ thường có trình độ khoa học máy tính, và thường xuyên làm việc với cái gọi là "dữ liệu lớn" (Big Data).

4. Nhà nghiên cứu (Data Researchers)

Đó là những người áp dụng những kỹ năng được đào tạo trong khoa học cùng với các công cụ và kỹ thuật, số liệu. Một số có bằng tiến sĩ, và các ứng dụng sáng tạo các công cụ toán học mang lại những hiểu biết và sản phẩm có giá trị.




Sơ đồ cấu trúc về Khoa học dữ liệu

Dựa trên cấu trúc này, chúng ta bắt đầu phân tích và tìm hiểu về vai trò của toán học và Thống kê , tầm quan trọng, định hướng của chúng cho hoạt động Khoa học Dữ liệu.

# I. KIẾN THỨC CƠ BẢN VỀ THỐNG KÊ CHO DATA ANALYST

## 1.1. Định nghĩa về Thống kê:

Khi nói về Statistics thì điều gì sẽ xuất hiện trong suy nghĩ của bạn đầu tiên?


Đối với hầu hết mọi người, nó nói đến các sự kiện, dữ liệu số học như số liệu về tỉ lệ thất nghiệp, giá nông sản, số lượng các kết hôn và ly hôn, ... Dưới đây là 2 định nghĩa phổ biến của Thống Kê:

Thống kê là những dữ kiện hay dữ liệu, xuất hiện dưới dạng số(numerical) hoặc không phải dạng số (nonnumerical), được tổ chức và tóm tắt, để cung cấp thông tin hữu ích và dễ tiếp cận cho một chủ đề cụ thế nào đó.

Thống kê là khoa học về tổ chức và tóm tắt thông tin dạng số hoặc không phải dạng số.

Nhiệm vụ của các nhà thống kê (statistician) là phân tích dữ liệu nhằm mục đích tổng quát hóa (generalization) và đưa ra các kết luận.

Ví dụ :Một nhà phân tích chính trị (political analyst) có thể sử dụng một phần dữ liệu từ bỏ phiếu của người dân để dự đoán xem ai là người có khả năng đắt cử mà không cần phải có toàn bộ dữ liệu.

Theo nghiên cứu thì các bạn đừng nên dịch từ Generalization ra tiếng việt, bạn có thể hiểu nó là hành vi đưa ra một nhận định, kết luận cho đại diện cho toàn bộ tập dữ liệu (population) từ những dữ liệu mẫu được thu thập (sample), vì trong nhiều trường hợp thu thập toàn bộ dữ liệu cần thiết là việc không khả thi và tốn kém.

Nếu phải bắt buộc dịch chúng tôi sẽ dùng từ Khái quát hóa.

## 1.2. Các loại Thống kê:

Statistics được chia làm 2 loại chính: Descriptive Statistics (Thống kê mô tả) và Inferential Statistics (Thống kê suy luận)

Thống kê mô tả bao gồm các phương pháp tổ chức và tóm tắt thông tin.

Thống kê mô tả bao gồm việc xây dựng đồ thị, bảng số liệu và tính toán các chỉ số mô tả (descriptive measures) như: trung bình (mean), độ biến thiên (variation), bách phân vị (percentiles).

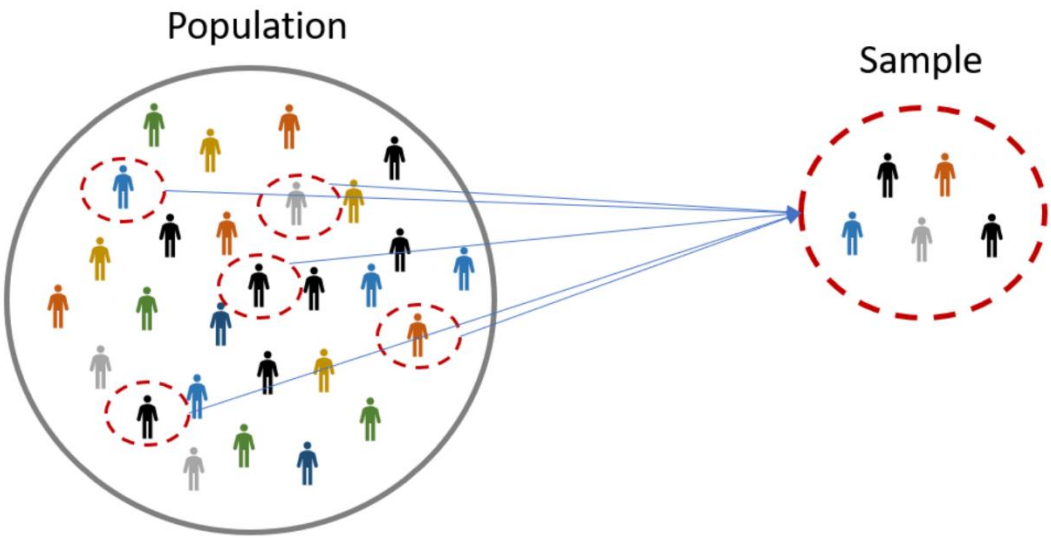
Trước khi tìm hiểu về Inferential Statistics thì chúng ta sẽ đi qua khái niệm về Population và Sample, Census trước, có thể các bạn đã nghe 3 từ này ở đâu đó rồi.

Population : Tập hợp tất cả các cá nhân hoặc đối tượng đang được xem xét trong một nghiên cứu thống kê.

Mẫu(sample) : Phần Population mà thông tin được lấy từ đó.

Population là toàn bộ phần tử hay đối  tượng được cho là có mặt trong nghiên cứu của bạn không phải dịch ra là Population đâu  , trong khi đó Sample chỉ là một phần trong đó, chúng tôi hay gọi là một tập con (subset).

Ví dụ sau đây sẽ giúp bạn hiểu rõ 2 khái niệm này




Population và mẫu

Giả sử bạn đang thực hiện nghiên tiền bầu cử, thì việc phỏng vấn toàn bộ người trong độ tuổi đi bầu (Population) là bất khả thi, mất rất nhiều thời gian và rất đắt đỏ.

Nên thay vào đó chúng ta sẽ chọn ra khoảng vài nghìn cử tri (Sample) để thực hiện lấy ý kiến .

Từ tập dữ liệu Sample này các nhà thống kê sẽ thực hiện Thống kê suy luận để đưa ra kết luận đại diện cho toàn bộ người trong độ tuổi đi bầu hay Population đấy các bạn. Thống kê suy luận bao gồm các phương pháp rút ra và đo lường độ tin cậy của các kết luận về tổng thể dựa trên thông tin thu được từ một mẫu của tổng thể.

Thống kê mô tả bao gồm các phương pháp nhằm rút ra và đo lường mức độ tin cậy của các kết luận về toàn bộ tập dữ liệu Population dựa trên tập dữ liệu mẫu Sample thu được.Thống kê mô tả và thống kê suy luận có mối liên hệ chặt chẽ với nhau.

Bạn gần như phải luôn luôn sử dụng các kỹ thuật của thống kê mô tả để tổ chức, tóm tắt các thông tin thu được từ tập Sample trước khi thực hiện thống kê suy luận.

Hơn nữa nhờ thống kê mô tả cho bạn thông tin về đặc điểm, tính chất của Sample, giúp bạn hiểu rõ dữ liệu hơn, dẫn đến việc lựa chọn các phương pháp phân tích suy luận phù hợp.

Nó giống như việc thăm khám, chẩn đoán bệnh xong thì mới cho thuốc được vậy.

### 1.3. Phân loại các nghiên cứu Thống kê:

Nếu như mục đích của nghiên cứu là kiểm tra và khám phá thông tin, những đặc điểm đặc biệt, thông tin hữu ích có trong bản thân dữ liệu thôi thì đây là nghiên cứu thiên hướng mô tả (Descriptive Statistics).

Các Reports và Dashboards mà các bạn Data Analyst xây dựng cho công ty là một ví dụ.

Đây là một ví dụ cho Thống kê mô tả vì nó chỉ là một bảng tóm tắt kết quả mà thôi, không có bất kì suy luận nào.

| Ticket | Votes | Percentage |
|---|---|---|
| Truman–Barkley (Democratic) | 24,179,345 | 49.7 |
| Dewey–Warren (Republican) | 21,991,291 | 45.2 |
| Thurmond–Wright (States Rights) | 1,176,125 | 2.4 |
| Wallace–Taylor (Progressive) | 1,157,326 | 2.4 |
| Thomas–Smith (Socialist) | 139,572 | 0.3 |

Kết quả bỏ phiếu tổng thống mỹ năm 1948

Tuy nhiên nếu như dữ liệu thu thập được là một Sample thuộc Population nào đấy, sau đó sử dụng chúng để đưa ra kết luận cho Population thì nó là Thống kê suy luận.

Một nhóm nghiên cứu đã tiến hành thí nghiệm về tốc độ giữa báo và sư tử bằng cách đo thời gian hoàn thành 1 km của 20 con báo và 20 con sư tử, kết luận cho thấy báo chạy nhanh gấp đôi sư tử.

Đây là một nghiên cứu thống kê suy luận vì thứ nhất chúng ta dùng tập mẫu 20 con báo và 20 con sư tử (Sample) để đưa ra kết luận về toàn bộ cá thể báo, sư tử (Population), rõ ràng việc lấy mẫu toàn bộ sư tử và báo là bất khả thi, thứ hai sau khi thí nghiệm chúng ta đã đưa ra kết luận mang tính suy luận về tốc độ của 2 loài này.

### 1. 4. Quá trình phát triển của Thống kê:

Theo lịch sử, thống kê mô tả có trước thống kê suy luận, điều tra Population đã có trước đây rất lâu từ thời La Mã, qua nhiều thế kỉ những ghi chép về các số liệu sinh, tử, hôn nhân, thuế suất dẫn đến sự phát triển tự nhiên của thống kê mô tả.

Trong khi đó thống kê suy luận chỉ mới phát triển gần đây, bước tiến lớn nhất bắt đầu từ nghiên cứu của Karl Pearson (1857–1936) và Ronald Fisher (1890–1962) đã xuất bản những phát hiện của họ vào những năm đầu thế kỉ XX. Sau đó thống kê suy luận đã được áp dụng trong hầu hết các lĩnh vực của cuộc sống.

Các hiểu biết về thống kê sẽ giúp bạn nhận định xem những điều bạn đọc trên báo chí hay Internet có chính xác hay không.

Như ví dụ ở trên nhóm nghiên cứu sinh thực nghiệm trên 40 sư tử và báo để đưa ra kết luận về vài chục ngàn cá thể báo, sư tử trên thế giới có chính xác hay không ?

### 1.5. Nghiên cứu quan sát vs Thí nghiệm thiết kế:

Bên cạnh việc phân loại các nghiên cứu thuộc thống kê mô tả hay suy luận chúng ta còn phải phân biệt chúng là Nghiên cứu quan sát hay Thí nghiệm được thiết kế.

Trong nghiên cứu quan sát các nhà nghiên cứu chỉ cần quan sát các đặc điểm và ghi nhận số liệu đo lường của tập Sample. Trong Designed Experiment các nhà nghiên cứu sẽ thực hiện các liệu pháp (treatment) và kiểm soát đối tượng thí nghiệm (có tác động lên tập mẫu đó các bạn) sau đó mới ghi lại các đặc điểm và số liệu đo lường.

Observational Study chỉ cho chúng ta thấy được mối liên kết (association) trong khi Designed Experiment lại cho chúng ta thấy được mối quan hệ nguyên nhân, kết quả, chúng tôi sẽ đưa ra một vài ví dụ cho bạn dễ hiểu hơn

Observational Study: Các nhà nghiên cứu muốn xem xét giả thuyết học thêm giờ buổi tối có khiến trẻ có thành tích tốt hơn hay không ? Họ đã lấy ý kiến trên 100 trẻ và cho kết quả là ... Đây là nghiên cứu quan sát vì người thực hiện không tác động vào đối tượng nghiên cứu

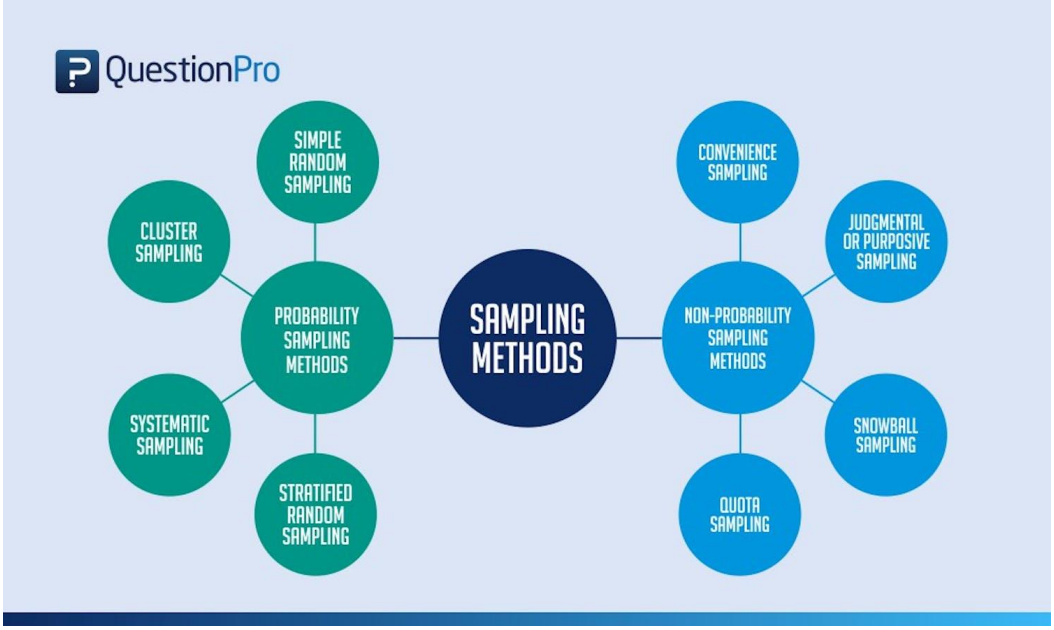
Designed Experiment: Các nhà nghiên cứu muốn kiểm tra một giả thiết là giảm độ sáng màn hình giúp tăng tuổi thọ của pin, họ đã tập hợp 30 cái laptop lại là, 15 trong số đó hạ độ sáng màn hình xuống mức 3, 15 cái kia giữ ở cài đặt mặc định, ... Rõ ràng họ đã tác động vào vật thể quan sát và điều chỉnh nó (tăng giảm độ sáng màn hình) nên đây là một nghiên cứu được thiết kế.

Tham khảo: Introductory Statistics (2017) - Pearson

## II. KỸ THUẬT LẤY MẪU TRONG THỐNG KÊ

### 2.1. Phân biệt Census, Sampling và Experimentation

Nếu bạn đang muốn thực hiện một nghiên cứu Thống kê và dữ liệu bạn cần không có sẵn (trong ngân hàng dữ liệu, các nghiên cứu trước đó, ...) thì bạn phải thực hiện lấy dữ liệu toàn diện hay còn gọi là Census. Tuy nhiên như ở bài viết trước đó chúng tôi đã có nói việc ghi nhận dữ liệu của toàn bộ Population là không khả thi, tốn nhiều tiền bạc và thời gian.




Nếu các bạn chưa rõ Sample và Population là gì có thể tham khảo nội dung của phần 1.

Điều tra Population — nghĩa là, bằng cách thu thập thông tin cho toàn bộ Population quan tâm

Thay vào đó 2 phương pháp khác là Sampling (lấy mẫu) và Experimentation (thí nghiệm) được sử dụng nhiều hơn vì tính khả thi của chúng.

Nếu bạn thấy phương pháp lấy mẫu phù hợp với bài nghiên cứu thì bước kế tiếp là chọn loại lấy mẫu. Bởi vì tập Sample sẽ được dùng để suy ra những kết luận cho tập lớn Population, vì thế chúng nên là những Sample có tính đại diên (representative) cho số đông.

Ví dụ: lấy cân nặng trung bình của các cầu thủ bóng đá chuyên nghiệp làm giá trị đại diện cho cân nặng trung bình của tất cả nam giới trường thành sẽ không phù hợp. Cũng tương tự như vậy lấy thu nhập trung bình đầu người của TP Hồ Chí chúng tôi đại diện cho GDP cả nước sẽ không thỏa đáng.

Để thấy được mức độ ảnh hưởng của việc chọn mẫu (Sampling) không có tính đại diện cho số đông, ta cùng xem một ví dụ thú vị sau đây. Trước cuộc bầu cử tổng thống Mỹ năm 1936, Tạp chí Literacy Digest đã tiến hành thăm do quần chúng, nhóm thực hiện khảo sát đã hỏi một tập mẫu (Sample) người dân xem liệu họ sẽ bỏ phiếu cho Franklin D. Roosevelt ứng viên đảng Dân chủ hay Alfred Landon ứng viên đảng cộng hòa.

Dự trên kết quả khảo sát, tạp chí dự đoán Landon sẽ giành chiến thắng. Tuy nhiên khi bầu cữ diễn ra, kết quả cho thấy Roosevelt đã thắng lợi áp đảo với hơn 60% phiếu bầu. Chuyện gì đã xảy ra thế ?

Sample là những người có sở hữu ô tô hoặc điện thoại, vào thời điểm năm 1936 họ được liệt vào nhóm có điều kiện và những người này thì có xu hướng ủng hộ đảng cộng hòa

Tỉ lệ phản hồi thấp (chỉ có 25% người được thăm dò trả lời), điều này dẫn đến sự thiên vị (bias) trong kết quả (đa số những người trả lời có xu hướng ủng hộ Landon)

Hầu hết các thủ tục lấy mẫu hiện đại đều áp dụng Probability Sampling - Lấy mẫu ngẫu nhiên. Với phương pháp này, người ta sẽ sử dụng các công cụ ngẫu nhiên như: tung đồng xu, tham khảo bảng số ngẫu nhiên (random table), hoặc công cụ chọn số ngẫu nhiên (random number generator - https://www.random.org/) giúp họ chọn ra một phần tử ngẫu nhiên đưa vào tập Sample thay vì để con người quyết định một cách cảm tính.

Nếu bạn đang thắc mắc là nếu lỡ xui chúng tôi chọn ngẫu nhiên trúng tập Sample không mang tính đại diện thì sao ? Thì bạn thắc mắc đúng rồi đấy, nó không giúp bạn loại bỏ hết được nguy cơ nhưng sẽ giúp bạn hạn chế phần nào, trong phần sắp tới chúng tôi sẽ giới thiệu với bạn các phương pháp lấy mẫu ngẫu nhiên phổ biến.

## 2.2. Lấy mẫu ngẫu nhiên đơn giản

Bạn có thể hiểu đơn giản là mẫu được chọn một cách ngẫu nhiên, bất kì phần tử nào cũng có xác suất được chọn như nhau, giống như việc bạn tung đồng xu lên thì xác suất mặt ngửa là 50% và mặt sấp cũng là 50% => xác suất bằng nhau.

***Lấy mẫu ngẫu nhiên đơn giản : Một quy trình lấy mẫu mà mỗi mẫu có thể có của một kích thước nhất định đều có khả năng là mẫu thu được như nhau.***

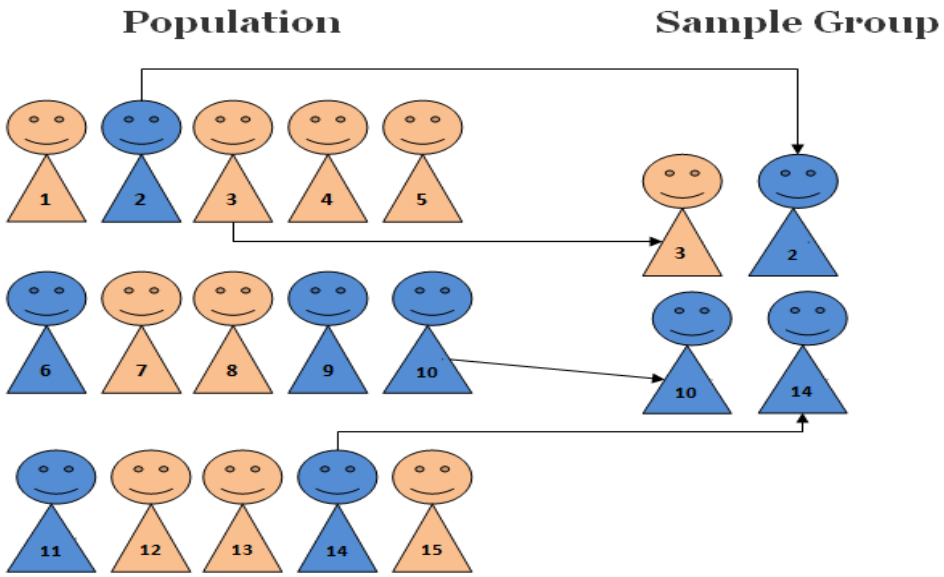
***Mẫu ngẫu nhiên đơn giản : Một mẫu thu được bằng cách lấy mẫu ngẫu nhiên đơn giản***.

Có 2 loại Simple Random Sampling:

Simple random sampling with replacement (SRSWR): một item có thể được chọn nhiều lần vì chọn xong rồi chúng tôi lại bỏ vào chỗ cũ (replace)

Simple random sampling without replacement (SRS): một item được chọn nhiều nhất là 1 lần, chọn xong thì lấy hẳn ra luôn.

Nếu như chúng ta nói là dùng phương pháp Simple Random Sampling và không mô tả thêm gì, bạn có thể hiểu nó là SRS .




### 2.3. Lấy mẫu ngẫu nhiên có hệ thống

Simple random sampling là phương pháp lấy mẫu đơn giản và dễ hiểu, cứ chọn ngẫu nhiên thôi, tuy nhiên trong nhiều trường hợp đặc thù thì phương pháp này sẽ bộc lộ điểm yếu, ví dụ khi một tập con trong dữ liệu bắt buộc xuất hiện hay dữ liệu có tính phân tán về mặt địa lý. Ví dụ như khi bạn chọn đáp viên từ một Population đến từ 3 tỉnh thành để làm khảo sát, để có tính khách quan bạn được yêu cầu phải chọn sao cho tất cả 3 tỉnh thành đều có người được chọn. Từ đó, các nhà nghiên cứu cho ra đời phương pháp chọn mẫu hệ thống

| bước | Sự miêu tả |
|---|---|
| Bước 1 | Chia kích thước dân số cho kích thước mẫu và làm tròn kết quả xuống số nguyên gần nhất, m. |
| Bước 2 | Sử dụng một bảng số ngẫu nhiên hoặc một thiết bị tương tự để có được một số, k, giữa 1 và m. |
| Bước 3 | Chọn vào mẫu những thành viên của tổng thể được đánh số k, k + m, k + 2m, . . . . |

Bây giờ chúng ta sẽ cùng đi vào một ví dụ cho dễ hiểu hơn . Giáo sư Professor Hassett muốn chọn ra 15 học sinh trong tổng số 728 sinh viên của chúng tôi sử dụng phương pháp lấy mẫu hệ thống.

Bước 1: Population size = 728, sample size = 15, m = 728/15 = 48 (đã làm tròn)

Bước 2: Chọn một số bất kì ngẫu nhiên k từ 1->m tức là 1->48 ấy, bạn cứ chọn số nào cũng được , ví dụ chọn k = 22

Bước 3: Sau khi đã có k và m rồi bạn cứ áp dụng công thức lấy các số thức k, k+m, k+2m, ... chúng ta sẽ có bảng sau:

| | | | | |
|---|---|---|---|---|
| 22 | 166 | 310 | 454 | 598 |
| 70 | 214 | 358 | 502 | 646 |
| 118 | 262 | 406 | 550 | 694 |

Các sinh viên có số thứ tự trong bảng trên sẽ được chọn tham gia khảo sát. Tóm gọn lại phương pháp này sẽ cắt sinh viên thành 48 phần và mỗi phần chúng tôi chọn 1 người ra để tăng độ phủ (coverage), thay vì nếu chọn theo phương pháp đơn giản, các phần tử được chọn sẽ không may tụ họp chung một chỗ, mà ở quá gần nhau thì có khả năng là cùng đặc tính, cùng sở thích ... giống như việc bạn ngồi gần bạn bè trên giảng đường vậy.

## 2.4. Lấy mẫu cụm

Phương pháp lấy mẫu theo cụm đặc biệt hữu ích khi dữ liệu có tính phân tán về mặt địa lý.

| bước | Sự miêu tả |
|---|---|
| Bước 1 | Chia dân số thành các nhóm (cụm). |
| Bước 2 | Lấy một mẫu ngẫu nhiên đơn giản của các cụm. |
| Bước 3 | Sử dụng tất cả các thành viên của các cụm thu được trong Bước 2 làm mẫu. |

Ví dụ bạn chia nhóm người khảo sát ra theo tỉnh họ sinh sống (cluster), sau đó mỗi tỉnh chọn ra ngẫu nhiên 2 người để tham gia khảo sát.

Hình bên dưới là ví dụ của việc lấy khảo sát của người dân theo từng cụm bàn cờ, bạn sẽ nhận ra được việc chia nhóm theo cluster giúp bạn có cái nhìn khái quát hơn, thay vì cứ chọn ngẫu nhiên mà vô tình 90% lại rớt hết vào khu nhà giàu trung lưu hay khu thu nhập thấp, dẫn đến cách đánh giá sai lệch.


## 2.5. Lấy mẫu phân tầng

Lấy mẫu phân tổ - Statified Sampling là phương pháp lấy mẫu đáng tin cậy hơn cluster sampling, tập Population sẽ được chia thành các nhóm (tổ) gọi là starta (số nhiều), sau đó với mỗi stratum (số ít) chúng tôi sẽ lấy mẫu ngẫu nhiên ở trên đó. Tuy nhiên hơi khác so với cluster sampling thì số lượng sample lấy ở mỗi stratum sẽ phụ thuộc vào độ lớn của nó, hay gọi là propotional allocation, giống như lương nhiều thì đóng thuế nhiều vậy.

| bước | Sự miêu tả |
|---|---|
| Bước 1 | Chia dân số thành các tiểu quần thể (tầng). |
| Bước 2 | Từ mỗi tầng, lấy một mẫu ngẫu nhiên đơn giản có kích thước tỷ lệ thuận với kích thước của tầng; nghĩa là, cỡ mẫu cho một tầng bằng tổng cỡ mẫu nhân với cỡ tầng chia cho quy mô dân số. |
| Bước 3 | Sử dụng tất cả các thành viên thu được trong Bước 2 làm mẫu. |

Bây giờ chúng ta sẽ cùng đi vào một ví dụ cụ thể luôn . Giả sử chúng ta có một tập population lớn 2000 người được chia thành các strata có độ lớn như sau: 400, 600, 800, 200. Hãy sử dụng phương pháp lấy mẫu phân tổ để lấy Sample 10 người.

Bước 1: Chia nhỏ nhóm theo các đặt điểm như mức thu nhập, độ tuổi, ... Thật may mắn chúng ta đã được đề bài chia sẵn rồi nên bước này bỏ qua .

Bước 2: Xác định mỗi stratum chúng ta sẽ lấy bao nhiêu item theo công thức sau:

sample_size x (stratum_size/population_size) = 10*(400/2000) = 2

Sau đó bạn sẽ chọn ngẫu nhiên một số bất kì stratum, bạn có thể dùng radom.org để random s. Chi tiết ở bảng bên dưới .

| Stratum | Size | Numbered | Sample size | Sample |
|---|---|---|---|---|
| #1 | 400 | 1–400 | 2 | 166, 264 |
| #2 | 600 | 401–1000 | 3 | 454, 511, 620 |
| #3 | 800 | 1001–1800 | 4 | 1246, 1420, 1759, 1793 |
| #4 | 200 | 1801–2000 | 1 | 1938 |

Tổng hợp các số được chọn làm Sample ở bước 2 thôi

## 2.6. Lấy mẫu nhiều tầng

Đây là phương pháp áp dụng việc sử dụng nhiều phương pháp lấy mẫu: ngẫu nhiên, phân tổ, phân cụm, ... thường được sử dụng trong nhiều nhiên cứu có quy mô lớn. chúng tôi sẽ chia sẽ chi tiếc lựa chọn cách lấy mẫu nào cho phù hợp trong các bài viết tiếp theo

## III. CÁC KHÁI NIỆM VỀ THỐNG KÊ MÔ TẢ

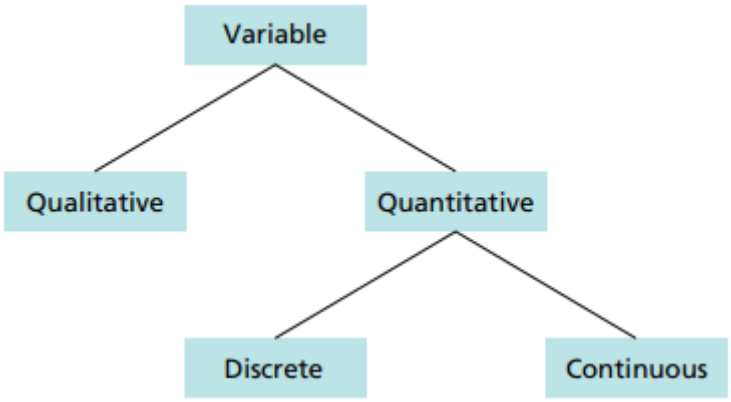
### 3.1. Variables & Data

Trước khi đi vào mô tả dữ liệu thì chúng ta cần nắm rõ một vài khái niệm cơ bản trước.

Variable - Biến số: một đặc điểm thay đổi từ người này sang người khác hay vật nay sang vật khác, ví dụ như chiều cao, cân nặng, số lượng cặp sinh đôi, giới tính, tình trạng hôn nhân và màu mắt. Từ biến của nó chính là biến đổi đó các bạn, gọi tắc là biến giống như hàm số gọi ngắn gọn là hàm. Chúng ta có các loại biến sau:

Qualitative - Biến định tính: Biến số mà giá trị của nó ko ở dạng số như tình trạng hôn nhân Có hoặc Không, Giới tính là Nam và Nữ, đây không phải là dữ liệu dạng số. Một tên gọi khác của nó là Categorical Variable.

Quantitative - Biến định lượng: Biến số mà giá trị nó ở dạng số. Nhưng trong nhánh này sẽ chia làm 2 dạng là Discrete (Rời rạc) và Continous (Liên tục), biến rời rạc là biến có giá trị nguyên và số lượng của chúng có thể đếm được ví dụ như Số lượng cặp sinh đôi, số lượng trẻ em cả nước, trong khi đó biến liên tục tồn tại dưới dạng khoảng, và giá trị của biến có thể là bất kì giá trị nào trong khoảng đó ví dụ chiều cao của người Việt trong khoảng từ 50->2200m.




Variable

Các giá trị của biến bất kì gọi là Data, một giá trị nằm trong Data gọi là Observation (chiều cao của Linh là 1m7). Tập hợp dữ liệu của nhiều biến được gọi là Dataset. Tương tự với biến chúng ta sẽ có qualitative data, quantitative data, discrete data, và continuous data.

Tại sao bạn cần phải xác định loại dữ liệu ? xác định được loại dữ liệu cho phép bạn lựa chọn đường phương pháp thống kê phù hợp, không phải tất cả phương pháp đều phù hợp với mọi loại dữ liệu, nên việc xác định đúng sẽ giúp bạn chuẩn bị tốt nhất cho những bước tiếp theo.

Bây giờ chúng ta sẽ cùng đi vào các bước tiếp theo, chúng tôi sẽ chia sẽ loại dữ liệu và mô tả các xử lý chúng.

## 3.2. Tổ chức Dữ liệu định tính

Việc đầu tiên bạn làm trong phần mô tả này là tổ chức chúng thành các bảng, biểu đồ hoặc đồ thị để nắm được những ý chính của dữ liệu, nói cho dễ hiểu là bạn tóm tắt nó lại cho dễ hiểu đó.

> **Frequency Distribution of Qualitative Data**
> A frequency distribution of qualitative data is a listing of the distinct values
> and their frequencies.

### *3.2.1 Frequency Table*

Bước một là bắt đầu với bảng tần suất - Frequency Table, chúng tôi xin phép dùng tiếng Anh luôn . Bảng này chỉ dùng được với dữ liệu định tính thôi , cho biết tần suất của mỗi giá trị của mỗi cột dữ liệu, ví dụ trong cột giới tính thì Nam xuất hiện bao nhiêu lần, và tương tự với Nữ, khá là dễ hiểu nha.

Bước 1: Lấy ra danh sách unique values - giá trị không bị trùng của tập dữ liệu, ở ví dụ bên dưới bạn có thể thấy chúng ta có 3 unique party là Democratic, Republican và Other, tương tự với giới tính thì chỉ có 2 unique values là Nam và Nữ

Bước 2: Đếm số lần xuất hiện của mỗi unique values này

| | | | | |
|---|---|---|---|---|
| Democratic | Other | Democratic | Other | Democratic |
| Republican | Republican | Other | Other | Republican |
| Republican | Republican | Republican | Democratic | Republican |
| Republican | Democratic | Democratic | Other | Republican |
| Democratic | Democratic | Republican | Democratic | Democratic |
| Republican | Republican | Other | Other | Democratic |
| Republican | Democratic | Republican | Other | Other |
| Republican | Republican | Republican | Democratic | Republican |

Dữ liệu gốc

| Party | Tally | Frequency |
|---|---|---|
| Democratic | 𝍸 𝍸 III | 13 |
| Republican | 𝍸 𝍸 𝍸 III | 18 |
| Other | 𝍸 IIII | 9 |
| | | 40 |

Frequency Table

### *3.2.2 Relative-Frequency Distributions*

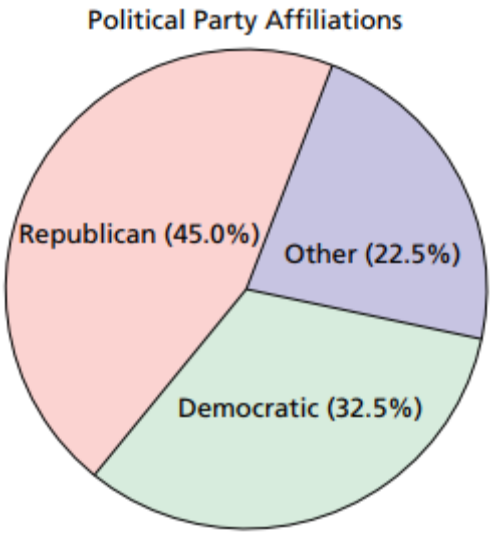
Tương tự như Frequency Table, Relative frequency sẽ hiển thị phần trăm trên tổng số thay vì là số lần xuất hiện. Bạn chỉ cần lấy frequency chia cho tổng số dòng dữ liệu có trong bảng.

| Party | Relative frequency | |
|---|---|---|
| Democratic | 0.325 | ← *13/40* |
| Republican | 0.450 | ← *18/40* |
| Other | 0.225 | ← *9/40* |
| | 1.000 | |

Relative frequency Table

### *3.2.3 Pie Charts*

Một phương pháp khác để tóm tắt dữ liệu là vẻ biểu đồ, một trong số đó là biểu đồ hình tròn. Thông thường chúng tôi sẽ vẻ chúng khi muốn thuyết trình hay trình bày với sếp, thay vì nhìn vào Frequency Table thì ta có thể chọn cách vẻ biểu đồ để người đối diện nắm bắt được thông tin nhanh và trực quan hơn.

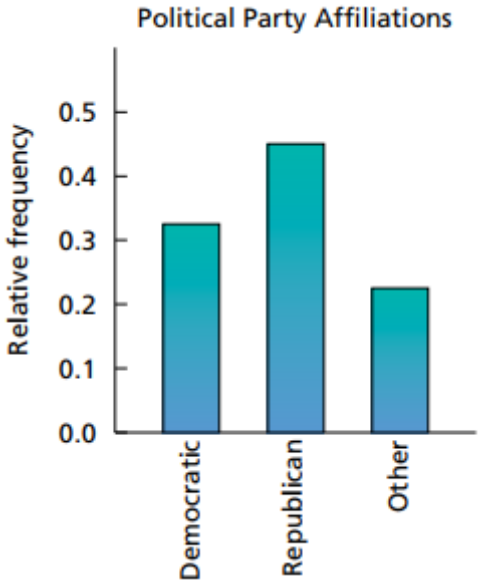



Pie Chart

Bạn có thể thấy dữ liệu của Pie chart giống relative frequency table đúng ko ạ, chính là nó đấy, Theo nghiên cứu thì Chart ở trên còn thiếu con số frequency [45% (13)] là hoàn hảo, vừa có số đếm vừa có phần trăm.

### *3.2.4 Bar Charts*

Bên dưới là biểu đồ cột thể hiện cùng nội dung với biểu đồ hình tròn bên trên. Cá nhân chúng tôi thấy chúng tôi thường dùng Bar chart trong trường hợp muốn sắp xếp thứ tự trên trục x, ví dụ trục x sẽ là các thứ trong tuần từ thứ hai đến chủ nhật, thay vì dùng pie chart phải xem thuận hay ngược kim đồng hồ sẽ khá khó khăn thì bar chart rất rõ ràng trực quang.




## 3.3. Tổ chức Dữ liệu định lượng

Đối với dữ liệu định lượng trước tiên chúng ta sẽ nhóm dữ liệu lại thành các class - nhóm hay lớp (còn được gọi là categories hay là bins) và sau đó làm việc với chúng như dữ liệu định tính. Có 3 nguyên tắc chung để bạn nhóm dữ liệu định lượng thành class:

Số lượng class phải không quá nhiều cũng không quá ít, quá ít sẽ dẫn tới việc bỏ qua các đặc điểm khác biệt giữa các class, quá nhiều thì lại làm cho chúng tôi khó quan sát và đôi khi không thấy được tính tương đồng.

Mỗi điểm dữ liệu (observation) phải thuộc một class duy nhất.

Trong trường hợp khả thi, tất cả các lớp nên có cùng số phần tử, trong trường hợp lý tưởng bạn.

Một số phương pháp được sử dụng để nhóm dữ liệu thành các class: single-value grouping, limit grouping, và cutpoint grouping.

### *3.3.1 Phương pháp phân lớp*

*3.3.1.1 Single-Value Grouping*

Nghe tên chắc bạn đã đoán ra được cách phân class này rồi đúng hông, phương pháp này xem mỗi điểm dữ liệu là một class, nó chỉ thích hợp với tập dữ liệu có ít unique value (giá trị không trùng lặp) và ở dạng rời rạc (discrete) không phải continous (liên tục). Ví dụ như điểm số từ 0 đến 10, chỉ có 10 unique value thôi, khá ít và dễ đếm.

| Number of TVs | Frequency | Relative frequency |
|---|---|---|
| 0 | 1 | 0.02 |
| 1 | 16 | 0.32 |
| 2 | 14 | 0.28 |
| 3 | 12 | 0.24 |
| 4 | 3 | 0.06 |
| 5 | 2 | 0.04 |
| 6 | 2 | 0.04 |
| | 50 | 1.00 |

Số lượng TV của hộ gia đình

*3.3.1.2 Limit Grouping*

Trong trường hợp dữ liệu quá nhiều thì chúng ta sẽ không sử dụng phương pháp Single Point được, vừa không thể thống kê được chúng, bạn có thể tưởng tượng bản tần suất trên dài vài chục trang thì không thể nói là bạn đang summarize dữ liệu được, chúng ta sẽ dùng phương pháp Limit grouping, tạo ra các khoản dữ liệu để phân lớp. Chỉ dùng cho dữ liệu dạng rời rạc, có thể đếm được

| Days to maturity | Tally | Frequency | Relative frequency |
|---|---|---|---|
| 30–39 | III | 3 | 0.075 |
| 40–49 | I | 1 | 0.025 |
| 50–59 | 𝍸 III | 8 | 0.200 |
| 60–69 | 𝍸 𝍸 | 10 | 0.250 |
| 70–79 | 𝍸 II | 7 | 0.175 |
| 80–89 | 𝍸 II | 7 | 0.175 |
| 90–99 | IIII | 4 | 0.100 |
| | | 40 | 1.000 |

Ngày đáo hạn của các khoản vay ngắn hạn

*3. 3.1.3 Cutpoint Grouping*

Tương tự với Limit grouping nhưng sử dụng với dữ liệu dạng liên tục - continous.

| Weight (lb) | Frequency | Relative frequency |
|---|---|---|
| 120–under 140 | 3 | 0.081 |
| 140–under 160 | 9 | 0.243 |
| 160–under 180 | 14 | 0.378 |
| 180–under 200 | 7 | 0.189 |
| 200–under 220 | 3 | 0.081 |
| 220–under 240 | 0 | 0.000 |
| 240–under 260 | 0 | 0.000 |
| 260–under 280 | 1 | 0.027 |
| | 37 | 0.999 |

Thống kê cân nặng

Chúng tôi xin tổng hợp lại phần lựa chọn phương pháp phân class như sau

| Phương pháp | Khi nào sử dụng |
|---|---|
| Single-value grouping | Dữ liệu rời rạc, ít unique value |
| Limit grouping | Dữ liệu rời rạc,ở dạng số nguyên, nhiều unique value |
| Cutpoint grouping | Dữ liệu liên tục |

### *3.3.2 Biểu đồ*

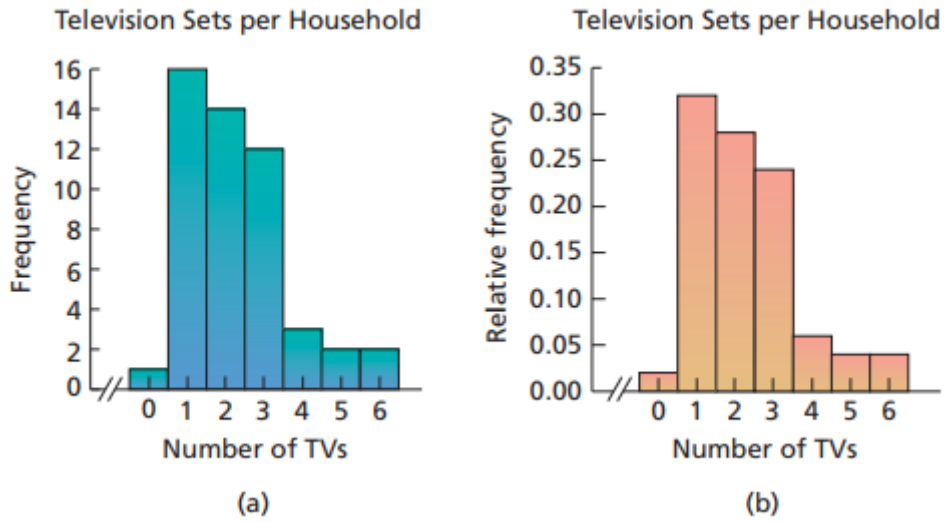
Cũng giống như dữ liệu định tính, chúng ta có thể biểu diễn dữ liệu định lượng dưới dạng biểu đồ. Có 3 phương pháp phổ biến là histograms, dotplots, và stem-and-leaf

#### *3.3.2.1 Histogram*

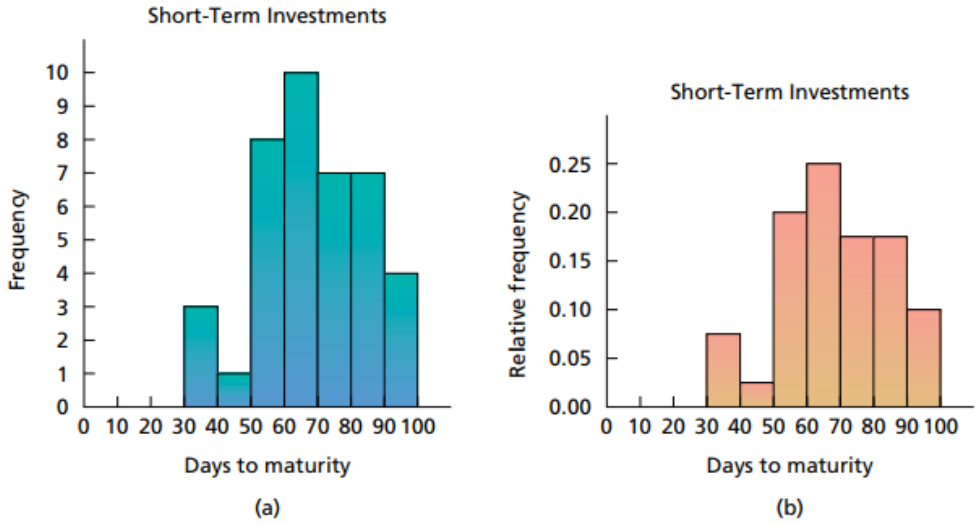
Histogram biểu diễn class trên trục x và tần suất (relative frequencies, percents) ở trục y, nói đơn giản bạn sẽ vẽ barchart với dữ liệu tần suất nhưng thay vì cách xa nhau thì chúng sẽ được đặt sát lại và sắp xếp theo độ lớn tăng dần.

Với single point grouping thì bạn đặt tên class ngay trung tâm của mỗi cột

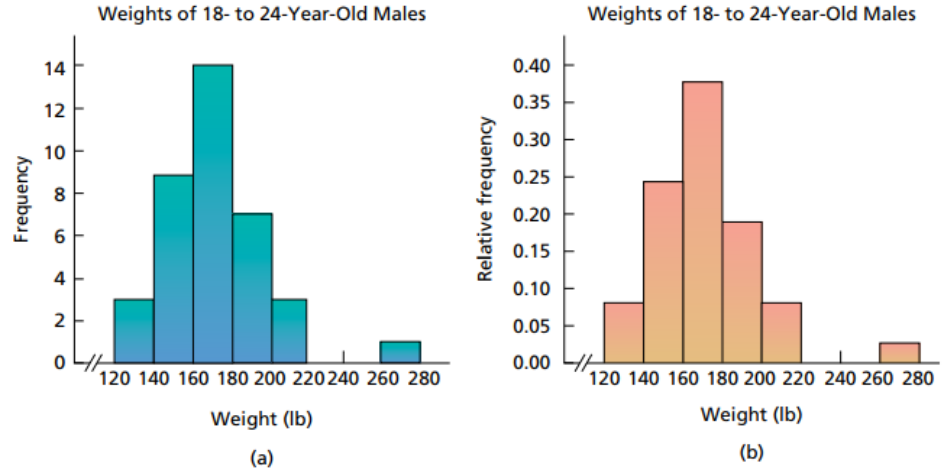
Với limit và cutpoint grouping bạn sẽ điền chặn trên và chặn dưới của group ở 2 bên cột




Hình 1




Hình 2

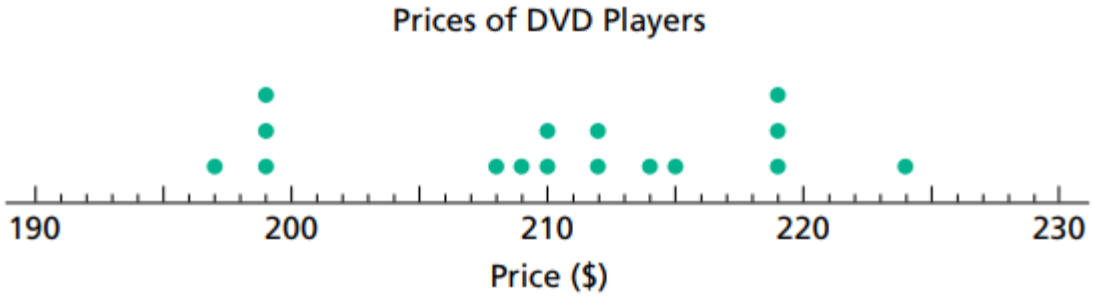



Hình 3

Histogram có thể giúp bạn nhận diện các giá trị outliers (quả nhỏ hoặc quá lớn so với phần đông dữ liệu)

*3.3.2.1 Dotplots*

Một dạng biểu diễn hình học cho dữ liệu định lượng nữa là Dotplot, mỗi observation sẽ được biểu diễn thành 1 điểm tương ứng với trục hoành, nếu có 2 giá trị bằng nhau thì chúng sẽ xếp chồng lên nhau. Dotplot thường được sử dụng với tập dữ liệu nhỏ vừa phải, nhìn vào đây bạn sẽ dễ nhận thấy các cụm dữ liệu hay outliers.




*3.3.2.2 Stem-and-Leaf Diagrams - Stemplot*

Cá nhân chúng tôi thấy biểu đồ này khá là giống Histogram nhưng thay vì hiển thì chiều dài cột thì ở đây sẽ hiển thị cụ thể số liệu.

Days to maturity for 40 short-term investments

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 70 | 64 | 99 | 55 | 64 | 89 | 87 | 65 |
| 62 | 38 | 67 | 70 | 60 | 69 | 78 | 39 |
| 75 | 56 | 71 | 51 | 99 | 68 | 95 | 86 |
| 57 | 53 | 47 | 50 | 55 | 81 | 80 | 98 |
| 51 | 36 | 63 | 66 | 85 | 79 | 83 | 70 |

Input Data

| Stems | Leaves |
|---|---|
| 3 | 6 8 9 |
| 4 | 7 |
| 5 | 0 1 1 3 5 5 6 7 |
| 6 | 0 2 3 4 4 5 6 7 8 9 |
| 7 | 0 0 0 1 5 8 9 |
| 8 | 0 1 3 5 6 7 9 |
| 9 | 5 8 9 9 |

Stemplot

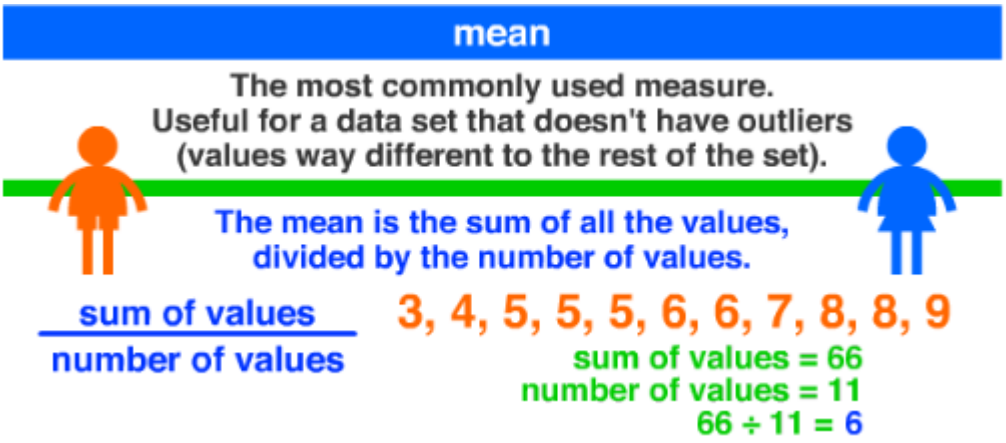
Nếu bạn lật ngang lại thì giống biểu đồ histogram đúng ko ạ, chúng tôi chia dữ liệu ra làm 2 phần stems (thân cây) và leaves (lá), như dữ liệu ở trên thì stems là hàng chục của số và leaves là hàng đơn vị, nếu dữ liệu có 3 chữ số thì stems là hàng trăm và chục, hàng đầu tiên biểu diễn các số sau 36, 38, 39. Với biểu đồ này bạn nên sử dụng chúng trong trường hợp lượng dữ liệu nhỏ. Bản thân chúng tôi thì khá ít dùng loại này

### 3.4. Measures of Center

Các chỉ số thể hiện giá trị trung tâm , giá trị tiêu biểu hay bạn có thể gọi là giá trị đại diện cho phần đông dữ liệu, có nhiều cách chọn giá trị đại diện ví dụ như: Mean - Trung bình, Median - Trung vị, Mode. Các chỉ số này được gọi là Mesuares of Central Tendency hay Measures of Center với mục đích chọn ra giá trị tiêu biểu, đủ điều kiện đại diện cho phần lớn các giá trị có trong tập dữ liệu của bạn. Theo kinh nghiệm của chúng tôi thì với mỗi trường hợp khác nhau sử dụng chỉ số này sẽ tốt hơn chỉ số khác, chúng tôi sẽ hướng dẫn các tip sử dụng ở phần bên dưới.

#### *3.4.1 Mean - Trung bình*

Chắc các bạn không xa lạ với chỉ số Mean này như chiều cao trung bình của Nam, Nữ ở Việt Nam lần lượt là 168cm và 156cm. Cách tính rất đơn giảm bạn sum tất cả các giá trị lại và chia cho số lượng giá trị ( thường kí hiệu là n), lưu ý nhỏ là trung bình không phải là giá trị nằm ở trung tâm các bạn . Mean chỉ sử dụng được với dữ liệu định lượng




Mean

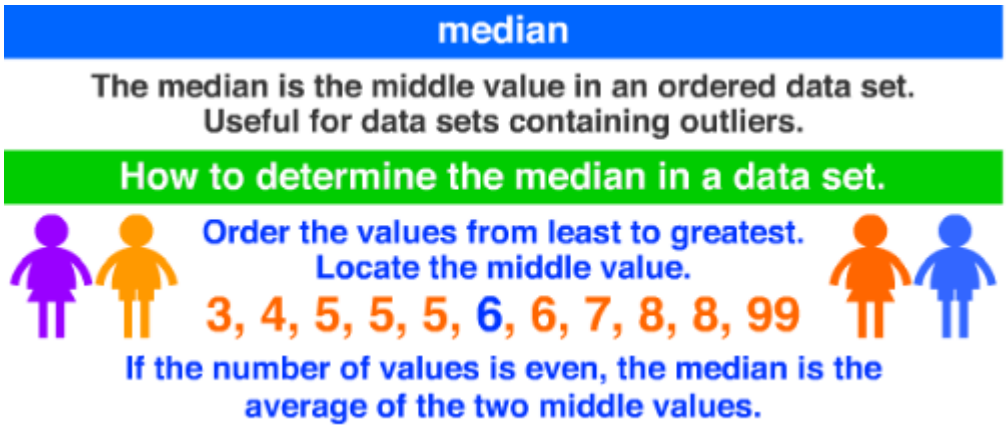
#### *3.4.2 Median - Trung vị*

Trung vị là giá trị nằm ở trung tâm thật sự, khác với trung bình ở phía trên, đầu tiên bạn sẽ sắp xếp dữ liệu theo thứ tự tăng dần sau đó chọn ra giá trị trung tâm bằng công thức (n+1)/2 với n là tổng số lượng dữ liệu bạn có.

Như hình bên dưới bạn sẽ thấy n = 11 nên suy ra vị trí của median sẽ là (11+1)/2 =6, vị trí thứ 6 lại trùng hợp có giá trị bằng 6, chúng ta nói median của tập dữ liệu là 6. Vậy nếu trong trường hợp vị trí chia ra bị lẻ thì sao ví dụ bạn có n = 10 thì vị trí

của median sẽ là (10+1)/2 = 5.5 lúc này bạn sẽ có median bằng giá trị tại vị trí số 5 và số 6 cộng lại chia đôi.

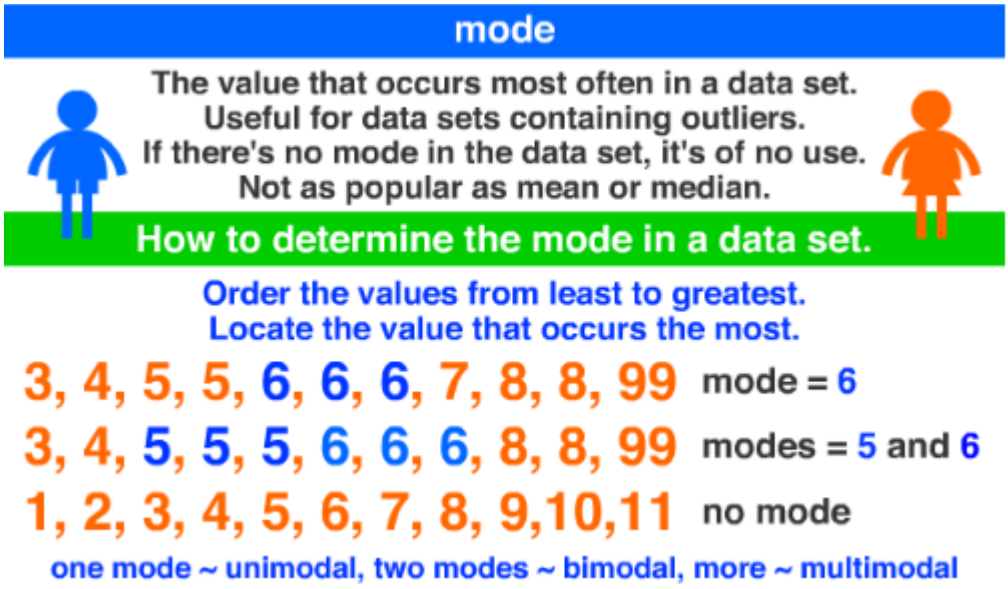
Median chính là giá trị ở vị trí trung tâm nên nó cũng sẽ chia dữ liệu của bạn thành 2 phần bằng nhau về số lượng, bên trái median (giá trị từ 3 đến 5 như hình dưới) là 50% số lượng dữ liệu (tức là 5 số) và bên phải cũng tương tự như vậy 50% số lượng dữ liệu (tức là 5 số), lưu ý là số lượng nha các bạn. Median chỉ sử dụng được với dữ liệu định lượng




Median

### *3.4.3 Mode - Yếu vị*

Phần trước các bạn đã tìm hiểu Frequency Table rồi đúng ko ạ, để tính được mode đầu tiên bạn sẽ tính tần suất xuất hiện của mỗi giá trị, Mode chính là giá trị có tần suất xuất hiện nhiều nhất, nếu 2 giá trị có cùng tần suất và cùng lớn nhất thì chúng ta có 2 Mode, nếu như không có giá trị nào xuất hiện hơn 1 lần thì tập dữ liệu ấy không có Mode. Mode sử dụng được với cả dữ liệu định tính và định lượng.




Mode

### *3.4.4 Mean vs Median vs Mode*

Chúng ta sẽ cùng xem xét trường hợp nào thì sử dụng giá trị nào để đạt được hiệu quả biểu đạt tốt nhất .

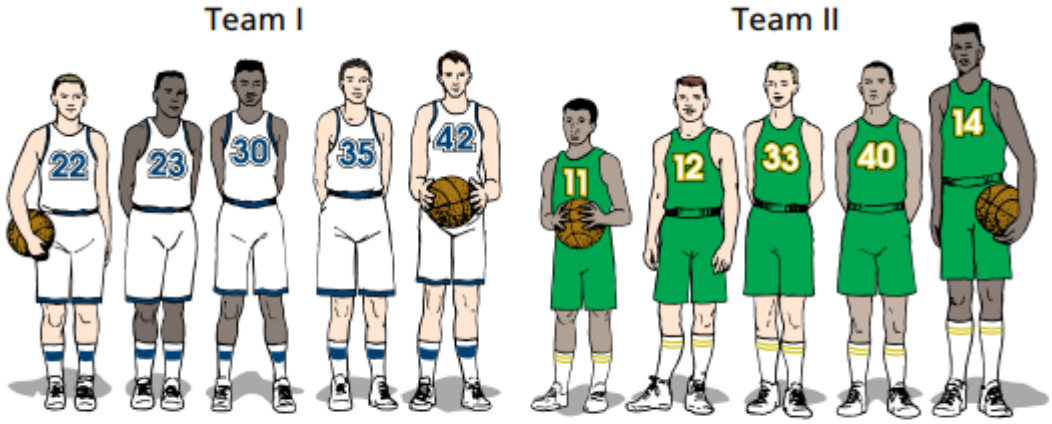
Bạn dễ dàng nhận thấy Mean sẽ bị ảnh hưởng nếu dữ liệu chứa nhiều giá trị quá lớn hoặc quá bé (outliers) trong khi Median thì không. Ví dụ như dãy số 1,1,1,2,100 dãy số này có Mean = 21 và median = 1, 1 chắc hẳn sẽ đại diện cho dữ liệu tốt hơn con số 21 rất nhiều, chúng tôi khuyên bạn hãy sử dụng Median khi dữ liệu của bạn có nhiều giá trị outliers quá lớn hoặc quá bé ở hai đầu các trường hợp còn lại hãy sử dụng Mean.

Trong các thống kê Population trung bình ở mỗi tiểu bang ở Mỹ người ta đã dùng con số Median thay vì Mean, mặc dù họ vẫn dùng từ "trung bình" để người khác dễ hiểu nhưng thật sự nó được tính từ Median để hạn chế sự ảnh hưởng của các tiểu bang có Population quá đông như California, Texas hay quá ít như Alaska, Vermont, tương tự với số người trên một km2.

Mode lại hoàn toàn mang một ý nghĩa ... không giống với Median và Mean khi chúng cố gắng tìm ra giá trị ở vị trí trung tâm. Thì giá trị xuất hiện nhiều nhất Mode lại không chắc chắn phải nằm ở trung tâm. chúng tôi đã từng dùng Mode trong bài toán xác định khung thời gian một user online nhiều nhất trong ngày, và sau đó phân loại họ.

## 3.5 Measures of Variation

Chúng ta đã đi qua khái niệm giá trị trung tâm và xuất hiện nhiều nhất, tuy nhiên sẽ có trường hợp 2 tập dữ liệu có chung Mean, Median, Mode nhưng vẫn có sự khác biệt, như trường hợp chiều cao của 10 cầu thủ bóng rổ dưới đây.




| | Team I | | | | | Team II | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Feet and inches | 6' | 6'1" | 6'4" | 6'4" | 6'6" | 5'7" | 6' | 6'4" | 6'4" | 7' |
| Inches | 72 | 73 | 76 | 76 | 78 | 67 | 72 | 76 | 76 | 84 |

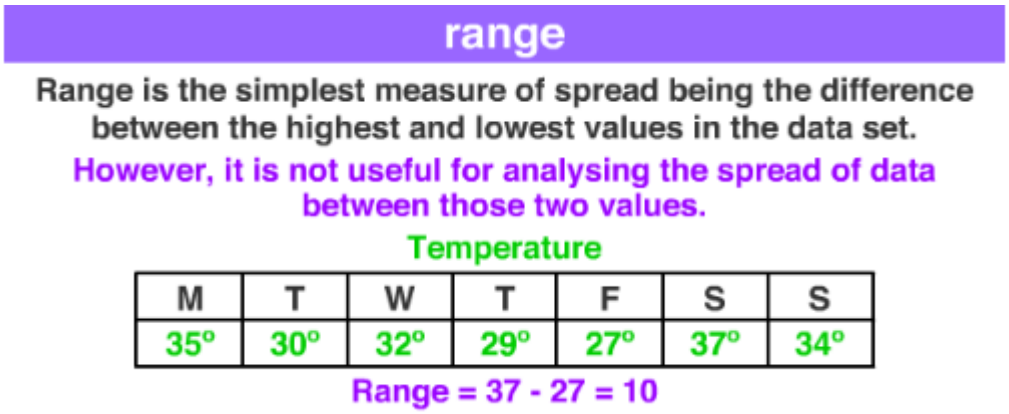
Five starting players on two basketball teams

Hai team cho chung chỉ số chiều cao trung bình là 75 inches, Median là 76 inches và mode là 76 inches. Sự khác biệt khá rõ ràng ở đây là Team 1 có chiều cao đồng đều hơn team 2, để mô tả sự khác biệt này người ta sử dụng các chỉ số đo sự

thay đổi, biến thiên của dữ liệu được biết đến với tên gọi là measures of variation hay measures of spread. Các chỉ số Measure of Variation phổ biến nhất là : range, quartiles, deciles, percentiles, the five number summary, standard deviation

### *3.5.1 Range*

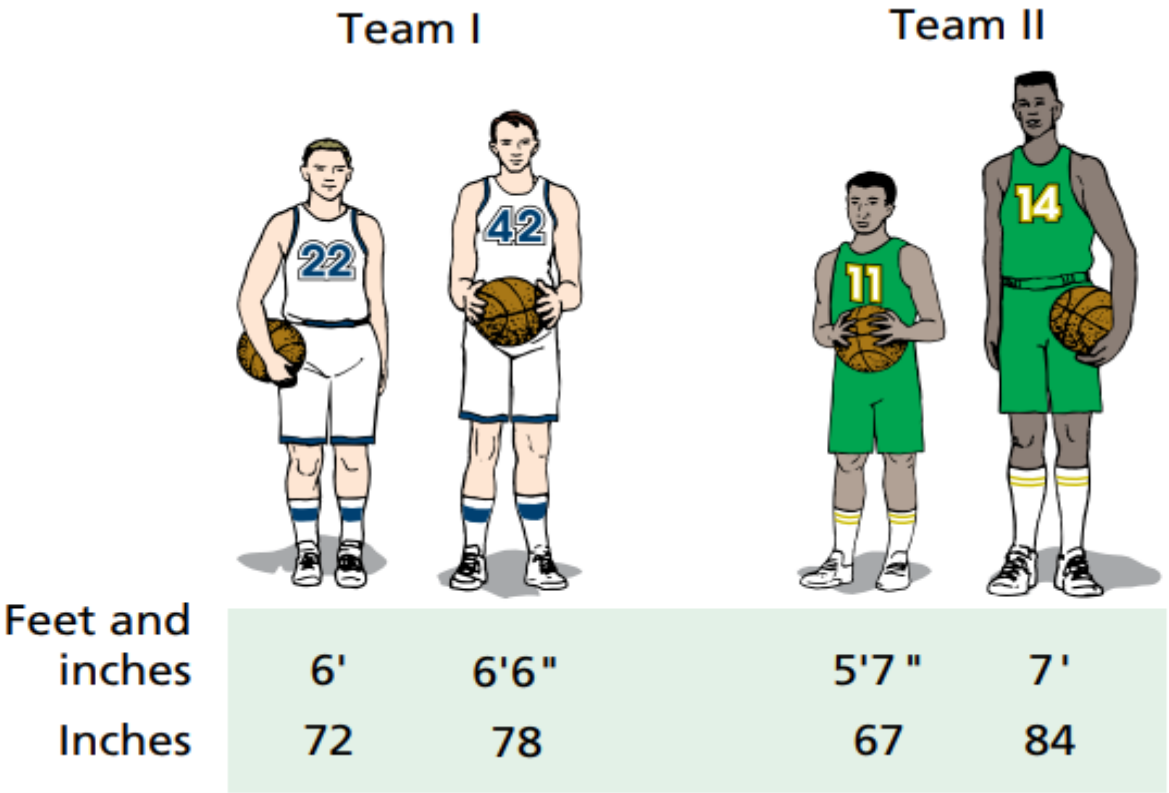
Range là hiệu số giữa giá trị lớn nhất và nhỏ nhất (range = max-min)




Range

Team I: Range = 78 - 72 = 6 inches

Team II: Range = 84 - 67 = 17 inches



### *3.5.2 Standard Deviation*

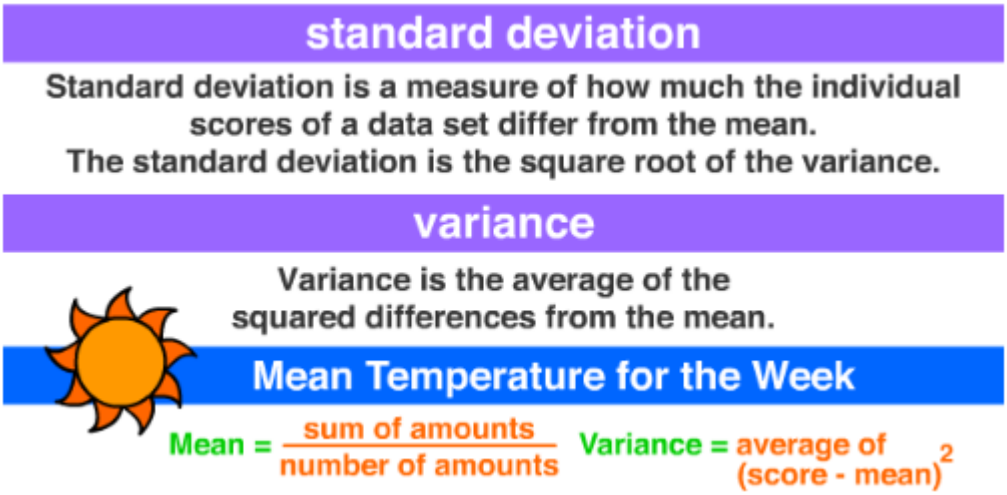
Standard Deviation (Độ lệch chuẩn - Std) cho ta biết được khoảng cách trung bình (độ phân tán) của các điểm dữ liệu so với giá trị trung bình (Mean). Nếu độ lệch chuẩn thấp tức là dữ liệu có tính biến động thấp và ngược lại, ví dụ mã cổ phiếu mà bạn đang xem xét mua vào có độ lệch chuẩn lớn (Std), cho thấy biên độ giao động xung quanh mức giá trung bình rất lớn, nếu mua bạn mua thì có khả năng sẽ lời rất nhiều hoặc lỗ rất nhiều so với các mã cổ phiếu có Std thấp.

Độ lêch chuẩn được tính bằng căn bậc hai của Phương sai - Variance. Cách tướng phương sai như sau:

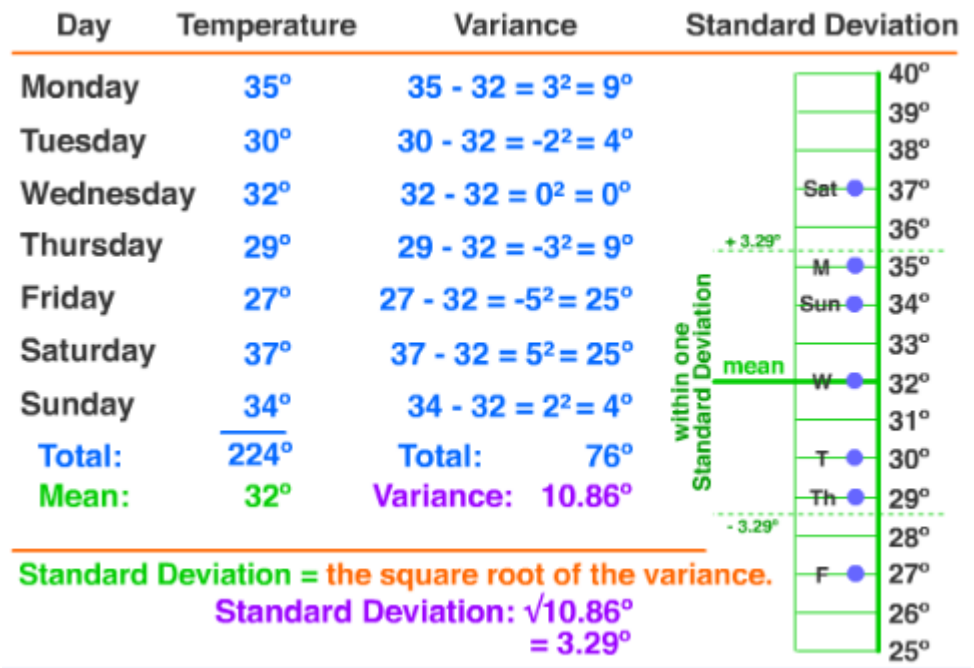
Tìm giá trị trung bình - Mean

Với mỗi điểm dữ liệu bạn lấy giá trị đó trừ đi Mean và bình phương chúng lên

Cộng tất cả kết quả từ bước trước chia cho số lượng dữ liệu khảo sát (n)




Bên dưới là tính toán chi tiết Độ lệch chuẩn của nhiệt độ  trong ngày



Standard Deviation

Bạn có thắc mắc giống như chúng tôi lý do tại sao Variance lại phải bình phương lên không, thay vì cứ lấy trị tuyệt đối của mỗi điểm dữ liệu trừ đi Mean rồi lấy trung bình ra Độ lệch chuẩn là xong ? Câu trả lời là phép tính bình phương sẽ "nhấn mạnh" các giá trị cách xa điểm Mean, nếu giá trị đang xét cách Mean 2 đơn vị thì phương sai là 4 tuy nhiên nếu cách 5 đơn vị thì phương sai lại tăng lên đến 25, một con số rất lớn, hiểu một cách đơn giản nếu dữ liệu của bạn chứa rất nhiều oulier - cách xa điểm mean về cả 2 phía quá bé hoặc quá lớn thì phương sai của bạn sẽ cực lớn, dẫn đến Std cũng sẽ lớn nốt, bình phương làm nổi bật các giá trị oulier. Và vì phương sai đã bình phương rồi nên để trở về đơn vị cũ buộc bạn phải căn bậc hai phương sai ra Độ lệch chuẩn để dễ so sánh với dữ liệu gốc ban đầu.

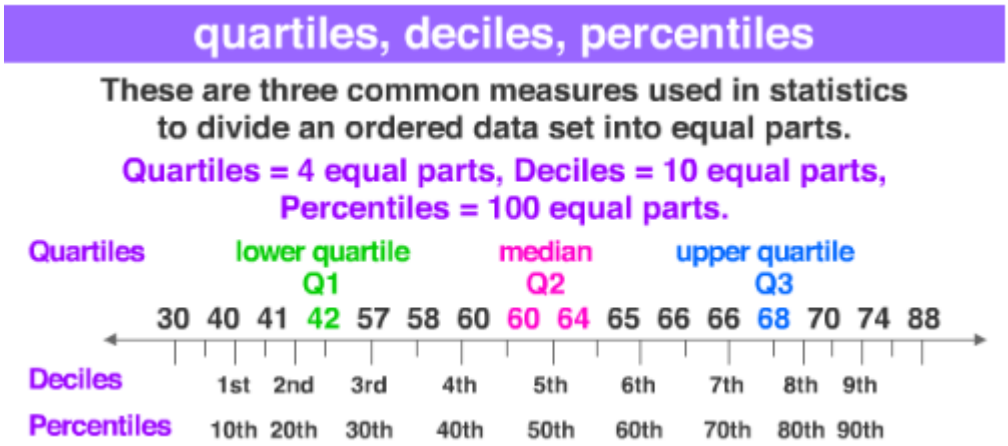
### *3.5.3 Quartiles, Deciles, Percentiles*

Bạn đã cùng chúng tôi tìm hiểu qua Median, là điểm nằm ở trung tâm chia dữ liệu ra làm 2 phân có số lượng bằng nhau, Quartiles, Deciles, Percentiles cũng tương tự như vậy, lưu ý cần phải sắp xếp dữ liệu tăng dần trước khi tính toán:

Quartiles: chia dữ liệu ra 4 phần bằng nhau

Deciles: chia dữ liệu thành 10 phần bằng nhau

Percentiles: chia dữa liệu thành 100 phần bằng nhau, Median chính là Percentitle(0.5) hay đọc là Percentitle 50%




Lý do tại sao phải chia dữ liệu ra như vậy ? chúng tôi sẽ đưa ra một ví dụ cho bạn dễ hiểu khi sếp bạn yêu cầu tính toán thời gian tối đa để giao một đơn hàng là bao nhiêu để sếp biết mà trao đổi với đối tác. Bạn không thể dùng thời gian giao hàng lớn nhất của tháng trước theo đúng ý sếp được vì lý do các đơn hàng ấy thường là có vấn đề: hàng thất lạc, nhà cung cấp giao thiếu phải giao thêm cho đủ, .... hay nói chính xác chúng là ouliers.

Bạn nảy ra một ý kiến sao chúng tôi không sắp xếp thời gian giao hàng của tất cả đơn hàng theo thứ tự tăng dần từ dưới lên trên,và lấy con số ở mức 90% số lượng dữ liêu, 10% còn lại bạn cho chúng là outliers, lưu ý ở đây giúp chúng tôi là 90% số lượng nha, nếu dữ liệu bạn có 100 dòng thì mức 90% dữ liệu là ở dòng thứ 90 từ dưới

đếm lên, con số đó chính là Percentile(0.9). Ví dụ chúng tôi có P(0.9) = 3.5 ngày thì chúng tôi sẽ nói với sếp, 90% đơn hàng của chúng ta được giao sớm hơn hoặc bằng 3.5 ngày, sếp cứ yên tâm dùng số này.

Với Quartile bạn sẽ có 4 phần nên Q(1) = Percentile(0.25) ->(25%), ... Decile 10 phần nên D1  = Percentile(0.1). Vì thế chúng tôi hay dùng Percentile hơn vì nó chi tiết nhất và 2 cái trên thì đều có thể quy ra Percentitle được.