Organizing, Visualizing and Describing Data

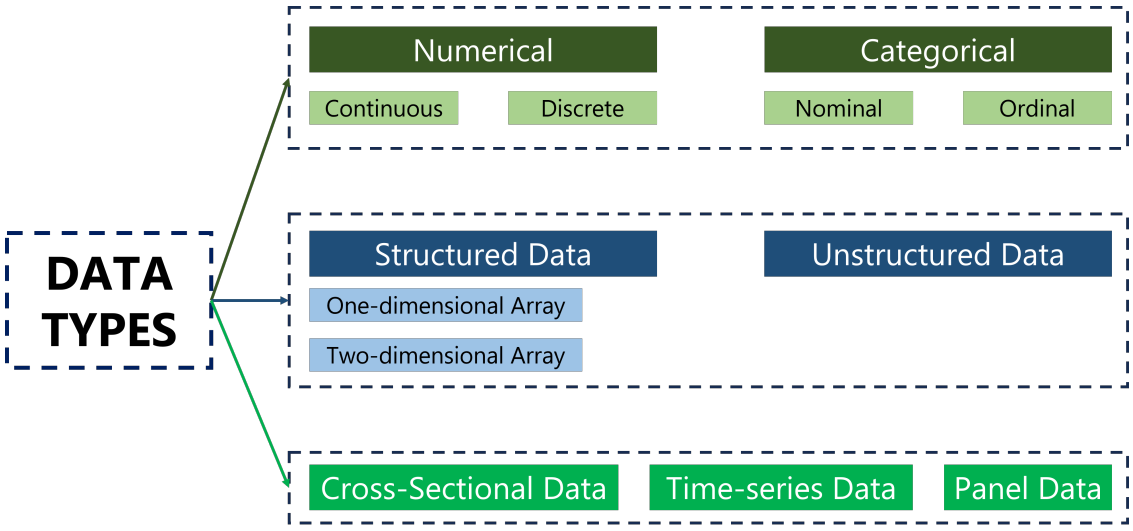
August 9, 2023

[ ]:
```
from IPython.display import Image
```

1 Data Types

[ ]:
```
# Phân loại dữ liệu
Image(filename = "Pictures/01.png")
```

[ ]:



2 Organizing Data for Quantitative Analysis

Phân tích định lượng và mô hình hóa thường yêu cầu dữ liệu đầu vào phải được xử lý và ở một định dạng nhất định

Dựa trên số lượng biến, dữ liệu thô (`raw data`) có thể được tổ chức bởi một trong hai dạng điển hình

1. Mảng một chiều (`one-dimensional array`)
    - Là dạng tổ chức dữ liệu đơn giản nhất, thể hiện tập hợp dữ liệu thuộc cùng một kiểu dữ liệu - phù hợp với biểu diễn dữ liệu của một biến duy nhất

- Tổ chức lại dữ liệu giúp lưu giữ các thông tin quan trọng bên cạnh thông tin từ thống kê mô tả (`descriptive statistics`), ví dụ như sự xuất hiện của các xu hướng tăng/giảm hay việc liệu các mẫu hình xuất hiện một cách có hệ thống theo thời gian hay không...

2. Mảng hai chiều (`two-dimensional rectangle array`, hay `data table`)
   - Là dạng tổ chức dữ liệu phổ biến nhất, đáp ứng yêu cầu đầu vào cho các tính toán bằng máy móc hoặc trình bày dữ liệu trực quan phục vụ nhu cầu sử dụng của con người
   - Mỗi cột trình bày một biến, trong khi mỗi hàng trình bày một quan sát

3 Summarizing Data

3.1 Using Frequency Distribution

Bảng phân phối tần suất (`frequency distribution`), hay bảng một chiều (`one-way table`) là một bảng được xây dựng bằng cách đếm số quan sát của một biến theo các giá trị hoặc nhóm riêng biệt, hoặc bằng cách đếm số quan sát của một biến số thuộc các tập hợp số hình thành từ trước (`interval/bucket`)

[ ]:
```
# Bảng phân phối tần suất với biến định tính
Image(filename = "Pictures/02.png")
```

[ ]:

Sector (Variable)	Absolute Frequency	Relative Frequency
Industrials	73	15.2%
Information Technology	69	14.4%
Financials	67	14.0%
Consumer Discretionary	62	12.9%
Health Care	54	11.3%
Consumer Staples	33	6.9%
Real Estate	30	6.3%
Energy	29	6.1%
Utilities	26	5.4%
Materials	26	5.4%
Communication Services	10	2.1%
Total	**479**	**100.0%**

[ ]:
```
# Bảng phân phối tần suất với biến liên tục
Image(filename = "Pictures/03.png")
```

[ ]:

Return Bin (%)	Absolute Frequency	Relative Frequency (%)	Cumulative Absolute Frequency	Cumulative Relative Frequency (%)
−5.0 to −4.0	1	0.08	1	0.08
−4.0 to −3.0	7	0.56	8	0.64
−3.0 to −2.0	23	1.83	31	2.46
−2.0 to −1.0	77	6.12	108	8.59
−1.0 to 0.0	470	37.36	578	45.95
0.0 to 1.0	555	44.12	1,133	90.06

3.2 Using Contingency Table

- Bảng tương quan (`contingency table`, hay `two-way table`) là bảng biểu thể hiện phân phối tần suất đồng thời của ít nhất hai biến định tính, và được sử dụng để tìm kiếm mối quan hệ giữa các biến (`finding patterns between the variables`)
- Bảng tương quan được xây dựng bằng cách liệt kê tất cả giá trị của một biến định tính trên một hàng và thực hiện điều tương tự với một biến định tính khác trên một cột
- Số giá trị của một biến trong một bảng tương quan là hữu hạn, đồng thời biến này cũng có thể mang ý nghĩa phân hạng hoặc không
- Dữ liệu được trình bày trong bảng tương quan có thể là tần suất tuyệt đối, hoặc tần suất tương đối dựa trên tổng số quan sát hoặc tổng số quan sát biên

[ ]:
```
# Bảng tương quan, thể hiện tần suất tuyệt đối
Image(filename = "Pictures/04.png")
```

[ ]:

	Market Capitalization Variable (3 Levels)			
Sector Variable (5 Levels)	**Small**	**Mid**	**Large**	**Total**
Communication Services	55	35	20	110
Consumer Staples	50	30	30	110
Energy	175	95	20	290
Health Care	275	105	55	435
Utilities	20	25	10	55
Total	**575**	**290**	**135**	**1,000**

[ ]:
```
# Bảng tương quan, thể hiện tần suất tương đối theo tổng số quan sát
Image(filename = "Pictures/05.png")
```

[ ]:

	Market Capitalization Variable (3 Levels)			
Sector Variable (5 Levels)	**Small**	**Mid**	**Large**	**Total**
Communication Services	5.5%	3.5%	2.0%	11.0%
Consumer Staples	5.0%	3.0%	3.0%	11.0%
Energy	17.5%	9.5%	2.0%	29.0%
Health Care	27.5%	10.5%	5.5%	43.5%
Utilities	2.0%	2.5%	1.0%	5.5%
Total	**57.5%**	**29.0%**	**13.5%**	**100%**

4 Data Visualization

*Trực quan hóa (**`visualization`**) là việc trình bày dữ liệu dưới dạng biểu đồ và/hoặc đồ thị nhằm gia tăng sự hiểu biết và đạt được cái nhìn sâu sắc về dữ liệu*

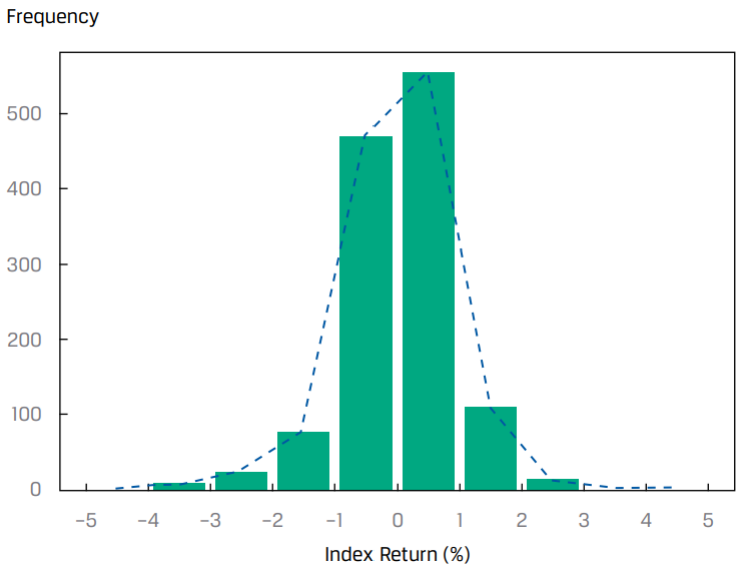
4.1 Histogram and Frequency Polygon

`Histogram` là đồ thị thể hiện phân phối của dữ liệu số thông qua độ dài của các thanh hoặc các cột để biểu thị tần suất tuyệt đối của từng nhóm giá trị (`interval`)

Đa giác tần suất (`Frequency Polygon`) là một đa giác được tạo bởi trục hoành và một đường nối các điểm biểu diễn tần số của các nhóm giá trị khác nhau.

[ ]:
```
# Histogram
Image(filename = "Pictures/06.png")
```

[ ]:

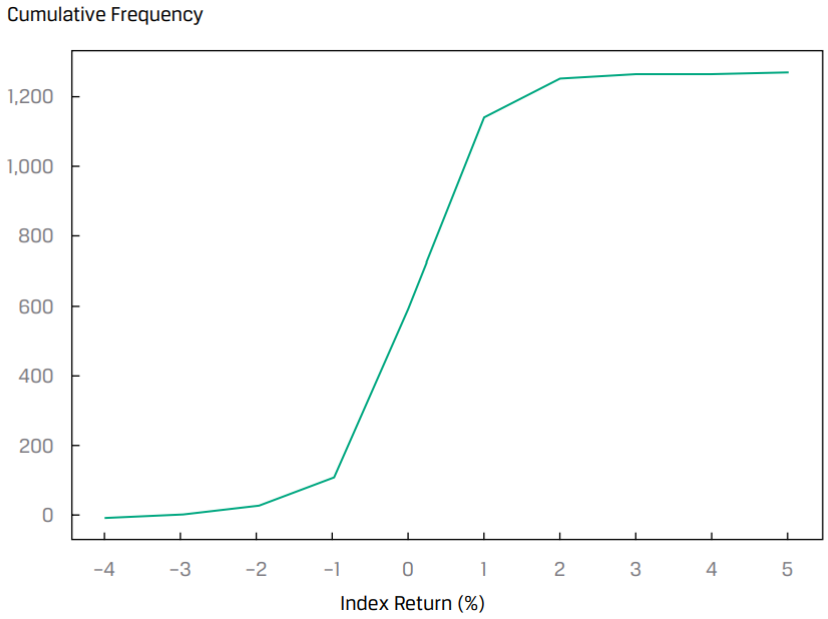


4.2 Cummulative Absolute Frequency Distribution

Đồ thị phân phối tích lũy có thể được hình thành dựa trên tần suất tích lũy tuyệt hoặc tần suất tích lũy tương đối. Đồ thị này cho biết số lượng hoặc tỷ lệ số lượng quan sát nhỏ hơn một giá trị cụ thể

[ ]:
```
# Cummulative Absolute Frequency Distribution
Image(filename = "Pictures/07.png")
```

[ ]:



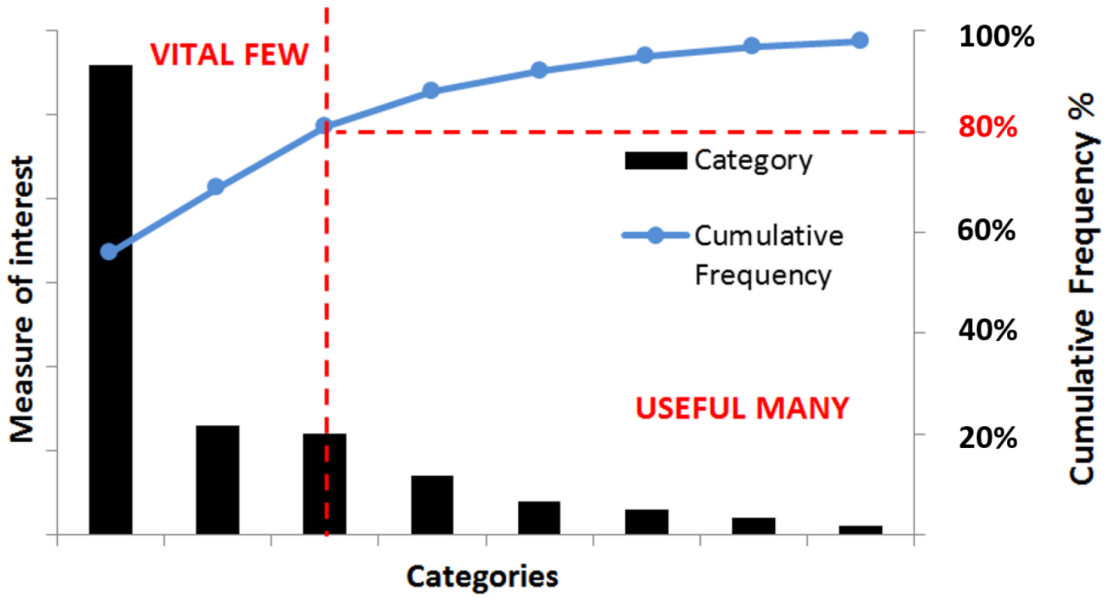
4.3 Pareto Chart

`Pareto Chart` là một dạng kết hợp giữa biểu đồ cột và biểu đồ đường. Trong đó:

- Các nhóm giá trị (`interval`) được sắp xếp theo thứ tự tần suất giảm dần trên trục hoành (do đó chiều cao của các cột giảm dần theo chiều dương)
- Đường tần suất tích lũy thể hiện cho tần suất tích lũy của các nhóm giá trị, theo thứ tự từ trái sang phải trong biểu đồ (do đó độ dốc của đường giảm dần theo chiều dương)

[ ]:
```
# Pareto Chart
Image(filename = "Pictures/08.png")
```

[ ]:

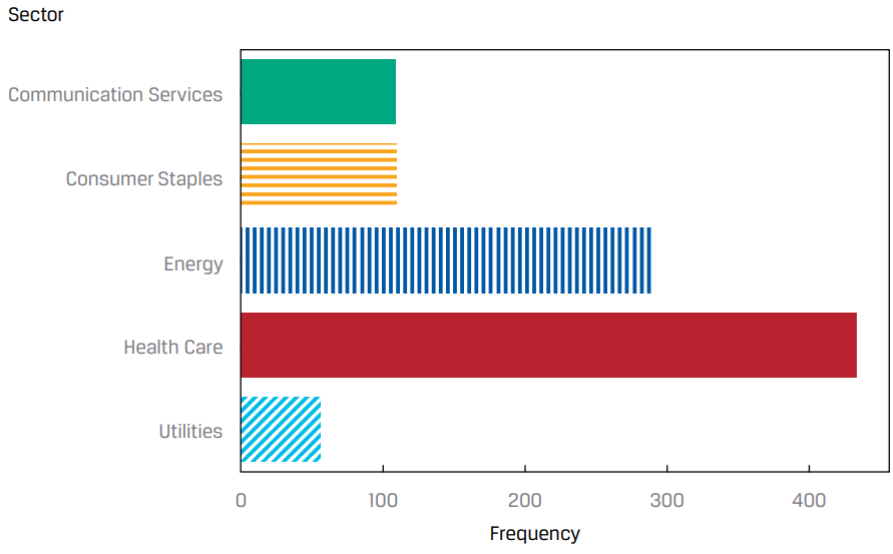


4.4 Bar Chart

Biểu đồ thanh (`bar chart`) là một dạng biểu đồ phù hợp để thể hiện tần suất xuất hiện (các) giá trị của một biến danh nghĩa (`categorical`), trong đó mỗi thanh đại diện cho một giá trị danh nghĩa riêng biệt và chiều dài của thanh cho biết tần suất xuất hiện của giá trị đó

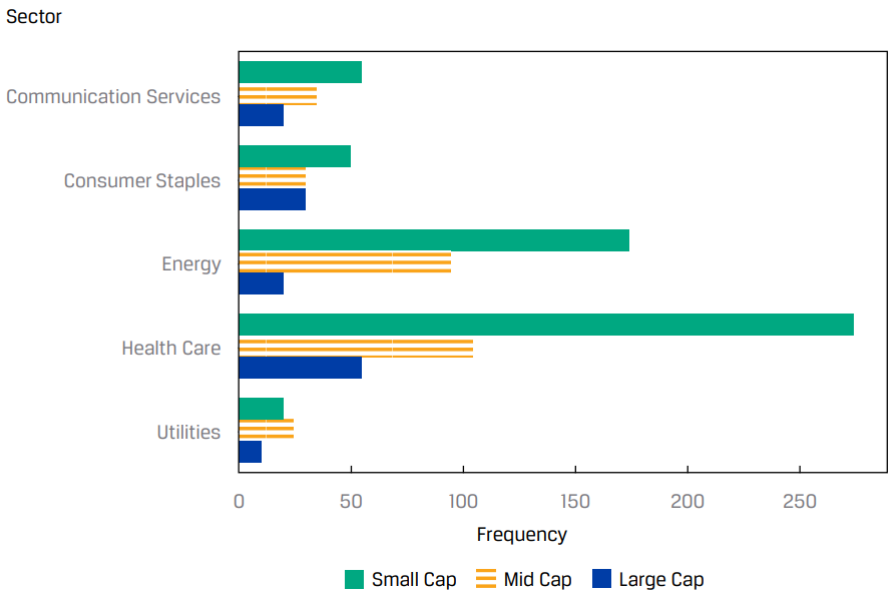
Trong trường hợp cần trực quan hóa dữ liệu với hai biến danh nghĩa, chúng ta cần sử dụng biểu đồ thanh theo cụm (`clustered bar chart`/`grouped bar chart`) hay biểu bồ cột chồng (`stacked bar chart`) nhằm trình bày được tần suất kết hợp giữa hai biến này

[ ]:
```
# Basic Bar Chart
Image(filename = "Pictures/09.png")
```

[ ]:


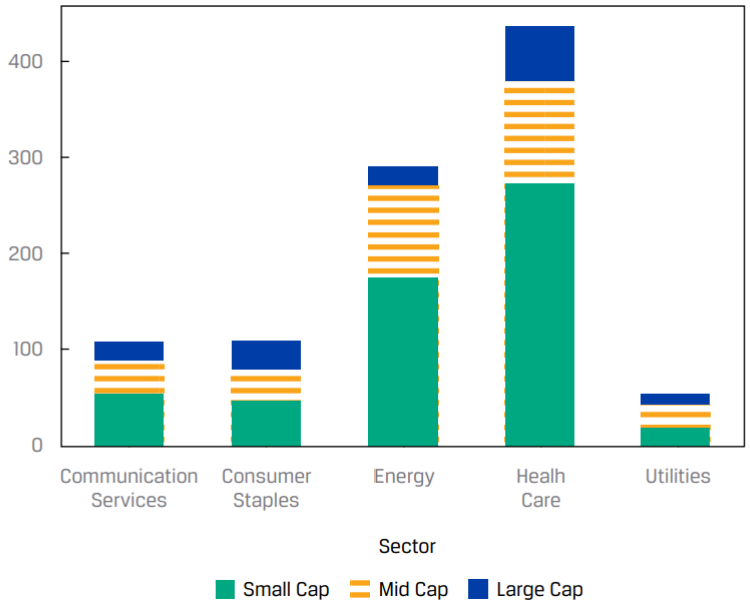
[ ]:
```
# Clustered Bar Chart
Image(filename = "Pictures/10.png")
```

[ ]:



[ ]:
```
# Stacked Bar Chart
Image(filename = "Pictures/11.png")
```

[ ]:



4.5 Tree-Map

`Tree-map` là một dạng biểu đồ được sử dụng để trực quan hóa dữ liệu danh nghĩa, bên cạnh biểu đồ cột

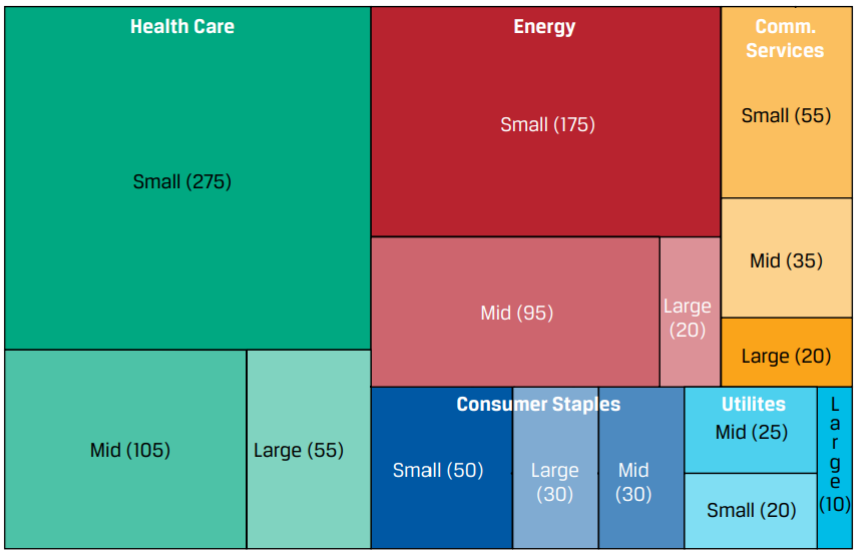
`Tree-map` bao gồm một tập hợp các hình chữ nhật với màu sắc khác nhau, trong đó mỗi hình chữ nhật đại diện cho một giá trị danh nghĩa và diện tích của hình chữ nhật tỷ lệ thuận với tần suất của giá trị tương ứng.

Lưu ý rằng, `tree-map` có thể mô tả dữ liệu của nhiều biến danh nghĩa đồng thời, trong đó biến thứ n được biểu diễn bằng cách chia tách các hình chữ nhật sẵn có (từ việc biểu diễn biến thứ $n-1$) thành các "hình chữ nhật con"

`Tree-map` sẽ trở nên khó đọc, nếu nó phải biểu diễn nhiều hơn 3 biến định tính.

[ ]:
```
# Tree-Map
Image(filename = "Pictures/12.png")
```

[ ]:



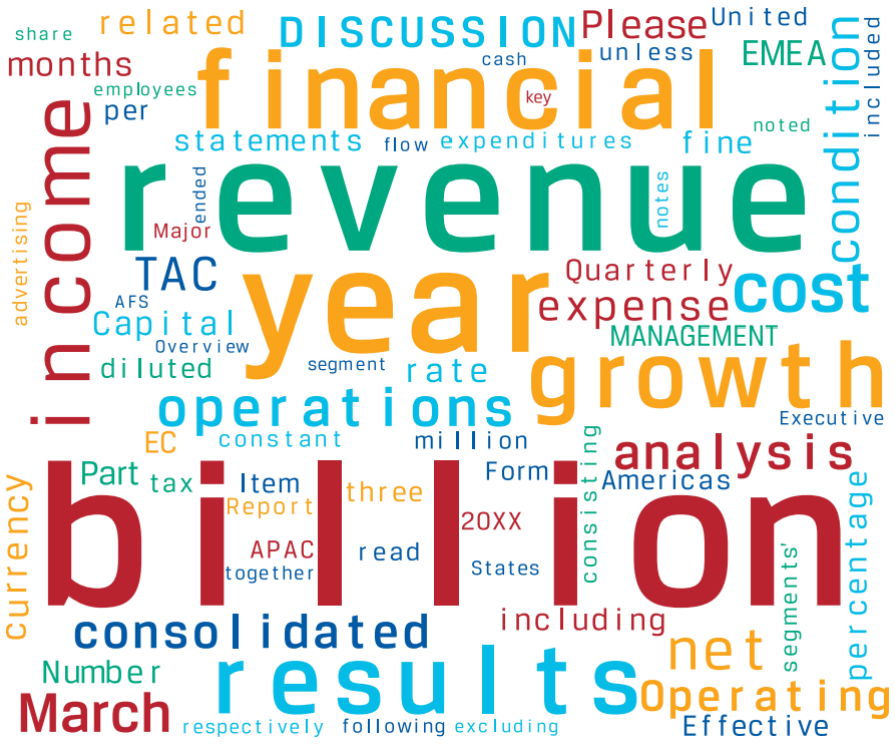
4.6 Word Cloud/Tag Cloud

`Word Cloud` là dạng biểu đồ mô tả tần suất của dữ liệu phi cấu trúc, cụ thể là văn bản. Trong đó:

- Kích thước của từ thể hiện cho tần suất xuất hiện của bản thân từ đó trong văn bản, và chúng ta có thể nhanh chóng nắm bắt chủ đề và/hoặc nội dung chính của văn bản đó.
- Màu sắc của từ thể hiện trạng thái của từ đó, ví dụ như tích cực hay tiêu cực, lạc quan hay bi quan, yêu hay ghét...

[ ]:
```
# Word Cloud
Image(filename = "Pictures/13.png")
```

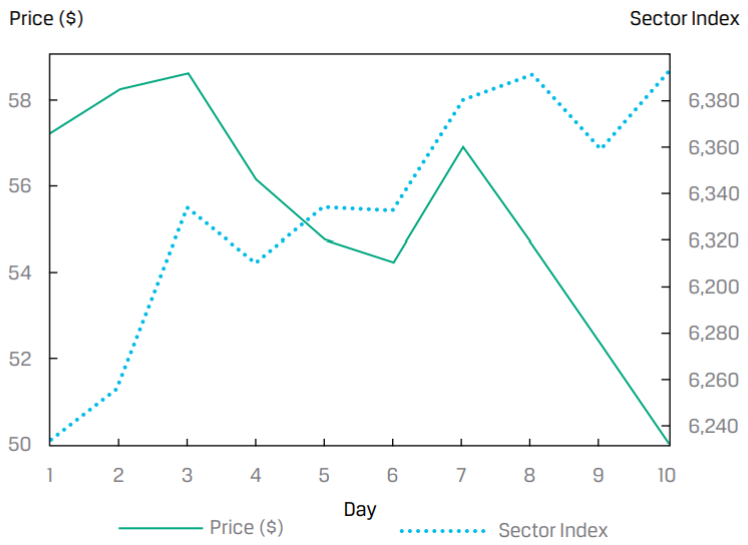
[ ]:



4.7 Line Chart

Biểu đồ đường được sử dụng để trực quan hóa các quan sát có thứ tự, đặc biệt thường được sử dụng để mô tả sự biến động của chuỗi thời gian

[ ]:
```
# Line Chart
Image(filename = "Pictures/14.png")
```

[ ]:


Các lợi ích của việc sử dụng biểu đồ đường:

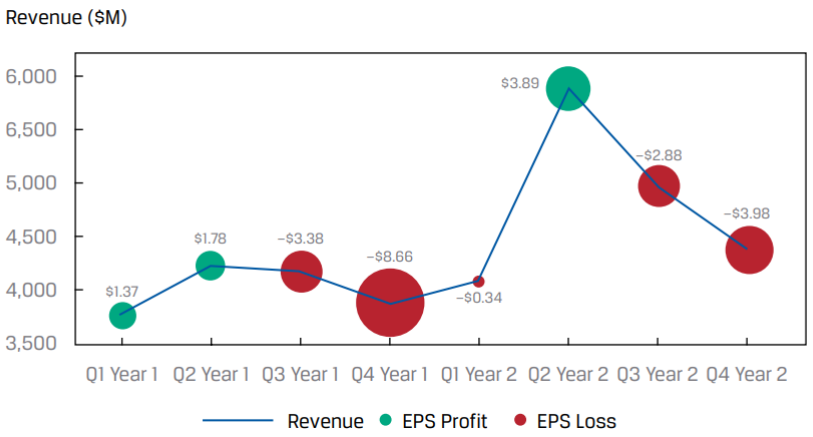
- Thể hiện các biến động và xu hướng cơ bản của dữ liệu một cách rõ ràng và đơn giản, giúp chúng ta dễ hình dung và dự báo chuỗi dữ liệu
- Phù hợp với việc trực quan hóa dữ liệu lớn
- Thích hợp với việc so sánh khi có thể trực quan hóa nhiều tập điểm dữ liệu khác nhau trên cùng một biểu đồ

4.8 Bubble Line Chart

Nếu vùng quan tâm của chúng ta nhiều hơn một đặc điểm (một biến), việc thể hiện tất cả các đặc điểm này trên một biểu đồ sẽ rất hữu ích khi giúp chúng ta có được cái nhìn tổng thể hơn về dữ liệu. Chúng ta có thể thay thế các điểm dữ liệu bằng các bong bóng với kích thước và/hoặc màu sắc khác nhau để thể hiện chiều thứ ba của dữ liệu và các thông tin bổ sung. Dạng biểu đồ này được gọi là `bubble line chart`

[ ]:
```
# Bubble Line Chart
Image(filename = "Pictures/15.png")
```

[ ]:

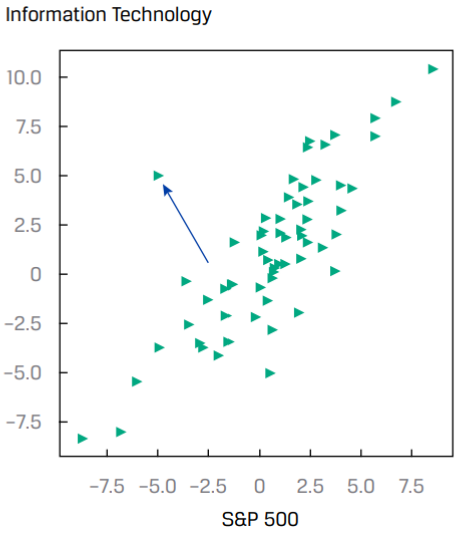


4.9 Scatter Plot

Biểu đồ phân tán (`scatter plot`) là một loại biểu đồ để trực quan hóa sự thay đổi chung của hai biến số. Nó là một công cụ hữu ích để thể hiện và cho cái nhìn tổng quan về các mối quan hệ tiềm năng giữa các biến

[ ]:
```
# Scatter Plot
Image(filename = "Pictures/16.png")
```

[ ]:

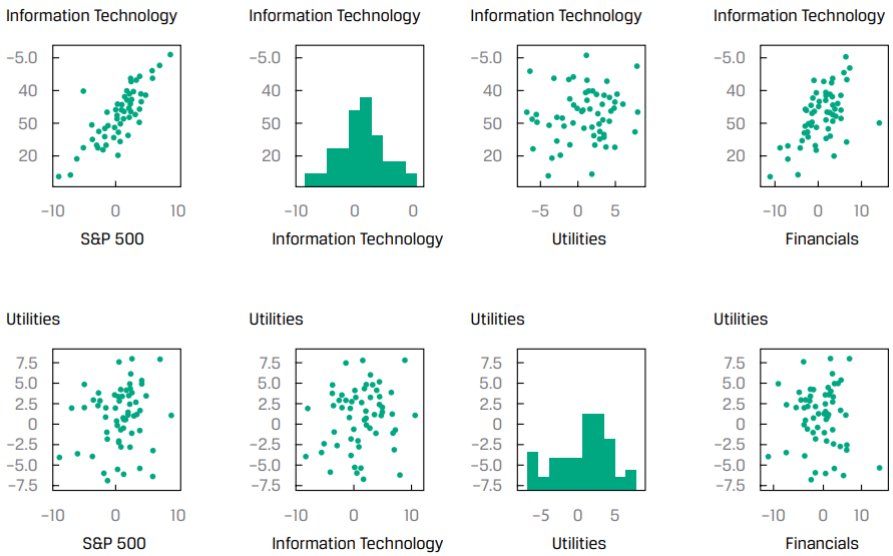


4.10 Scatter Plot Matrix

Ma trận biểu đồ phân tán (`scatter plot matrix`) bản chất là tập hợp của các biểu đồ phân tán giữa các cặp biến

[ ]:
```
# Scatter Plot Matrix
Image(filename = "Pictures/17.png")
```

[ ]:

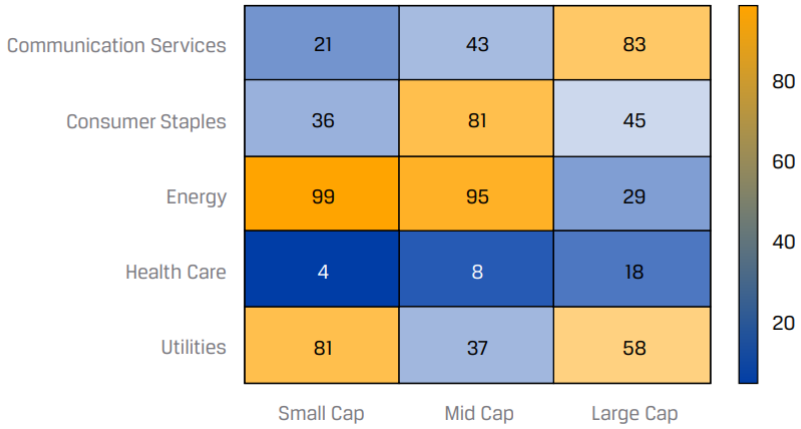


4.11 Heat Map

Biểu đồ nhiệt (`heat map`) là một dạng biểu đồ được sử dụng để tổ chức và tóm tắt dữ liệu dưới dạng bảng (tương tự `contingency table`), trong đó các ô trong bảng được mã hóa màu để thể hiện cho các giá trị lớn nhỏ khác nhau

[ ]:
```
# Heat Map
Image(filename = "Pictures/18.png")
```

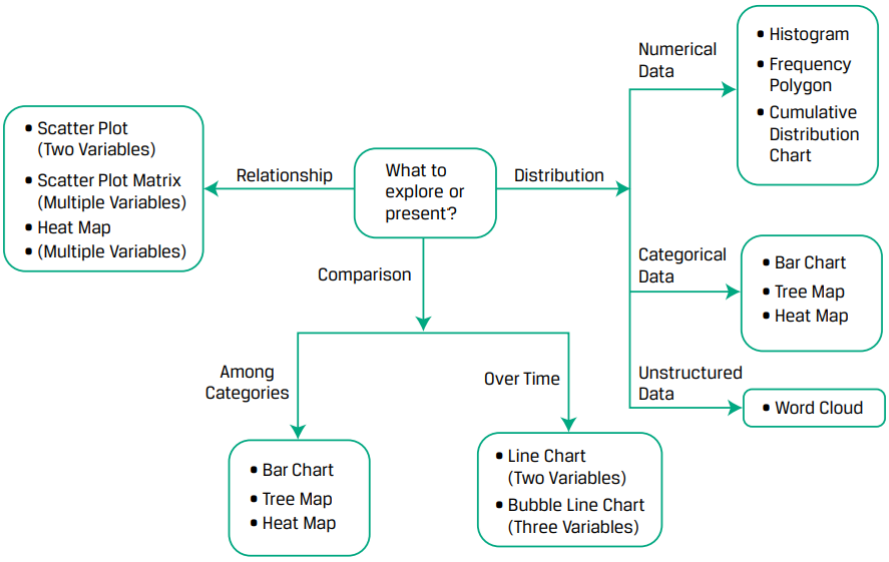
[ ]:



4.12 Guide to Selecting among Visualization Types

[ ]:
```
# Select among Visualization Types
Image(filename = "Pictures/19.png")
```

[ ]:



5 Measure of Central Tendency

Xác định xu hướng tập trung (`measure of central tendency`) xác định nơi dữ liệu được tập trung

Phép đo vị trí (`measure of location`) không chỉ bao gồm các phép đo trung tâm, mà còn quan tâm đến các phép đo minh họa vị trí và phân phối của dữ liệu

5.1 Arithmetic Mean

Trung bình cộng bằng tổng giá trị của các quan sát chia cho số quan sát

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

5.2 Trimmed Mean và Winsorized Mean

Trên thực tế, các giá trị ngoại lai có thể chỉ đại diện cho một vài giá trị hiếm trong tổng thể, nhưng nó cũng có thể là lỗi trong việc ghi chép dữ liệu hoặc giá trị của quan sát này được tạo ra từ một tổng thể khác so với tổng thể của mẫu.

Giá trị ngoại lai có thể khiến các xu hướng tập trung bị lệch. Trong trường hợp này, chúng ta cần:

1. Làm sạch dữ liệu
2. Lựa chọn biến khác với cùng mục đích, hoặc sử dụng các phép chuyển đổi biến (ví dụ như lấy logarit tự nhiên)
3. Nếu không thể sử dụng biến thay thế hoặc chuyển đổi biến, dưới đây là ba cách xử lý:
    - Không làm gì cả
    - Xóa bỏ tất cả các giá trị ngoại lai
    - Thay thế các giá trị ngoại lai bằng các giá trị khác

Ứng với cách thứ hai, chúng ta có thể sử dụng trung bình lược bỏ (`trimmed mean`). Đại lượng này được tính bằng cách loại bỏ một tỷ lệ phần trăm giá trị lớn nhất và giá trị nhỏ nhất và tính trung bình cộng các giá trị còn lại

Ứng với cách thứ ba, chúng ta có thể sử dụng trung bình gán (`winsorized mean`). Đại lượng này được tính bằng cách thay thế một tỷ lệ phần trăm giá trị nhỏ nhất bằng một giá trị cụ thể, và thực hiện điều tương tự với tỷ lệ phần trăm giá trị lớn nhất, sau đó tính trung bình cộng của mẫu thu được

5.3 Median

Trung vị (`median`) là giá trị nằm ở chính giữa chuỗi giá trị đã được sắp xếp tăng dần hoặc giảm dần. Trong một mẫu gồm n quan sát, trung vị bằng giá trị ở vị trí $\frac{n+1}{2}$ nếu n lẻ, và bằng trung bình hai giá trị ở vị trí $\frac{n}{2}$ và $\frac{n+1}{2}$ nếu n chẵn

5.4 Mode

`Mode` là giá trị có tần số cao nhất trong phân phối

Một phân phối có thể có một `mode`, nhiều `mode` hoặc thậm chí không có `mode` nào.

- Nếu phân phối chỉ có một `mode`, nó được gọi là `unimodal`
- Nếu phân phối có hai `mode`, nó được gọi là `bimodal`
- Nếu phân phối có ba `mdoe`, nó được gọi là `trimodal`
- Nếu tất cả giá trị trong tập dữ liệu là khác nhau, phân phối không có `mode` bởi không có giá trị nào xuất hiện nhiều lần hơn so với giá trị khác

Biến liên tục có thể không có `mode`. Tuy nhiên, nếu dữ liệu được nhóm vào các `interval`, chúng ta có thể tìm thấy một `interval` hoặc nhiều `intervals` có tần suất cao nhất (`modal interval`)

5.5 Weighted Mean

Định nghĩa trung bình trọng số (`weighted mean`) được phát triển từ phân tích danh mục - chúng ta cần quan tâm nhiều hơn đến trọng số khác nhau giữa các quan sát khác nhau

Với tập quan sát $x_1, x_2, ..., x_n$ và trọng số tương ứng là $w_1, w_2, ..., w_n$, trung bình trọng số được tính như sau:

$$\overline{x_w} = \sum_{i=1}^{n} w_i x_i$$

Trong đó, $\sum_{i=1}^{n} w_i = 1$

5.6 Geometric Mean

Trung bình nhân (`geometric mean`) thường được sử dụng để tính tỷ lệ thay đổi trung bình qua thời gian hoặc để tính tốc độ tăng trưởng của một biến. Trong đầu tư, chúng ta thường sử dụng trung bình nhân để tính tỷ suất sinh lợi trung bình của một tài sản/một danh mục, hoặc tính tỷ lệ tăng trưởng của một biến tài chính như lợi nhuận hoặc doanh thu

Công thức tính trung bình nhân của tập giá trị $x_1, x_2, ..., x_n$, trong đó $x_i 0 với $i = 1, 2, ..., n$ là

$$\overline{x_G} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

5.7 Harmonic Mean

Trung bình điều hòa (`harmonic mean`) thường được sử dụng trong đầu tư để tính chi phí trung bình

$$\overline{x_H} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

5.8 Q-A-G-H Inequality

Bất đẳng thức Q-A-G-H thể hiện mối quan hệ giữa trung bình cộng, trung bình nhân, trung bình toàn phương (`quadratic average`) và trung bình điều hòa: $\overline{x_Q} \geq \overline{x} \geq \overline{x_G} \geq \overline{x_H}$

$$\sqrt{\sum_{i=1}^{n} x_i^2} \geq \frac{\sum_{i=1}^{n} x_i}{n} \geq \sqrt[n]{\prod_{i=1}^{n} x_i} \geq \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

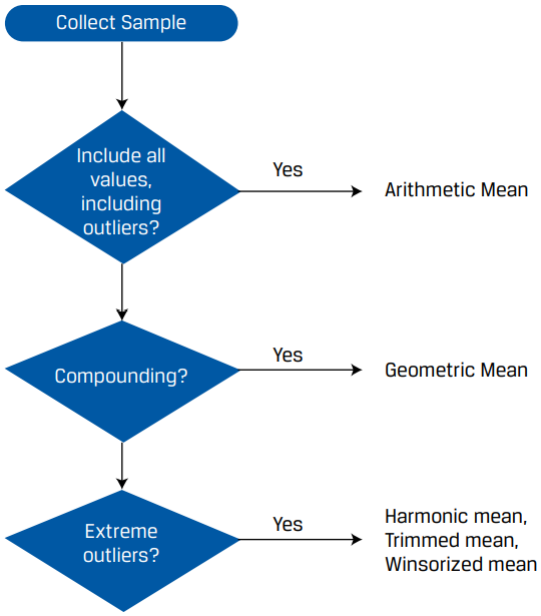
Note. *Công thức trung bình toàn phương*

$$\overline{x_Q} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

5.9 Deciding Which Central Tendency Measure to Use

[ ]:
```
# Which mean to use in what circumstances
Image(filename = "Pictures/20.png")
```

[ ]:



6 Quantiles

Những người làm thống kê sử dụng thuật ngữ phân vị (`quantile` hay `fractile`) để chỉ các điểm cắt phân phối thành các khoảng liên tục với xác suất bằng nhau

6.1 Quartiles, Quintiles, Deciles, Percentiles

Chúng ta biết rằng trung vị chiai phân phối thành hai nửa có xác suất như nhau. Chúng ta có thể định nghĩa các phân vị mà chúng chia phân phối thành các phần nhỏ hơn. Cụ thể

- Tứ phân vị (`Quartiles`): gồm 3 điểm, chia phân phối thành 4 phần có xác suất bằng nhau

Độ trải giữa (`Interquartile` - IQR) được tính bằng chênh lệch giữa tứ phân vị thứ ba và tứ phân vị thứ nhất: $IQR = Q_3 - Q_1$

- Ngũ phân vị (`Quintiles`): gồm 4 điểm, chia phân phối thành 5 phần có xác suất bằng nhau
- Thập phân vị (`Deciles`): gồm 9 điểm, chia phân phối thành 10 phần có xác suất bằng nhau

- Bách phân vị (`Percentiles`): gồm 99 điểm, chia phân phối thành 100 phần có xác suất bằng nhau

Khi xử lý dữ liệu thực tế, chúng ta thường phải ước lượng giá trị của một bách phân vị nào đó. Giả sử chúng ta cần ước lượng giá trị của bách phân vị thứ y ($P_y) của một phân phối mà mẫu của nó có n quan sát. Quy trình xác định bách phân vị này gồm 3 bước:

1. Sắp xếp lại dữ liệu thành chuỗi theo thứ tự tăng dần
2. Xác định vị trí tương đối của phân vị (L_y) bằng công thức

$$L_y = (n+1)\frac{y}{100}$$

3. Xác định giá trị P_y. Ở đây có hai trường hợp
   - Nếu L_y là số nguyên, P_y chính là giá trị thứ L_y trong chuỗi
   - Nếu L_y không là số nguyên, P_y được xác định thông qua hai giá trị thứ $[L_y]$ và $([L_y]+1)$ trong chuỗi bằng phương pháp nội suy tuyến tính

$$P_y = X_{[L_y]} \times (1 - (L_y - [L_y])) + X_{[L_y]+1} \times (L_y - [L_y])$$

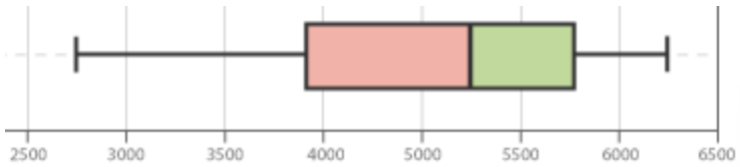
6.2 Box and Whisker Plot

Một cách để biểu diễn độ phân tán của dữ liệu thông qua tứ phân vị là sử dụng biểu đồ hộp. Biểu đồ này có hai cách vẽ:

- Không biểu diễn các giá trị ngoại lai: hộp có biên dưới và biên trên lần lượt là Q_1 và Q_3, râu dưới và râ trên đại diện cho giá trị lớn nhất và giá trị nhỏ nhất, đồng thời trung vị được biểu diễn bởi vạch bên trong hộp
- Biểu diễn các giá trị ngoại lai: hộp có biên trên và biên dưới lần lượt là Q_1 và Q_3, rào dưới bằng Q_1 trừ đi 1.5 lần IQR, rào trên bằng Q_3 cộng với 1.5 lần IQR, các giá trị ngoại lai nằm bên ngoài hai rào được biểu diễn bằng các điểm chấm, trung vị được biểu diễn bởi vạch bên trong hộp

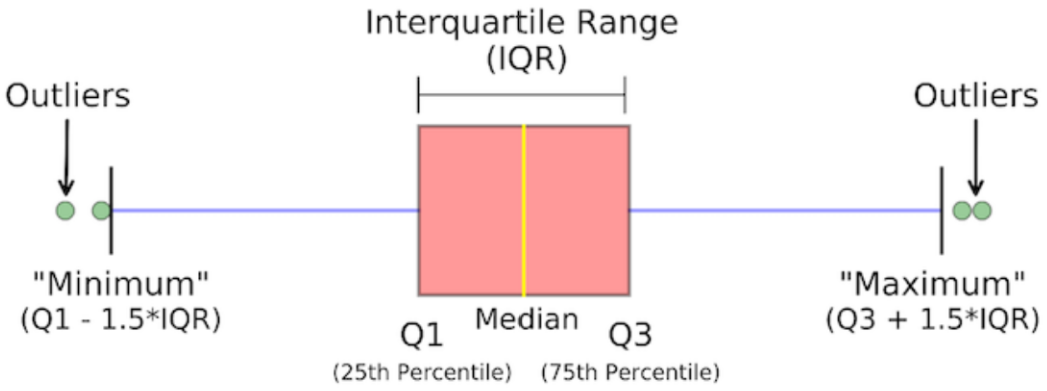
[ ]:
```
# Box & Whisker Plot without fences
Image(filename = "Pictures/21.png")
```

[ ]:



[ ]:
```
# Box & Whisker Plot with fences
Image(filename = "Pictures/22.png")
```

[ ]:



6.3 Quantiles in Investment Practice

Phân vị được sử dụng trong đánh giá hiệu suất danh mục cũng như việc nghiên cứu và phát triển chiến lược đầu tư

Hiệu suất của các nhà quản lý quỹ thường được chuẩn hóa bởi phân vị tương ứng với hiệu suất của họ so với các nhà quản lý tương đương

Phân vị cũng thường được ứng dụng trong nghiên cứu đầu tư. Việc chia dữ liệu theo các phân vị dựa trên một số đặc điểm giúp nhà phân tích đánh giá tác động của đặc điểm đó tới một số lượng thuộc tính khác. Ví dụ, các nghiên cứu tài chính thực nghiệm xếp hạng các công ty dựa trên giá trị thị trường VCSH và phân nhóm dựa trên thập phân vị. Thập phân vị đầu tiên chứa danh mục đầu tư của các công ty có giá trị thị trường nhỏ nhất và thập phân vị thứ mười chứa danh mục của những công ty lớn nhất. Việc xếp hạng công ty như thế này cho phép các nhà phân tích so sánh hiệu suất của các công ty nhỏ so với các công ty lớn

7 Measures of Dispersion

7.1 The Range

Khoảng giá trị (`Range`) là khoảng cách giữa giá trị lớn nhất và giá trị nhỏ nhất của dữ liệu

$$Range = Max - Min$$

7.2 The Mean Absolute Deviation

Độ lệch trung bình tuyệt đối (MAD) bằng trung bình giá trị tuyệt đối độ lệch của tất cả các quan sát với giá trị trung bình của dữ liệu

$$MAD = \frac{\sum_{i=1}^{n} |X_i - \overline{X}|}{n}$$

MAD là một cách giải quyết vấn đề về dấu của các độ lệch (rõ ràng chúng ta có $\sum_{i=1}^{n} (X_i - \overline{X}) = 0$)

7.3 Sample Variance and Sample Deviation

Một cách tiếp cận khác để giải quyết vấn đề về dấu của các độ lệch chính là sử dụng bình phương các độ lệch. Với cách này, chúng ta có các đại lượng là **phương sai mẫu** và **độ lệch chuẩn mẫu**

7.3.1 Phương sai mẫu

$$s^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$$

7.3.2 Độ lệch chuẩn mẫu

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}}$$

Ở đây, phương sai mẫu s^2 và độ lệch chuẩn mẫu s là các *ước lượng không chệch* của phương sai tổng thể σ^2 và độ lệch chuẩn tổng thể σ

8 Downside Deviation and and Coefficient of Variation

8.1 Downside Deviation

Phương sai/Độ lệch chuẩn tỷ suất sinh lợi của tài sản thường được xem như thước đo rủi ro của tài sản. Hai đại lượng này tính đến cả các giá trị lớn hơn và nhỏ hơn giá trị trung bình (nghĩa là xét cả rủi ro tăng giá và rủi ro giảm giá). Tuy nhiên, các nhà đầu tư thường chỉ quan tâm đến rủi ro giảm giá (`downside risk`). Điều đó có nghĩa là phân tích viên cần phát triển các thước đo rủi ro giảm giá

Trên thực tế, chúng ta thường lo lắng về các giá trị nhỏ hơn một mức B nào đó hơn là giá trị trung bình $\overline{X}$. Độ lệch chuẩn dưới (`target downside deviation` hay `target semideviation`) là một thước đo phân tán cho các quan sát ở bên dưới mức mục tiêu

$$s_{2nd} = \sqrt{\sum_{i=1}^{n} \frac{\left(Min(X_i - B, 0)\right)^2}{n-1}}$$

8.2 Coefficient of Variation

Trong một số trường hợp, chúng ta khó có thể sử dụng dộ lệch chuẩn để giải thích mức bộ biến đổi tương đối giữa các bộ dữ liệu khác nhau, bởi vì giá trị trung bình có sự khác biệt rõ rệt và/hoặc có sự khác biệt về đơn vị đo. Đó là lý do chúng ta cần sử dụng các thước đo phân tán tương đối (`relative dispersion`) - đại lượng mô tả mức độ phân tán so với một tham chiếu, để giải thích trong các trường hợp này

Hệ số biến thiên (`Coefficient of Variation` - CV) là một thước đo phân tán tương đối tiêu biểu

$$CV = \frac{s}{\overline{X}}$$

Lấy một ví dụ, khi các quan sát là tỷ suất sinh lợi, hệ số biến thiên đo lường mức rủi ro trên mỗi đơn vị lợi nhuận. Tuy nhiên, không phải lúc nào thống kê này cũng có ý nghĩa, cụ thể khi $\overline{X}$ âm