Statistical Measures of Asset Returns

August 9, 2023

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

1 Measures of Central Tendency and Location

(xem thêm Prerequisite 2 / Quantitative)

Thước đo xu hướng tập trung của dữ liệu (`measures of central tendency`) chỉ ra nơi dữ liệu có xu hướng tập trung

Thước đo vị trí của dữ liệu (`measures of location`) không chỉ đo lường xu hướng tập trung của dữ liệu, mà còn mô tả nhiều phương diện khác về vị trí và/hoặc phân phối của dữ liệu

1.1 Measures of Central Tendency

Các thước đo xu hướng tập trung của dữ liệu thường gặp bao gồm:

1. Trung bình cộng (`The Arithmetic Mean`)
2. Trung vị (`The Median`)
3. Mode (`The Mode`)

Trong dữ liệu, có thể có các ngoại lai gây ảnh hưởng đến các giá trị kể trên. Phân tích viên có thể xử lý vấn đề này theo 3 cách: xóa bỏ các giá trị ngoại lai, thay thế các giá trị ngoại lai, hoặc không làm gì cả

1.2 Measures of Location

Phân vị (`Quantile`) là thước đo thường được sử dụng để nghiên cứu về vị trí của dữ liệu. Các phân vị phổ biến có thể kể đến là:

1. Tứ phân vị (`Quartile`)
2. Ngũ phân vị (`Quintile`)
3. Thập phân vị (`Decile`)
4. Bách phân vị (`Percentile`)

Biểu đồ hộp (`Box and Whisker Plot`) thường được sử dụng để mô tả phân phối của dữ liệu

2 Measures of Dispersion

(xem thêm Prerequisite 2 | Quantitative)

Sự phân tán của dữ liệu mô tả phân phối dữ liệu xung quanh nơi dữ liệu tập trung. Giá trị này đại diện cho **sự rủi ro** và/hoặc ** sự không chắc chắn**

Thước đo phân tán có thể được phân thành hai loại: thước đo phân tán tuyệt đối (`measures of absolute dispersion`) và thước đo phân tán tương đối (`measures of relative dispersion`)

Các thước đo phân tán thường gặp có thể kể đến là:

1. Khoảng dữ liệu (`Range`)
2. Trung bình độ lệch tuyệt đối (`Mean Absolute Deviation` - MAD)
3. Phương sai (`Variance`) và độ lệch chuẩn (`Standard Deviation`)
4. Độ lệch dưới chuẩn (`Downside Deviation` hay `Target Semideviation`)
5. Hệ số biến thiên (`Coefficient of Variation`)

3 Measures of Shape of a Distribution

Các giá trị trung bình và phương sai là chưa đủ để mô tả sự phân phối của một biến liên tục, ví dụ như việc chúng ta không thể xác định được độ lệch cụ thể của từng quan sát là dương hay âm. Do đó, ngoài việc quan tâm đến các thước đo xu hướng tập trung, vị trí và độ phân tán, chúng ta cần nghiên cứu thêm các đặc điểm quan trọng khác của phân phối (ví dụ như sự đối xứng hay độ nhọn của dữ liệu)

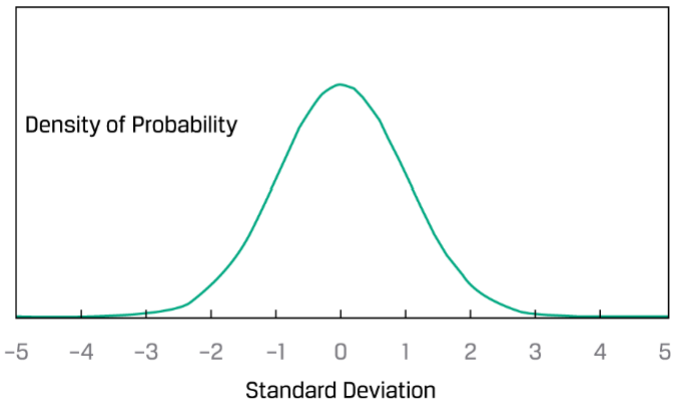
3.1 Normal Distribution

Phân phối chuẩn là một trong những phân phối dữ liệu thông dụng và quan trọng nhất; đóng vai trò trung tâm trong việc lựa chọn danh mục đầu tư dựa trên mô hình trung bình - phương sai (`mean - variance model`). Các đặc điểm quan trọng của phân phối chuẩn bao gồm:

- Phân phối đối xứng, có hình chuông
- Trung bình, trung vị và mode bằng nhau
- Hoàn toàn xác định bằng hai yếu tố: trung bình và phương sai của dữ liệu

[ ]:
```
# Normal Distribution
Image(filename = "Pictures/01.png")
```

[ ]:



Tuy nhiên, việc xác định hình dạng của các phân phối khác thường yêu cầu nhiều thông tin hơn.

3.2 Skewness

Một phân phối không đối xứng là một phân phối nghiêng (`skewed distribution`). Đối với phân phối chỉ có một mode, nếu:

- Phân phối lệch dương: $mode < median < mean$
- Phân phối lệch âm: $mode > median > mean$

Độ nghiêng (`Skewness`) được sử dụng để đo lường độ nghiêng của phân phối. Nó được tính bằng trung bình lập phương độ lệch với giá trị trung bình (nhằm giữ được dấu của độ lệch) và được chuẩn hóa bằng lập phương của độ lệch chuẩn mẫu để đảm bảo rằng đại lượng này vô hướng

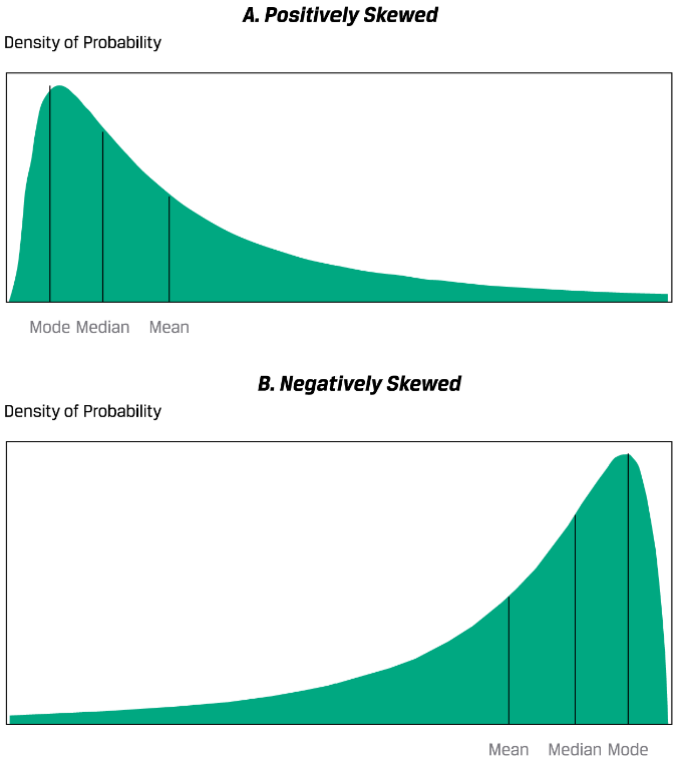
Nếu một phân phối bị lệch dương, điều đó có nghĩa là trung bình lớn hơn trung vị và hơn một nửa độ lệch mang dấu âm. Tuy nhiên tổng lập phương các độ lệch lại mang dấu dương cho thấy rằng các độ lệch dương dễ mang các giá trị cực đoan hơn so với các độ lệch âm. Cách giải thích tương tự cũng được áp dụng nếu phân phối bị lệch âm

Với mẫu lớn ($n > 100$), độ nghiêng mẫu có thể được xác định gần đúng bởi công thức

$$Skewness \approx \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^3}{ns^3}$$

[ ]:
```
# Positively and Negatively Skewed Distributions
Image(filename = "Pictures/02.png")
```

[ ]:



3.3 Kurtosis

Một lý do khác của sự khác biệt giữa một phân phối nào đó so với phân phối chuẩn chính là xu hướng tương đối của nó tạo ra các độ lệch lớn hơn so với giá trị trung bình. Rõ ràng rằng, xác suất xảy ra độ lệch cực lớn so với giá trị trung bình cao hơn hàm ý mức độ rủi ro cao hơn

Độ nhọn (`Kurtosis`) là thước đo trọng số kết hợp của các đuôi trong một phân phối so với phần còn lại của phân phối đó, nghĩa là tỷ lệ của tổng xác suất nằm ngoài một mức độ lệch chuẩn nào đó so với trung bình (ví dụ như xác suất nằm ngoài 2.5 độ lệch chuẩn so với trung bình)

Với mẫu lớn ($n > 100$), độ nhọn mẫu và độ nhọn mở rộng mẫu có thể được tính gần đúng như sau:

$$K \approx \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^4}{ns^4}$$

$$K_E = K - 3$$

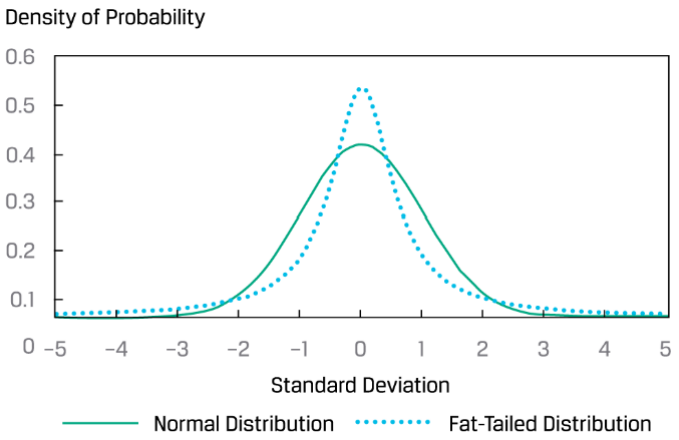
Có ba dạng độ nhọn thường gặp:

- Leptokurtic (fat-tailed): phân phối nhọn hơn phân phối chuẩn, có nghĩa là xác suất xuất hiện các giá trị xung quanh trung bình và các độ lệch cực đoan cao hơn so với phân phối chuẩn. Giá trị `Kurtosis` ứng với phân phối này lớn hơn 3.0

- Platykurtic (thin-tailed): phân phối ít nhọn hơn phân phối chuẩn, có nghĩa là xác suất xuất hiện các giá trị xung quanh trung bình và các độ lệch cực đoan thấp hơn so với phân phối chuẩn. Giá trị `Kurtosis` ứng với phân phối này nhỏ hơn 3.0
- Mesokurtic: phân phối có độ nhọn tương tự với phân phối chuẩn. Giá trị `Kurtosis` ứng với phân phối này xấp xỉ 3.0

[ ]:
```
# Leptokurtic Distribution
Image(filename = "Pictures/03.png")
```

[ ]:



4 Correlation between Two Variables

4.1 Scatter Plot

Biểu đồ phân tán là một công cụ hữu ích để mô tả các mối quan hệ tiềm năng giữa hai biến

Biểu đồ phân tán mang đến nhiều thông tin có giá trị bất chất cấu trúc đơn giản của bản thân nó. Thứ nhất, nó biểu diễn các mối quan hệ tiềm năng giữa các biến (tuyến tính, phi tuyến tính hoặc không có quan hệ rõ ràng). Thứ hai, độ mạnh của mối quan hệ đó được xác định bằng mức độ tập trung của các điểm dữ liệu xung quanh một đường được vẽ dựa trên việc quan sát. Thứ ba, biểu đồ phân tán giúp chúng ta dễ xác định được khoảng giá trị và các ngoại lai tồn tại trong dữ liệu và từ đó đưa ra phương án xử lý phù hợp

4.2 Covariance and Correlation

4.2.1 Definition

Hiệp phương sai mẫu (`sample covariance` - s_{XY}) là một thước đo mô tả tương quan thay đổi của hai biến

$$s_{XY} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

Hệ số tương quan (`correlation` - r_{XY}) là một thước đo chuẩn hóa mô tả tương quan biến đổi của

hai biến trong cùng một mẫu. Hệ số tương quan mẫu được tính bằng cách lấy hiệp phương sai chia cho tích độ lệch chuẩn mẫu của hai biến

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

Hệ số tương quan mẫu thể hiện mức độ của mối quan hệ tuyến tính giữa hai biến ngẫu nhiên

4.2.2 Properties of Correlation

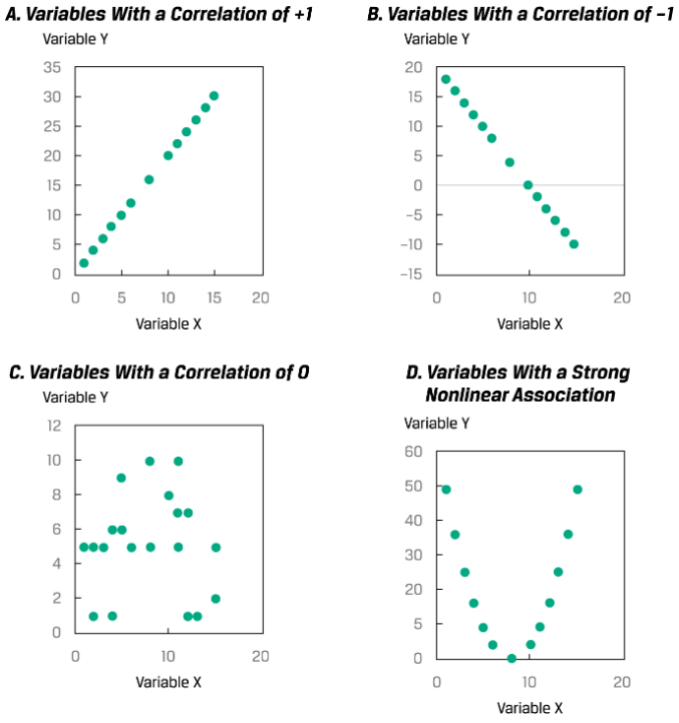
1. Với hai biến ngẫu nhiên X, Y, khoảng giá trị của hệ số tương quan luôn nằm trong đoạn từ -1 đến 1

$$-1 \leq r_{XY} \leq 1$$

2. Hệ số tương quan bằng 0 hàm ý rằng không có bất kỳ mối tương quan tuyến tính nào giữa các biến
3. Hệ số tương quan dương hàm ý rằng tồn tại mối tương quan tuyến tính đồng biến, hệ số này nhận giá trị bằng 1 khi có tương quan tuyến tính đồng biến tuyệt đối
4. Hệ số tương quan âm hàm ý rằng tồn tại mối tương quan tuyến tính nghịch biến, hệ số này nhận giá trị bằng -1 khi có tương quan tuyến tính nghịch biến tuyệt đối

[ ]:
```
Image(filename = "Pictures/04.png")
```

[ ]:



4.2.3 Limitations

1. Hệ số tương quan không thể xác định các tương quan phi tuyến tính giữa các biến (ví dụ như tương quan bậc 2, lũy thừa, logarit, nghịch đảo…)
2. Hệ số tương quan trở nên không đáng tin cậy nếu tồn tại ngoại lại ở ít nhất một trong hai biến
3. Mối tương quan không hàm ý quan hệ nhân quả giữa các biến (đó có thể chỉ là sự ngẫu nhiên!)

Thuật ngữ **tương quan giả mạo** (`spurious correlation`) được sử dụng để mô tả các hiện tượng sau:

- Mối tương quan giữa hai biến phản ánh mối quan hệ may rủi trong một tập dữ liệu cụ thể (ví dụ như xác suất trời mưa và xác suất chỉ số chứng khoán tăng điểm trong phiên)
- Mối tương quan gây ra bởi một phép tính kết hợp từng biến trong hai biến với một biến thứ ba (ví dụ như việc cổ tức và tổng tài sản của công ty có mối tương quan thấp, nhưng việc chuẩn hóa hai biến này cho vốn hóa thị trường khiến tương quan giữa chúng trở nên cao hơn)
- Tương quan phát sinh không phải bởi mối quan hệ trực tiếp giữa các biến mà từ mối quan hệ của chúng với biến thứ ba (ví dụ như việc chiều cao có tương quan dương với vốn từ vựng, nhưng mối quan hệ cơ bản ở đây có thể kể đến là tương quan giữa tuổi đời với lần lượt các biến chiều cao và vốn từ vựng)

Nhà đầu tư phải thực sự thận trọng trong việc xây dựng chiến lược đầu tư dựa trên các mối tương quan cao. Các tương quan giả mạo có thể gợi ý các chiến lược sinh lãi cao trên giấy tờ nhưng lại không hiệu quả trên thực tế

Việc sử dụng các hình ảnh, biểu đồ (ví dụ như biểu đồ phân tán) có thể khiến chúng ta đưa ra các phán đoán vô thức về giá trị nhân quả giữa các biến