Lecture 2

Part 1 Descriptive Statics, Summary Measures, and Data Visualization

Arithmetic Mean

$$\text{Arithmetic Mean} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- The arithmetic mean is the sum of all data points divided by the number of points.
- It is the most commonly used mean and represents a balance point of the data.

Example:

- For the data set $\{3, 5, 7, 9\}$, the arithmetic mean is $\frac{3+5+7+9}{4} = 6$.

Geometric Mean

$$\text{Geometric Mean} = \left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}}$$

- The geometric mean is the nth root of the product of all data points.
- It is useful for data that are multiplicative or vary exponentially.

Example:

- For the data set $\{1, 3, 9\}$, the geometric mean is $(1 \times 3 \times 9)^{\frac{1}{3}} = 3$.

Harmonic Mean

$$\text{Harmonic Mean} = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$

- The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals.
- It is useful for rates and ratios, such as speed or density.

Example:

- For the data set $\{2, 3, 4\}$, the harmonic mean is $\frac{3}{\frac{1}{2}+\frac{1}{3}+\frac{1}{4}} \approx 2.77$.

Harmonic Mean

The harmonic mean can be used to calculate the average speed when the distances are the same but the speeds differ. For example, if for a certain journey, the speed is 60 mph for the first half and 30 mph for the second half (with both distances being equal), the average speed is the harmonic mean of the two speeds, which is 40 mph.

Assume the distance to be a, then the average speed is:

$$\frac{\text{the distance}}{\text{the time spent}} = \frac{a}{\frac{0.5a}{30} + \frac{0.5a}{60}} = \frac{2}{\frac{1}{30} + \frac{1}{60}} = 40$$

Variance

- What we just calculated is known as the **variance**,which measures the spread or variability of a given distribution of data.
- Population variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

- Sample variance:

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

Why Sample Variance has $n-1$ on denominator?

Recall the terms:

$$\begin{aligned}\mu &= \text{population mean}\\ \sigma^2 &= \text{population variance}\\ \bar{X} &= \text{sample mean}\\ s^2 &= \text{sample variance}\\ X_1, X_2, \cdots, X_n &= \text{sample values}\end{aligned}$$

Why Sample Variance has $n - 1$ on denominator?

Let's start from this...

Flip a coin: if it lands heads up, you win \$1; if it lands tails up, you win \$0. Flip the coin ten times, how much do you think you can win?

What about if you flip it 100 or 1000 times? Do you think it will converge to a certain value as the number of flips increases?

Law of Large Numbers

Law of large numbers:

If you repeat an experiment independently a large number of times and average the result, what you obtain should be close to the expected value.
In other words, if you take an infinite number of samples, the mean of the sample statistics will increasingly approach the population parameter.

Link to proof if you are interested

Why Sample Variance has $n-1$ on denominator?

From the Law of large numbers we know: If we sample infinitely many times and calculate the average of mean and variance we get:

$$\mathbb{E}[\bar{X}] = \mu$$

Then what about

$$\mathbb{E}[\sum_{i=1}^{n}(X_i - \bar{X})^2]$$

Note: $\mathbb{E}[\cdot]$ is a sign that denotes mathematical expectations.

Recall the population variance formula:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

Sample variance:

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

Rewrite $\mathbb{E}[\sum_{i=1}^{n}(X_i - \bar{X})^2]$ as:

$$\mathbb{E}[\sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X})^2]$$

Expand the squared term:

$$\mathbb{E}[\sum_{i=1}^{n}((X_i - \mu)^2 + 2(X_i - \mu)(\mu - \bar{X}) + (\mu - \bar{X})^2)]$$

Simplify:

$$\mathbb{E}[\sum_{i=1}^{n}(X_i - \mu)^2] + \mathbb{E}[\sum_{i=1}^{n} 2(X_i - \mu)(\mu - \bar{X})] + \mathbb{E}[\sum_{i=1}^{n}(\mu - \bar{X})^2]$$

The middle term equals zero because $\mathbb{E}[\bar{X}] = \mu,$ so:

$$\mathbb{E}[\sum_{i=1}^{n}(X_i - \mu)^2] + \mathbb{E}[n(\mu - \bar{X})^2]$$

which is

$$\mathbb{E}[\sum_{i=1}^{n}(X_i - \mu)^2] + n\mathbb{E}[(\mu - \bar{X})^2]$$

We know that $\mathbb{E}[\sum_{i=1}^{n}(X_i - \mu)^2] = n\sigma^2$ and $\mathbb{E}[(\mu - \bar{X})^2] = \frac{\sigma^2}{n},$ so:

$$n\sigma^2 + n(\frac{\sigma^2}{n}) = n\sigma^2 + \sigma^2$$

Therefore:

$$\mathbb{E}[\sum_{i=1}^{n}(X_i - \bar{X})^2] = (n-1)\sigma^2$$

If we use n in the denominator:

$$\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2] = \frac{n-1}{n}\sigma^2$$

This is biased because it's less than the true population variance σ^2.

But if we use $(n-1)$ in the denominator:

$$\mathbb{E}[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2] = \sigma^2$$

This gives us an unbiased estimator of the population variance.

Standard Deviation

- Since we squared the observations when calculating the variance, the units for variance is the "original units" squared.
- The standard deviation is defined as the square-root of the variance.
- Conveys same information as variance, but is now in the "original units" of measure.

Standard Deviation

- Population standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}$$

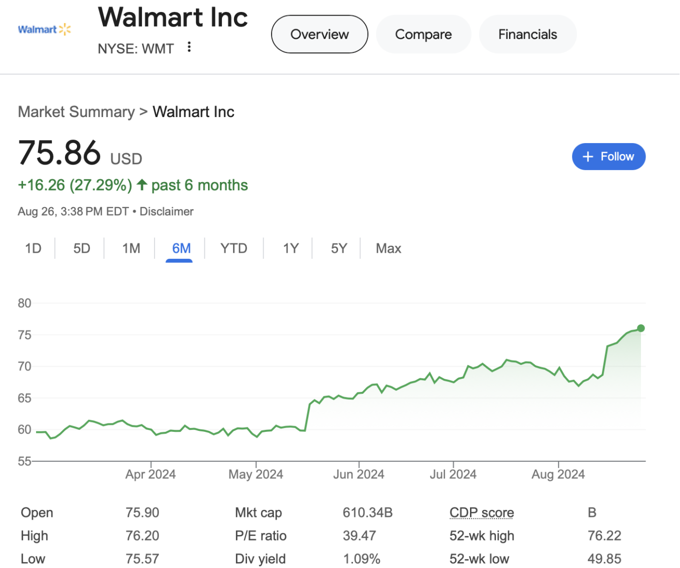
- Sample standard deviation:

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

Application in Finance

- Understanding variability is very important in finance.
- Variance (or standard deviation) of an investment can be used as a measure of risk, e.g., on profits/return.
- Larger variance implies larger risk.
- Usually, the higher the expected rate of return, the higher the risk.

Walmart



Walmart Inc

NYSE: WMT

Overview

Compare

Financials

Market Summary > Walmart Inc

75.86 USD

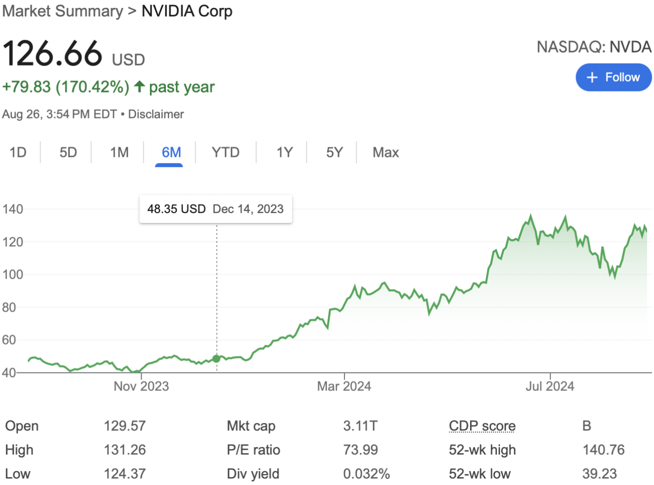
+16.26 (27.29%) ↑ past 6 months

Aug 26, 3:38 PM EDT • Disclaimer

+ Follow

Open	75.90	Mkt cap	610.34B	CDP score	B
High	76.20	P/E ratio	39.47	52-wk high	76.22
Low	75.57	Div yield	1.09%	52-wk low	49.85

NVIDIA



Market Summary > NVIDIA Corp

126.66 USD

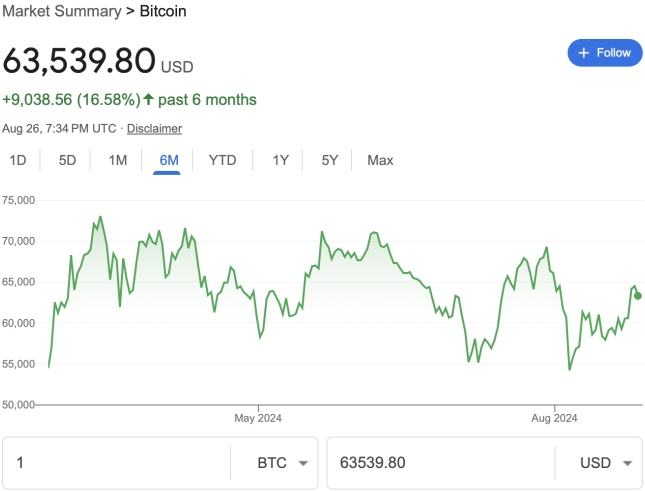
+79.83 (170.42%) ↑ past year

Aug 26, 3:54 PM EDT • Disclaimer

NASDAQ: NVDA

Open	129.57	Mkt cap	3.11T	CDP score	B
High	131.26	P/E ratio	73.99	52-wk high	140.76
Low	124.37	Div yield	0.032%	52-wk low	39.23

Bitcoin



Market Summary > Bitcoin

63,539.80 USD

+9,038.56 (16.58%)↑ past 6 months

Aug 26, 7:34 PM UTC · Disclaimer

1 BTC

63539.80 USD

Application in Finance

Rates of returns for two stocks X and Y.

X	8.3	-6.2	20.9	-2.7	33.6	42.9	24.4	5.2	3.1	30.5
Y	12.1	-2.8	6.4	12.2	27.8	25.3	18.2	10.7	-1.3	11.4

$\bar{X} = 16\%$ and $s_X^2 = 280.34\%^2$.
$\bar{Y} = 12\%$ and $s_Y^2 = 99.37\%^2$.

Application in Finance: Sharpe Ratio

- It is a financial metric used to evaluate the risk-adjusted return of an investment or portfolio.
- The ratio helps investors understand how much excess return they are receiving for the extra volatility they endure for holding a riskier asset.

Sharpe Ratio

$$\text{Sharpe Ratio} = \frac{R_p - R_f}{\sigma_p}$$

- R_p: Return of a portfolio (portfolio here means combination of financial assets).
- R_f: Risk-free rate (typically the return of a government bond).
- σ_p: Standard deviation of the portfolio's excess return (a measure of risk).

Sharpe Ratio

- High Sharpe Ratio: Indicates that the investment is providing a higher return for the risk taken. It is generally seen as a sign of good investment performance.
- Low or Negative Sharpe Ratio: Indicates that the returns are not sufficient to justify the risk taken, or the investment is underperforming relative to a risk-free investment.

Sharpe Ratio

Rates of returns for two stocks X and Y, assume a risk-free rate (CDs or bond rate) of 3%.

X	8.3	-6.2	20.9	-2.7	33.6	42.9	24.4	5.2	3.1	30.5
Y	12.1	-2.8	6.4	12.2	27.8	25.3	18.2	10.7	-1.3	11.4

$X - R_f$	5.3	-9.2	17.9	-5.7	30.6	39.9	21.4	2.2	0.1	27.5
$Y - R_f$	9.1	-5.8	3.4	9.2	24.8	22.3	15.2	7.7	-4.3	8.4

$\bar{X} - R_f = 13\%$ and $s^2_{X-R_f} = 280.34\%^2$.
$\bar{Y} - R_f = 9\%$ and $s^2_{Y-R_f} = 99.37\%^2$.

Question: Why s^2 remains the same?

Sharpe Ratio

The standard deviation is $s_{X-R_f} = 16.74\%$ and $s_{Y-R_f} = 9.97\%$.

So we have the Sharpe Ratios:

- For stock X:

$$\frac{\bar{X} - R_f}{s_{X-R_f}} = \frac{13\%}{16.74\%} = 0.78$$

- For stock Y:

$$\frac{\bar{Y} - R_f}{s_{Y-R_f}} = \frac{9}{9.97} = 0.90$$

Coefficient of Variation

The **coefficient of variation** is a measure of variability defined as the standard deviation divided by the mean.

- Population:

$$CV = \frac{\sigma}{\mu}$$

- Sample:

$$cv = \frac{s}{\bar{X}}$$

- Takes into account scale or magnitude of your data.

Covariance

Covariance is a measure of the linear relationship between two variables and describes how they move in relates to one another.

- Why might this be important?
- Example: We have two stocks, X and Y. Let's suppose that X moves exactly with the market,whereas Y moves opposite to the market.

Covariance

- Scenario 1: You invest all your money in stock X. The market does really well. Stock X pays off \$1000.
- Scenario 2: You invest all your money in stock Y. The market crashes. Stock Y pays off \$1000.
- Covariance is very important in forming a portfolio.

Covariance

- Population covariance:

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)$$

- Sample covariance:

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

Covariance

Another example.

- Let's say we're interested in how closely fathers' heights are related to sons' heights.
- The logical thing to do is to calculate the covariance of fathers' and sons' heights, right?
- Suppose we do that and find the covariance to be $20cm^2$.

Covariance

- What if we're also interested in the relationship between the number of years of education of mothers and daughters?
- Once again, suppose we calculate the covariance of mothers' and daughters' years of education, and find it to be 3.5 years^2.
- Now, are fathers' and sons' heights more closely related to each other than mothers' and daughters' years of education?

Correlation Coefficient

- **Correlation** is also a measure of the linear relationship between two variables.
- Population correlation coefficient:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- Sample correlation coefficient:

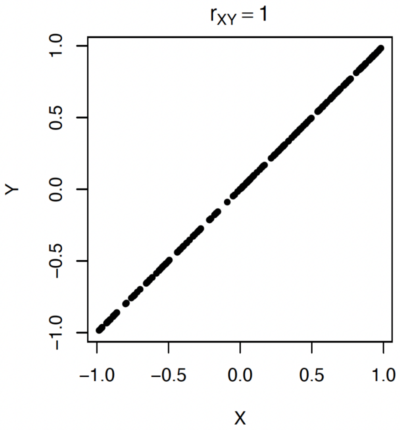
$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

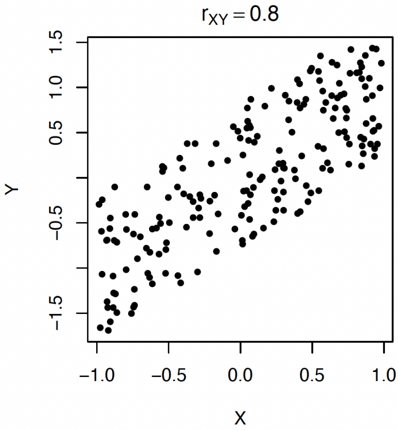
Correlation Coefficient

- A correlation coefficient standardizes the covariance of two variables such that its value will lie between -1 and 1 (inclusive).
- The interpretation is more intuitive than covariance,and direct comparisons between variables in different units of measure can be conducted.
- So, we could compare the correlation between fathers' and sons' heights to the correlation between mothers' and daughters' years of education.

Scatter Plots

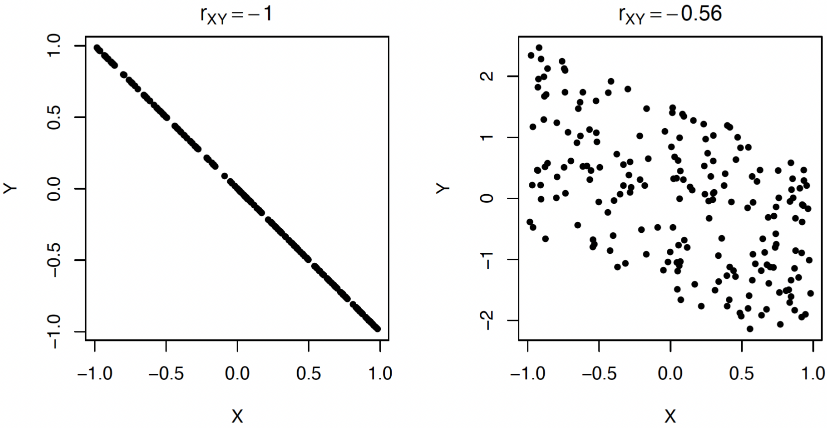
If $0 < r_{XY} \leq 1$ then we have a positive linear relationship, with the strength dependent on how close r_{XY} is to 1:





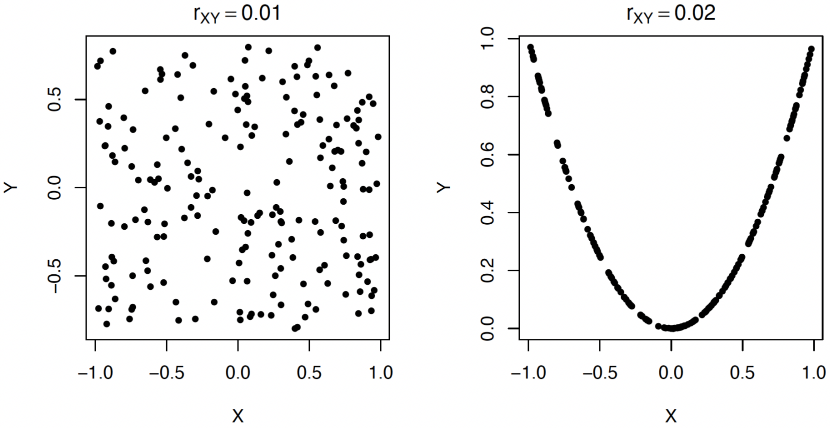
Scatter Plots

If $-1 \leq r_{XY} < 0$ then we have a negative linear relationship, with the strength dependent on how close r_{XY} is to -1:



Scatter Plots

If $r_{XY} \approx 0$ then we have no linear relationship:



Correlation and Causation

Note that correlation does not imply causation.

Examples of variables that could be correlated but no causal relationship:

- Number of ice cream sales and the rate of drowning deaths.