Data and financial modelling

Lianghai Xiao

https://github.com/styluck/mat_fin

Outline

- **Organizing, Visualizing, and Describing Data**
- Probability Concepts
- Common Probability Distributions
- Sampling and Estimation
- Hypothesis Testing
- Introduction to Linear Regression

组织

ORGANIZING DATA

- The term **data** encompasses **information** in any form.
- We may classify data types from three different perspectives:
  - **Numerical** versus **categorical**.
  - **Time series** versus **cross-sectional**.
  - **Structured** versus **unstructured**.

Numerical and Categorical Data

数值

- **Numerical data**, or **quantitative data**, are values that can be counted or measured.

标签

- **Categorical data**, or **qualitative data**, consist of labels that can be used to classify a set of data into groups. Categorical data may be nominal or ordinal.

mutual funds -> 共同基金 -> 公募基金

Nominal data and Ordinal data

- **Nominal data** are labels that cannot be placed in order logically.
  - Fixed-income mutual funds may be classified as corporate bond funds, municipal bond funds, international bond funds, and so on.
- **Ordinal data** can be ranked in a logical order. Every item is assigned to one of multiple categories based on a specific characteristic, then these categories are ordered with respect to that characteristic.

  capitalization 市值

  - The ranking of 1,000 small-cap growth stocks by performance may be done by assigning the number 1 to the 100 best-performing stocks, the number 2 to the next 100 best-performing stocks, and so on through the number 10 for the 100 worst-performing stocks.

Time Series and Cross-Sectional Data

时间 序列

- A **time series** is a set of observations taken periodically, most often at equal intervals over time. Daily closing prices of a stock over the past year and quarterly earnings per share of a company over a five-year period are examples of time series data.

截面

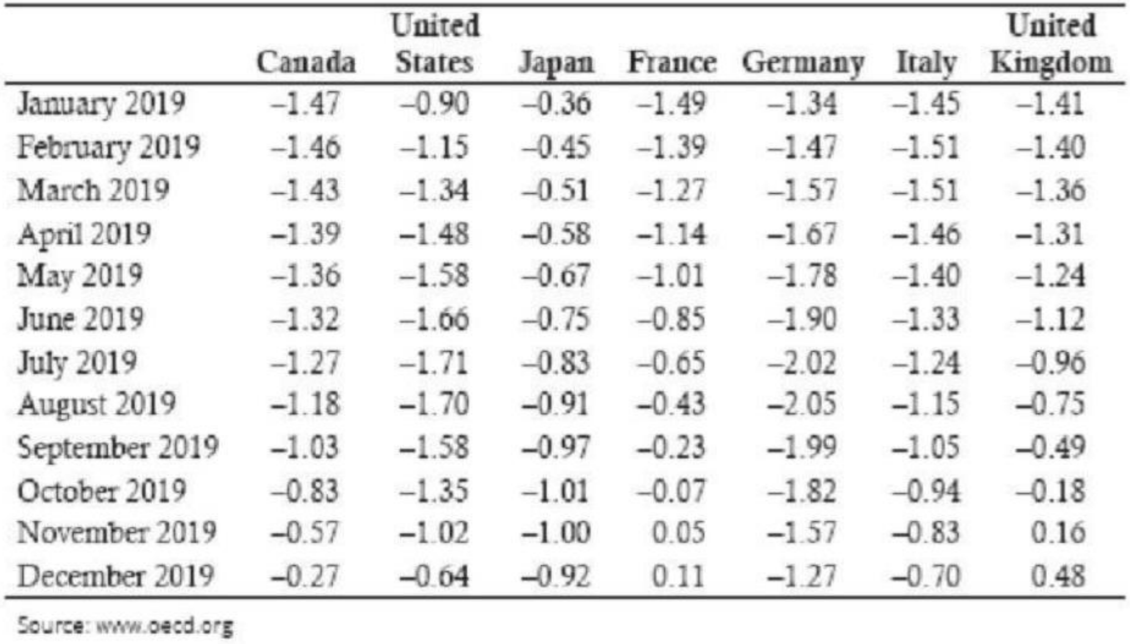
- **Cross-sectional** data refers to a set of comparable observations all taken at one specific point in time. Today's closing prices of the 30 stocks in the Dow Jones Industrial Average and fourth-quarter earnings per share for 10 health care companies are examples of cross-sectional data.

面板数据

Panel data

- Time series and cross-sectional data may be combined to form panel data. Panel data are often presented in tables.

Figure 2.1: OECD Composite Leading Indicators, Year-on-Year Growth Rate



	Canada	United States	Japan	France	Germany	Italy	United Kingdom
January 2019	–1.47	–0.90	–0.36	–1.49	–1.34	–1.45	–1.41
February 2019	–1.46	–1.15	–0.45	–1.39	–1.47	–1.51	–1.40
March 2019	–1.43	–1.34	–0.51	–1.27	–1.57	–1.51	–1.36
April 2019	–1.39	–1.48	–0.58	–1.14	–1.67	–1.46	–1.31
May 2019	–1.36	–1.58	–0.67	–1.01	–1.78	–1.40	–1.24
June 2019	–1.32	–1.66	–0.75	–0.85	–1.90	–1.33	–1.12
July 2019	–1.27	–1.71	–0.83	–0.65	–2.02	–1.24	–0.96
August 2019	–1.18	–1.70	–0.91	–0.43	–2.05	–1.15	–0.75
September 2019	–1.03	–1.58	–0.97	–0.23	–1.99	–1.05	–0.49
October 2019	–0.83	–1.35	–1.01	–0.07	–1.82	–0.94	–0.18
November 2019	–0.57	–1.02	–1.00	0.05	–1.57	–0.83	0.16
December 2019	–0.27	–0.64	–0.92	0.11	–1.27	–0.70	0.48

Source: www.oecd.org

Structured and Unstructured Data

- Time series, cross-sectional, and panel data are examples of **structured data**—they are organized in a defined way.
  - Market data, such as security prices
- **Unstructured data** refers to information that is presented in a form with no defined structure.
  - Management's commentary in company financial statements is an example of unstructured data

管理层致词

公司财报

Frequency and related distributions

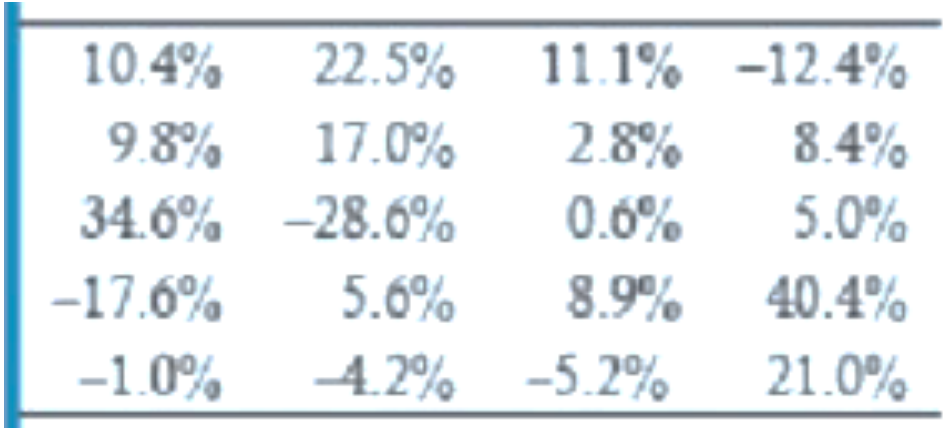
- A **frequency distribution** is a tabular presentation of statistical data that aids the analysis of large data sets. Frequency distributions summarize statistical data by assigning them to specified groups, or intervals.

Frequency and related distributions

- Procedure to construct a frequency distribution:
  - **Define the intervals.**
    - to which data measurements (observations) will be assigned.
    - Intervals must be **mutually exclusive** so that each observation can be placed in only one interval,
    - The total set of intervals should cover the **total range of values** for the entire population.
  - **Tally the observations.** After the intervals have been defined, the observations must be tallied or assigned to their appropriate interval.
  - **Count the observations**.The **absolute frequency**, or simply the frequency is the actual number of observations that fall within a given interval.

Example: Constructing a frequency distribution

- Use the data in Table A to construct a frequency distribution for the returns on Intelco's common stock.



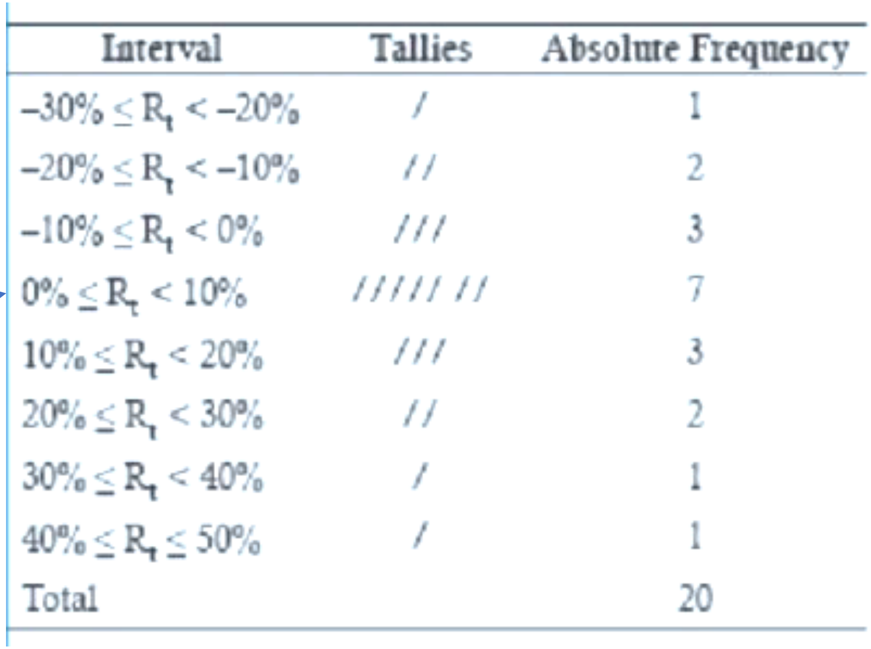
10.4%	22.5%	11.1%	–12.4%
9.8%	17.0%	2.8%	8.4%
34.6%	–28.6%	0.6%	5.0%
–17.6%	5.6%	8.9%	40.4%
–1.0%	–4.2%	–5.2%	21.0%

Example: Constructing a frequency distribution

- **Step 1: Defining the interval.** the range of returns is 69.0% (-28 6% to 40.4%) Using a return interval of 1% would result in 69 separate intervals, which in this case is too many. So let’s use eight non-overlapping intervals with a width of 10%. The lowest return intervals will be -30% ≤ R < -20%, and the intervals will increase to 40%≤R≤50%.

Example: Constructing a frequency distribution

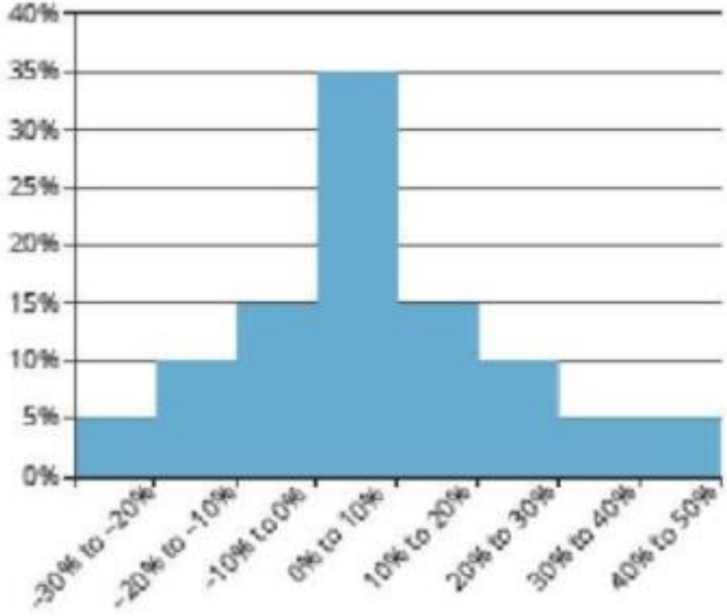
- Step 2-3: Tally the observations and count the observations within each interval.

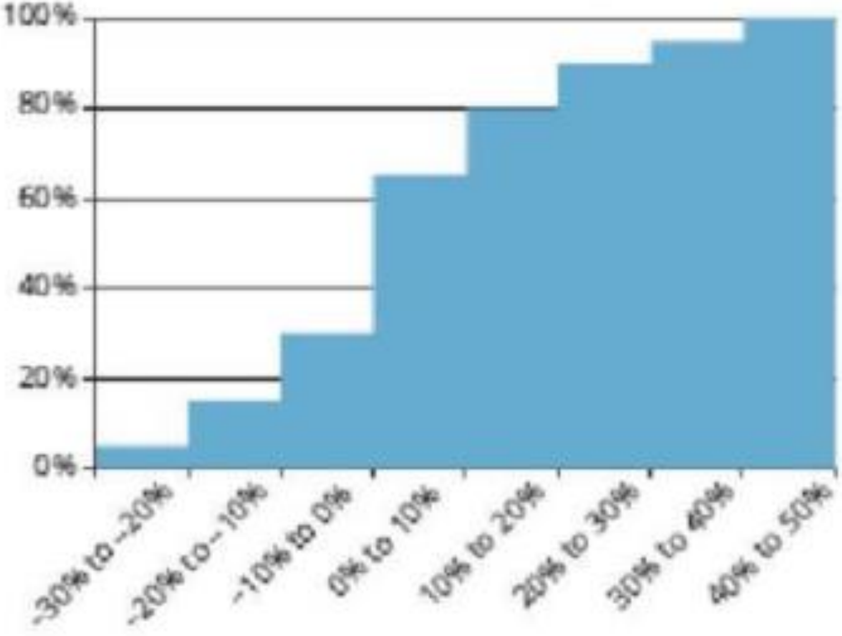


Interval	Tallies	Absolute Frequency
$-30\% \leq R_t < -20\%$	/	1
$-20\% \leq R_t < -10\%$	//	2
$-10\% \leq R_t < 0\%$	///	3
$0\% \leq R_t < 10\%$	///// //	7
$10\% \leq R_t < 20\%$	///	3
$20\% \leq R_t < 30\%$	//	2
$30\% \leq R_t < 40\%$	/	1
$40\% \leq R_t \leq 50\%$	/	1
Total		20

modal interval: the interval with the greatest frequency

Example: Constructing a frequency distribution





VISUALIZING DATA

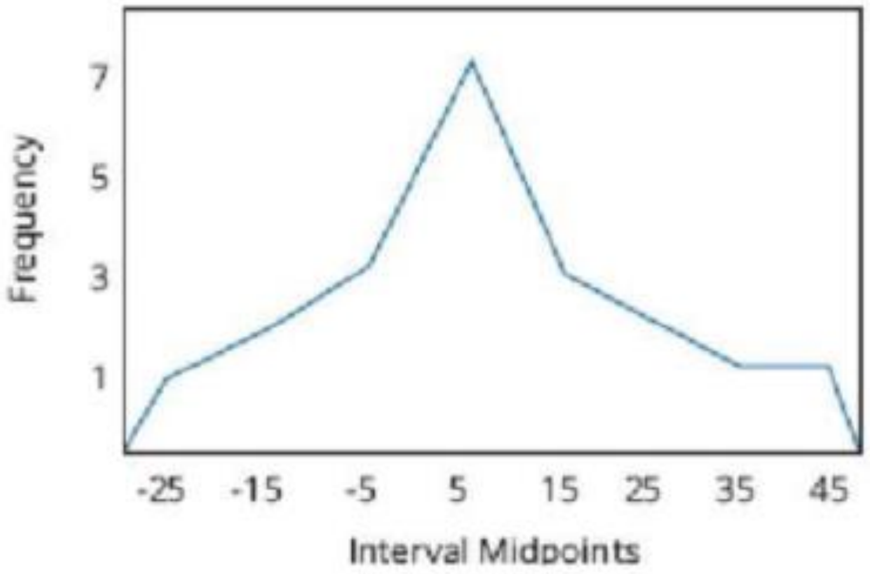
数据可视化

直方图

- A **histogram** is the graphical presentation of the absolute frequency distribution. A histogram is simply a bar chart of continuous data that has been classified into a frequency distribution. The attractive feature of a histogram is that it allows us to quickly see where most of the observations are concentrated.

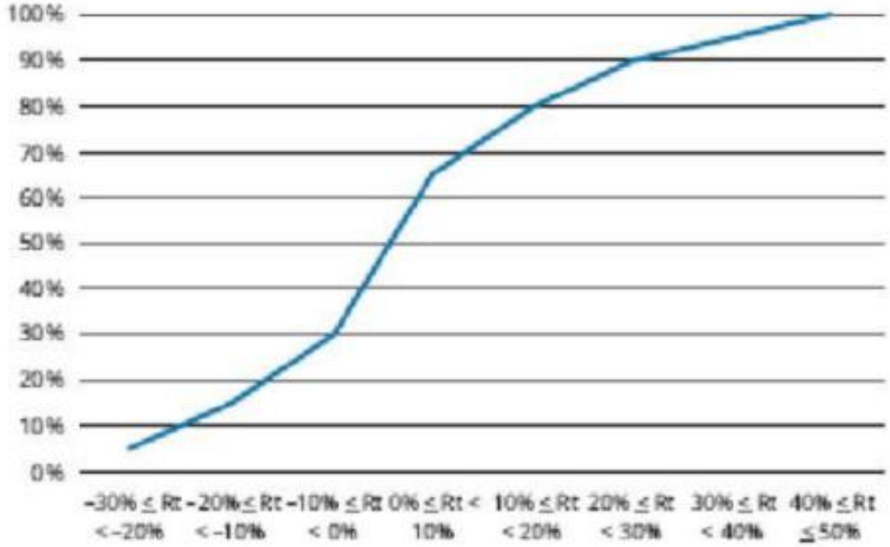
VISUALIZING DATA

- To construct a **frequency polygon**, successive frequencies at the midpoints of the intervals are joined with line segments.



VISUALIZING DATA

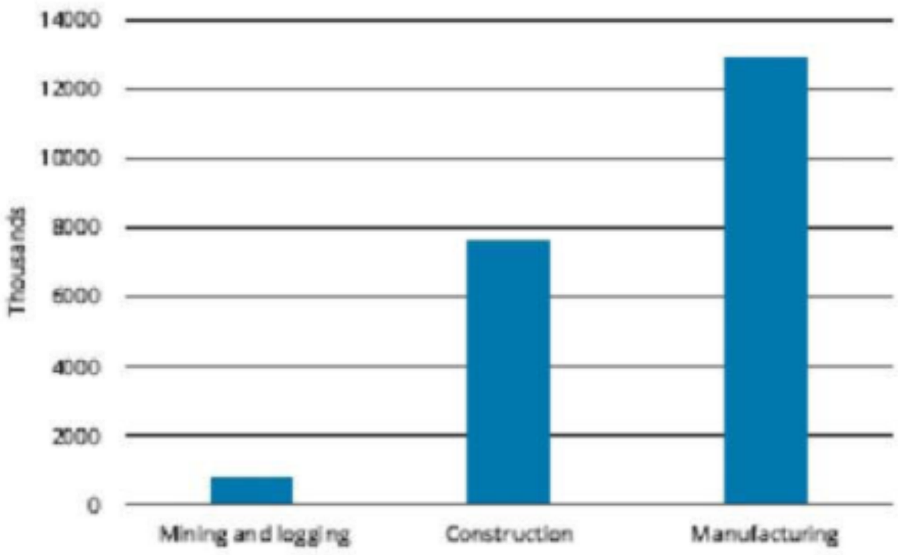
- A **cumulative frequency distribution** chart displays either the cumulative absolute frequency or the cumulative relative frequency.



VISUALIZING DATA

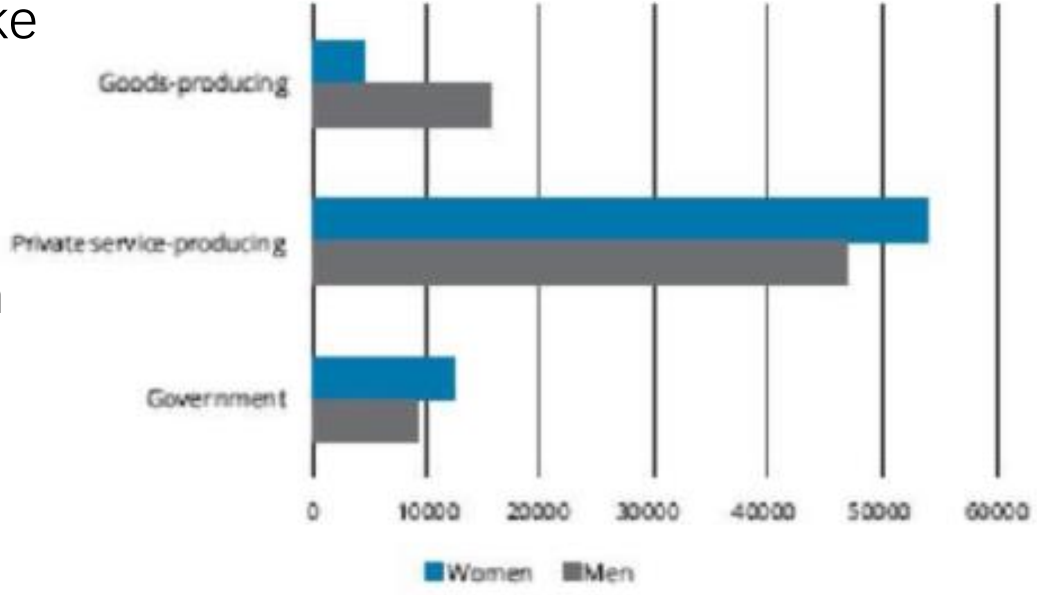
- The histogram shown earlier is an example of **a bar chart**. In general, bar charts are used to illustrate relative sizes, degrees, or magnitudes. The bars can be displayed vertically or horizontally. Figure shows a bar chart of employment in goods-producing industry groups in the United States. From this chart, we can see that the construction industries employ about 10 times as many people as the mining and logging industries and that manufacturing payrolls are a bit less than twice as large as construction payrolls.

柱状图



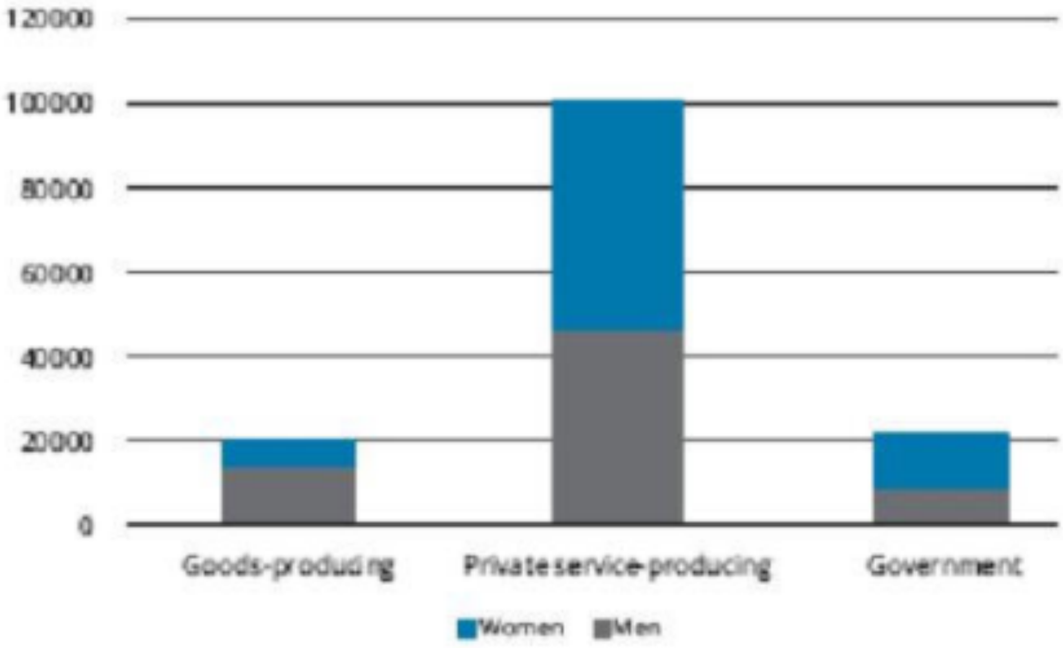
VISUALIZING DATA

- A **grouped bar chart** or **clustered bar chart** can illustrate two categories at once, much like a data table. Figure displays the number of men and women employed in three segments of the U.S. economy. Here we can see that more men than women are employed in the goods producing industries, but more women than men are employed in the service-producing industries and government.



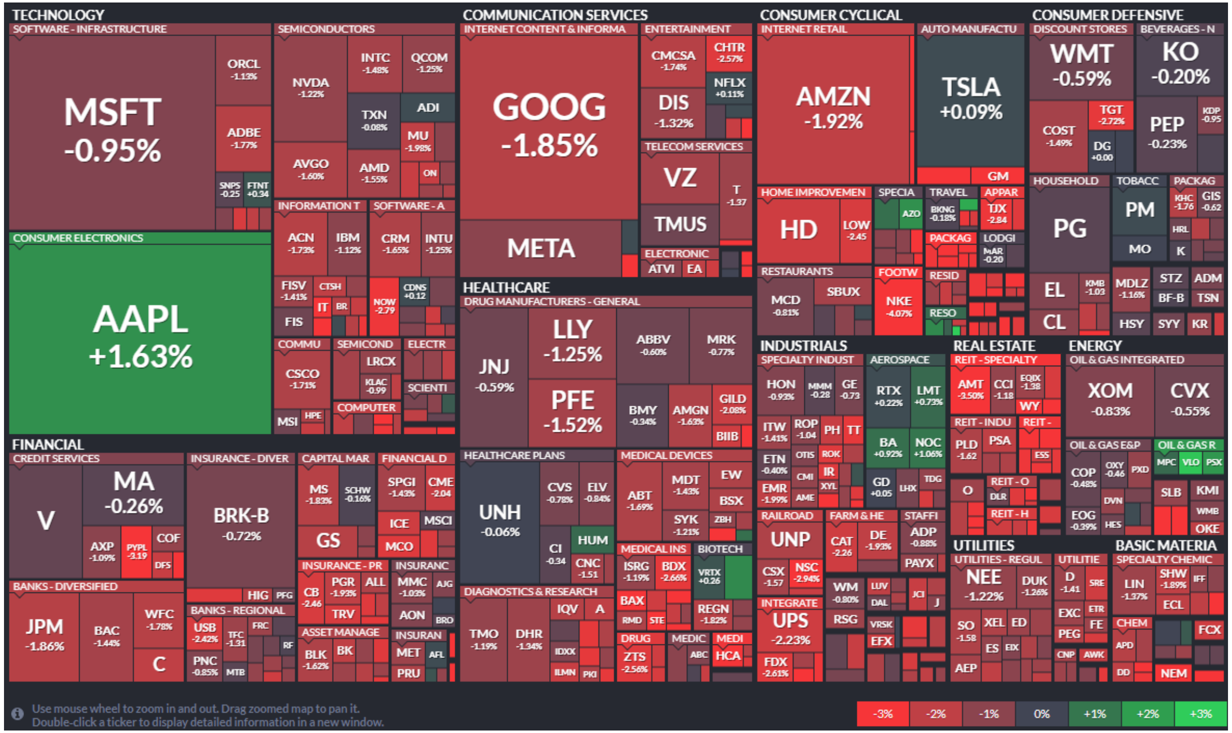
VISUALIZING DATA

- Another way to present two categories at once is with a stacked bar chart, as shown in Figure. In a stacked bar chart, the height of each bar represents the cumulative frequency for a category (such as goods-producing industries) and the colors within each bar represent joint frequencies (such as women employed in government). From this stacked bar chart, we can see the size of the private service-producing sector relative to the other two sectors.



VISUALIZING DATA

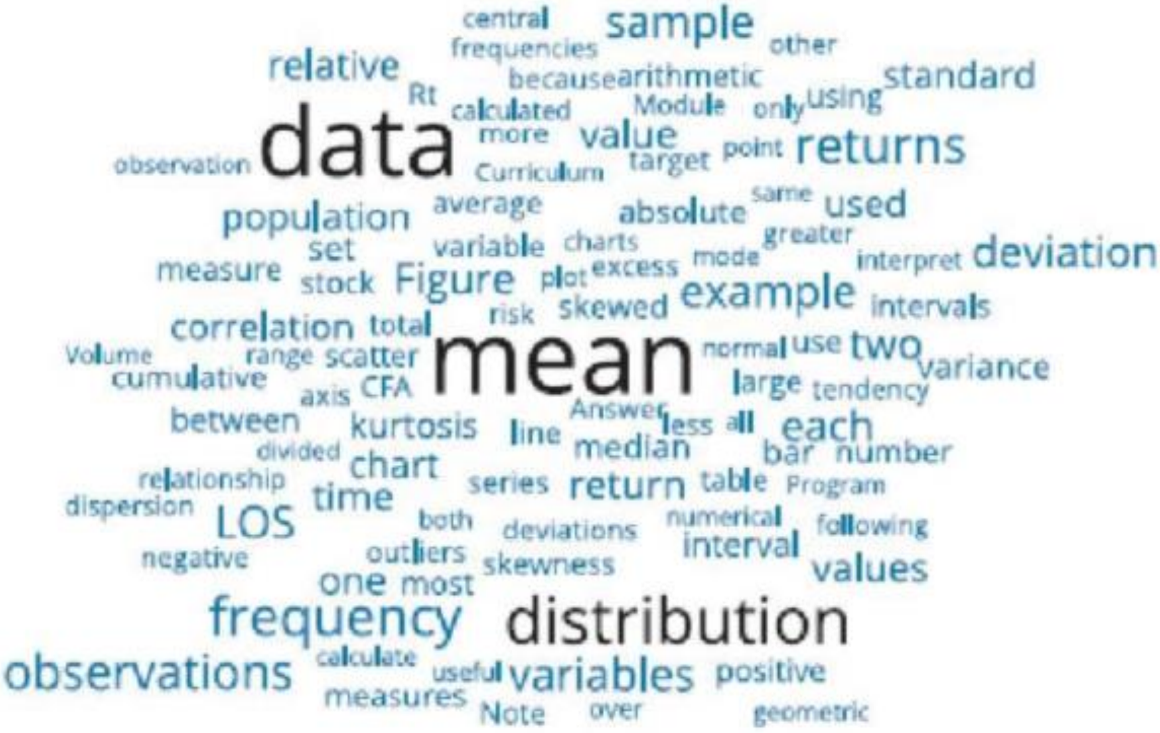
- A **tree map** is another method for visualizing the relative sizes of categories. A common example of a treemap that is well suited to its data is the stock heat map generated by FinViz. The small rectangles are individual company ticker symbols, and the rectangular groups are the sectors



Use mouse wheel to zoom in and out. Drag zoomed map to pan it.
Double-click a ticker to display detailed information in a new window.

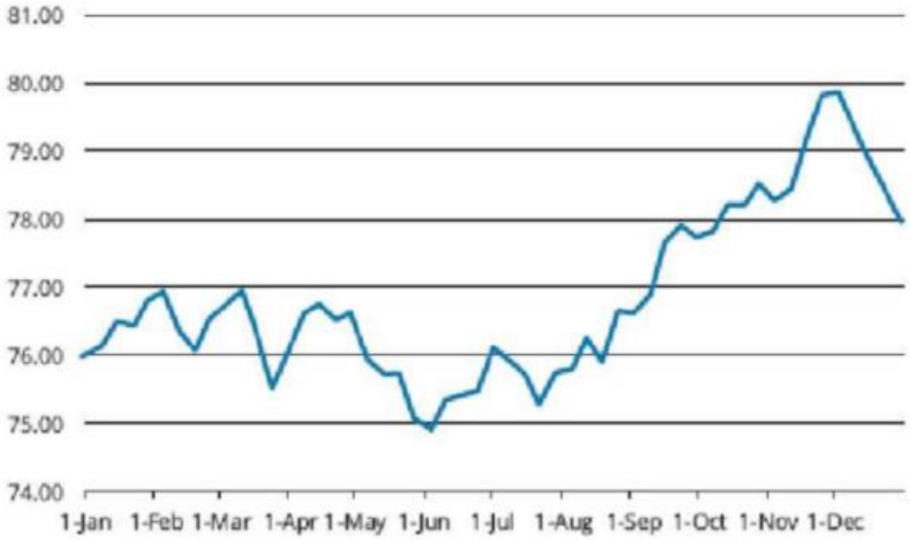
VISUALIZING DATA

- When analyzing text, a useful visualization technique is a **word cloud**. A word cloud is generated by counting the uses of specific words in text data. It displays frequently occurring words, in type sizes that are scaled to the frequency of their use. Figure is an example of a word cloud generated from this reading. From this word cloud, we can easily see two of the major concepts this reading addresses: types of data and definitions of the mean.



VISUALIZING DATA

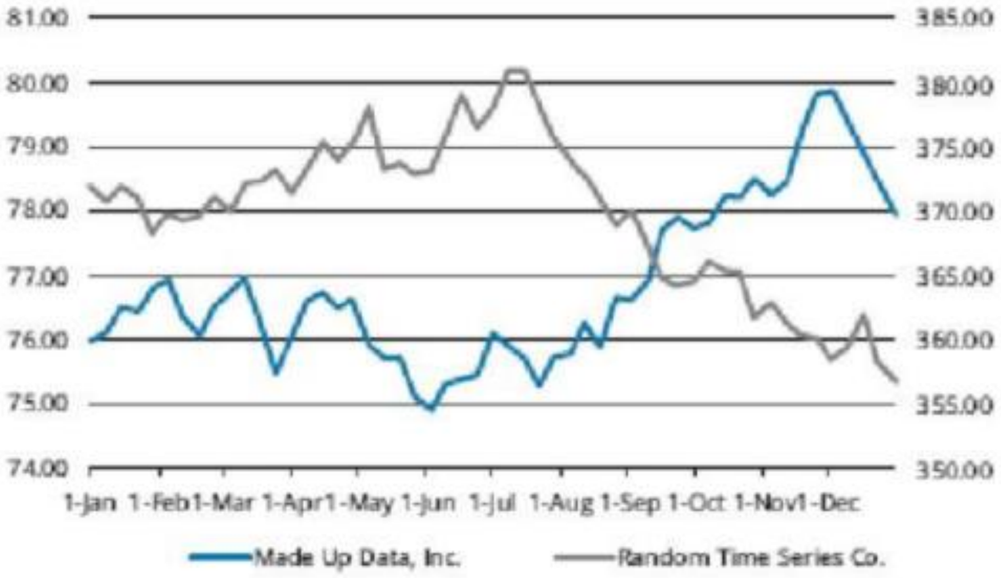
- We have already seen some examples of line charts. Line charts are particularly useful for illustrating time series data, such as securities prices. Figure is a line chart of weekly closing prices for a hypothetical stock.



VISUALIZING DATA

- Multiple time series can be displayed on a line chart if their scales are comparable. It is also possible to display two time series on a line chart if their scales are different, by using left and right vertical axes as shown in Figure. This is one way of showing changes in two variables over time relative to each other.

Figure 2.15: Dual-Scale Line Chart



How to select among visualization types

- **Relationships**. Scatter plots, scatter plot matrices, and heat maps.
- **Comparisons**. Bar charts, tree maps, and heat maps for comparisons among categories; line charts, dual-scale line charts, and bubble line charts for comparisons over time.
- **Distributions**. Histograms, frequency polygons, and cumulative distribution charts for numerical data; bar charts, tree maps, and heat maps for categorical data; and word clouds for text data.

中心 趋势

MEASURES OF CENTRAL TENDENCY

- **Measures of central tendency** identify the center, or average, of a data set. This central point can then be used to represent the typical, or expected, value in the data set.
- To compute the **population mean**, all the observed values in the population are summed (EX) and divided by the number of observations in the population, N. Note that the population mean is unique in that a given population only has one mean. The population mean is expressed as:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

MEASURES OF CENTRAL TENDENCY

- The **sample mean** is the sum of all the values in a sample of a population, EX, divided by the number of observations in the sample, n. It is used to make inferences about the population mean. The sample mean is expressed as:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

MEASURES OF CENTRAL TENDENCY

均值
中位数
众数

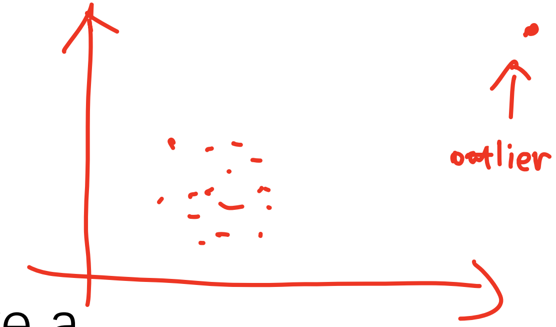
算术平均值

- **arithmetic means**: The population mean and sample mean are both examples of arithmetic means.
- The arithmetic mean is the only measure of central tendency for which the sum of the deviations from the mean is zero. This property can be expressed as follows:

$$\text{sum of mean deviations} = \sum_{i=1}^{n}(X_i - X) = 0$$

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

The outliers



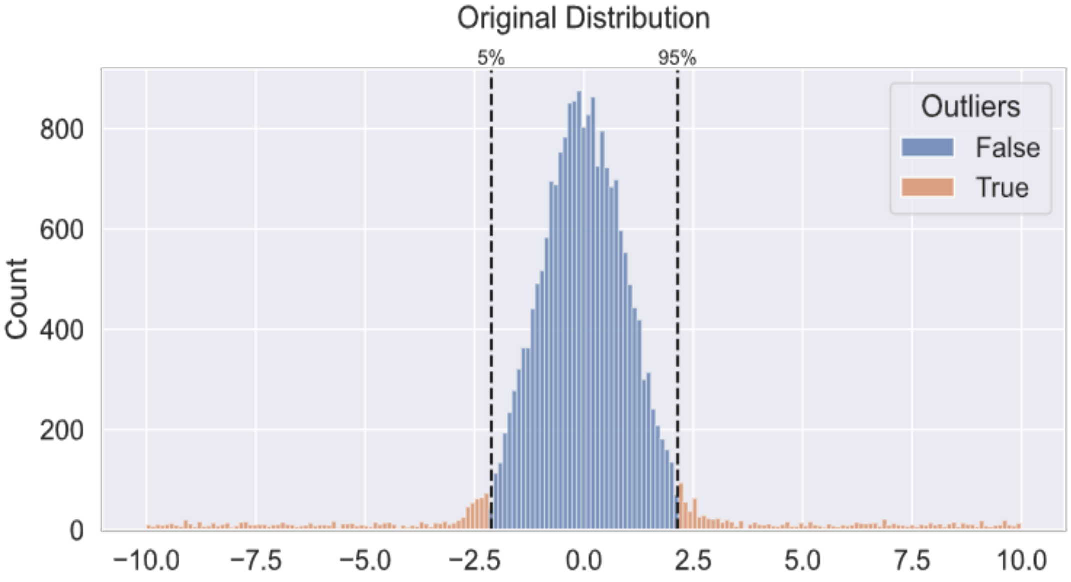
异常值

- Unusually large or small values, **outliers** can have a disproportionate influence on the arithmetic mean. The mean of 1, 2, 3, and **50** is 14 and is not a good indication of what the **individual data values really are**. On the positive side, the arithmetic mean uses all the information available about the observations. The arithmetic mean of a sample from a population is the best estimate of both the true mean of the sample and of the value of a single future observation.

Trimmed mean and winsorized mean

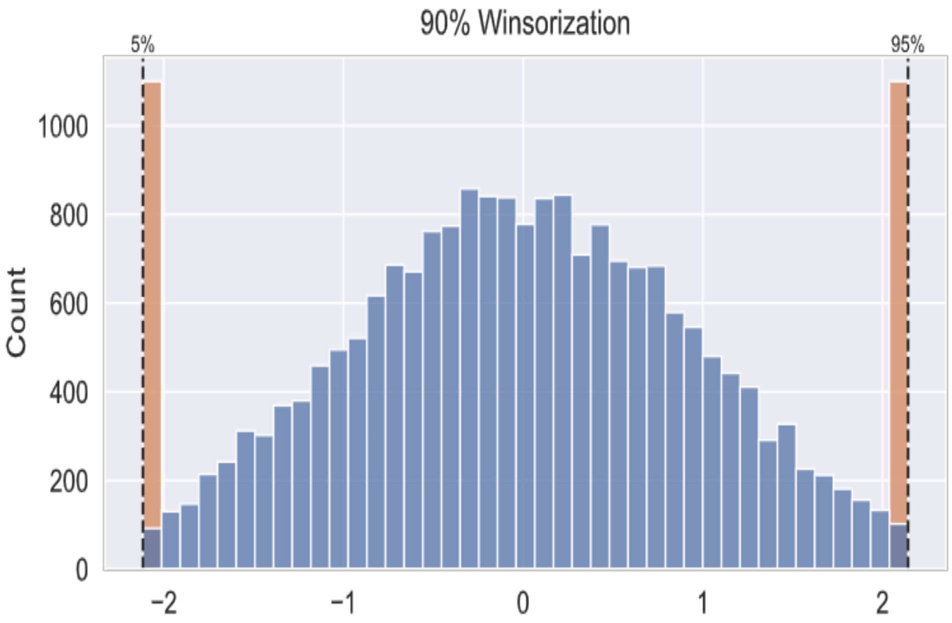
裁剪 均值

- **trimmed mean**: In some cases, a researcher may decide that outliers should be excluded from a measure of central tendency. One technique for doing so is to use a trimmed mean. A trimmed mean excludes a stated percentage of the most extreme observations. A 1% trimmed mean, for example, would **discard** the lowest 0.5% and the highest 0.5% of the observations.



Trimmed mean and winsorized mean

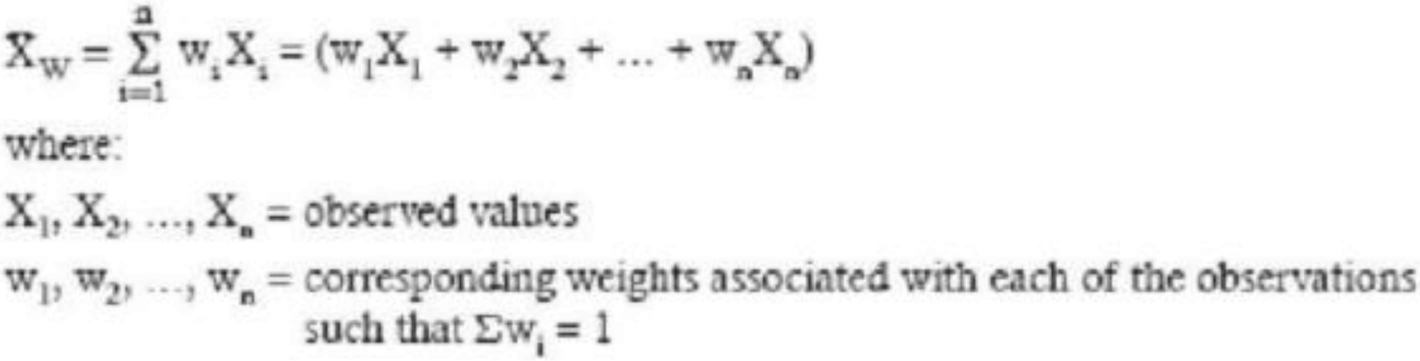
- **winsorized mean**: Another technique is to use a winsorized mean. Instead of discarding the highest and lowest observations, we substitute a value for them. To calculate a 90% winsorized mean, for example, we would determine the 5th and 95th percentile of the observations, **substitute** the 5th percentile for any values lower than that, substitute the 95th percentile for any values higher than that, and then calculate the mean of the revised data set.



Weighted mean

加权平均

- The computation of a weighted mean (or weighted average) recognizes that different observations may have a disproportionate influence on the mean. The weighted mean of a set of numbers is computed with the following equation:



$$X_W = \sum_{i=1}^{n} w_i X_i = (w_1X_1 + w_2X_2 + \dots + w_nX_n)$$

where:

$X_1, X_2, \dots, X_n$ = observed values

$w_1, w_2, \dots, w_n$ = corresponding weights associated with each of the observations such that $\Sigma w_i = 1$

Example: Weighted mean as a portfolio return

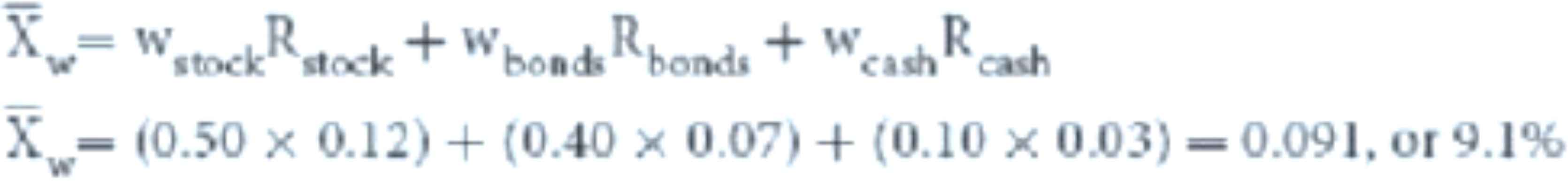
投资组合

- A portfolio consists of 50% common stocks, 40% bonds, and 10% cash. If the return on common stocks is 12%, the return on bonds is 7%, and the return on cash is 3%, what is the portfolio return?

$50\% \times 12 + 40\% \times 7\% + 10\% \times 3 = 8.1\%$

Example: Weighted mean as a portfolio return

- A portfolio consists of 50% common stocks, 40% bonds, and 10% cash. If the return on common stocks is 12%, the return on bonds is 7%, and the return on cash is 3%, what is the portfolio return?



$$\overline{X}_w = w_{stock}R_{stock} + w_{bonds}R_{bonds} + w_{cash}R_{cash}$$

$$\overline{X}_w = (0.50 \times 0.12) + (0.40 \times 0.07) + (0.10 \times 0.03) = 0.091, \text{ or } 9.1\%$$

The median

- The **median** is the midpoint of a data set when the data is arranged in ascending or descending order. Half the observations lie above the median and half are below. To determine the median, arrange the data from the highest to the lowest value, or lowest to highest value, and find the middle observation.

Mode

众数

- The **mode** is the value that occurs most frequently in a data set. A data set may have more than one mode or even no mode. When a distribution **has one value** that appears most frequently, it is said to be **unimodal**. When a set of data has **two or three values that occur most frequently**, it is said to be **bimodal** or **trimodal**, respectively.

Geometric mean

几何 平均

- The geometric mean is often used when calculating investment returns over multiple periods or when measuring compound growth rates. The general formula for the geometric mean, G, is as follows:

$$G = \sqrt[n]{X_1 \times X_2 \times \ldots \times X_n} = (X_1 \times X_2 \times \ldots \times X_n)^{1/n}$$

- The geometric mean return (R_G) can be computed using the following equation:

$$(1+S_3) = \sqrt[3]{((1+S_1)(1+1y1y)(1+2y1y))}$$

$$1 + R_G = \sqrt[n]{(1+R_1) \times (1+R_2) \times \ldots \times (1+R_n)}$$

Harmonic mean

- A harmonic mean is used for certain computations, such as the average cost of shares purchased over time. The harmonic mean is calculated as

$$\frac{N}{\sum_{i=1}^{N} \frac{1}{X_i}}$$

- where there are N values of X_j.

MEASURES OF LOCATION AND DISPERSION

- **Quantile** is the general term for a value at or below which a stated proportion of the data in a distribution lies. Examples of quantiles include the following:
  - **Quartile**. The distribution is divided into quarters.
  - **Quintile**. The distribution is divided into fifths.
  - **Decile**. The distribution is divided into tenths.
  - **Percentile**. The distribution is divided into hundredths (percents)

90 quantile

The range

范圍

- The **range** is a relatively simple measure of variability, but when used with other measures, it provides extremely useful information. The range is the distance between the largest and the smallest value in the data set, or:
  - range = maximum value - minimum value

The mean absolute deviation

- The **mean absolute deviation** (MAD) is the average of the absolute values of the deviations of individual observations from the arithmetic mean:

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

存在 outlier 数据：MAD

The sample variance

- The sample variance, s^2, is the measure of dispersion that applies when we are evaluating a sample of n observations from a population. The sample variance is calculated using the following formula:

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

→ Normal Distribution

The sample variance

- The denominator for s^2 is n - 1, one less than the sample size n. Based on the mathematical theory behind statistical procedures, the use of the entire number of sample observations, n, instead of n -1 as the divisor in the computation of s^2, will systematically **underestimate** the population variance, particularly for small sample sizes.
- This systematic underestimation causes the sample variance to be a **biased estimator** of the population variance.

The sample standard deviation

- The sample standard deviation is the square root of the sample variance. The sample standard deviation, s, is calculated as:

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

$\$K$

$\mu = \$\overline{K}$

$\sigma^2 = L$

$\sigma = \$\sqrt{L}$

$\$\overline{k} + \$\sqrt{L}$

Relative dispersion

- **Relative dispersion** is the amount of variability in a distribution relative to a reference point or benchmark.
- Relative dispersion is commonly measured with the **coefficient of variation (CV)**, which is computed as:

$$CV = \frac{s_x}{\bar{X}} = \frac{\text{standard deviation of x}}{\text{average value of x}}$$

- CV measures the amount of dispersion in a distribution relative to the distribution's mean.

Example: Coefficient of variation

- You have just been presented with a report that indicates that the mean monthly return on Tbills is 0.25% with a standard deviation of 0.36%, and the mean monthly return for the S&P 500 is 1.09% with a standard deviation of 7.30%. Your unit manager has asked you to compute the CV for these two investments and to interpret your results.

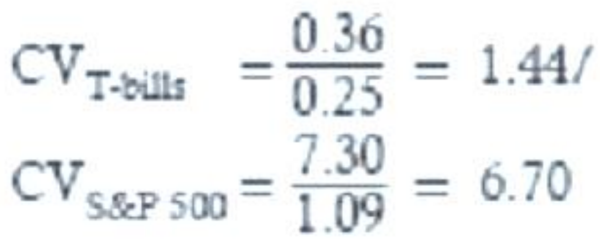
$$CV = \frac{\text{standard deviation}}{\text{average}}$$

$$CV_{Tbill} = \frac{0.36\%}{0.25} \doteq 1.44$$

$$CV_{S\&P500} = \frac{7.30\%}{1.09\%} = 6.70$$

Example: Coefficient of variation

- You have just been presented with a report that indicates that the mean monthly return on T-bills is 0.25% with a standard deviation of 0.36%, and the mean monthly return for the S&P 500 is 1.09% with a standard deviation of 7.30%. Your unit manager has asked you to compute the CV for these two investments and to interpret your results.



$$CV_{\text{T-bills}} = \frac{0.36}{0.25} = 1.44/$$

$$CV_{\text{S\&P 500}} = \frac{7.30}{1.09} = 6.70$$

- These results indicate that there is less dispersion (risk) per unit of monthly return for T-bills than for the S&P 500 (1.44 versus 6.70).

Downside risk 下行风险

- One measure of downside risk is **target downside deviation**, which is also known as **target semideviation**.
- Calculating target downside deviation is similar to calculating standard deviation,
- But only include deviations from the target value in our calculation if the outcomes are below that target.

$$s_{target} = \sqrt{\frac{\sum\limits_{\text{all } X_i < B}^{n} (X_i - B)^2}{n-1}}$$

$B=\bar{x}$

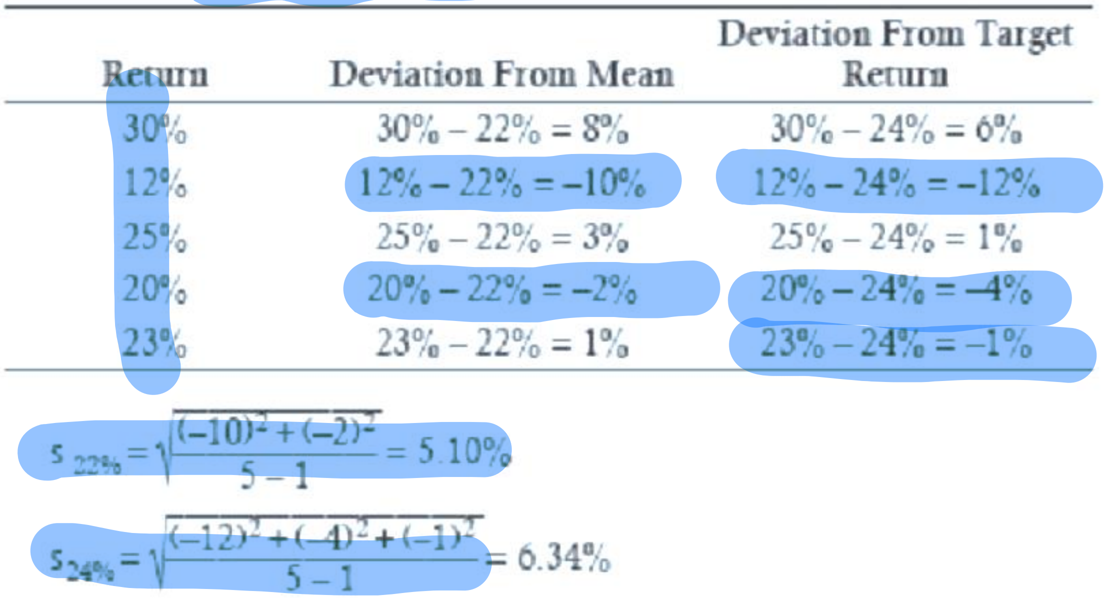
where B is the target.

Example: Target downside deviation

- Calculate the target downside deviation based on the data in the preceding examples, for a target return equal to the mean (22%), and for a target return of 24%.

Example: Target downside deviation

- Calculate the target downside deviation based on the data in the preceding examples, for a target return equal to the mean (22%), and for a target return of 24%.



Return	Deviation From Mean	Deviation From Target Return
30%	30% – 22% = 8%	30% – 24% = 6%
12%	12% – 22% = –10%	12% – 24% = –12%
25%	25% – 22% = 3%	25% – 24% = 1%
20%	20% – 22% = –2%	20% – 24% = –4%
23%	23% – 22% = 1%	23% – 24% = –1%

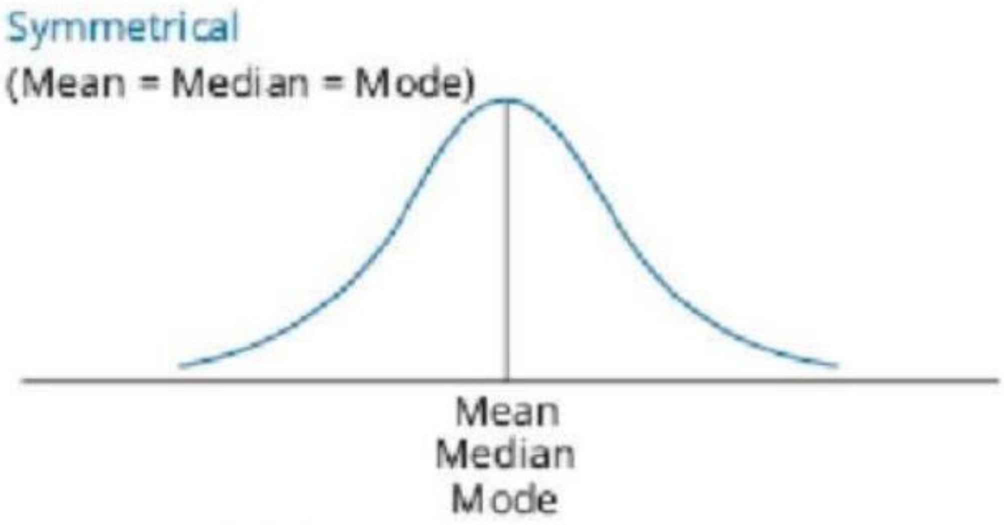
$$s_{22\%} = \sqrt{\frac{(-10)^2 + (-2)^2}{5-1}} = 5.10\%$$

$$s_{24\%} = \sqrt{\frac{(-12)^2 + (-4)^2 + (-1)^2}{5-1}} = 6.34\%$$

SKEWNESS

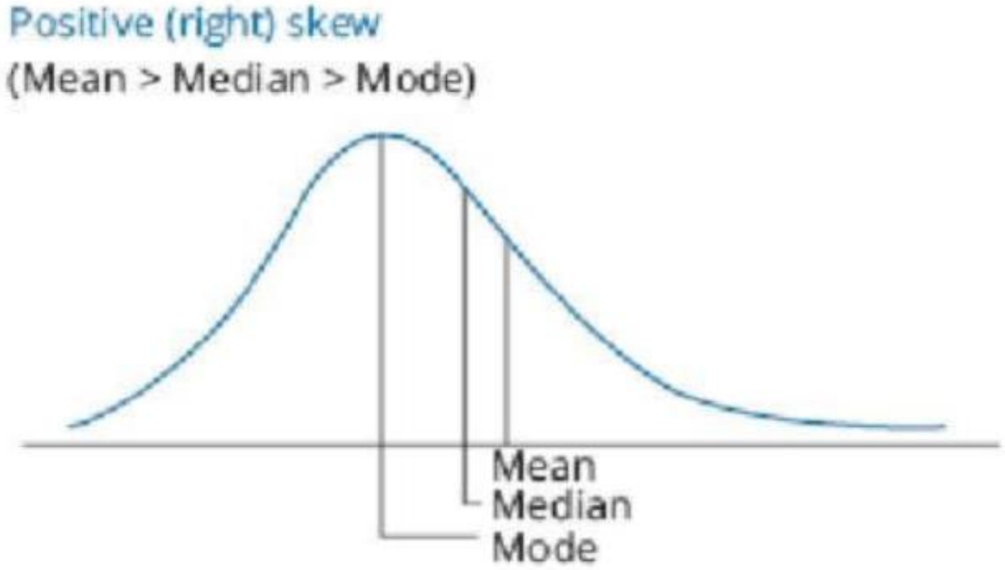
- A distribution is **symmetrical** if it is shaped identically on both sides of its mean. Distributional symmetry implies that intervals of losses and gains will exhibit the same frequency.
- **Skewness**, or skew, refers to the extent to which a distribution is not symmetrical.
- Nonsymmetrical distributions may be either positively or negatively skewed and result from the occurrence of **outliers** in the data set.
- Outliers are observations extraordinarily far from the mean, either above or below

SKEWNESS



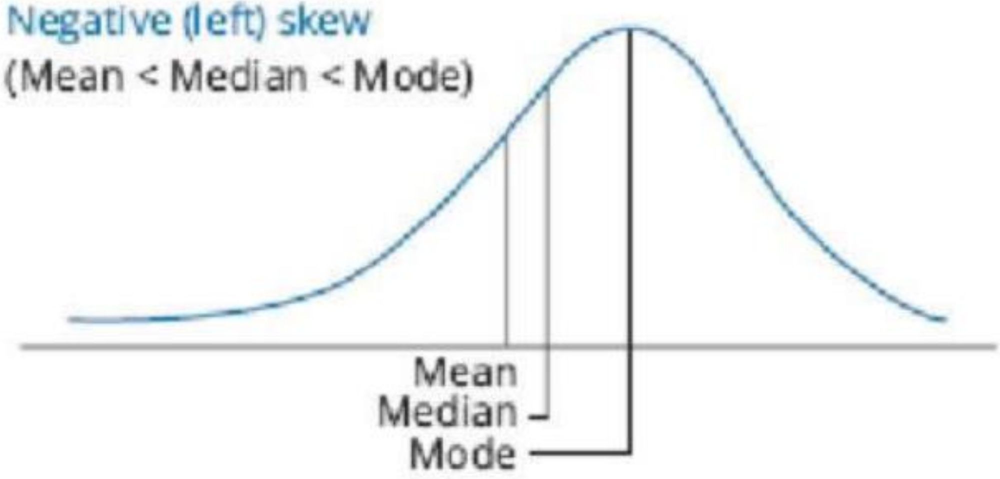
SKEWNESS

- A **positively skewed** distribution is characterized by outliers greater than the mean (in the upper region, or right tail). A positively skewed distribution is said to be **skewed right** because of its relatively **long upper** (right) tail.



SKEWNESS

- A **negatively skewed** distribution has a disproportionately large amount of outliers less than the mean that fall within its lower (left) tail. A negatively skewed distribution is said to be **skewed left** because of its long lower tail.



Sample skewness

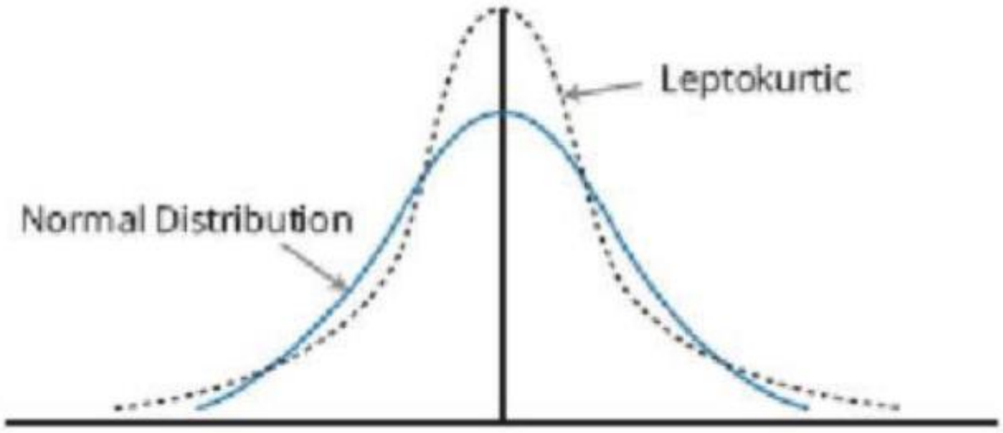
- Sample skewness is equal to the sum of the cubed deviations from the mean divided by the cubed standard deviation and by the number of observations. Sample skewness for large samples is computed as:

$$\text{sample skewness} = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - X)^3}{s^3}$$

- where s = sample standard deviation.

Kurtosis

- **Kurtosis** is a measure of the degree to which a distribution is more or less peaked than a normal distribution. **Leptokurtic** describes a distribution that is more peaked than a normal distribution, whereas **platykurtic** refers to a distribution that is less peaked, or flatter than a normal distribution. A distribution is **mesokurtic** if it has the same kurtosis as a normal distribution.



Kurtosis

- **Sample kurtosis** is measured using deviations raised to the fourth power:

$$\text{sample kurtosis} = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - X)^4}{s^4}$$

- Where s = sample standard deviation

Covariance

- **Covariance** is a measure of how two variables move together. The calculation of the **sample covariance** is based on the following formula:

$$s_{X,Y} = \frac{\sum_{i=1}^{n}\{[X_i - X][Y_i - Y]\}}{n-1}$$

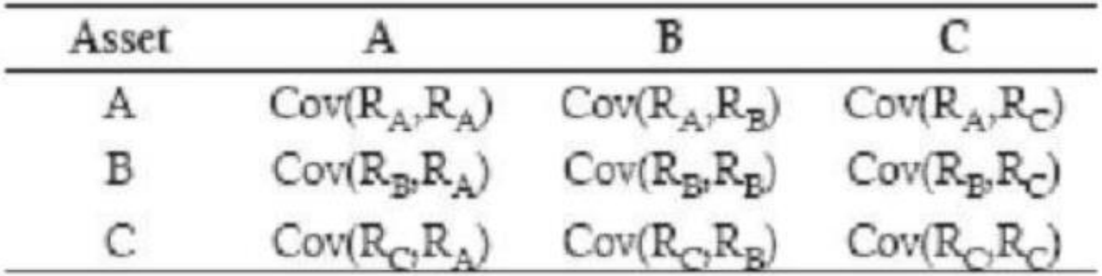
- where
  - X_i = an observation of variable X
  - Y_i = an observation of variable Y
  - X = mean of variable X
  - Y = mean of variable Y
  - n = number of periods

Covariance

- Covariance may range from **negative infinity** to **positive infinity**.
- A **positive covariance** indicates that when one random variable is above its mean, the other random variable tends to be above its mean as well.
- A **negative covariance** indicates that when one random variable is above its mean, the other random variable tends to be below its mean.

Covariance matrix

- A covariance matrix shows the covariances between returns on a group of asset



Asset	A	B	C
A	Cov(R_A,R_A)	Cov(R_A,R_B)	Cov(R_A,R_C)
B	Cov(R_B,R_A)	Cov(R_B,R_B)	Cov(R_B,R_C)
C	Cov(R_C,R_A)	Cov(R_C,R_B)	Cov(R_C,R_C)

- Note that the diagonal terms are the variances of each asset’s returns, i.e., $Cov(R_A, R_A) = Var(R_A)$

Correlation coefficient

- A standardized measure of the linear relationship between two variables is called the correlation coefficient, or simply correlation. The correlation between two variables, X and Y, is calculated as:

$$\rho_{XY} = \frac{s_{XY}}{s_X s_Y} \text{ which implies,}$$

$$s_{XY} = \rho_{XY} s_X s_Y$$

Correlation coefficient

- The properties of the correlation of two random variables, X and Y, are summarized here:
  - Correlation measures the strength of the linear relationship between two random variables.
  - Correlation has no units.
- The correlation ranges from -1 to +1. That is, $-1 \leqslant \rho_{XY} \leqslant +1$

Correlation coefficient

- If p_{XY} = 1.0, the random variables have **perfect positive correlation**. This means that a movement in one random variable results in a proportional positive movement in the other relative to its mean.
- If p_{XY} = -1.0, the random variables have **perfect negative correlation**. This means that a movement in one random variable results in an exact opposite proportional movement in the other relative to its mean.
- If p_{XY} = 0, there is **no linear relationship** between the variables, indicating that prediction of Y cannot be made on the basis of X using linear methods.

Example: Correlation

- The variance of returns on stock A is 0.0028, the variance of returns on stock B is 0.0124, and their covariance of returns is 0.0058. Calculate and interpret the correlation of the returns for stocks A and B.

Example: Correlation

- The variance of returns on stock A is 0.0028, the variance of returns on stock B is 0.0124, and their covariance of returns is 0.0058. Calculate and interpret the correlation of the returns for stocks A and B.
- First, it is necessary to convert the variances to standard deviations:
  - $s_A = (0.0028)^{1/2} = 0.0529$
  - $s_B = (0.0124)^{1/2} = 0.1114$
- Now, the correlation between the returns of stock A and stock B can be computed as follows:
  - $\rho_{AB} = \frac{0.0058}{0.0529*0.1114} = 0.9842$
- The fact that this value is close to +1 indicates that the linear relationship is not only positive but very strong.

Spurious correlation

- **Spurious correlation** refers to correlation that is either the result of chance or present due to changes in both variables over time that is caused by their association with a third variable.
- The correlation between the age of each year's Miss America and the number of films Nicholas Cage appeared in that year is 87%. This seems a bit random. The correlation between the U.S. spending on science, space, and technology and suicides by hanging, strangulation, and suffocation over the 1999-2009 period is 99.87%. Impressive correlation, but both variables increased in an approximately linear fashion over the period.

REVISION

- **a. identify and compare data types.**
- We may classify data types from three different perspectives: numerical versus categorical, time series versus cross sectional, and structured versus unstructured.
- A time series is a set of observations taken at a sequence of points in time. Cross-sectional data are a set of comparable observations taken at one point in time. Time series and cross-sectional data may be combined to form panel data

REVISION

- **b. calculate and interpret measures of central tendency.**
- The **arithmetic mean** is the average:

$$X = \frac{\sum_{i=1}^{n} X_i}{n}$$

- The **geometric mean** is used to find a compound growth rate:

$$G = \sqrt[n]{X_1 \times X_2 \times \ldots \times X_n}$$

- The **weighted mean** weights each value according to its influence:

$$X_W = \sum_{i=1}^{n} w_i X_i$$

REVISION

- **c. calculate and interpret measures of central tendency.**
- The **harmonic mean** can be used to find an average purchase price, such as dollars per share for equal periodic investments:

$$X_H = \frac{N}{\sum_{i=1}^{N} \frac{1}{X_i}}$$

- The **median** is the midpoint of a data set when the data are arranged from largest to smallest.
- The **mode** of a data set is the value that occurs most frequently

REVISION

- **d. evaluate alternative definitions of mean to address an investment problem.**
- **Arithmetic mean** is used to estimate expected value, value of a single outcome from a distribution.
- **Geometric mean** is used calculate or estimate periodic compound returns over multiple periods.
- A **trimmed mean** omits outliers and a **winsorized mean** replaces outliers with given values, reducing the effect of outliers on the mean in both cases.

REVISION

- **e. calculate and interpret measures of dispersion.**
- The range is the difference between the largest and smallest values in a data set.
- Mean absolute deviation (MAD) is the average of the absolute values of the deviations from the arithmetic mean:

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

REVISION

- **e. calculate and interpret measures of dispersion.**
- The **sample variance** is defined as the mean of the squared deviations from the arithmetic mean or from the expected value of a distribution:

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

REVISION

- **e. calculate and interpret measures of dispersion.**
- **Standard deviation** is the positive square root of the variance and is frequently used as a quantitative measure of risk.
- The **coefficient of variation** for sample data, $CV = \frac{s}{\overline{\boldsymbol{X}}}$, is the ratio of the standard deviation of the sample to its mean (expected value of the underlying distribution)

REVISION

- **f. calculate and interpret target downside deviation.**
- **Target downside deviation** or semideviation is a measure of downside risk. Calculating target downside deviation is similar to calculating standard deviation, but in this case, we choose a target against which to measure each outcome and only include outcomes below that target when calculating the numerator.
- The formula for target downside deviation is:

$$s_{target} = \sqrt{\frac{\sum_{\text{all } X_i < B}^{n} (X_i - B)^2}{n - 1}}$$

where B is the target value

RIVISION

- **g. interpret skewness.**
- Skewness describes the degree to which a distribution is not symmetric about its mean. A right-skewed distribution has positive skewness. A left-skewed distribution has negative skewness.
- For a positively skewed, unimodal distribution, the mean is greater than the median, which is greater than the mode.
- For a negatively skewed, unimodal distribution, the mean is less than the median, which is less than the mode.

REVISION

- **h. interpret kurtosis.**
- Kurtosis measures the peakedness of a distribution and the probability of extreme outcomes (thickness of tails):
- Excess kurtosis is measured relative to a normal distribution, which has a kurtosis of 3.
- Positive values of excess kurtosis indicate a distribution that is leptokurtic (fat tails, more peaked), so the probability of extreme outcomes is greater than for a normal distribution.
- Negative values of excess kurtosis indicate a platykurtic distribution (thin tails, less peaked)

REVISION

- **i. interpret correlation between two variables.**
- Correlation is a standardized measure of association between two random variables. It ranges in value from -1 to +1 and is equal to

$$\frac{cov_{A,B}}{\sigma_A \sigma_B}$$

- Scatterplots are useful for revealing nonlinear relationships that are not measured by correlation.
- Correlation does not imply that changes in one variable cause changes in the other. Spurious correlation may result by chance or from the relationships of two variables to a third variable.