soleadea

LEVEL 1: QUANTITATIVE METHODS

Reading 2 (2nd out of 7): BASICS OF STATISTICS

Difficulty: easy

Benchmark Study Time: 4.5h

2022

THIS E-BOOK:

- is a selective summary of the corresponding Reading in your CFA® Program Curriculum,
- provides place for your own notes,
- helps you structure your study and revision time!

How to use this e-book to maximize your knowledge retention:

1. **Print** the e-book in duplex and bind it to keep all important info for this Reading **in one place.**
2. **Read** this e-book, best twice, to grasp the idea of what this Reading is about.
3. **Study** the Reading from your curriculum. **Here add** your notes, examples, formulas, definitions, etc.
4. **Review** the Reading using this e-book, e.g. write your summary of key concepts or revise the formulas at the end of this e-book (if applicable).
5. **Done?** Go to your study plan and change the Reading's status to **green**:
   (it will make your Chance-to-Pass-Score™ grow ☺).
6. **Come back** to this e-book from time to time to **regularly review for knowledge retention!**

NOTE: While studying or reviewing this Reading, you can use the tables at the end of this e-book and mark your study/review sessions to hold yourself accountable.

STATISTICAL METHODS

Definitions

Statistics is a science that deals with quantitative methods of studying the regularities of large-scale phenomena.

Statistics:

- descriptive statistics,
- statistical inference.

Descriptive statistics deals with the synthesis, processing, and presentation of data.

Statistical inference is aimed at making generalizations concerning a population based on its sample.

Population consists of all elements of a group. A descriptive characteristic of a population is called a **parameter**.

Sample is a subset of a population. A sample is usually selected randomly and a **random sample** is another key term in statistics. A sample is described by **sample statistics**, for example a sample mean.

Scales

Statistics uses a variety of numerical data which we obtain and measure. For correct analysis, we need to know what scale we're dealing with. Measurement scales can be ranked from the weakest to the strongest:

1. A nominal scale → different categories but not ranked, e.g. three groups: women AND men AND children – none of the groups is better than the other.
2. An ordinal scale → different categories, ranked but values cannot be added or subtracted, e.g. companies divided into three groups according to growth prospects: high, medium, low.
3. An interval scale → values can be added or subtracted, e.g. the Celsius scale.
4. A ratio scale → a true zero point exists, e.g. amount of money, rate of return.

The stronger the scale, the more we know about the relationships between observations. Each scale has all the "good" features of the proceeding scales.

HERE KNOWLEDGE RETENTION HAPPENS | WRITE: notes, examples, formulas, definitions, relations, etc.

Summarizing data

Data can be summarized using frequency distributions. A frequency distribution groups observations into intervals, also called classes or bins. In this method, data are usually presented in a table, which makes it easier to work with a large number of observations.

Frequency distribution is made up of intervals that include observations within a set range of values.

Absolute frequency is the number of observations falling in an interval.

Relative frequency is absolute frequency divided by the number of all observations.

Cumulative relative frequency is the sum of the relative frequencies of consecutive intervals.

When constructing a frequency distribution:

- remember to sort the data in ascending order first,
- each interval should have the same width equal to the range divided by the assumed number of intervals,
- regarding interval end points if e.g. there are three intervals **A**: <0,1), **B**: <1,2), and **C**: <2,3> - "0" is in interval **A**, "1" is in interval **B**, and both "2" and "3" are in interval **C**.

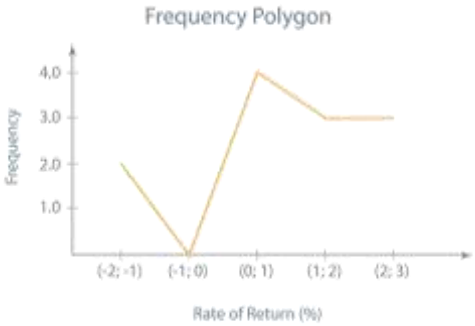
Graphic representation of data

A graphic representation of data is very useful as a quick way of getting acquainted with statistical observations.

Tools for graphic presentation of data:

- the histogram,
- the frequency polygon.

A **histogram** is a bar chart in which each bar represents an interval. The height of each bar indicates the frequency for an interval. Using a histogram we can quickly and easily find out which interval includes the most observations.



To draw a **frequency polygon**, we need to plot the midpoints of all intervals on the x-axis and the respective absolute frequencies on the y-axis. Then, we need to connect the points with a straight line. A frequency polygon provides us with information similar to what we can find in a histogram.

HERE KNOWLEDGE RETENTION HAPPENS | WRITE: notes, examples, formulas, definitions, relations, etc.

MEASURES OF CENTRAL TENDENCY

Measures of central tendency and measures of location help to determine the similarities between the elements of a dataset. Types of measures of central tendency:

- the arithmetic mean,
- the median,
- the mode,
- the weighted average mean,
- the geometric mean,
- the harmonic mean.

Arithmetic mean

The arithmetic mean, also called the mean or the average, is defined as the sum of all observations divided by the number of observations.

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

Where:

- $\overline{X}_a$ – arithmetic mean,
- X_i – observation 'i',
- n – number of observations.

A **population mean** and a **sample mean** are calculated as arithmetic means. In the case of a population mean, we sum up all of the observations and divide the result by the number of observations in the population. In the case of a sample mean, we take the corresponding data for the sample.

Median

The median is the value of the middle element of a set. Therefore, a median divides a dataset into two equal parts. Half of the observations are smaller than the median and the other half is greater. When there is an even number of elements in a set, the median is the arithmetic mean of the two neighboring middle numbers. When there is an odd number, the median is the element in the middle. To find the median, we have to sort the data in ascending order.

Mode

The mode is the most frequently occurring element of a set. A set may have no mode at all, one mode, or several modes. If each observation is different, there is no mode in the set. If we are able to identify one most frequent value, we're dealing with unimodal distribution and the dataset has one mode. If, for example, our set has two most frequent values, we're dealing with bimodal distribution and the dataset has two modes.

HERE KNOWLEDGE RETENTION HAPPENS | WRITE: notes, examples, formulas, definitions, relations, etc.

Weighted mean

The weighted mean is often used in portfolio analysis.

$$\overline{X}_w = \sum_{i=1}^{n} w_i \times X_i$$

Where:

- $\overline{X}_w$ – weighted mean,
- X_i – observation 'i',
- w_i – weight,
- n – number of observations.

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

Geometric mean

The geometric mean is often used to compute the average rate of return over a series of periods or to calculate the growth rate.

$$\overline{X}_g = \sqrt[n]{X_1 \times X_2 \times \ldots \times X_n}$$

In the case of rate of return or growth rate: $\overline{R}_g = \sqrt[n]{(1 + R_1) \times (1 + R_2) \times \ldots \times (1 + R_n)} - 1$

Where:

- $\overline{X}_g, \overline{R}_g$ – geometric mean,
- X_i, R_i – observation 'i',
- n – number of observations.

Harmonic mean

The harmonic mean can be used e.g. to determine the average purchase price paid for stocks bought in several months for the same monthly budget OR to calculate the average time necessary for a given product to be produced.

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

Where:

- $\overline{X}_h$ – harmonic mean,
- X_i – observation 'i',
- n – number of observations.

HERE KNOWLEDGE RETENTION HAPPENS | WRITE: notes, examples, formulas, definitions, relations, etc.

MEASURES OF LOCATION

Measures of location are a broader concept than measures of central tendency. The latter, as the name suggests, refers to the middle (or center) of data. Measures of location provide us with information on observations in different locations, not only in the center. Therefore, measures of central tendency are a sub-type of measures of location.

Measures of location:

- percentiles,
- quartiles,
- quintiles,
- deciles.

Quartiles divide the distribution into quarters, quintiles into fifths, deciles into tenths, and percentiles into hundredths.

As in the case of a median, to find any of the measures of location, we have to sort the data in ascending order.

Percentile

A given percentile is a value below which a given percentage of observations is located. Percentiles divide a set into a hundred parts. The **location of a percentile**:

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

Where:

- y – percentile,
- L_y – location of the percentile,
- n – number of observations.

Quartiles

Quartiles divide a dataset into four parts. A quartile represents a quarter of a dataset, so there are four quartiles. 1st quartile is 25th percentile, 2nd quartile is 50th percentile, and so on.

Quintiles

Quartiles divide a dataset into five parts. 1st quintile is 20th percentile, 2nd quintile is 40th percentile, and so on.

Deciles

Deciles divide a dataset into ten parts. 1st decile is 10th percentile, 2nd decile is 20th percentile…

HERE KNOWLEDGE RETENTION HAPPENS | WRITE: notes, examples, formulas, definitions, relations, etc.

MEASURES OF DISPERSION

Measures of dispersion tell us about how observations are dispersed around the mean.

Range

Range is defined as the difference between the largest and the smallest value in a dataset.

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

Mean absolute deviation (MAD)

Mean absolute deviation is the arithmetic mean of the absolute values of the deviations of the individual elements of the dataset around the mean.

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

Where:

- $\overline{X}$ – sample mean,
- X_i – observation 'i',
- n – number of observations.

Variance

Variance is the mean of the squared deviations of individual observations around the mean. Variance is difficult to interpret but it's often easier to work with during complicated computations.

Population variance:

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

Where:

- σ^2 – population variance,
- μ – population mean,
- X_i – observation 'i',
- N – size of the population.

HERE KNOWLEDGE RETENTION HAPPENS | WRITE: notes, examples, formulas, definitions, relations, etc.

Sample variance:

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

Where:

- s^2 – sample variance,
- $\overline{X}$ – sample mean,
- X_i – observation 'i',
- n – number of observations in the sample.

Standard deviation

Standard deviation is the square root of variance.

Population standard deviation:

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

Where:

- σ – population standard deviation,
- μ – population mean,
- X_i – observation 'i',
- N – size of the population.

Sample standard deviation:

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

Where:

- s – sample standard deviation,
- $\overline{X}$ – sample mean,
- X_i – observation 'i',
- n – number of observations in the sample.

HERE KNOWLEDGE RETENTION HAPPENS | WRITE: notes, examples, formulas, definitions, relations, etc.

STATISTICS – OTHER CONCEPTS

Chebyshev's inequality

There are certain relationships between the arithmetic mean, which is a measure of central tendency, and standard deviation, i.e. a measure of dispersion. One of the general principles that tell us about these relationships is Chebyshev's inequality. Chebyshev's inequality tells us that – for every $k > 1$– at least:

$$1 - \frac{1}{k^2}$$

observations occur within **±k** standard deviations of the mean. Chebyshev's inequality holds for **any distribution with finite variance**.

Coefficient of variation

If we want to compare the dispersion of two datasets that differ in the volume, standard deviation may not be the right measure. In this case, we need a relative measure of dispersion to standardize the dispersion of observations. One such measure is the coefficient of variation.

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

Where:

- CV – coefficient of variation,
- $\overline{X}$ – sample mean,
- s – sample standard deviation.

Sharpe ratio

Sharpe ratio is commonly used in the assessment of the efficiency of investment portfolio management. Thanks to the construction of the Sharpe ratio, we can find out about how many units of profit there are per one unit of risk.

$$S_R = \frac{\overline{R}_p - \overline{R}_f}{s_p}$$

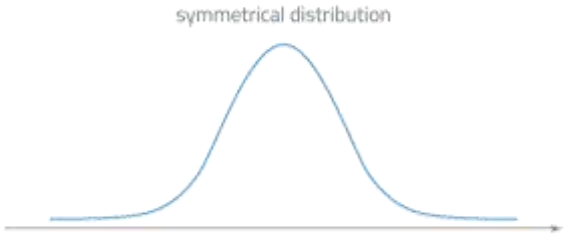
Where:

- $\overline{R}_p$ – mean return on the portfolio,
- $\overline{R}_f$ – mean return on a risk-free asset,
- s_p – standard deviation of return on the portfolio.

HERE KNOWLEDGE RETENTION HAPPENS | WRITE: notes, examples, formulas, definitions, relations, etc.

Skewness and Kurtosis

Skewness

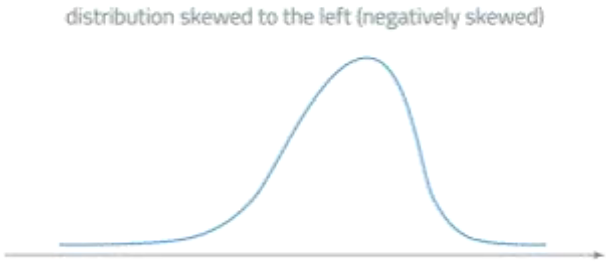


Skewness is a measure of the asymmetry of a distribution. It tells us how observations are distributed around the mean.

symmetrical distribution →mode = median = mean

If most observations lie to the left of the mean, we say that the distribution is skewed to the right or positively skewed. This kind of distribution has a long tail on its right side.

positively skewed distribution →mode < median < mean



When most observations lie to the right of the mean, we say that the distribution is skewed to the left or negatively skewed. A distribution with a negative skew has a long tail on its left side.

negatively skewed distribution →mean < median < mode

HERE KNOWLEDGE RETENTION HAPPENS | WRITE: notes, examples, formulas, definitions, relations, etc.

Sample skewness:

$$S_K = \left(\frac{n}{(n-1)\times(n-2)}\right)\times\sum_{i=1}^{n}\frac{(X_i-\overline{X})^3}{s^3}$$

Where:

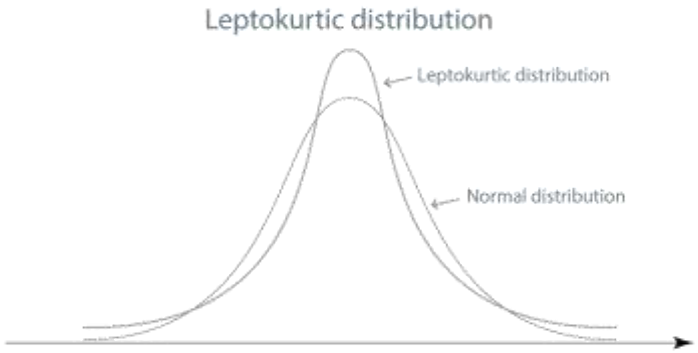
- S_K – sample skewness,
- n – number of observations in the sample,
- $\overline{X}$ – sample mean,
- s – sample standard deviation.

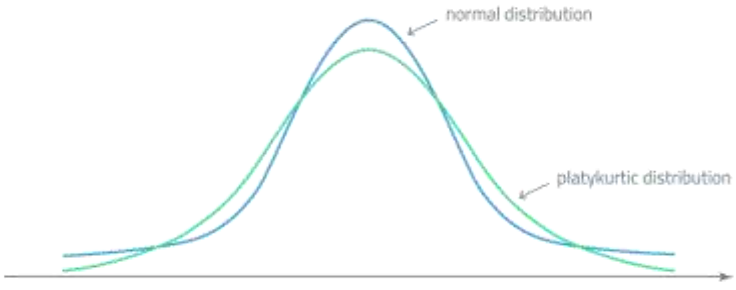
Kurtosis

Kurtosis is a measure of the peakedness of a distribution or, in other words, a measure of the concentration of results. It tells us how strongly the observations are clustered around the mean. Types of distribution:

- leptokurtic,
- platykurtic,
- mesokurtic.

If observations for a distribution are more concentrated around the mean than in the case of a normal distribution, such a distribution is called **leptokurtic**. A leptokurtic distribution is peaked much more than normal because many observations are located close to the mean. Also, this type of distribution is characterized by the so-called fat tails. This means that the probability of extreme results (namely very low or very high ones) is higher than in the case of a normal distribution.





A distribution is flattened, or **platykurtic**, when observations are less concentrated around the mean than in the case of a normal distribution. Distributions similar to a normal distribution are called **mesokurtic** distributions.

HERE KNOWLEDGE RETENTION HAPPENS | WRITE: notes, examples, formulas, definitions, relations, etc.

Types of kurtosis:

- kurtosis,
- excess kurtosis.

The **kurtosis of a normal distribution is 3**. For platykurtic distributions, kurtosis is less than 3 and for leptokurtic ones, greater than 3.

Excess kurtosis is the value by which the kurtosis of a distribution differs from the kurtosis of a normal distribution. Therefore, the **excess kurtosis of a normal distribution is 0**. For leptokurtic distributions, excess kurtosis is greater than 0. The excess kurtosis of platykurtic distributions is less than 0.

Sample excess kurtosis (you don't need to know the formula in your exam):

$$K_E = \left(\left(\frac{n \times (n+1)}{(n-1) \times (n-2) \times (n-3)} \right) \times \sum_{i=1}^{n} \frac{(X_i - \overline{X})^4}{s^4} \right) - \frac{3 \times (n-1)^2}{(n-2) \times (n-3)}$$

Where:

- K_E – sample excess kurtosis,
- n – number of observations in the sample,
- $\overline{X}$ – sample mean,
- s – sample standard deviation.

Arithmetic mean vs Geometric mean: Application

Practitioners assume that the arithmetic mean doesn't reflect the reality well when we deal with historical returns. The arithmetic mean should be used when we invest the same amount of money in each examined period. However, in most cases, we don't do that. We usually invest what we have available from the previous period. So, the invested amount differs from period to period because it includes past losses and gains. When this is the case, we should rather use the geometric mean, which better illustrates the average growth of our assets period by period.

However, if we want to determine future returns, the arithmetic mean can be useful. We will use the arithmetic mean if, for example, we calculate the expected rate of return for the next investment period and we assume that two equally probable scenarios for that period may occur. If this is the case, it doesn't make sense to use the geometric mean and the arithmetic mean seems a better choice.

Remember: The geometric mean equals the arithmetic mean only if the rates of return (observations) are equal period by period. In all other cases, the geometric mean is always lower than the arithmetic mean and the greater the variation of return (variation of observations) for each period, the greater the difference.

HERE KNOWLEDGE RETENTION HAPPENS | WRITE: notes, examples, formulas, definitions, relations, etc.

Summarizing key concepts:

☐ Statistics – basic definitions, scales, etc.

My summary:

☐ Measures of central tendency

My summary:

☐ Measures of location

My summary:

☐ Measures of dispersion

My summary:

☐ Chebyshev's inequality

My summary:

☐ Coefficient of variation

My summary:

☐ Skewness

My summary:

☐ Kurtosis

My summary:

☐ Arithmetic mean vs Geometric mean

My summary:

Reviewing formulas:

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

Write down the formula:

$$\overline{X}_w = \sum_{i=1}^{n} w_i \times X_i$$

Write down the formula:

$$\overline{X}_g = \sqrt[n]{X_1 \times X_2 \times \ldots \times X_n}$$

Write down the formula:

$$\overline{R}_g = \sqrt[n]{(1 + R_1) \times (1 + R_2) \times \ldots \times (1 + R_n)} - 1$$

Write down the formula:

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

Write down the formula:

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

Write down the formula:

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

Write down the formula:

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

Write down the formula:

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

Write down the formula:

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

Write down the formula:

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

Write down the formula:

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

Write down the formula:

$$S_R = \frac{\bar{R}_p - \bar{R}_f}{s_p}$$

Write down the formula:

$$S_K = \left(\frac{n}{(n-1) \times (n-2)}\right) \times \sum_{i=1}^{n} \frac{(X_i - \bar{X})^3}{s^3}$$

Write down the formula:

Keeping myself accountable:

TABLE 1 | STUDY

When you sit down to study, you may want to **try the Pomodoro Technique** to handle your study sessions: study for 25 minutes, then take a 5-minute break. Repeat this 25+5 study-break sequence all throughout your daily study session.

✓ Tick off as you proceed.

POMODORO TIMETABLE: study-break sequences (25' + 5')													
date		date		date		date		date		date		date	
25'		25'		25'		25'		25'		25'		25'	
5'		5'		5'		5'		5'		5'		5'	
25'		25'		25'		25'		25'		25'		25'	
5'		5'		5'		5'		5'		5'		5'	
25'		25'		25'		25'		25'		25'		25'	
5'		5'		5'		5'		5'		5'		5'	
25'		25'		25'		25'		25'		25'		25'	
5'		5'		5'		5'		5'		5'		5'	

TABLE 2 | REVIEW

Never ever neglect revision! Though it's not the most popular thing among CFA candidates, regular revision is what makes the difference. If you want to pass your exam, **schedule & do your review sessions.**

REVIEW TIMETABLE: When did I review this Reading?													
date		date		date		date		date		date		date	
date		date		date		date		date		date		date	