LEVEL 1: QUANTITATIVE METHODS

Reading 5 (5th out of 7): SAMPLING & ESTIMATION

Difficulty: **medium**

Benchmark Study Time: **3h**

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.

INTRODUCTION TO SAMPLING

Definitions

Population consists of all elements of a group. It can be perceived as a set of all the members of the group we're interested in.

Sample is a subset of a population. A sample is usually selected randomly. A sample is described by sample statistics, for example a sample mean.

Simple random sample (random sample) is a subset of elements selected randomly from a population in such a way that each element of the population has an equal chance of being included in the sample.

Simple random sampling is a process of drawing a simple random sample from a population.

In order to ensure the existence of certain groups in a sample, **stratified random sampling** is used. In stratified random sampling, the population is divided into subpopulations (strata) based on one or more criteria. Then, from each stratum, with the aid of simple random sampling, we select a number of elements proportional to the relative size of each stratum in the population and get them together.

Sample statistics are numerical characteristics of a sample and are used to estimate population parameters. A parameter is a real value that describes the selected characteristic of an entire population. The examples of sample statistics are sample mean or sample standard deviation. The observed sample statistic usually differs from the actual value of a parameter. To measure the difference, we use **sampling error**. Sampling error is the difference between the observed value of the sample statistic and the population parameter which is estimated using this sample statistic.

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

Types of data

Types of data:

- cross-sectional data,
- time-series data.

Cross-sectional data are data showing the values of properties of different things at a specific point in time, for example share prices listed on the Warsaw Stock Exchange on a given day.

Time-series data is a set of observations at different times. One example would be a change in the price of a listed company's shares for a certain period of time.

Biases associated with sampling

Types of biases:

- data-mining bias – instead of looking for patterns in data sets, we try to adjust our model to the data that we explore,
- sample selection bias – we systematically exclude from the data set some information that is relevant to the whole sample,
- survivorship bias – we analyze only the current, most up to date data (e.g. exclude companies that ceased to exist and estimate the average return of an industry),
- look-ahead bias – when we have to process data that aren't available on the test day,
- time-period bias, e.g. we use a short time series of data and the results may not be significant in a longer period.

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

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

THE CENTRAL LIMIT THEOREM

Definition

Assumptions:

A population is characterized by finite variance and is described by any distribution.

The central limit theorem:

If we draw samples of the same size from the population, the distribution of the sample mean calculated from these samples will be approximately normal if the size of the samples is large enough ($n \geq 30$).

Standard error

The standard deviation of the distribution of the sample mean, which is equal to the square root of the variance of the distribution of the sample mean, is known as the standard error of the sample mean and is given by the following formulas:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

OR

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

the higher the sample size → the lower the standard error

the higher the population standard deviation or sample standard deviation → the higher the standard error

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

ESTIMATORS

Introduction

Statistical inference can be divided into two parts:

- hypothesis testing,
- estimation.

Hypothesis testing can help us answer questions such as:

Whether the population mean is equal to 10% or not?

Estimation helps us answer the following question:

What is the value of a population parameter with a given probability?

Thanks to estimation we can for example state that with a 95% probability the population mean will be within a given interval.

Estimator

An estimator is a formula used to estimate the value of a parameter of a distribution. For example, if we want to estimate the average return on a population of bonds, we can draw a sample of bonds and compute the sample mean. The sample mean will be an estimator that we will use to estimate the value of the population mean.

To allow us to estimate the parameters of a population correctly, estimators should be:

- unbiased,
- efficient,
- consistent.

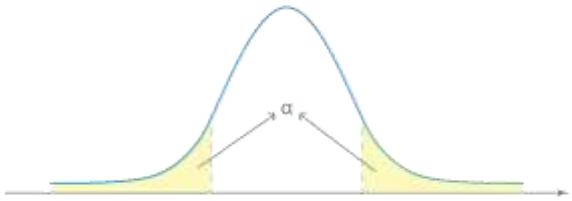
An estimator is **unbiased** when its expected value equals the parameter it is intended to estimate.

An estimator is **efficient** when its variance is the smallest among the unbiased estimators of the same population parameter.

An estimator is **consistent** when the probability that an estimate will be close to the population parameter gets bigger as the sample size increases.

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

Confidence intervals for population mean



A confidence interval is a range for which one can assert with a given probability $1 - \alpha$, called the degree of confidence, that it will contain the parameter it is intended to estimate. This interval is often referred to as the $100 \times (1 - \alpha)$ % confidence interval for the parameter. Alpha is the probability of an error, namely that the parameter is not within the confidence interval.

There are two types of estimation that we can perform:

- point estimation,
- interval estimation.

We use a point estimate when we need a single value as an estimate of a population parameter. For example, the point estimate of the population mean is the sample mean. When we use a point estimate, we must simply find the correct estimator of the parameter. In point estimation, we seek to arrive at a concrete number, i.e. the value of the parameter, while in the case of interval estimation we need to find a range of values that we expect to include the parameter with a particular degree of confidence. This range of values is called the confidence interval.

For the confidence interval:

$$\text{lower confidence limit} = (\text{point estimate}) - (\text{reliability factor}) \times (\text{standard error})$$

$$\text{upper confidence limit} = (\text{point estimate}) + (\text{reliability factor}) \times (\text{standard error})$$

Standard error

The standard deviation of a sample statistic, which is equal to the square root of the variance of the sample mean, is known as the standard error of the sample mean:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

OR

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

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

Reliability factor

The value of the **reliability factor** depends on the so-called degree of confidence. The greater the degree of confidence, i.e. the more we want to be sure that the parameter is within a given confidence interval, the greater the reliability factor. What is more, the greater the reliability factor, the wider the confidence interval.

the greater the degree of confidence	the greater the reliability factor	the wider the confidence interval
the greater the population standard deviation or the sample standard deviation	the greater the standard error	the wider the confidence interval
the greater the sample size	the lower the standard error	the narrower the confidence interval

Algorithm of interval estimation

1. Draw a sample of size n from a population.
2. Calculate the sample mean.
3. Assume a certain degree of confidence – for example 99% or 95%.
4. Calculate alpha based on the value of the degree of confidence. For example, if the value of the degree of confidence is 99%, then the value of alpha is 1%.
5. Check the value of the reliability factor in tables, assuming alpha calculated in step 4.
6. Calculate the standard error.
7. Compute the lower and the upper confidence limit.

Appropriate tests

Sampling from:	Statistic for Small Sample	Statistic for Large Sample
Normal distribution with known variance	z-statistic	z-statistic
Normal distribution with unknown variance	t- statistic	t-statistic or z-alternative
Nonnormal distribution with known variance	no statistics available	z-statistic
Nonnormal distribution with unknown variance	no statistics available	t-statistic or z-alternative

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

Z-statistic is based on the normal distribution and is used when we know the population variance.

Z-alternative is also based on the normal distribution but we use it when the population variance is unknown. For both z-statistic and z-alternative, we look up the value of the reliability factor in the tables for the normal distribution. We can use **z-alternative only if the sample is large** because we know from the central limit theorem that for large samples the sample mean is approximately normally distributed.

T-DISTRIBUTION

The t-distribution is symmetric and bell-shaped, like the normal distribution, and is fully described by degrees of freedom.

There is a different t-distribution for each number of degrees of freedom.

Compared to the normal distribution, the t-distribution is flatter around the mean and has fatter tails.

$$\text{mean of t-distribution} = 0$$

$$\text{variance of t-distribution} = \frac{\text{df}}{\text{df} - 2} \text{ for df} > 2$$

T-statistic is based on Student's t-distribution. This distribution is particularly useful in hypothesis testing and constructing confidence intervals when the sample is small. This distribution is defined by only one parameter called degrees of freedom.

Degrees of Freedom (df)

The concept of degrees of freedom is present in many important distributions. DF is a number of independent observations used to estimate a statistic. In the case of a confidence interval, the number of degrees of freedom is equal to sample size minus 1.

Confidence interval

A $100 \times (1 - a)\%$ confidence interval for the population mean is given by:

$$\overline{X} \pm t_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$$

Where:

- n – sample size,
- $(n - 1)$ – number of degrees of freedom,
- $t_{\frac{\alpha}{2}}$ – reliability factor (based on t-distribution),
- $\frac{s}{\sqrt{n}}$ – standard error.

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

Properties of t-distribution

T-distribution:

1. is symmetric.
2. is fully defined by a single parameter called degrees of freedom, which is equal to the number of observations minus 1.
3. has fatter tails than the standard normal distribution.
4. the higher the number of degrees of freedom and, as a consequence, the higher the sample size, the more the t-distribution resembles the normal distribution.

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

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

Summarizing key concepts:

☐ Definitions: population, sample, simple random sample, simple random sampling, stratified random sampling, sample statistics, sampling error

My summary:

☐ Types of data: cross-sectional data vs time-series data

My summary:

☐ Biases associated with sampling

My summary:

☐ The Central Limit Theorem

My summary:

☐ Standard error

My summary:

☐ Estimator (unbiased, efficient, consistent)

My summary:

☐ Confidence intervals for population mean

My summary:

☐ Sampling – choosing appropriate tests

My summary:

☐ T-Distribution

My summary:

☐ Degrees of Freedom

My summary:

Reviewing formulas:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

OR

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

Write down the formula:

$$\text{lower confidence limit} = (\text{point estimate}) - (\text{reliability factor}) \times (\text{standard error})$$

$$\text{upper confidence limit} = (\text{point estimate}) + (\text{reliability factor}) \times (\text{standard error})$$

Write down the formula:

$$\bar{X} \pm t_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$$

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	