UNIVERSITY OF RUHUNA
DEPARTMENT OF MATHEMATICS
BACHELOR OF SCIENCE (GENERAL) DEGREE (LEVEL II)
APPLIED MATHEMATICS
IMT 224β/AMT 224β:APPLIED STATISTICS

Population vs. Sample Variance and Standard Deviation

Variance and standard deviation are widely used measures of dispersion of data. Variance is defined and calculated as the average squared deviation from the mean. Standard deviation is calculated as the square root of variance or in full definition, standard deviation is the square root of the average squared deviation from the mean.

Describing vs. forecasting in statistics

In general statistics performs two main tasks. Its goal is either to **describe** something that has already happened or already exists (descriptive statistics), or to **estimate** something that has not happened yet or is not fully known (inferential statistics).

Descriptive statistics deals with the problem how to effectively look at data we already have. Inferential statistics (the estimating and forecasting part of statistics) deals with the problem of not having all the data.

Population vs. sample

The primary task of inferential statistics (or estimating or forecasting) is making an opinion about something by using only an **incomplete sample of data**.

In statistics it is very important to distinguish between **population** and **sample**. A **population** is defined as all members (e.g. occurrences, prices, annual returns) of a specified group. Population is the whole group.

A **sample** is **a part of a population** that is used to describe the characteristics (e.g. mean or standard deviation) of the whole population. The size of a sample can be less than 1%, or 10%, or 60% of the population, but it is never the whole population.

Population vs. sample variance and standard deviation

When calculating variance and standard deviation, it is important to know whether we are calculating them for the whole population using all the data, or we are calculation them using only a sample of data. In the first case we call them **population variance** and **population standard deviation**. In the second case we call them **sample variance** and **sample standard deviation**.

Example 1

What is the standard deviation of last years returns of the 12 funds I have invested in?

There is no estimating or forecasting in this task. I am only interested in the 12 funds I have invested in and I dont care about the thousands of other funds which exist in the world. My population is only these 12 funds. I have all the data available, as it is very easy to find these 12 funds performance data.

I take the performance of each of the 12 funds in the last year, calculate the mean, then the deviations from the mean, square the deviations, sum the squared deviations up, divide by 12 (the number of funds), and get the variance. Then the square root of variance is the standard deviation. In this case, because I have the data for the whole population available, I call them **population variance** and **population standard deviation**.

Example 2

What is the standard deviation of last years returns of equity funds in the world?

Compared to calculating standard deviation of concretely specified 12 funds, I now want to know the standard deviation of returns of all equity funds in the world. My population is now much larger than in the previous example. There are thousands of equity funds in the world. Some of them probably aren't on the Bloomberg, don't have a website, and don't publish their performance. In short, I have no chance that I could get the data for all the funds. And even if I could, it would take a long time and cost a lot of money to get all the data.

Contrary to the previous example, I now don't have all the data available and I will have to **estimate** the **population's standard deviation** from a **sample**.

Estimating population standard deviation from a sample

So how will I do it? I will try to collect the data for some of the equity funds these funds will be my sample. It is not necessary (and probably not possible) to collect the data for all the funds in the world (the population). I must only make sure that my sample is large enough. While having the data for 5 funds would probably be insufficient to estimate standard deviation for the whole population, 100 funds data can be enough and still very realistic to get.

Taking the data for these 100 funds I calculate the variance and standard deviation in the same way as in example 1 with my 12 funds.

The difference in calculation: population vs. sample variance

There is only one little difference in the calculation of variance and it is at the very end of it. When I calculate **population variance**, I then divide the sum of squared deviations from the mean by the number of items in the population. When I calculate **sample variance**, I divide it by the number of items in the sample less one.

As a result, the calculated sample variance (and therefore also the standard deviation) will be slightly higher than if we would have used the population variance formula. The purpose of this little difference it to get a better and unbiased estimate of the populations variance.

****