Data reduction and descriptive statistics

Business Mathematics (BK/IBA)
Quantitative Research Methods I (EBE)

August 2015

1 Introduction

Economics, marketing, finance, and most sciences in general, are based on empirical data, on facts so to say. We make observations on prices, sales, and many more phenomena that we wish to study. Let us take an example: the sales of ice cream of a particular vendor near the beach. During summer, he has his business for 72 days. Eager to know how his business is flourishing, he carefully writes down his sales (in euro) every day. At the end of the summer, he has a record of 72 numbers, looking like 320, 410, 176, etc. These are the data. But data make no real knowledge. If we ask, how were your sales, he can do no better than recite the set of 72 numbers. Impatiently, we will interrupt him, and insist: "yeah, I mean, on the average". Or: "in total".

This is an important observation: there are situations in which we want to condense data into more meaningful summaries that are easier to communicate or understand. For obvious reasons, the process of constructing such (numerical) summaries is often referred to as data reduction. It yields summaries that are known as descriptive statistics.

The average (or mean) is the most important of such descriptive statistics. It expresses the central tendency of the data set. There are, however, other measures of central tendency, such as the median value. All such statistics are examples of measures of location.

Further, a central value is not always enough to describe the phenomenon under study. If we want to convey more information about the data, we could, e.g., compare a measure of spread or variation (variability). How constant are the sales? Are they fluctuating wildly, or only mildly? Thus, we may be interested in a second descriptive statistic, one that reflects the degree of variation. Such indicators are often referred to as measures of dispersion. The most important members of this family are the variance and the standard deviation.

All such summary indicators are examples of descriptive statistics: measures that describe a data set in concise terms. Whenever people speak of "statistics", they often have this in mind. Indeed, the national and supranational statistical bureaus, such as the CBS in The Netherlands, Eurostat in the EU, and the United Nations Statistics Division collect statistics in this descriptive, data-summarizing sense.

The term statistics, however, has more meanings. Later in your study, you will do a course on statistics, but then, we will address the subject of inferential statistics, which studies how to infer properties of a large or even infinite population on the basis of knowledge of only a sample of data. In the present course, we will not address inferential statistics, but only discuss descriptive statistics.

2 Measures of location

We first discuss the measures of location, also known as the measures of central tendency. Such measures are often used to give a one-number summary of "how big". For instance, if we ask how many hamburgers are sold daily by McDonald's, we want an indication like 6.5 million, acknowledging that the number will vary day by day.

2.1 The mean

The most well-known measure of location is the average, also known as the mean, or more precisely, the arithmetic mean. For a dataset of length n with observed numerical values $x_1, x_2, \ldots, x_n$, the mean $\bar{x}$ is calculated by

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right) \tag{1}$$

Using the summation notation, this can be written concisely as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

The symbol $\bar{x}$ is pronounced "x bar".

The book by Sydsæter and Hammond (fourth edition) introduces the mean briefly on page 56. It uses a slightly different notation

$$\mu_x = \frac{1}{T}\sum_{i=1}^{T} x_i \tag{3}$$

The individual data point is in both notations written as x_i, but we could obviously also have used x_j or z_k. The number of data points is written here as n, and by the book as T. This is also just a choice, and not in any way fundamental. The last difference is the symbol used at the lefthand-side of the definition. We use $\bar{x}$, the book uses μ_x. This is not merely an unimportant difference. In our later course on statistics, we will use both notations to represent different but related concepts, $\bar{x}$ referring to the sample mean, and μ_x to the population mean. Basically, whenever no context is given, we do not know if the data set represents a sample or a population, so both notations are right, and we just use

the word "mean" to refer to it. In our example on Hamburger sales, suppose that we have collected data on sales on 34 days randomly selected over the year. In that case, the data set $x_1, x_2, \ldots, x_{34}$ is a sample, so the notation $\bar{x}$ is more appropriate. In business and economics, we will often have samples: on sales, investments, employment, prices, etc. If we want to find out the average price of a beer in Dutch bars, we will make a sample of n prices $p_1, \ldots, p_n$, and calculate the sample mean $\bar{p}$ by the formula given, which should in that case be rephrased as

$$\bar{p} = \frac{1}{n} \sum_{i=1}^{n} p_i \tag{4}$$

2.2 Mean versus expected value

The discussion on sample versus population touches on another distinction, between mean and expected value. In fact, the expected value is in some imprecise texts considered to be identical to the mean, but there is a crucial difference. The expected value is defined on probability distributions, and it is equal to the population mean. The sample mean is instead defined for the data observed in a sample.

In practice, the word expectation is unfortunately used by some books to refer to the mean of a set of observed values. This is not correct, for theoretical reasons (the expectation is defined for probability distributions, not for data), in a numerical sense (the expected value of a die is 3.5, whereas the mean after 1000 observations may well be 3.52), and even for linguistic reasons (the 3.52 was not expected, but calculated from observations).

2.3 Properties of the mean

The mean is a numeric property of a data set. It can also be conceived as a function of the data set to a single number:

$$f : (x_1, x_2, \ldots, x_n) \to \frac{1}{n} (x_1 + x_2 + \cdots + x_n) \tag{5}$$

or in vector-notation

$$f : \mathbf{x} \to \frac{1}{n} \sum_{i=1}^{n} x_i \tag{6}$$

The mean satisfies some convenient properties. Most importantly, it is a linear function of its arguments. This can be formulated mathematically as follows:

$$f(a\mathbf{x}) = af(\mathbf{x}) \tag{7}$$

and

$$f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y}) \tag{8}$$

In words, if we scale all arguments up or down by a factor a, the mean will also scale up down by a factor a. It implies that the mean can be calculated in an arbitrary unit, because changes of unit will rescale the value of the mean. If the average price of beer is 2.02 euro, it is also 202 cent.

The mean also satisfies the property of additivity. That means that the sum of the means of two data sets is equal to the mean of the sum of the data points.

Another property of the mean is that it has a unit of measurement (cm, sec, dollar) that is equal to that of the data points. So, if x is in euro/year, $\bar{x}$ is in euro/year as well.

2.4 Other measures of central tendency

Besides the (arithmetic) mean, additional measures of central tendency are used. The median is an important one, as is the geometric mean. We will postpone the discussion of such alternative measures of central tendency to the course on statistics.

3 Measures of dispersion

Next, we discuss the measures of dispersion. Dispersion refers to how scattered the data points are around their mean. If all data points have the same value, there is no dispersion. Dispersion is thus measured around the mean (or another central value). A natural starting point therefore is to consider the deviation of each data point from the mean, so $x_i - \bar{x}$, and to calculate the mean value of this series of deviations. However, that will not work. The reason is that the mean deviation is always 0. See the exercises for a proof.

The problem is caused by the fact that positive deviations ($x_i > \bar{x}$) cancel the effect of negative deviations ($x_i < \bar{x}$), while we of course need non-cancelling deviations. There are two main ways to make sure this does not happen:

- by looking at the absolute value of each deviation;
- by looking at the square of each deviation.

The first option yields the mean absolute deviation (MAD), which we will not further discuss. The second option yields the variance, which we will discuss below.

3.1 The variance

The variance s_x^2 of a data set of length $n > 1$ is defined as

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{9}$$

It is clearly a formula that aims to calculate a sort of mean squared deviation from the mean.

There is one subtle issue: while we would expect to divide by n to find the mean square deviation, we actually divide by $n - 1$. There are theoretical reasons for this that will be discussed in the later course on statistics. Further, in some cases, a division by n is actually performed. If you use software (or a calculator with a σ-button), always check what happens. Some programs (or calculators) offer both options, indicated for instance as σ_{n-1} and σ_n.

3.2 Properties of the variance

What happens when we scale the data set, for instance, if we move from meter to centimeter? Such a change of data can be represented as $y = ax$, with a the scaling factor. Under this change, the variance will transform as

$$\begin{aligned} s_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(ax_i - a\bar{x})^2 \\ &= \frac{a^2}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= a^2 s_x^2 \end{aligned} \tag{10}$$

In words, if we change all data by a factor 100 (as we do when moving from meter to centimeter), the variance will change by a factor $100^2 = 10,000$, while the mean will move by a factor 100.

This is, of course, related to the fact that the variance has a unit that is the square of the unit of the data. If x is in meter, s_x^2 is in square meter.

Above, we already discussed the issue of $\bar{x}$ versus μ. For the variance, a similar phenomenon can be observed. Here the book discusses the variance as σ_x^2 (or $\sigma_{x,x}$; see below), while we use s_x^2 (or perhaps $s_{x,x}$). For now, we keep s_x^2, but come back to this in the course on statistics.

3.3 The standard deviation

Because the variance has a unit that is the square of that of the data, one often uses the square root of the variance, also known as the standard deviation s_x:

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{11}$$

The standard deviation has simpler scaling properties: it just scales, like the mean, linearly with the data.

3.4 The coefficient of variation

In some cases it is convenient to have a measure of dispersion that is independent of the scale of the data. The coefficient of variation (or variation coefficient) is a measure that satisfies this criterion. It is defined as

$$CV_x = \frac{s_x}{\bar{x}} \tag{12}$$

It can be interpreted as a relative standard deviation, relative to the mean of the data. It has no unit, and if we change the unit of measurement, the coefficient

of variation will not change. It is easy to prove this, again using $a = ax$:

$$\begin{aligned} CV_y &= \frac{s_y}{\bar{y}} \\ &= \frac{as_x}{a\bar{x}} \\ &= \frac{s_x}{\bar{x}} \\ &= CV_x \end{aligned} \tag{13}$$

Notice that the coefficient of variation is only defined when $\bar{x} \neq 0$.

3.5 Other measures of dispersion

Again, there is a wide choice of alternative measures of dispersion. We do not discuss these here, but will mention a few in the statistics course later on.

4 Measures of association

Central tendency and dispersion apply to single data vectors. In some cases, it is informative to summarize to which extent one data vector varies with another data vector. For instance, if we have observations on house prices and floor areas, we can compute the mean and the variance price and the mean and the variance floor area, but probably are also interested to what extent big houses are more expensive. Probably, they are, but there are exceptions, so the association is not perfect. Can we develop an indicator of association?

We mention two widely-used indicators of association: the covariance and the correlation coefficient.

These indicators are only meaningful for related (or paired) data vectors. We cannot compute the covariance of the prices of beer in 50 bars in the UK and 50 bars in the US, because there is no sense of "pairing" between a bar in the UK and a bar in the US. But we can study the covariance of the price of beer in a certain bar in the UK in 50 years and in a certain bar in the US in the same 50 years. The observation in the UK in 1971 can then be paired with that in the US is the same year. Thus, covariance will indicate to what extent price increases in one country will be associated with prices increases in the other country. Obviously, the idea of pairing means that the two data sets must have the same length.

4.1 The covariance

The covariance is an indicator of two data vectors, say $\mathbf{x}$ and $\mathbf{y}$. We therefore add a double subscript to it and write $s_{x,y}$:

$$s_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{14}$$

The covariance measures to which extent the two data sets "run together".

Above, we discussed that the variance (and the standard deviation) can be calculated with a denominator n or $n-1$, and that calculators and software

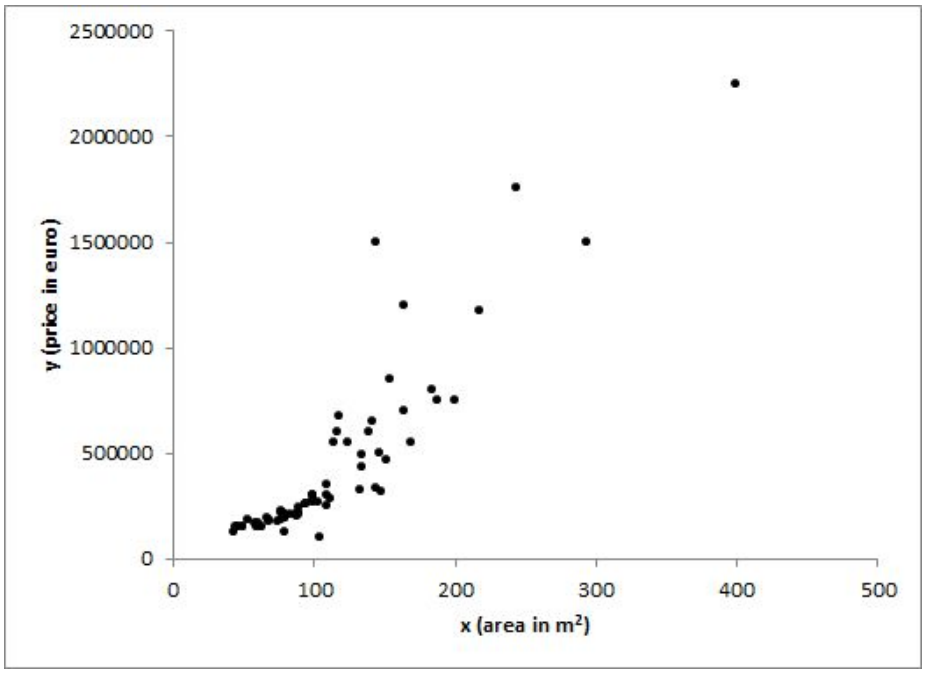

Figure 1: Scatterplot of the house prices and floor areas of a sample of 71 houses in Amstelveen in 2014.

sometimes offer both options, but also that it is not always clear which formula is implemented. The same remark applies for the covariance. We define it for this moment with $n-1$, but we will come back in the statistics course on the form with n.

The covariance has a rather complicated unit: the product of the units of the two underlying data vectors. In the example of house prices in euro and floor areas in square meter, the covariance is measured in euro times square meter. Do not confuse this with euro *per* square meter!

The similarity of the term "covariance" to "variance", the similarity of the formulas, and the similarity of the symbols suggest that we can try to calculate the covariance of a data set with itself, so $s_{x,x}$. We find

$$s_{x,x} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x}) = s_x^2 \tag{15}$$

Thus the covariance of a data set with itself is simply the variance of that data set. Of course, that one is never negative. The covariance of two different data sets, however, can be positive, negative, or zero. If it is positive, the two run together. If it is negative, the two run in opposite way. And if it is zero, the two are said to be unrelated.

4.2 The correlation coefficient

A final data summary indicator is the "relative covariance": the correlation coefficient. It is defined as

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y} \tag{16}$$

The correlation coefficient is scale-free: it has no unit, and it is bounded between -1 and 1. A value of 1 means that the two data sets are perfectly linearly correlated. For instance, if x is a series of temperature data in Celsius. and y the same series in Fahrenheit, the correlation coefficient will be 1. A value between 0 and 1 means that there is some correlation, but not a perfect linear correlation. For instance, there is quite a good linear correlation ($r = 0.9$) between the floor area of a sample of houses and the prices of theses houses. However, the floor area is not the only thing that matters, because other factors (age, location, presence of a garage) have an influence as well. A correlation coefficient smaller than 1 can also indicate a non-linear relationship. For instance, if y depends on x in a quadratic way, the data points will lie on a perfect parabola, but the correlation coefficient will be less than 1 because a parabola is not a straight line.

The correlation coefficient of a data vector with itself is 1. A value of -1 indicates a perfect linear relationship in the opposite way. A value of 0 indicates that the variables are not related. Note well that the value of the correlation coefficient does not indicate an empirically important relation.

4.3 Other measures of association

A final word: there are other indicators for association besides those mentioned. All such indicators have special properties, which make them superior in certain cases and inferior in other cases. There is no one-size-fits-all indicator for centrality, dispersion, and association.