Summary statistics

Understanding summary statistics

Getting a grasp of the variables in your dataset is the very first thing to do when you want to model their effects. Let's look at the most important measures that can summarise your variables.

Once you have gathered a dataset, the next thing you will do is look at what you are dealing with: what can we learn from the variables before we start modelling them? There are a couple of approaches you can use to answer this question. Most notably, many people are drawn to plotting and we will look into that later. First, we will look at summary statistics, which can give us an insight into the distribution of values and other particularities such as outliers.

The most commonly used and well-known summary statistic is the average, or (sample) mean. It distils the overall trend that is present in a set of numerical datapoints. If our dataset has size n, then we calculate the mean as:

$$\overline{x} = \frac{1}{n}\Sigma_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

The mean is a measure of central tendency. Another well-known measure is the median. The median of a dataset, ordered from low to high, is:

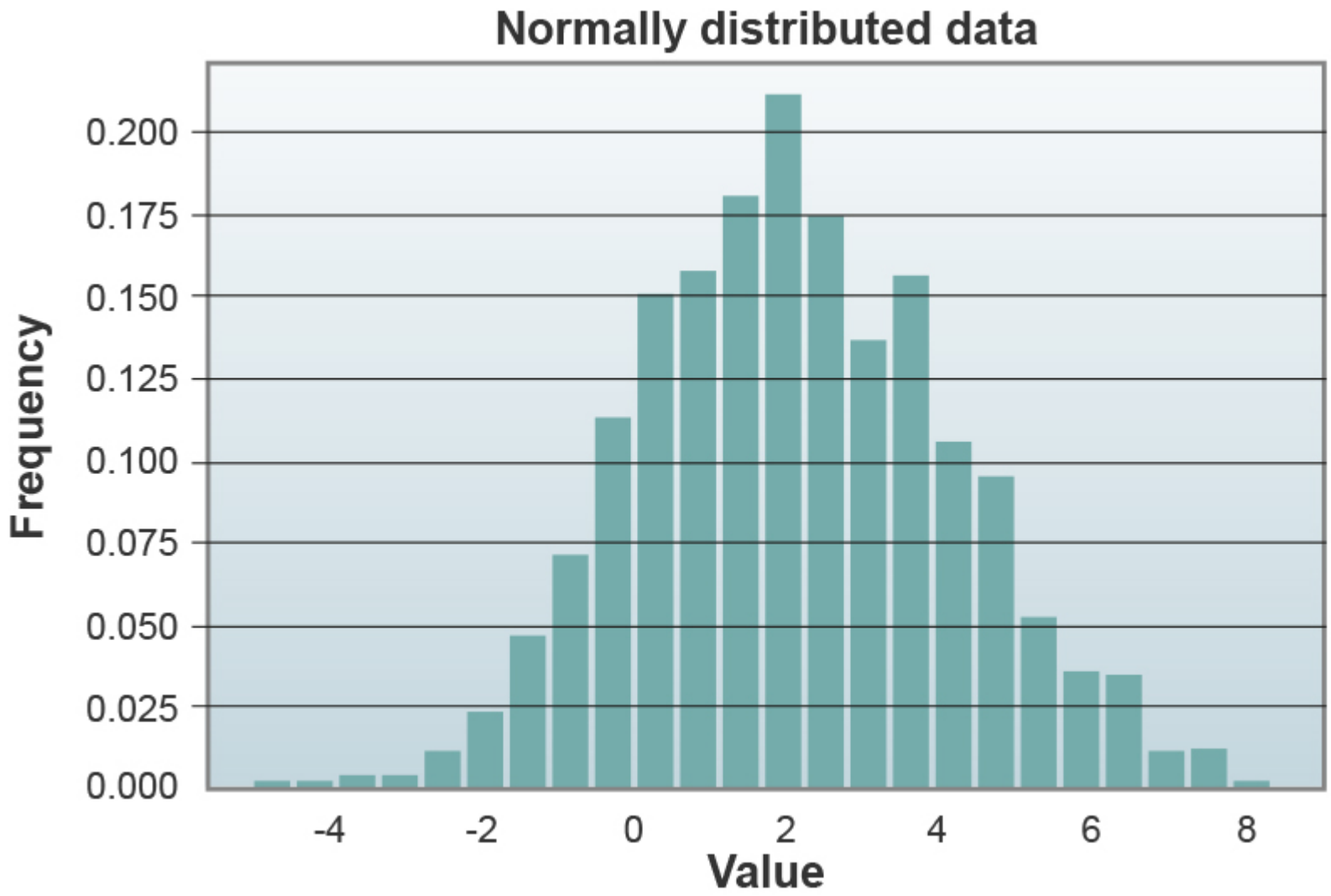
$$\tilde{x}_{0.5} = \begin{cases} x_{\frac{n+1}{2}}, n \text{ is odd} \\ \frac{1}{2} \cdot (x_{\frac{n}{2}} + x_{\frac{n}{2}+1}), n \text{ is even} \end{cases}$$

It is the middle observation of the dataset. In this case, 50% of all the data has a value lower and 50% of the data has a value higher than the median. We say that the median is the value for the 50% quantile. We can also calculate the 25% and 75% quantile $\tilde{x}_{0.75}$ and $\tilde{x}_{0.25}$ respectively, for instance, where 25% (75%) of the data has a value lower and 75% (25%) has a value higher than that statistic. The 25% (75%) percentile is calculated as the median of the lower (upper) part of the dataset in case n is even, and the median of the lower (upper) part including the median in case n is uneven.

Although both measures are used for getting a grasp of the central tendency, they might differ quite significantly. For example, take a look at the following histograms:

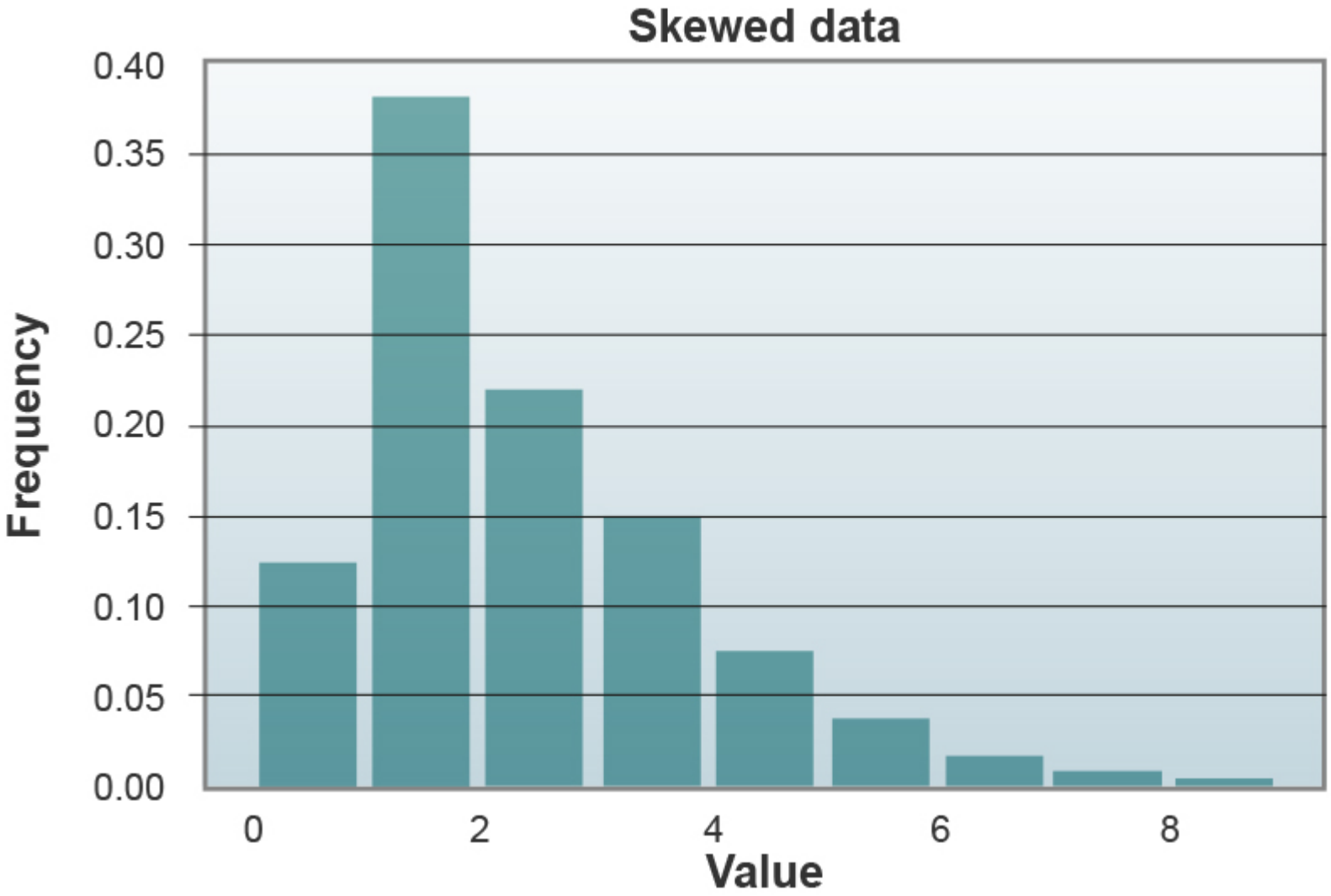
Histogram 1

Mean: 2.148 Median: 2.0339



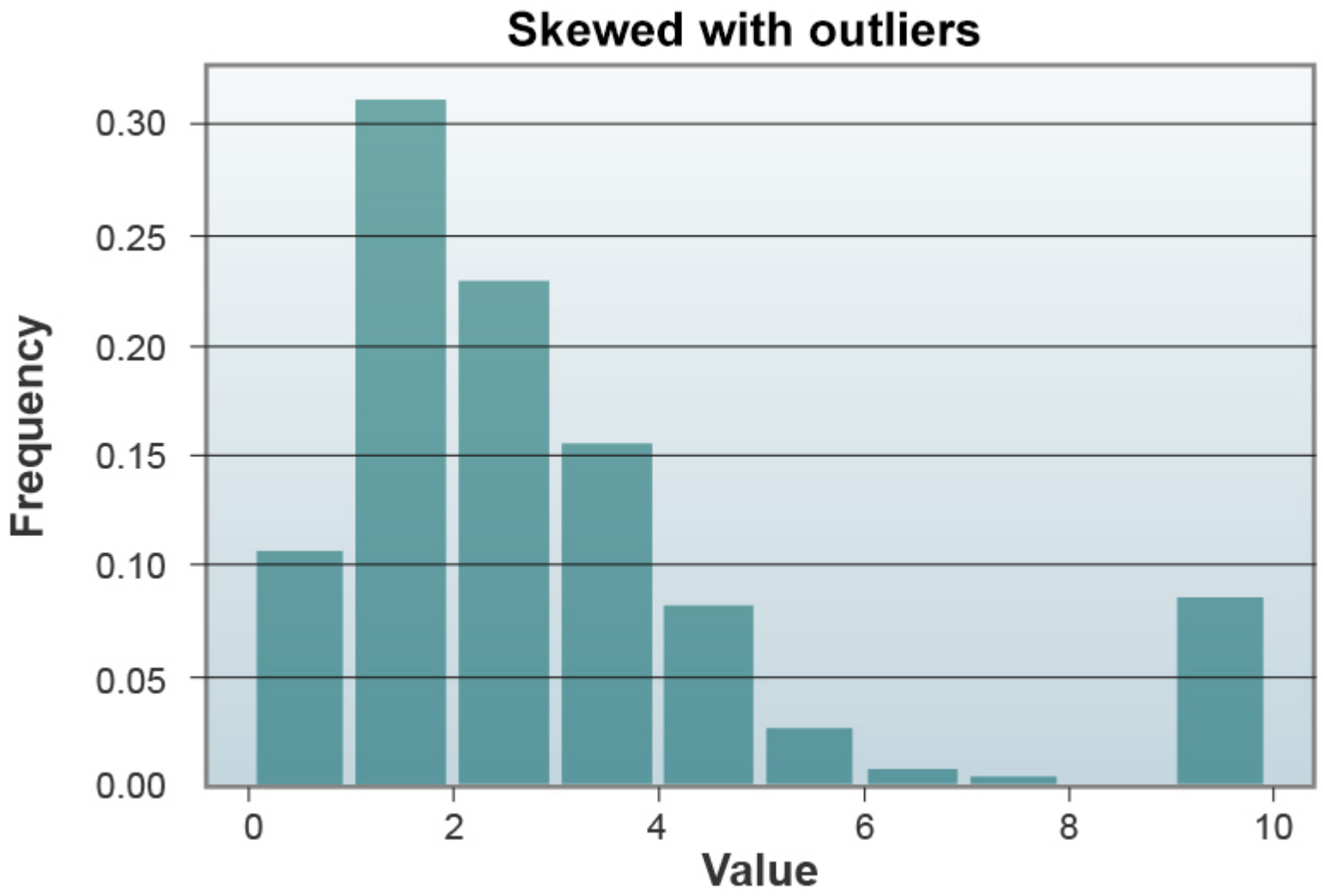
Histogram 2

Mean: 1.88 Median: 1.0



Histogram 3

Mean: 2.488 Median: 2.0



We see that for the first dataset, which is symmetric, the median and mean are relatively similar. If we have skewed data, such as in figure 2, we see that the median is lower and more reflective of the central tendency. Skewness indicates that more values are either below the mean like in the figure, which is confusingly called right skewed, or above the mean which is called left skewed.

When we have serious outliers, like in the third figure, we see that again the median is more reflective of the actual central tendency, as the mean is quickly drawn towards the outliers.

To get a better view on the meaning of the central tendency, we use measures for dispersion indicating to what extend the data is distributed around the central tendency.

The most well-known one is probably the (sample) standard deviation:

$$s = \sqrt{\frac{1}{n-1}\Sigma_{i=1}^{n}(x_i - \overline{x})^2}$$

It expresses the spread of the variables around the mean. Often the standard deviation is used to express when a certain observation is an outlier. By indicating whether an observation is 2 or 3 standard deviations removed from the mean, we get an idea of how extraordinary its appearance is.

We can also calculate differences at different points in our set of observations. First, there is the range:

$$R = x_n - x_1$$

This is the difference between the first and last observation.

Next, we can also use the interquartile range, defined as:

$$d_Q = \tilde{x}_{0.75} - \tilde{x}_{0.25}$$

All these measures can give you a good idea of how a continuous variable is distributed. The story for categorical values is different. We mainly have to resort to frequency tables. These will show whether certain values are particularly overrepresented. Frequencies can easily be converted into proportions, which can make the importance of certain values even more visible quickly.