Overview: Descriptive Statistics

What are descriptive statistics?

At the simplest level, descriptive statistics **summarise and describe relatively basic but essential features of a quantitative dataset** – for example, a set of survey responses. They provide a snapshot of the characteristics of your dataset and allow you to better understand, roughly, **how the data are "shaped"** (more on this later). For example, a descriptive statistic could include the **proportion** of males and females within a sample or the **percentages** of different age groups within a population.

Another common descriptive statistic is the humble **average** (which in statistics-talk is called the **mean**). For example, if you undertook a survey and asked people to rate their satisfaction with a particular product on a scale of 1 to 10, you could then calculate the average rating. This is a very basic statistic, but as you can see, it **gives you some idea of how this data point is shaped**.

What about inferential statistics?

Now, you may have also heard the term **inferential statistics** being thrown around, and you're probably wondering how that's different from descriptive statistics. Simply put, **descriptive statistics describe and summarise the sample itself**, while inferential statistics use the data from a sample to **make inferences or predictions about a population**.

Put another way, descriptive statistics help you **understand your dataset**, while inferential statistics help you **make broader statements about the population**, based on what you observe within the sample.

Why do descriptive statistics matter?

While descriptive statistics are relatively simple from a mathematical perspective, they play a **very important role in any research project**.

The reason for this is that descriptive statistics help you, as the researcher, **comprehend the key characteristics of your sample** without getting lost in vast amounts of raw data. In doing so, they provide a **foundation for your quantitative analysis**. Additionally, they enable you to quickly identify **potential issues** within your dataset – for example, suspicious outliers, missing responses and so on. Just as importantly, descriptive statistics **inform the decision-making process** when it comes to choosing which inferential statistics you'll run.

Long story short, it's **essential that you take the time to dig into your descriptive statistics** before looking at more "advanced" inferentials. It's also worth noting that, depending on your research aims and questions, **descriptive stats may be all that you need in any case**. So, don't discount the descriptives!

The "Big 7" descriptive statistics

Beyond the counts, proportions and percentages we mentioned earlier, we have what we call the "Big 7" descriptive. These can be divided into two categories – measures of central tendency and measures of dispersion.

Measures of central tendency

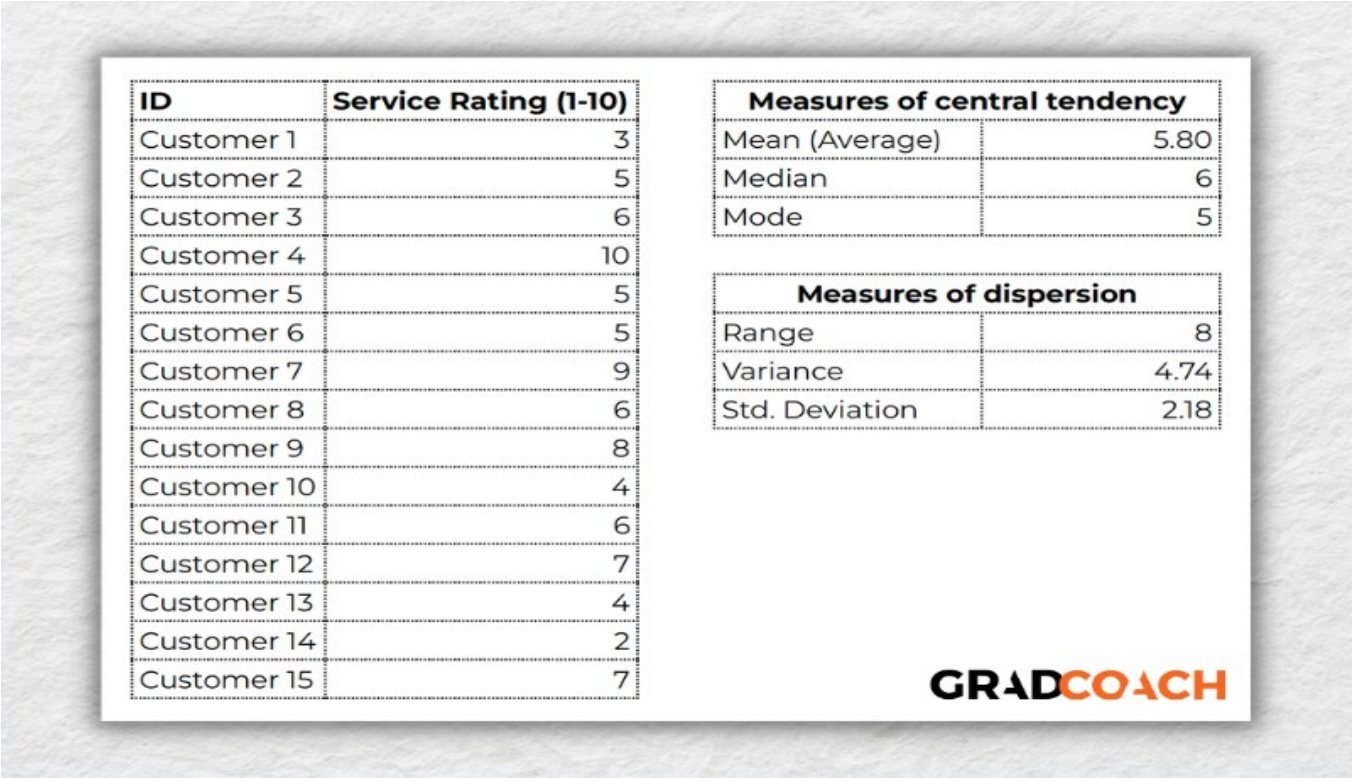
True to the name, measures of central tendency **describe the centre or "middle section"** of a dataset. In other words, they provide some indication of **what a "typical" data point looks like** within a given dataset. The three most common measures are:

The **mean**, which is the mathematical average of a set of numbers – in other words, the sum of all numbers divided by the count of all numbers.

The **median**, which is the middlemost number in a set of numbers, when those numbers are ordered from lowest to highest.

The **mode**, which is the most frequently occurring number in a set of numbers (in any order). Naturally, a dataset can have one mode, no mode (no number occurs more than once) or multiple modes.

To make this a little more tangible, let's look at a sample dataset, along with the corresponding mean, median and mode. This dataset reflects the service ratings (on a scale of 1 – 10) from 15 customers.



ID	Service Rating (1-10)
Customer 1	3
Customer 2	5
Customer 3	6
Customer 4	10
Customer 5	5
Customer 6	5
Customer 7	9
Customer 8	6
Customer 9	8
Customer 10	4
Customer 11	6
Customer 12	7
Customer 13	4
Customer 14	2
Customer 15	7

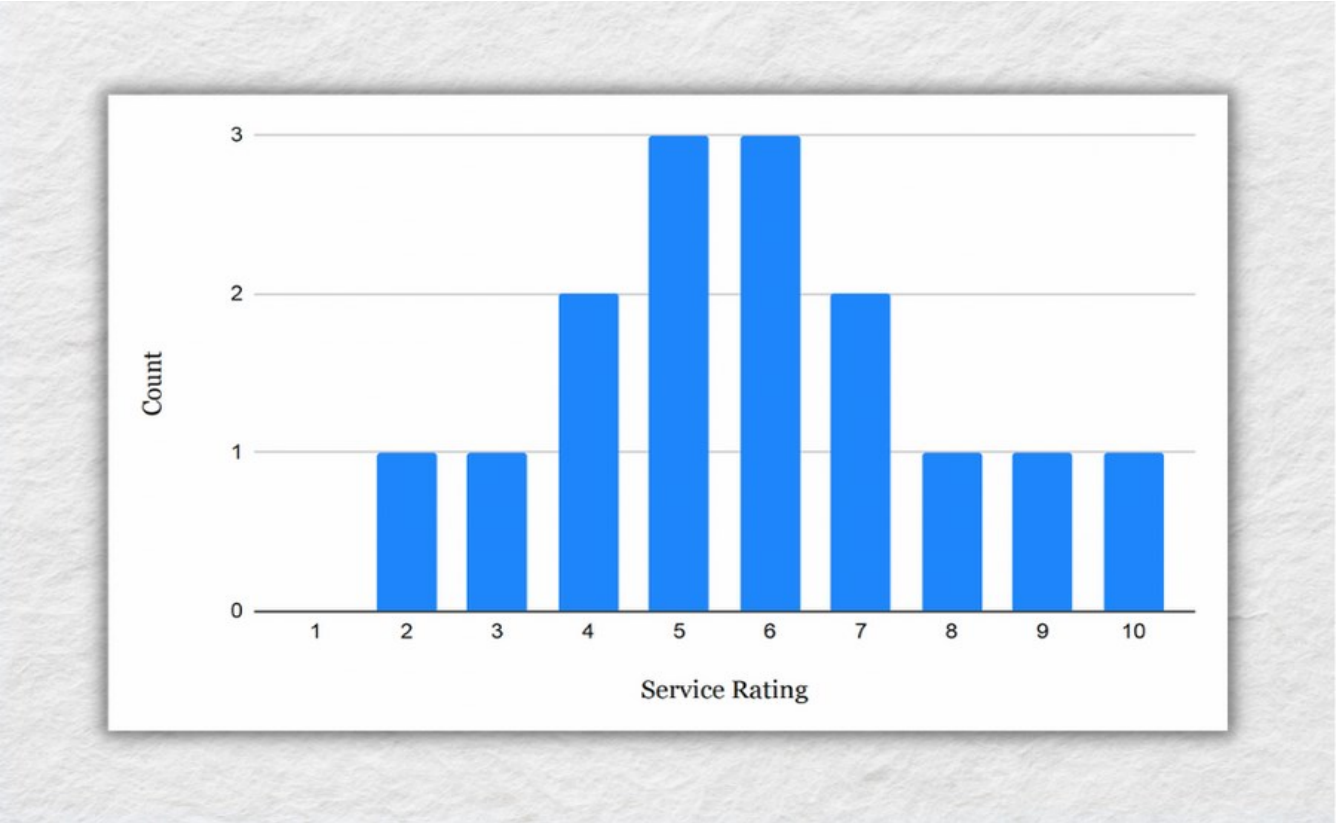
Measures of central tendency	
Mean (Average)	5.80
Median	6
Mode	5

Measures of dispersion	
Range	8
Variance	4.74
Std. Deviation	2.18

As you can see, the **mean of 5.8** is the average rating across all 15 customers. Meanwhile, **6 is the median**. In other words, if you were to list all the responses in order from low to high, Customer 8 would be in the middle (with their service rating being 6). Lastly, the number **5 is the most frequent rating** (appearing 3 times), making it the mode.

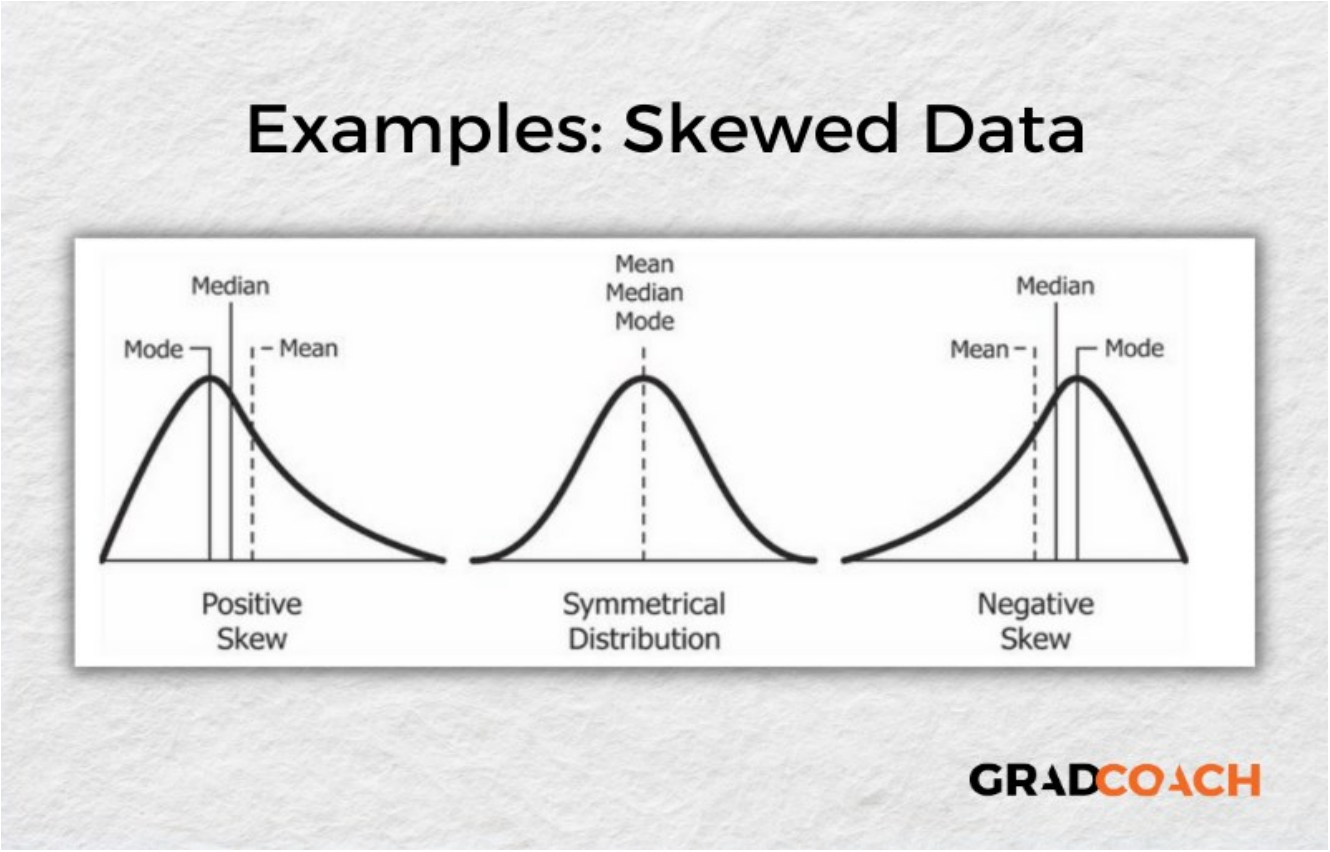
Together, these three descriptive statistics give us a **quick overview of how these customers feel about the service levels** at this business. In other words, most customers feel rather lukewarm and there's certainly room for improvement. From a more statistical perspective, this also means that **the data tend to cluster around the 5-6 mark**, since the mean and the median are fairly close to each other.

To take this a step further, let's look at the **frequency distribution of the responses**. In other words, let's count how many times each rating was received, and then plot these counts onto a bar chart.



As you can see, the responses **tend to cluster toward the centre of the chart**, creating something of a bell-shaped curve. In statistical terms, this is called a **normal distribution**.

As you delve into quantitative data analysis, you'll find that **normal distributions are very common**, but they're certainly not the only type of distribution. In some cases, the **data can lean toward the left or the right** of the chart (i.e., toward the low end or high end). This lean is reflected by a measure called **skewness**, and it's important to pay attention to this when you're analysing your data, as this will have an impact on what types of inferential statistics you can use on your dataset.



Measures of dispersion

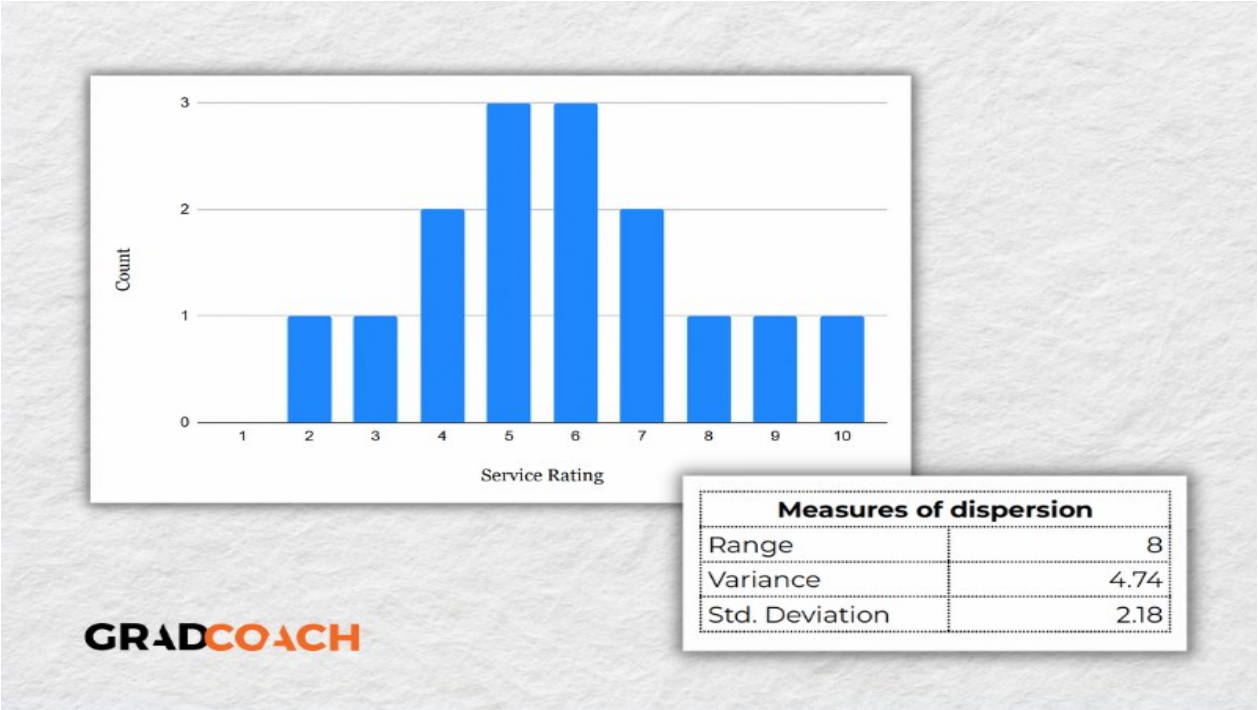
While the measures of central tendency provide insight into how "centred" the dataset is, it's also important to understand **how dispersed that dataset is**. In other words, to what extent the data cluster toward the centre – specifically, the mean. In some cases, the majority of the data points will sit very close to the centre, while in other cases, they'll be scattered all over the place. Enter the measures of dispersion, of which there are three:

Range, which measures the **difference between the largest and smallest number** in the dataset. In other words, it indicates how spread out the dataset really is.

Variance, which measures **how much each number in a dataset varies from the mean** (average). More technically, it calculates the average of the squared differences between each number and the mean. **A higher variance indicates that the data points are more spread out**, while a lower variance suggests that the data points are closer to the mean.

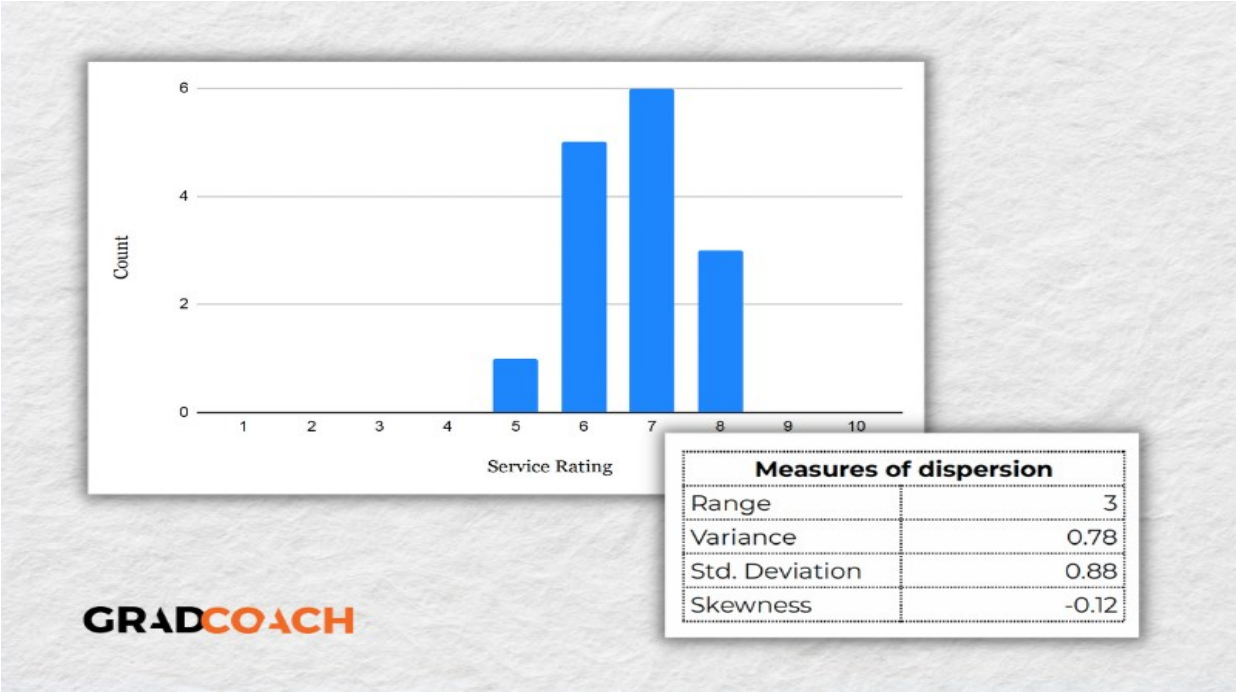
Standard deviation, which is **the square root of the variance**. It serves the same purposes as the variance, but is a bit easier to interpret as it **presents a figure that is in the same unit as the original data**.

Again, let's look at our sample dataset to make this all a little more tangible.



As you can see, the **range of 8** reflects the difference between the highest rating (10) and the lowest rating (2). The **standard deviation of 2.18** tells us that on average, results within the dataset are 2.18 away from the mean (of 5.8), reflecting a **relatively dispersed set of data**.

For the sake of comparison, let's look at another much more **tightly grouped (less dispersed) dataset.**



As you can see, all the ratings lay between 5 and 8 in this dataset, resulting in a **much smaller range, variance and standard deviation**. You might also notice

that the data are **clustered toward the right side of the graph** – in other words, the data are skewed. If we calculate the skewness for this dataset, we get a result of -0.12, confirming this right lean.

In summary, range, variance and standard deviation all **provide an indication of how dispersed the data are**. These measures are important because they **help you interpret the measures of central tendency within context**. In other words, if your measures of dispersion are all fairly high numbers, you need to **interpret your measures of central tendency with some caution**, as the results are not particularly centred. Conversely, if the data are all tightly grouped around the mean (i.e., low dispersion), the mean becomes a much more "meaningful" statistic).

Key Takeaways

We've covered quite a bit of ground in this post. Here are the key takeaways:

- Descriptive statistics, although relatively simple, are a **critically important** part of any quantitative data analysis.
- **Measures of central tendency** include the mean (average), median and mode.
- **Skewness** indicates whether a dataset leans to one side or another
- **Measures of dispersion** include the range, variance and standard deviation