

Introduction to data analysis

Michelle Arnetta

Summary

Data analysis is the process of inspecting and transforming data in order to extract useful insights. This guide will outline key techniques in descriptive and inferential statistics.

What is data analysis?

While it is often claimed that “data is the new gold,” the value of data primarily lies in how you can make sense of it. This prompts the need to learn **data analysis**, which is the process of inspecting and transforming data in order to extract useful insights.

This guide introduces you to data analysis. First, a distinction will be made between descriptive statistics and inferential statistics. Then on the descriptive statistics side, you will learn about measures of central tendency and dispersion, alongside data visualization. On the other hand, the inferential statistics section will include hypothesis testing, statistical relationships, and confidence intervals.

Distinguishing between descriptive statistics and inferential statistics

Descriptive statistics is concerned with summarizing and describing a set of data. For example, what is the “middle value” of the data, and how spread out is the data?

Before explaining **inferential statistics**, it is useful to establish what a **sample** and **population** is. A **population** is the totality of a group you want to study. Unfortunately, a population often contains too many individuals, so it is often inconvenient or sometimes even impossible to study every individual. Because of this, a **sample**, or a smaller subset of a population, is commonly taken for a statistical study.

Inferential statistics uses data samples to make inferences and predictions about the population that the sample was drawn from. This can be done by testing whether the values of two data samples are significantly different, examining whether two variables have a statistically significant relationship, and more that is beyond the scope of this study guide.

Descriptive statistics

Central tendency

Central tendency, often called the **average**, is the typical and middle value of a data set. The three most common measures of central tendency are **mean**, **median**, and **mode**.

Mean

Definition of mean

The **mean** is the sum of all values divided by the number of values.

The mean can be denoted by two symbols, depending on whether the data used is a sample or the whole population:

- \bar{x} = sample mean
- μ = population mean

Example 1

Many examples in this guide will use a very similar data set to each other, to show you that many different things can be done with even a single data set.

Cantor's Confectionery recorded a sample of the number of candies it had sold for the past 7 days. This can be summarized by the following list: {44, 32, 56, 76, 32, 32, 50}.

$$\bar{x} = \frac{44 + 32 + 56 + 76 + 32 + 32 + 50}{7} = 46$$

Median

Definition of median

The **median** is the middle value in a data set.

To find the **median**, the data set must first be arranged in numerical order.

- If the data set has an odd number of values, the **median** would be the value in the exact middle of the data set.
- If the data set has an even number of values, the **median** would be the two values in the middle of the data set, divided by two.

i Example 2

Cantor's Confectionery recorded a sample of the number of candies it had sold for the past 7 days. This can be summarized by the following list: $\{44, 32, 56, 76, 32, 32, 50\}$.

You can first order the list: $\{32, 32, 32, 44, 50, 56, 76\}$.

Notice that the list contains 7 elements, which is an odd number. So to find the middle element, you can divide the number of elements by 2 and round it up. In this case, $\frac{7}{2} = 3.5 \approx 4$.

The fourth element in the set, which is also the **median**, is 44.

i Example 3

Cantor's Confectionery recorded a sample of the number of candies it had sold for the past 6 days. This can be summarized by the following list: $\{44, 32, 56, 76, 32, 32\}$.

You can first order the list: $\{32, 32, 32, 44, 56, 76\}$.

Notice that the list contains 6 elements, which is an even number. So to find the two middle elements, you can divide the number of elements by 2. In this case, $\frac{6}{2} = 3$. Then the two middle elements would be the third and fourth elements.

The third element is 32, whereas the fourth element is 44. To find the median, you can sum these elements together and divide them by 2. So the **median** is $\frac{32+44}{2} = 38$.

Mode

i Definition of mode

The **mode** is the value that most frequently occurs in the data set.

i Example 4

Cantor's Confectionery recorded a sample of the number of candies it had sold for the past 7 days. This can be summarized by the following list: {44, 32, 56, 76, 32, 32, 56}. To find the mode, you can count how many times each unique value occurs in the data set. Here is a frequency table demonstrating that:

Unique Value	Frequency
32	3
44	1
56	2
76	1

Since 32 is the value that occurs in the data set for the highest number of times, it is the **mode**.

Why does this matter?

Calculating the central tendency can be important when you want to have a value that **summarizes the whole data set**, especially when you are making quick comparisons between two data sets. However, restricting yourself to finding only one measure of central tendency can lead to a misrepresentation of the data set.

For example, the **mean itself can be unreliable** because it risks being influenced by **outliers**, which are extreme values in the data set. If there exist extremely large values in the data set, the mean might be “pulled” higher than the actual center of the data. Similarly, if there exist extremely small values in the data set, the mean might be “pulled” lower than the actual center of the data. This is where comparing the mean to the mode and median would come in handy; if the mean differs greatly from the other measures of central tendency, this might indicate the presence of an outlier!

There are also cases where **finding the median and mode alone might be unhelpful**. In some data sets, there might not be any repeating values, so there is no point in attempting to find a mode. Other data sets may have unevenly weighted sampling. For instance, in a data set like {1, 1, 1, 50, 100}, the median of this data set would be 1, but it would be unintuitive to say that it accurately captures the center of the data.

Dispersion

The **dispersion** of a data set, also known as its **spread** or **variability**, refers to how scattered the data points are around the average. Four common measures of dispersion are **range**, **interquartile range**, **standard deviation**, and **variance**.

Range

Definition of range

The **range** is the difference between the **maximum** (highest value) and **minimum** (lowest value) of a data set.

Example 5

Cantor's Confectionery recorded a sample of the number of candies it had sold for the past 7 days. This can be summarized by the following list: {44, 32, 56, 76, 32, 32, 50}. You can first find the highest and lowest values from the list. The maximum is 76, whereas the minimum is 32.

To find the difference between these two values, you can **subtract the maximum by the minimum**. From this, the **range** is $76 - 32 = 44$.

Interquartile range

Definition of interquartile range

The **interquartile range** (often shortened to **IQR**) is the difference between the **upper quartile** and **lower quartile** of a data set.

- The **upper quartile** is the median of the data set's upper half.
- The **lower quartile** is the median of the data set's lower half.

In other words, it measures the dispersion of the middle 50% of the data set.

i Example 6

Cantor's Confectionery recorded a sample of the number of candies it had sold for the past 7 days. This can be summarized by the following list: $\{44, 32, 56, 76, 32, 32, 50\}$. You can first order the list: $\{32, 32, 32, 44, 50, 56, 76\}$. Then you can split the list into two subsets: the upper half and the lower half. Since 40 is the value in the exact middle, it will not be included in the subsets.

After splitting, the two subsets are $\{32, 32, 32\}$ and $\{50, 56, 76\}$. The medians of these subsets, respectively, are 32 and 56.

To find the difference between these two values, you can **subtract the upper quartile by the lower quartile**. From this, the **interquartile range** is $56 - 32 = 24$.

i Example 7

Cantor's Confectionery recorded a sample of the number of candies it had sold for the past 6 days. This can be summarized by the following list: $\{40, 32, 56, 76, 32, 32\}$. You can first order the list: $\{32, 32, 32, 40, 56, 76\}$. Then you can split the list into two subsets: the upper half and the lower half.

After splitting, the two subsets are $\{32, 32, 32\}$ and $\{40, 56, 76\}$. The medians of these subsets, respectively, are 32 and 56.

To find the difference between these two values, you can **subtract the upper quartile by the lower quartile**. From this, the **interquartile range** is $56 - 32 = 24$.

Standard deviation and variance

Variance is the average squared distance from the mean, while the **standard deviation** is the square root of the variance. Different notations and formulas are used for standard deviations and variances of a sample versus a population.

	Standard deviation	Variance	Notes
Population	$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$	$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$	μ is the population mean, and N is the size of the population
Sample	$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$	$s = \frac{\sum(x_i - \bar{x})^2}{n-1}$	\bar{x} is the population mean, and n is the size of the sample

i Example 8

Cantor's Confectionery recorded a sample of the number of candies it had sold for the past 7 days. This can be summarized by the following list: $\{40, 32, 56, 76, 32, 32, 50\}$. To find the **variance**, you can begin by finding the sample mean \bar{x} . In Example 1, you have already calculated that $\bar{x} \approx 38.4$.

Then you can calculate the **squared difference of each value from the sample mean** ($x_i - \bar{x}$):

- $(40 - 38.4)^2 = 2.56$
- $(32 - 38.4)^2 = 40.96$
- $(56 - 38.4)^2 = 309.76$
- $(76 - 38.4)^2 = 1413.76$
- $(32 - 38.4)^2 = 40.96$
- $(32 - 38.4)^2 = 40.96$
- $(50 - 38.4)^2 = 134.56$

To find $\sum(x_i - \bar{x})$, you can sum all the squared differences above.

$$\sum(x_i - \bar{x}) = 2.56 + 40.96 + 309.76 + 1413.76 + 40.96 + 40.96 + 134.56 = 1983.52$$

Finally, because the variance is $\frac{\sum(x_i - \bar{x})^2}{n-1}$, you can divide the newly calculated $\sum(x_i - \bar{x})$ by $n - 1$. Since the sales were recorded over 7 days, $n = 7$. As such, $n - 1 = 7 - 1 = 6$. Putting this all together,

$$s^2 = \frac{1983.52}{6} \approx 330.6$$

If you want to find the **standard deviation**, you can recall that the standard deviation is the **square root of the variance**. So, calculating the standard deviation,

$$s = \sqrt{330.6} \approx 18.2$$

For more information on standard deviation and variance, please read [Guide: Expected value, variance, standard deviation](#).

Why does this matter?

Measures of dispersion are useful for getting a **general overview of how spread out a data set is**, which again can be used for quick comparisons between multiple data sets. These measures become much more important later on when you dive into **inferential statistics**. For example, certain statistical tests involve comparing between multiple data sets, and to ensure reliable test results, the variances must remain similar across all the data sets being compared.

Similarly to the mean, **the variance and standard deviation are influenced by outliers**, possibly causing the value to drastically increase. As such, finding the interquartile range can be especially useful; because it measures the spread in the middle 50% of the data set, it will not be affected by outliers.

Data visualization

Data visualization involves representing data in a visual way, often highlighting patterns and outliers that would not have been visible in the raw data set. This study guide will cover four common data visualization types: bar charts, line graphs, histograms, pie charts, and scatter plots.

Bar charts

Bar charts use rectangular bars to represent data, making it useful for comparing data across multiple categories.

i Example 9

Cantor's Confectionery recorded a sample of the number of candies it had sold across different categories in the past week:

- 23 blueberry candies
- 49 grape candies
- 57 chocolate candies
- 16 orange candies

An employee wanted to quickly compare candy sales across different categories, so they decided to plot this data in the form of a bar chart: PICTURE HERE LATER

Note that bar charts do not have to be vertical. Bar charts are usually made to be vertical by default, but sometimes a horizontal format might be useful so that long label names do not overlap with each other.

Line graphs

Line graphs represent data values as a series of points on a graph, which are connected by line segments. This is useful for identifying trends and tracking changes across the x-axis (typically over time).

i Example 10

Cantor's Confectionery recorded a sample of the number of candies it had sold for the past 7 days. This can be summarized by the following list: {40, 32, 56, 76, 32, 32, 50}.

An employee wanted to see how the sales fluctuated over time, so they decided to plot this data in the form of a line graph: PICTURE HERE LATER

The number of sales do not appear to consistently increase or decrease, but the sales appear to peak on day 4, which may be worthy of further investigation.

Histograms

Histograms represent data in the form of bars stretching across a particular range of values. This is useful for visualizing the distribution of continuous data.

i Example 11

Cantor's Confectionery recorded the weights of their candies in the form of a frequency table:

Candy weights (in grams)	Frequency
$20 < x \leq 30$	2
$30 < x \leq 40$	3
$40 < x \leq 50$	5
$50 < x \leq 60$	4
$60 < x \leq 70$	1

An employee wanted to see how the weights data was distributed (in other words, the "shape" of the data), so they decided to plot this data in the form of a histogram:
PICTURE HERE LATER

While it is difficult to tell with such a small sample size, the weights data appears to be approximately normally distributed. This is because the data is symmetrical and forms a "bell shape" when plotted.

Pie charts

Pie charts illustrate data as sections of a circle or "slices of a pie", with each section representing a different category. This is useful for examining the relationship between different categories of data with the whole data set.

i Example 12

Cantor's Confectionery recorded a sample of the number of candies it had sold across different categories in the past week:

- 23 blueberry candies
- 49 grape candies
- 57 chocolate candies
- 16 orange candies

An employee wanted to visualize how each category compares to the whole and decided to plot this data in the form of a pie chart: PICTURE HERE LATER

As you can see, chocolate takes up most of the pie chart, followed by grape, blueberry, and orange.

Scatter plots

Scatter plots represent data in the form of points scattered in a graph, which is useful for identifying potential relationships between two variables. Moreover, scatter plots are often used in regression, which is covered in a later section in this guide.

i Example 13

Cantor's Confectionery recorded the data of candy prices, alongside each candy's number of sales. (This confectionery exclusively accepts a cryptocurrency named STARMASsTs).

The data is shown in the table below:

Candy Prices (in STAR- MASTs)	1.3	1.4	1.5	2.1	2.3	2.5	3.7	4.5	5.1	5.6
Num- ber of Sales	340	450	380	300	201	199	89	93	76	60

An employee wanted to see, at a glance, whether there was any correlation between candy prices and number of sales. They decided to plot this data in the form of a scatter plot: PICTURE HERE LATER

From this scatter plot, there appears to be an inverse trend: when the candy price increases, the number of sales decreases.

Inferential statistics

Hypothesis testing

Hypothesis tests help you to use data from a sample to test whether or not it is reasonable to believe a certain statistical characteristic is true for a whole population. A hypothesis test involves two hypotheses:

- Null hypothesis (H_0): This hypothesis represents the 'status quo' or no effect. It is always a statement of equality.
- Alternative hypothesis (H_1): This is the hypothesis that you are trying to test. It is always a statement of inequality.

(COPIED DIRECTLY FROM HYPOTHESIS TESTING GUIDE)

This hypothesis test will provide a **p-value**: the probability of obtaining a test result as extreme as the observed outcome if the null hypothesis is true. So the lower the p-value is, the more statistical evidence there is against the null hypothesis.

Before conducting the test, you should set a **significance level**, which is the level of certainty you want to test your hypothesis with. A commonly used significance level is 5% or 0.05. In the case of a 5% significance level, if you conduct a hypothesis test with under a 5% chance of obtaining the outcome if the null hypothesis is true, then you would consider it sufficient evidence to reject the null hypothesis.

i Example 14

Cantor's Confectionery has a candy machine that is supposed to produce 20 candies per bag. An employee suspects that the machine might be putting significantly more or less candies per bag than originally intended. With μ representing the population mean of candies per bag,

- $H_0 : \mu = 20$
- $H_1 : \mu \neq 20$

To test if the machine is working as it should be, the employee uses a two-tailed t-test with a significance level of 5%. The employee then compares the data set of candies per bag with a mean of 20 and obtains a high p-value > 0.05 .

This indicates that there is insufficient evidence for a statistically significant difference between the expected mean and the actual mean, so the candy machine is likely working as intended.

i Example 15

Cantor's Confectionery has conducted a marketing campaign and wants to test whether their sales have significantly increased after the marketing campaign. Let μ_1 be the population mean of sales before the marketing campaign, and μ_2 be the population mean of sales after the marketing campaign.

- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 < \mu_2$

To test this, they conducted a two-tailed two-sample t-test, with a significance level of 5%, between the sales data set before their campaign versus the sales data set after their campaign. From this test, they obtained a low p-value < 0.05 .

This indicates that there is sufficient evidence to suggest a statistically significant difference between the number of sales before and after the campaign, specifically with the number of sales increasing after the campaign was implemented.

For more information on hypothesis testing, please read [Guide: Introduction to hypothesis testing](#).

Correlation tests and regression analysis

A **correlation test** is a type of hypothesis test that determines whether the relationship between two or more variables is statistically significant. In testing for a linear relationship between two variables, the **Pearson correlation coefficient** (often denoted by r) is often used.

The value of this coefficient, which ranges from -1 to 1, indicates the strength and direction of the relationship between the variables. There are two possible directions for the relationship:

- Positive: when the value of one variable increases, the value of the other variable increases.
- Negative: when the value of one variable increases, the value of the other variable decreases.

While there are some gray areas, the value of the Pearson correlation coefficient is approximately interpreted in this way:

Value	Interpretation
0.80 to 1.00	Very strong positive correlation
0.60 to 0.79	Strong positive correlation
0.40 to 0.59	Moderate positive correlation
0.20 to 0.39	Weak positive correlation
0.00 to 0.19	Very weak positive correlation
-0.19 to 0.00	Very weak negative correlation
-0.39 to -0.20	Weak negative correlation
-0.59 to -0.40	Moderate negative correlation
-0.79 to -0.60	Strong negative correlation
-0.80 to -1.00	Very strong negative correlation

On the other hand, **regression analysis** uses a function to model the relationship between two or more variables. This would involve a **dependent variable** and one or more **independent variables**. To provide some reasoning for the naming, the dependent variable is the output of the regression function, so it is dependent on the values of any independent variables that are input into the function. So a **regression model** would allow you to predict the value of the dependent variable based on the value of the independent variable(s).

Regression analysis sometimes, but not always, deals with the linear relationship between two variables (one dependent variable and one independent variable). This would be known as **simple linear regression**, where the relationship is modeled as a linear function in the form of $y = mx + b$. The **coefficient of determination** (R^2 or r^2) indicates how well the model's independent variable explains the dependent variable, and the value of this ranges from 0 to 1. In simple linear regression specifically, the coefficient of determination is the squared value of the Pearson correlation coefficient.

Example 16

An employee from Cantor's Confectionery wishes to test if there is a statistically significant linear relationship between candy prices (independent variable) and number of sales (dependent variable), using the same data as Example 13. They decide to use simple linear regression for this purpose. This can be visualized below as a line of best fit overlaid on the scatter plot:

PICTURE HERE LATER

This can be modeled as the linear function $\text{sales} = -80.35(\text{price}) + 459.84$. The model's coefficient of determination is 0.8319, indicating that the price explains the number of sales well.

A Pearson correlation test indicated that the Pearson correlation coefficient is approximately -0.9121. This indicates a very strong negative relationship between the two variables. You can notice, also, that the coefficient of determination is the squared value of the Pearson correlation coefficient, as $(-0.9121)^2 \approx 0.8319$.

Tip

Correlation does not equal causation. For example, although candy prices and number of sales are strongly correlated, this does not necessarily show that one causes the other. A classic example is that there is a positive correlation between the number of ice cream sales and crime rates. Even so, it is likely not the case that ice cream sales cause crime rates to rise, or crime rates cause ice cream sales to increase. Rather, it is much more likely that there is a third variable (namely, the hot weather) that causes both ice cream sales and crime rates to increase!

Tip

It is important to note that 10 data points makes up quite a small sample size, and that a bigger sample size might lead to a more accurate regression model!

For more information on regression, please read [Guide: Introduction to regression.]

Confidence interval

i Definition of confidence interval

Confidence interval consists of an upper bound and lower bound used to estimate a population parameter (such as the population mean) within a certain probability. For example, a 95% confidence interval implies that if the sampling process were repeated 100 times, then 95 of the intervals calculated from those samples would contain the true population parameter.

Here is the formula for confidence interval:

$$CI = \bar{x} \pm z \frac{s}{\sqrt{n}}$$

- \bar{x} = sample mean
- z = z-score or confidence level value
- s = sample standard deviation
- n = sample size

i Example 17

Cantor's Confectionery recorded a sample of the number of candies it had sold for the past 7 days. This can be summarized by the following list: {40, 32, 56, 76, 32, 32, 50}. You want to calculate a 95% confidence interval for the population mean (the mean of all Cantor's Confectionery sales per day). To do this, you would need to find the sample mean \bar{x} , z-score, sample standard deviation s , and sample size n .

In Example 1, you have already calculated that $\bar{x} \approx 38.4$. In Example 8, you have also calculated that $n = 7$ and $s \approx 18.2$. Now you must find the z-score. By using a z-table, you would be able to find that the z-score for a confidence level of 95% is 1.96.

So, the 95% confidence interval for the population mean is:

$$CI = 38.4 \pm 1.96 \frac{18.2}{\sqrt{7}}$$

The upper bound is $38.4 + 1.96 \frac{18.2}{\sqrt{7}} \approx 51.9$, and the lower bound is $38.4 - 1.96 \frac{18.2}{\sqrt{7}} \approx 24.9$.

(IM CONCERNED THIS MAY BE WAY ABOVE THE DIFFICULTY LEVEL, NOT SURE WHAT TO DO WITH THIS EXAMPLE)

PMFs, PDFs, and CDFs

A **probability mass function (PMF)** is used for **discrete random variables**. It returns the probability that a discrete random variable X will take on a specific countable value x . So it can be expressed as $\mathbb{P}(X = x)$.

i Example 18

Cantor's Confectionery knows that the rate of customers entering the shop per hour can be modeled as $X \sim \text{Pois}(20)$.

An employee wants to find the probability of exactly 15 customers entering the shop in an hour. To find this, the employee could use a PMF, which can be represented as $\mathbb{P}(X = 15)$, approximately resulting in a probability of 0.052.

A **probability mass function (PDF)** is used for **continuous random variables**. It returns the probability that a continuous random variable X will take on a value that lies between a certain interval. So it can be expressed as $\mathbb{P}(a \leq X \leq b)$, where a is the lower bound of the interval and b is the upper bound of the interval.

i Example 19

Cantor's Confectionery knows that the lengths of their chocolate bars are normally distributed and can be modeled as $X \sim N(5.6, 1.44)$.

An employee wants to find the probability that a chocolate bar has a length between 3 to 5 inches. To find this, the employee could use a PDF, which can be represented as $\mathbb{P}(3 \leq X \leq 5)$, approximately resulting in a probability of 0.29.

A **cumulative distribution function (CDF)** is used for **both discrete and continuous random variables**. It returns the probability that a random variable X will take on a value that is less than or equal to a particular value x . So it can be expressed as $\mathbb{P}(X \leq x)$.

i Example 20

Cantor's Confectionery knows that the rate of customers entering the shop per hour can be modeled as $X \sim \text{Pois}(20)$.

An employee wants to find the probability of less than or equal to 15 customers entering the shop in an hour. To find this, the employee could use a PMF, which can be represented as $\mathbb{P}(X = 15)$, approximately resulting in a probability of 0.16.

Example 21

Cantor's Confectionery knows that the lengths of their chocolate bars are normally distributed and can be modeled as $X \sim N(5.6, 1.44)$.

An employee wants to find the probability that a chocolate bar has a length less than or equal to 5 inches. To find this, the employee could use a CDF, which can be represented as $\mathbb{P}(X \leq 5)$, approximately resulting in a probability of 0.31.

Tip

For continuous distributions, $\mathbb{P}(X = x) = 0$, so $\mathbb{P}(X \leq x) = \mathbb{P}(X < x)$.

Tip

In some cases, you might want to find the probability that a random variable X will take on a value that is greater than or equal to a particular value x . In other words, you want to find $\mathbb{P}(X \geq x)$. Since 1 represents the total probability of all outcomes, you can subtract the CDF result from 1 to get the desired result.

For continuous distributions, recall that $\mathbb{P}(X = x) = 0$, so:

$$\mathbb{P}(X \geq x) = 1 - \mathbb{P}(X < x) = 1 - \mathbb{P}(X \leq x)$$

For discrete distributions $\mathbb{P}(X \leq x) - \mathbb{P}(X = x) = \mathbb{P}(X \leq x - 1)$, so:

$$\mathbb{P}(X \geq x) = 1 - \mathbb{P}(X < x) = 1 - (\mathbb{P}(X \leq x) - \mathbb{P}(X = x)) = 1 - \mathbb{P}(X \leq x - 1)$$

For more information on PMFs, PDFs, and CDFs, please read [Guide: PMFs, PDFs, and CDFs](#).

(MIGHT BE NICE TO INTEGRATE INTERACTIVE FIGURES IN THIS SECTION)

Quick check problems

1. What is the median of this data set? $\{3, 1, 2, 4, 4\}$
2. What is the range of this data set? $\{16, 7, 900, 4, 3\}$.
3. A student wants to visualize their budget, seeing how the percentage of each spending category compares to the whole budget. What type of data visualization would suit this best?
4. You are given three statements below. Decide whether they are true or false.

