Statistics 101

Proudly built with R, Quarto and GitHub Copilot

Daniel Sánchez Pazmiño

Table of contents

Sessions 1-3: Introduction to Data and Statistics

Statistics is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions. Statistics is also a collection of tools and methods that you can use to get answers to important questions about data.

Thus, data is involved in all aspects of statistics. It is important to understand how data appears to us so that we can understand what happens under the hood of statistical analysis procedures.

Data & Tables

- What are data?

Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation. All the data collected in a particular study are referred to as the data set for the study. Often, data is presented in a table.

- What are variables, columns, rows, observations?

First, elements are the entities or things for which we collect data. A variable, often seen as a column, is a specific characteristic for the element. An observation, often seen as a row is a collection of variables for a given element.

For instance, see the following table.

Table 1: A fake table with fake data.

Name	Age	Height
John	20	175
Mary	19	165
Jane	21	170

Who are the elements? What are the variables? How many observations?

Elements are *not* the same as observations. Sometimes, there might be the same number of entities and observations, but this is not always the case. Consider the following table.

Table 2: A fake panel dataset with fake data.

Name	Age_2019	Age_2020	Height_2019	Height_2020
John	20	21	175	176
Mary	19	20	165	166
Jane	21	22	170	171

How many observations? How many entities?

The table above is known as a *wide form* table, where each of the elements has only one row and multiple columns for that row represent a variable for it, often at different points in time. In this case, we have the same number of elements and observations.

Wide forms are useful for some purposes, but not for others. For instance, it is not easy to compute the average age of the three people in the table above, as you would need to use

information for more tha one column. However, wide tables are often used to present data because they are most intuitive to *humans* reading the data.

The table below presents the exact same information, but in *long* form.

Table 3: A fake panel dataset with fake data.

Name	Age	Height	Year
John	20	175	2019
Mary	19	165	2019
Jane	21	170	2019
John	21	176	2020
Mary	20	166	2020
Jane	22	171	2020

Long form tables present elements with multiple observations, or rows. While long forms are not as intuitive to humans, they are much easier to work with for computers. It is important to note that the same information can be presented in both wide and long forms.

In statistical and data analytics computing, changing the format of data tables from wide to long or viceversa is often called *reshaping* or *pivoting* the data. Though the syntax of these processes is often difficult to memorise, the logic behind them is simple if you understand the difference between wide and long forms.

- What types of variables are there?

Variables can be classified into two types: qualitative and quantitative. Qualitative variables are those that describe an element in terms of some characteristic or attribute. Quantitative variables are those that describe an element in terms of some numerical measurement.

Table 4: A fake table with quantitative and qualitative variables.

Name	Grade	Color
John	20	Blue
Mary	19	Red
Jane	21	Green

- Under what forms datasets can be presented?

All the data collected about different elements on different variables for a particular study are called a dataset. Datasets can be divided in different formats. The most important for our purposes are *cross-sectional*, *panel* and *time series* datasets.

Cross-sectional datasets are those that contain information about different elements at the same point in time. For instance, the table below presents information about three people in 2019.

Table 5: A fake cross-sectional dataset with fake data in 2019.

Name	Age	Height
John	20	175
Mary	19	165
Jane	21	170

We also usually hear about *repeated cross-sectional datasets*, or *pooled cross-sections*. These are datasets that contain information about different elements at different points in time.

For instance, the long form table below presents information about three people in 2019 and 2020. Know that whenever time is involved, the *raw* data is usually in long form.

Table 6: A fake repeated cross-sectional dataset with fake data in 2019 and 2020.

Name	Age	Height	Year
John	20	175	2019
Mary	19	165	2019
Jane	21	170	2019
Karen	21	176	2020
Joseph	20	166	2020
Joan	22	171	2020

Notice the *Name* column above. The same people are not being observed every year. Pooled cross sections are useful to increase the amount of data available for our analysis, but they are not as useful as panel datasets.

Panel datasets are those that contain information about the same elements at different points in time. The table below presents information about three people in 2019 and 2020.

Table 7: A fake panel dataset with fake data in 2019 and 2020.

Name	Age	Height	Year
John	20	175	2019
Mary	19	165	2019
Jane	21	170	2019
John	21	176	2020
Mary	20	166	2020
Jane	22	171	2020

Look at the dataset above. Notice that the same people are being observed every year. This is what makes panel datasets so useful. We can observe the same person in different points time, which helps document the evolution of the variable of interest for that person. Panel data is very important data in the social sciences such as economics and education as it is the closes we can get to a controlled experiment.

Finally, we have time series datasets. These are datasets that contain information about the same element at different points in time. The table below presents information about the height of John in several different years.

Table 8: A fake time series dataset with fake data from 2015 to 2020

Name	Height	Year
John	175	2015
John	176	2016
John	177	2017
John	178	2018
John	179	2019
John	180	2020

Time series data is also important, especially for macroeconomics. You can think of time series data as a panel dataset with only one element, though it is possible that when you have many, many observations for different elements, the dataset is called a time series too. In here, we use time-series to only refer to datasets with one element but many observations in time.

- How to get data?

There are many ways to get data. The most common form for our purposes will be pulling data which was already collected by someone else. This is called *secondary data*.

It's important, however, to understand the processes of how secondary datasets are collected. For the social sciences, we often collect data from *observational studies*. These are studies where we observe the world as it is, without trying to change it. Surveys are a common form of observational studies. If we want to conclude anything about the world from observational studies, we have to be careful about *selection bias*, *ommitted variable bias*, among others.

Experimental studies are studies where we try to change the world in some way and observe the consequences of that change. For instance, we could try to change the way we teach a class and observe the consequences of that change. Experimental studies happen in *controlled* environments, where we can control the variables that are not of interest to us. Thus, randomisation is a key feature of experimental studies, as it allows us to isolate the effect of the variable of interest and thus avoid *selection bias* and *ommitted variable bias*.

Experimental studies are the most useful (and probably the only) way to analyse the impact of a change in a variable on another variable. However, they are not always possible. For instance, we cannot randomly assign people to different levels of education. Thus, we often have to rely on observational studies, but we have to be weary of *selection bias* and spuriuous correlations. We will talk about those things another day.

Descriptive Statistics

Most of the time, we will be interested in summarising the data we have. We will do that by calculating *descriptive statistics*. Imagine we have a dataset of 10,000 people, observing their GPA and their monthly wage. We could calculate the average GPA and the average wage. We could also calculate the median GPA and the median wage. We could also calculate the standard deviation of GPA and the standard deviation of wage. These are all descriptive statistics.

Table 9: A fake dataset with 10,000 observations of education and wage

id	Education.years	Wage
1	8	3137
2	19	8718
3	10	9361
4	4	8635
5	4	9232
6	8	7225
7	15	9485
8	5	7093
9	6	1623
10	7	2937

Notice how we observe `id`, an *identification variable*. Identification variables are variables that uniquely identify an observation. In this case, `id` is a variable that uniquely identifies a person. Identification variables are very important, as they allow us to link different datasets. For instance, we could have another dataset with the same people, but with their height and weight. We could then link the two datasets using the `id` variable. Joining datasets is an important skill in statistics.

Below we calculate some descriptive statistics for the dataset above.

Table 10: Descriptive statistics for the education and wage dataset

Education.years	Wage
Min. : 1.00	Min. : 1000
1st Qu.: 6.00	1st Qu.: 3304
Median :11.00	Median : 5632
Mean :10.62	Mean : 5548
3rd Qu.:16.00	3rd Qu.: 7779
Max. :20.00	Max. :10000

Many other descriptive statistics can be calculated for data. The most common ones are the mean, the median, the standard deviation, the minimum and the maximum, but others exist like the mode, the variance, the skewness, the kurtosis, etc.

Central Tendency Descriptive Statistics

The mean, the median and the mode are all measures of central tendency. They are all measures of the "middle" of the data. The mean is the average of the data. We calculate an average as follows:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where n is the number of observations and x_i is what we call the $i-th$ observation. It is simply the variable indexed by its position in the dataset. We always use it in summation notation (the $\sum$ sign above). Note how dividing everything by n is the same as multiplying everything by $\frac{1}{n}$.

The median is the middle observation of the data. When calculated by hand, we have to order the data from smallest to largest and then pick the middle observation. If there are an even number of observations, we take the average of the two middle observations.

For instance, consider a dataset as follows:

Table 11: Calculating a median by hand

x	sorted_x
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10

In the table above, we have a dataset with 10 observations. The median is the middle observation, which is 5.5 (the average of 5 and 6). The median is also known as the 50th percentile, as 50% of the data is below the median. So, in the dataset above, 50% of the data is below 5.5. It is also called the second quartile.

The median is a useful measure of central tendency, as it is not affected by outliers. For instance, consider the following dataset:

Table 12: Two variables, x and y, with the same mean, but different medians

x	y
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	1000

The average of x in the dataset above is 5.1, while the average of y is 115.11. These two datasets are very similar with the only difference that y has an outlier, which is 1,000. Outliers are

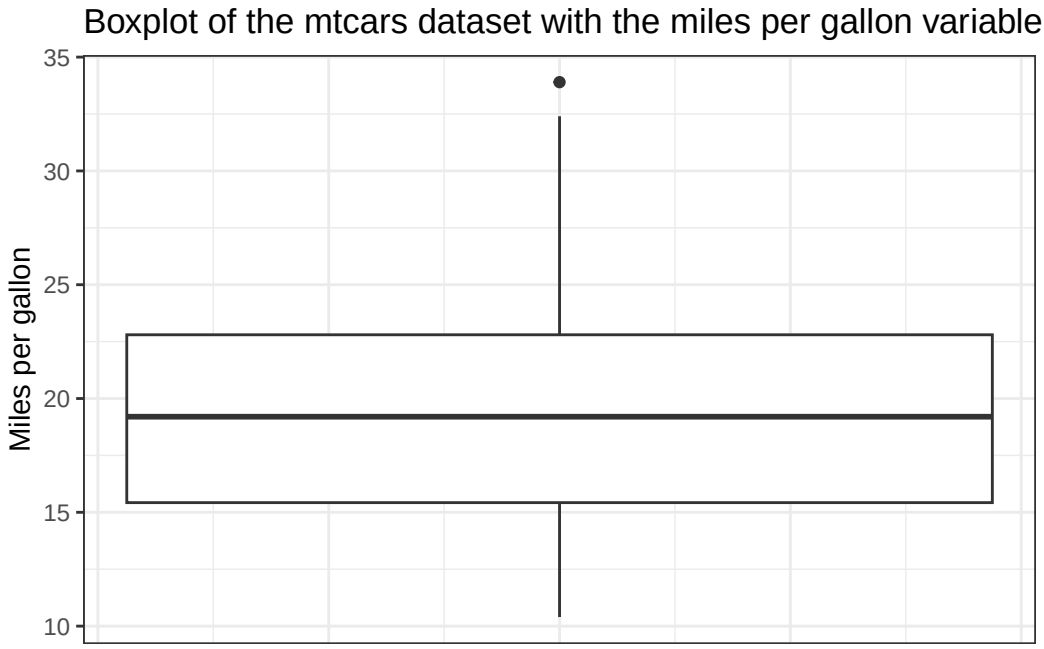
observations that are very far away from the rest of the data. However, the median of x is 5.5, while the median of y is also 5.5. Thus, the median is often less affected by outliers.

The *interpretation* is what is most important. Above, the median is 5.5. This means that 50% of the data have a value which is less than 5.5. Often we prefer to summarise the data by the median rather than the average, because we know there are a lot of outliers. An example is wages, where we know that most people make a relatively low wage, but there are a few people who make a lot of money. The average will be very high, but the median will be lower and will tell us a better idea of what is the typical wage, or the central tendency of the wage distribution.

Percentiles, though not precisely measures of central tendency, are also useful measures . The pth percentile is the value of the variable below which p% of the data is. For instance, the 25th percentile is the value of the variable below which 25% of the data is. The 25th percentile is also known as the first quartile. The 75th percentile is the value of the variable below which 75% of the data is. The 75th percentile is also known as the third quartile. It is a bit more difficult to calculate percentiles by hand, so we normally rely on software to calculate them for us.

You will often hear about the five-number summary of the data. The five-number summary is the minimum, the first quartile, the median, the third quartile and the maximum. It gives you a good idea of the distribution of the data. Further, you will also hear a lot about the IQR, or the interquartile range. The IQR is the difference between the third quartile and the first quartile. It is a measure of variability as well, which is very simple to calculate. Once you've obtained the values for the first and third quartiles, you simply subtract the first quartile from the third quartile. It is like the range, but it is not affected by outliers, thus, it will always be a better measure of variability than the range.

The IQR is always contained within the five number summary, as the five number summary presents the minimum, first quartile, median, third quartile, and the maximum. We can plot the IQR and the other information in the five number summary in graphs called boxplots. These are very useful graphs to summarise the data, but it is important to know how to interpret them. In a boxplot, the box represents the IQR, the line in the middle of the box represents the median, and the whiskers represent the minimum and the maximum. However, in some cases the whiskers do not represent the minimum and the maximum if there are outliers. In those cases, the whiskers represent one and a half times the IQR and the outliers are represented as points.



The final central tendency statistic is the mode. The mode is the most common observation of the data. However, it is not always possible to calculate the mode, as there might be more than one observation that is the most common. When there are two observations which are the most common, we say that the data is *bimodal*. When there are more than two observations which are the most common, we say that the data is *multimodal*. In multimodal data, the mode offers little information about the data.

Measures of Variability

The mean, the median and the mode are all measures of central tendency. However, they do not tell us anything about the variability of the data. There are three main measures of variability: the range and the variance. The standard deviation is the most commonly used measure of variability, but it is nothing but the square root of the variance, so it is the range and the variance that we will focus on.

The range is the difference between the maximum and the minimum of the data. It is the simplest measure of variability.

On the other hand, the variance is the average squared difference between each observation and the mean. As a shorthand, the variance is often called σ^2.

The formula for the variance is as follows:

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

where n is the number of observations, x_i is the ith observation and $\bar{x}$ is the mean of the data.

The formula for the variance is a bit complicated, but it is not too difficult to understand when we calculate it by hand for a simple dataset. Consider the following dataset:

Table 13: Calculating the variance by hand

x	mean_x	diff_with_mean	diff_sq
1	3	-2	4
2	3	-1	1
3	3	0	0
4	3	1	1
5	3	2	4

If we sum the squared differences and divide by the number of observations, we get the variance. The sum of the last column is 10, and we divide it by 5, the number of observations, to get 2. Thus, the variance is 2.

If you use a statistical software to calculate the variance, you will get that the variance of the dataset is 2.5. The reasons is because statistical software divides by $n-1$, not by n. We will see why when we talk about statistical inference.

Descriptive Statistics for Categorical Data

Categorical data cannot be summarised using the mean, the median, the standard deviation, the minimum and the maximum. For categorical data, we can calculate the frequency of each category. The frequency is computed by dividing the number of cases within each category by the total number of cases, as follows:

$$\text{Frequency} = \frac{\text{Number of cases in a category}}{\text{Total number of cases}}$$

Consider the following dataset of cars and their characteristics:

Table 14: The first 10 observations of the mtcars dataset

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4

As you see, the name of this dataset is `mtcars`. It is a famous dataset which is contained in the R Statistical Software. It is often used in the internet to illustrate examples of statistical computing, so it is useful to understand it. It contains information about 32 cars. The variables are as follows:

- `mpg`: Miles/(US) gallon
- `cyl`: Number of cylinders
- `disp`: Displacement (cu.in.)
- `hp`: Gross horsepower
- `drat`: Rear axle ratio
- `wt`: Weight (1000 lbs)
- `qsec`: 1/4 mile time
- `vs`: V/S, which means V-engine or straight engine.
- `am`: Transmission (0 = automatic, 1 = manual)
- `gear`: Number of forward gears
- `carb`: Number of carburetors

Table 15: Frequencies and percentages of each level of the variable vs

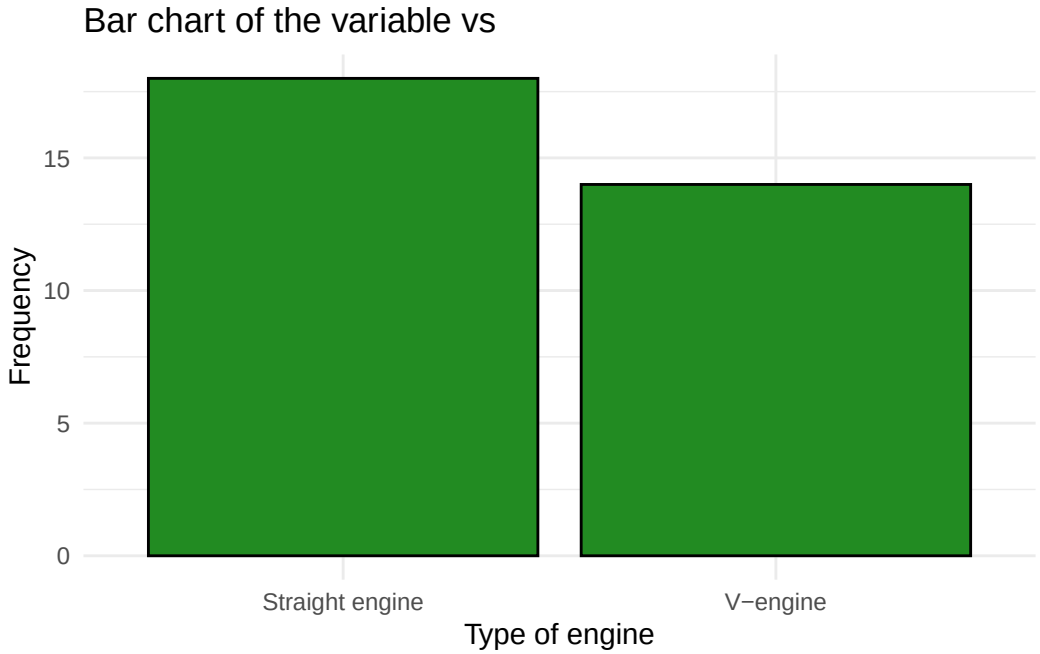
vs	n	percent
Straight engine	18	0.5625
V-engine	14	0.4375

The frequencies of the categories, or the *frequency distribution* is nothing but the number of observations in each category. The percentage is the frequency divided by the total number of observations. It's also often called the *relative frequency*.

Data Visualisation

Visualising Categorical Data

Categorical data can be visualised using bar charts. Bar charts are very simple to make. We just need to count the number of observations in each category and plot them.

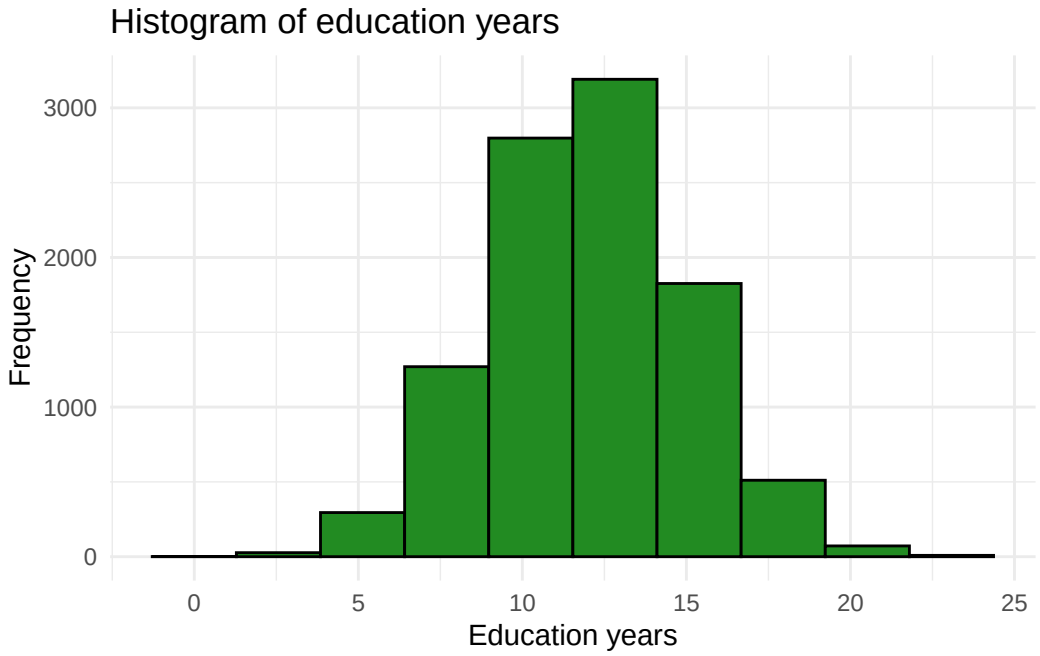


Other types of visualisations for categorical data include pie charts and donut charts, as well as waffle charts.

Visualising Numerical Data

Numerical data can be visualised using histograms. If we the education years for 10,000 people, we can make a histogram of the education years. The histogram will tell us about the *distribution* of education years in that dataset, i.e. how many people have 1 year of education, how many have 2 years, how many have 3 years, etc.

For that, we *group* the data in bins or intervals. How much people have 1-2 years of education? How much people have 3-5? And so on. We then plot the number of people in each bin on a graph.



It is often said that histograms inform us about the *frequency distribution* of a variable. Don't confuse this with the frequency distributions that categorical variables have, as in histograms we are not counting the number of observations in each category, we are just *binning* the data and then graphing the amount of observations in each bin.

There are other ways to visualise numerical data. An important one is the *cumulative* distribution. The cumulative distribution is the sum of the frequencies of each bin. For instance, if we have 10,000 observations, and we have 1,000 observations in the first bin, 2,000 in the second, 3,000 in the third, etc, the cumulative distribution will be 1,000 in the first bin, 3,000 in the second, 6,000 in the third, etc.

Below, we calculate the cumulative distribution of a dataset which contains the education years of people. We first make a cumulative distribution table, which contains the number of people in each bin and the running total.

Table 16: Cumulative distribution table

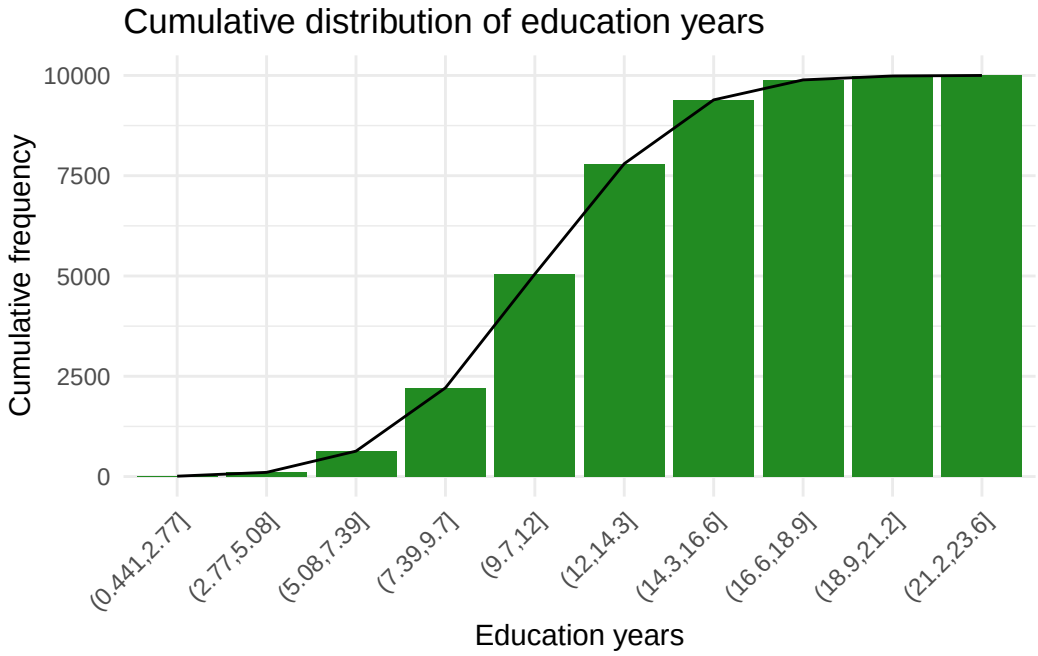
education_bin	n	percent	cumulative
(0.441,2.77]	9	0.0009	9
(2.77,5.08]	95	0.0095	104
(5.08,7.39]	529	0.0529	633
(7.39,9.7]	1578	0.1578	2211
(9.7,12]	2837	0.2837	5048
(12,14.3]	2752	0.2752	7800
(14.3,16.6]	1591	0.1591	9391
(16.6,18.9]	498	0.0498	9889
(18.9,21.2]	98	0.0098	9987
(21.2,23.6]	13	0.0013	10000

For instance, the first bin contains the number of people who report having 0.4 (less than six months) to 2.77 years of education. There are 9 people in this bin. So, the percent of the total number of observations is $9/10000 = 0.0009$. The cumulative column in that row takes a value of 9, as there are 9 who have less than 2.77 years of education.

The second bin contains the number of people who report having more than 2.77 to 5.13 years of education. There are 95 people in this bin. So, the percent of people who are in this bin is $95/10000 = 0.0095$. The cumulative column in that row takes a value of 104, as there are $9 + 95 = 104$ people who have less than 5.13 years of education. Thus, the cumulative distribution is the sum of the frequencies of each bin. You also see that there is a relative cumulative frequency. It is a big name but it is only the division of the cumulative frequency by the total number of observations.

Once we have the cumulative distribution table, we can plot it. We can plot the cumulative frequency against the education years. This will tell us how many people have less than 1 year of education, how many have less than 2 years, how many have less than 3 years, etc.

Further, often the relative cumulative frequency is mentioned. The relative cumulative frequency is the cumulative frequency divided by the total number of observations. It is the same as the cumulative frequency, but it is expressed as a percentage. For instance, if we have 10,000 observations, and we have 1,000 observations in the first bin, 2,000 in the second, 3,000 in the third, etc, the relative cumulative frequency will be 10% in the first bin, 30% in the second, 60% in the third, etc.



Here I graphed the cumulative distribution graph with bars and a line. Often, these graphs are presented with the line only. While these graphs might look scary, it's important to keep in mind what they are telling us. They are telling us how many observations are below a certain value. For instance, in the graph above, we can see that 50% of the observations are below 11.5 years of education. It is useful to keep the table in mind when looking at the graph.

Statistical Inference

Statistical inference is the process of making conclusions about a population based on a sample. In all of the examples above, we were working with fake data that I generated here and we assumed that it was the whole population.

But in real life, we don't have access to the whole population. For example, if we want to know the average education years for an entire country, it is probably impossible to survey all people living in the country to ask them what is their education level. How can we know the average education years for the whole country if we don't have access to the whole population?

In those cases, we take a *sample* of the population and make inferences about the population based on the sample. A sample is a subset of the population. For instance, if we take a sample of 100 people, we can calculate the average education years for those 100 people. If we believe that the sample is *representative* of the population, we can use the average education years of the sample to make a guess about the average education years of the population. The statistical agencies from countries, such as Statistics Canada, worry about creating surveys of

samples that are representative of the population. The most common surveys coming from these types of agencies are the census and the employment survey.

At some point, we will stop talking about descriptive statistics only and start worrying about their *accuracy*. If we take a sample of 100 people, and we calculate the average education years for those 100 people, how close is that average to the average of the population? While we don't observe the population, statistics and probability will allow us to make some predictions about this accuracy.

Differences between sample statistics and population parameters

Above, we calculated an average and a variance. We figured that we had access to the whole population. But in real life, we don't have access to the whole population. We only have access to a sample. So, we need to make a distinction between the statistics that we calculate from a sample and the parameters that we could calculate from a population if we had it.

Assuming we had access to the whole population, we could calculate the average education years of the population. We would call this the *population parameter*. We would denote it with the Greek letter mu (μ). However, if we have access to a sample, we can calculate the average education years of the sample. We would call this the *sample statistic*. We would denote it with the letter x-bar ($\bar{x}$).

There are no differences in the calculation of the average education years of the population and the average education years of the sample. The only difference is that we use different symbols to denote them. It is useful, anyway, to see how the notation varies when expressing the formula for this.

The formulas for the average education years of the population and of the sample is:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

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

Notice that the only change is that instead of using the lowercase n, we use uppercase N to denote the population size. N is the total number of observations in the population, while n is the total number of observations in the sample. So it is sensible to believe that N is bigger than n ($N > n$).

The variance at the population level is often denoted as σ^2, while at the sample level it is often denoted s^2. The variance does show a difference in calculation between the population parameter and the sample statistic. The population parameter is calculated as:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

While the sample statistic is calculated as:

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

The adjustment is that instead of dividing by N, we divide by $n-1$. This is called the *Bessel's correction*. It is a correction that is applied to the sample variance to make it an unbiased estimator of the population variance. We will not go into the details of this correction here. But it is important to know that the sample variance is calculated differently than the population variance.

Summary of Sessions 1-3

- Statistics is the science of collecting, analysing, and interpreting data.
- Data can be presented in tables which are long or wide. Long tables have multiple observations per element, whereas short or wide tables have one observation per element.
- Data can be categorical or numerical, and it can be presented in cross sectional, time-series, or panel format.
- Categorical data can be visualised using bar charts, pie charts, donut charts, and waffle charts.
- Numerical data can be visualised using histograms and cumulative distributions.
- Descriptive statistics is the branch of statistics that deals with describing data.
- The most important descriptive statistics of central tendency and variability are the mean, the median, the mode, the variance, and the standard deviation.
- The mean is the average of the data, the median is the middle value of the data, and the mode is the most frequent value of the data.
- The variance is the average squared distance from the mean, and the standard deviation is the square root of the variance.
- Inferential statistics is the branch of statistics that deals with making inferences about a population based on a sample.
- The difference of calculation at the sample vs. population level is the Bessel's correction, which is applied to the sample variance to make it an unbiased estimator of the population variance. Instead of dividing by N, we divide by $n-1$ in the sample variance.