How to Create Frequency Distributions

Distributions hold significant theoretical importance in this course. In this section, we will learn about frequency distributions, which is the basis for many of the concepts in this course.

Basics of Frequency Distributions

What is Frequency?

The _________ of a data point is how many times it shows up in the data set.

What is a Frequency Distribution?

A ______________________ is a table that lists either the raw data or classes (defined below) in the first column and corresponding frequency in the second column.

Before the advent of computers, frequency distributions were created manually. This involved arranging the data in ascending order and marking a tally for each occurrence of a data point. This process was prone to human errors such as miscounting or duplicating data.

With computers, creating frequency distributions is much faster and more accurate than manual methods. Tools like SPSS, MATLAB, and Excel offer advanced statistical capabilities. GeoGebra, however, is particularly useful in educational settings because of its simplicity and visualization tools, making it ideal for students learning this process.

Example 1

The data on the number of deaths directly caused by tornadoes in Tennessee is given in the table below. Create a Frequency Distribution for the number of deaths per year.

Copy Data to Clipboard

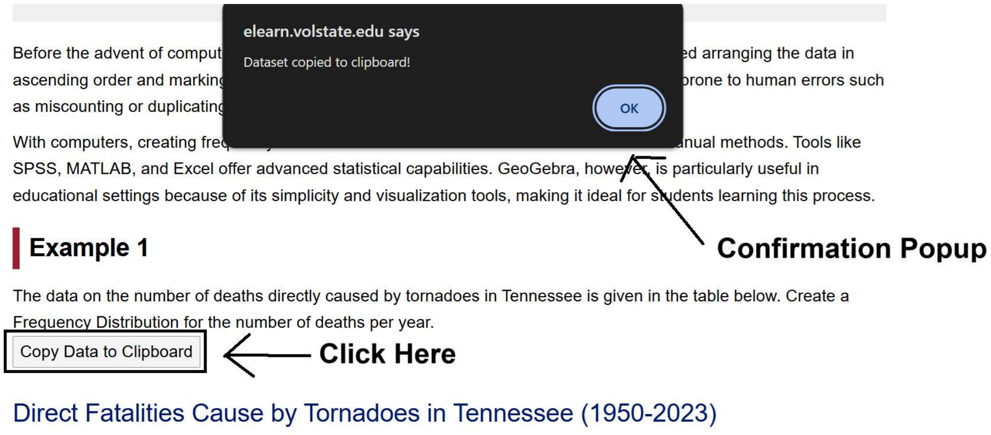
Direct Fatalities Caused by Tornadoes in Tennessee (1950-2023)

Year	Direct Fatalities	Year	Direct Fatalities	Year	Direct Fatalities	Year	Direct Fatalities
2023	17	2004	0	1985	0	1966	0
2022	0	2003	12	1984	1	1965	1
2021	4	2002	17	1983	0	1964	0
2020	27	2001	3	1982	0	1963	4
2019	0	2000	1	1981	0	1962	0
2018	1	1999	12	1980	0	1961	0

Year	Direct Fatalities	Year	Direct Fatalities	Year	Direct Fatalities	Year	Direct Fatalities
2017	0	1998	7	1979	2	1960	0
2016	2	1997	1	1978	0	1959	0
2015	2	1996	0	1977	0	1958	0
2014	2	1995	3	1976	0	1957	0
2013	0	1994	5	1975	3	1956	3
2012	3	1993	1	1974	47	1955	0
2011	32	1992	1	1973	1	1954	0
2010	1	1991	5	1972	0	1953	4
2009	2	1990	0	1971	3	1952	75
2008	31	1989	1	1970	3	1951	0
2007	0	1988	6	1969	0	1950	9
2006	34	1987	0	1968	4		

Solution

First, click on the **Copy Data to Clipboard** button and a popup should appear that indicates the copy was successful.



Before the advent of comput ed arranging the data in ascending order and marking prone to human errors such as miscounting or duplicating

elearn.volstate.edu says

Dataset copied to clipboard!

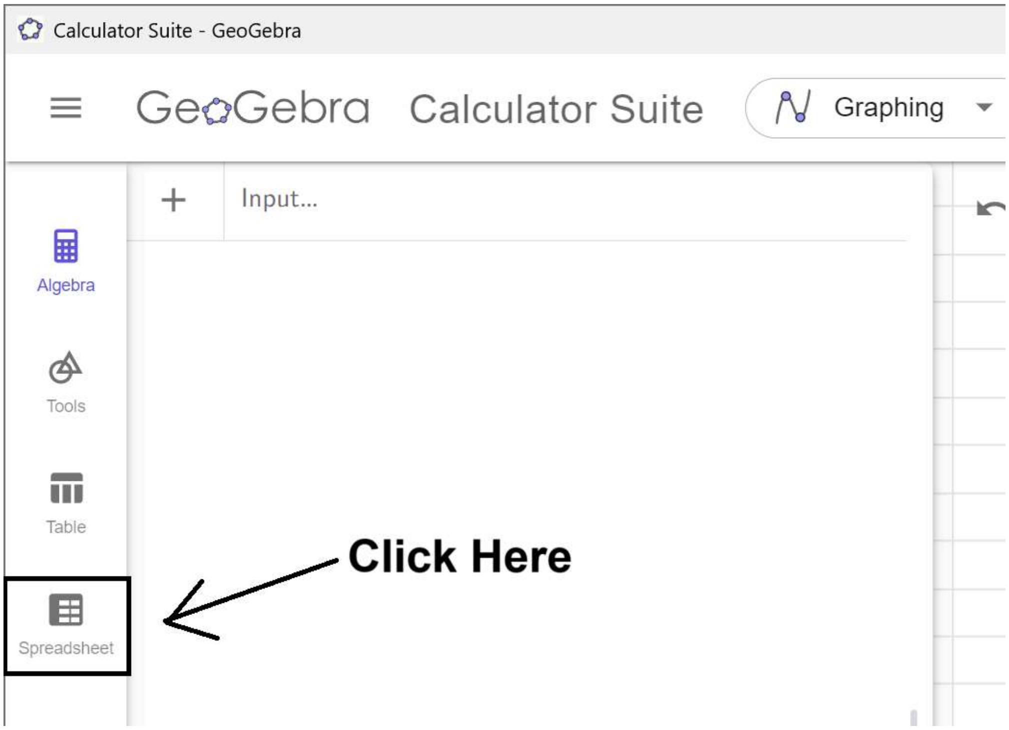
With computers, creating freq anual methods. Tools like SPSS, MATLAB, and Excel offer advanced statistical capabilities. GeoGebra, however, is particularly useful in educational settings because of its simplicity and visualization tools, making it ideal for students learning this process.

Example 1

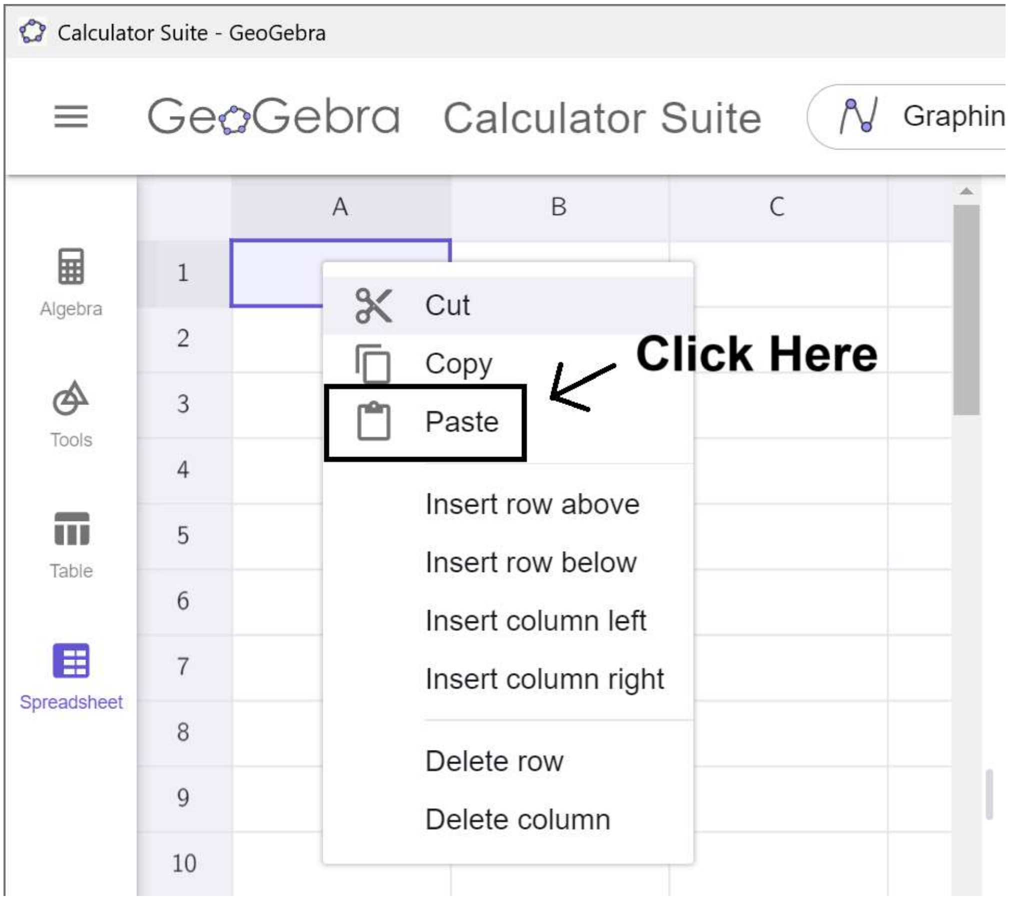
The data on the number of deaths directly caused by tornadoes in Tennessee is given in the table below. Create a Frequency Distribution for the number of deaths per year.

Direct Fatalities Cause by Tornadoes in Tennessee (1950-2023)

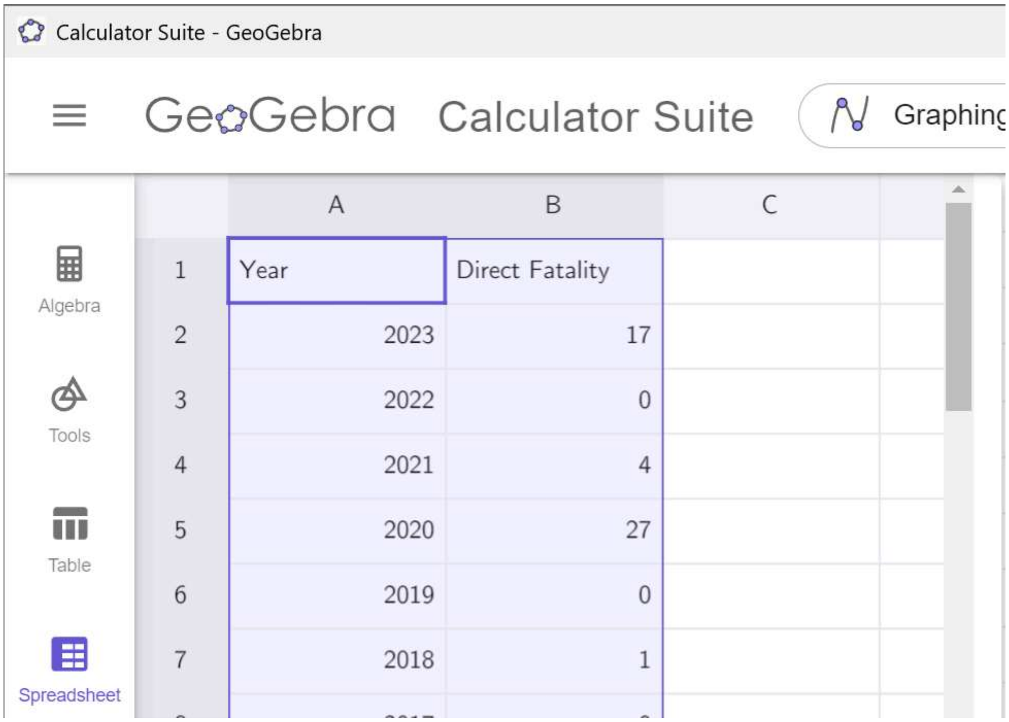
Next, open the Frequency Distribution Tool by clicking on this link or by going to the GeoGebra Tools module and clicking the link provided there. Once Frequency Distribution Tool is opened, click on the Spreadsheet tab.



Click on the A1 Cell, and right click to open up the Context Menu.



Select Paste from the Menu, and the data will copy into the spreadsheet. (This may take a moment to load. Please be patient.)

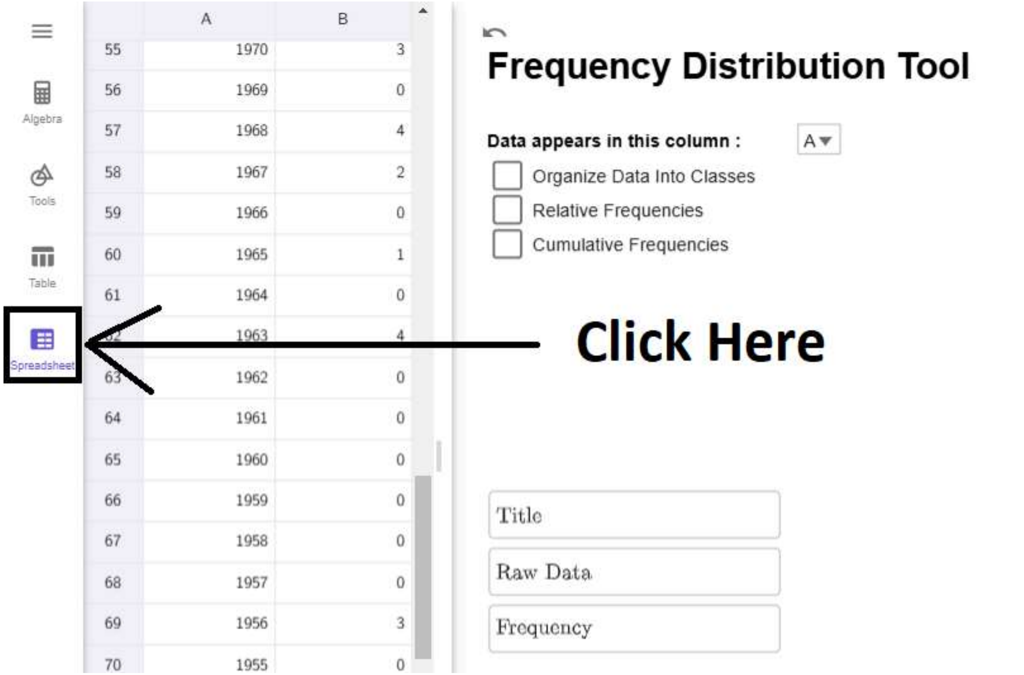


Calculator Suite - GeoGebra

GeoGebra Calculator Suite

	A	B	C
1	Year	Direct Fatality	
2	2023	17	
3	2022	0	
4	2021	4	
5	2020	27	
6	2019	0	
7	2018	1	

Notice that the data in which we are interested is in column B. Remember this! Click on the Spreadsheet tab to hide the data.



	A	B
55	1970	3
56	1969	0
57	1968	4
58	1967	2
59	1966	0
60	1965	1
61	1964	0
62	1963	4
63	1962	0
64	1961	0
65	1960	0
66	1959	0
67	1958	0
68	1957	0
69	1956	3
70	1955	0

Frequency Distribution Tool

Data appears in this column : A▾

□ Organize Data Into Classes

□ Relative Frequencies

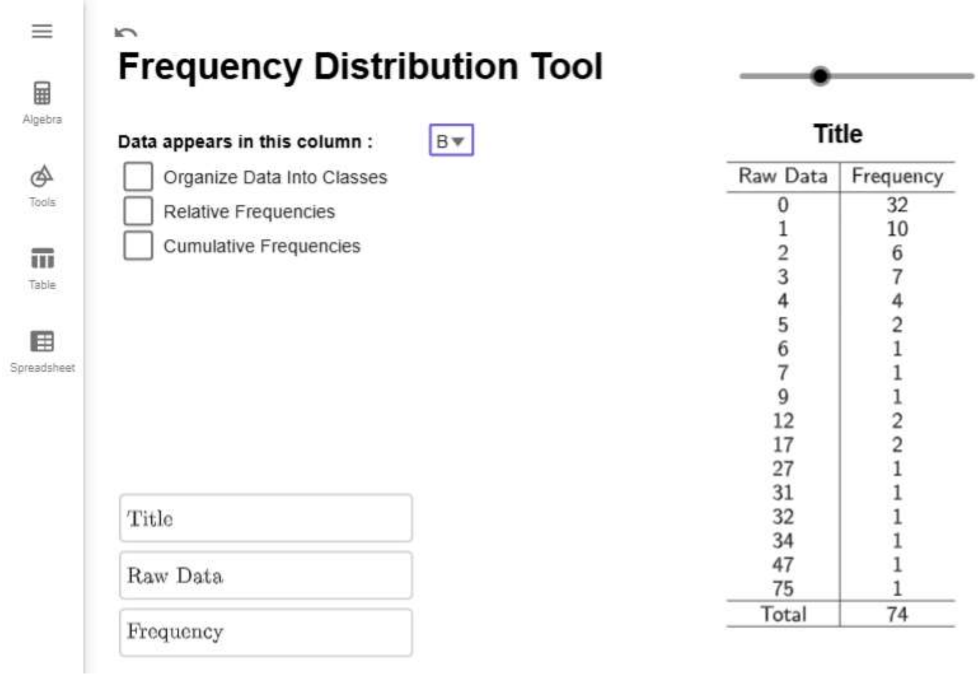
□ Cumulative Frequencies

Title

Raw Data

Frequency

Notice that the tool is set by default to column A. If we click on the dropdown box and select column B, we will get the correct frequency distribution.



Frequency Distribution Tool

Data appears in this column : B▼

□ Organize Data Into Classes
□ Relative Frequencies
□ Cumulative Frequencies

Title

Raw Data

Frequency

Title

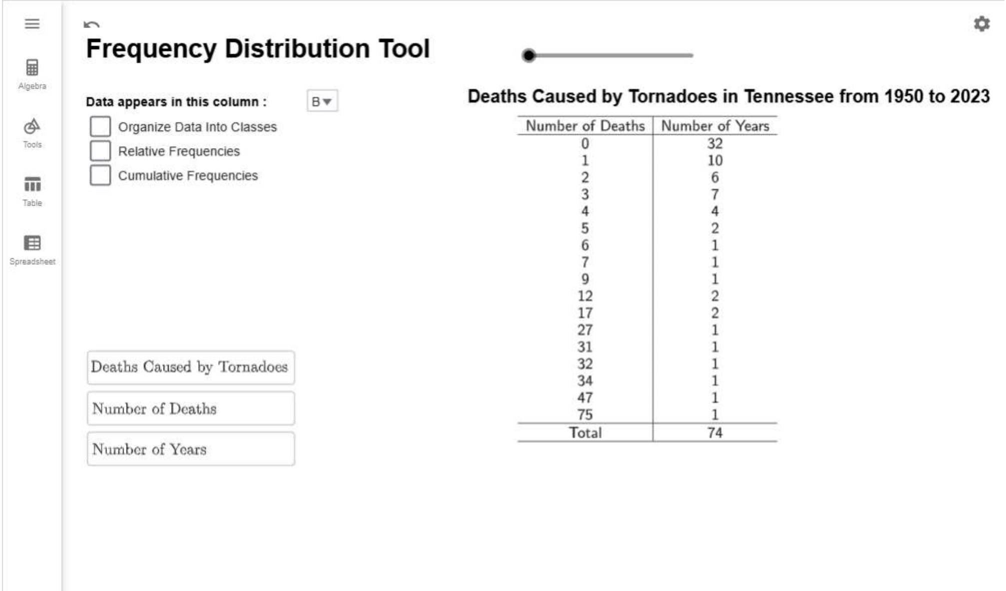
Raw Data	Frequency
0	32
1	10
2	6
3	7
4	4
5	2
6	1
7	1
9	1
12	2
17	2
27	1
31	1
32	1
34	1
47	1
75	1
Total	74

Interpretation

The numbers in the Data column represents the number of deaths reported. The numbers in the Frequency column represent the number of years from 1950 to 2023 that reported that many deaths. For example, 10 years between 1950 and 2023 reported 1 death caused by Tornadoes in Tennessee.

Note

The Title and Column headers as customizable. Just click on the corresponding TextBox in the lower right corner, delete the existing text, and enter what you want to appear there. There is also a slider above the title so you can adjust the horizontal positioning of the title.



Frequency Distribution Tool

Data appears in this column : B▼

□ Organize Data Into Classes
□ Relative Frequencies
□ Cumulative Frequencies

Deaths Caused by Tornadoes

Number of Deaths

Number of Years

Deaths Caused by Tornadoes in Tennessee from 1950 to 2023

Number of Deaths	Number of Years
0	32
1	10
2	6
3	7
4	4
5	2
6	1
7	1
9	1
12	2
17	2
27	1
31	1
32	1
34	1
47	1
75	1
Total	74

■

Frequency Distribution Classes

While the above example works well for small data sets, many frequency distributions group data into ranges instead of listing each unique data point individually. For example, we could group the number of years with 0 to 4 deaths, 5 to 9 deaths, 10 to 14 deaths, and so on, creating a row in the table for each of these ranges. These ranges are called _______ or **bins** in statistics.

If we created a new class for every 5 deaths, we would need 16 classes to cover all the data, from the class with 0 to 4 deaths up to the class with 75 to 79 deaths. However, this approach would lead to many classes with a frequency of 0, which makes the table less meaningful. To avoid this, we want the ranges to be large enough to capture meaningful data but not so small that most classes are empty.

So how many classes should we have? As the researcher, you get to decide, but a good rule of thumb is to aim for 5 to 20 classes. Once the number of classes is chosen, the formula below will help determine the range for each class, known as the **class width**.

Calculating Class Width?

The ___________ is the size of each class and determines how data is grouped. It is calculated using the formula

$$\text{class width } = \frac{\text{largest value } - \text{smallest value}}{\text{number of classes}}.$$

If the result is a decimal, always round __ to the next whole number to ensure all data points fit into the intervals. For example, if the calculated width is 9.38, round it up to 10. This ensures every data point fits into one of the classes without gaps.

What does this number mean?

The class width measures the difference between the smallest numbers in successive classes. For example, if our first class is from 0 to 4 deaths and our second class is from 5 to 9 deaths, then the class width is 5 since $5 - 0 = 5$.

Can I choose my own class width?

Yes! This formula isn't perfect, and occasionally excludes the maximal data point. Sometimes, you have a preferred width you would like to use. It is up to you! For academic purposes though, we will always use the formula unless a problem requires otherwise.

Example 2

Our data from Example 1 range from 0 deaths to 75 deaths. Find the class width that you will need if you want 6 classes.

Solution

Class widths, along with the minimum value of the data set, define what is known as the **boundary points** of the classes. Boundary points define where one class ends and another class begins.

How do I compute the Boundary Points?

The minimum value is typically the first boundary point. To find the other boundary points, keep adding the class width until you get a value that is larger than (but not equal to) the maximal value in our data set.

For example, if we wanted a class width of 5 for the classes of tornado deaths and we know the lowest value in our dataset is 0 and the largest value is 75, the boundary points would be

____

Can I use a number other than the minimum value for the first Boundary Point?

Yes! If you choose your own class size, typically you also choose the first boundary points as well. Just remember to keep adding the class width until you get a number larger than the largest value in the data set.

Example 3

Our data from Example 1 range from 0 deaths to 75 deaths. Suppose you want your boundary points to start at 7 and have class width 10. Find all the boundary points.

Solution

Calculating boundary points and determining the number of classes is straightforward. However, tallying data points in each class can be tedious and error-prone. Instead, we'll use our GeoGeobra tool to automate the tallying and format it into a table. The next example demonstrates this process.

Example 4

Using our data from Example 1, create a frequency distribution with 8 different classes. Write the resulting frequency distribution in the space below.

Solution

Before we close this section out, we will give a demonstration of creating a frequency distribution that starts at a value other than the minimum value of the data set.

Example 5

Use our data set from Example 1 to create a frequency distribution that starts at 7 and has a class width of 6.

Solution

⚠ Warning ⚠

Notice there are 13 classes, and each has a class width of 6. But, notice that the last class has a 0 in it. The best Frequency Distributions always have a non-zero frequency for the first and last class. Also, notice we missed all the years that had fewer than 7 deaths. So, while this is a correctly constructed Frequency Distribution, it is also a bit misleading.

That is why it is better to let the Lower Boundary be the ____ of the dataset. We get all the data, and our first and last classes will be a non-zero frequency ensuring that our table is just the right size for the number of classes or class width we desire.

How to Create Relative Frequency Distributions

An important concept related to frequency is **relative frequency**, and it is a foundational concept for several key topics in this course. We will use relative frequencies when discussing cumulative frequencies (discussed below), creating histograms, studying probability and probability distributions, and using statistical inference to discuss population proportions. By connecting data analysis to probability theory and statistical inference, relative frequency serves as a bridge between descriptive and inferential methods in a statistics course.

Relative Frequency Distributions

What is Relative Frequency?

__________________ has two interpretations.

- For **raw data**, relative frequency is the percentage of times that a particular value appears in a data set.
- For **data sorted into classes**, relative frequency is the percentage of data that appear in a given class.

In both cases, the formula for relative frequency is identical:

$$\text{relative frequency} = \frac{\{\text{frequency}\}}{n},$$

where n is the number of data points in the sample.

Notice that this will always give a percentage as a decimal, so we will always write 0.3947 instead of 39.47% in this course. Using percentages instead of decimals can cause errors in many of the formulas because most statistical formulas use relative frequencies rather than percentages to ensure accurate calculations.

⚠ Warning ⚠

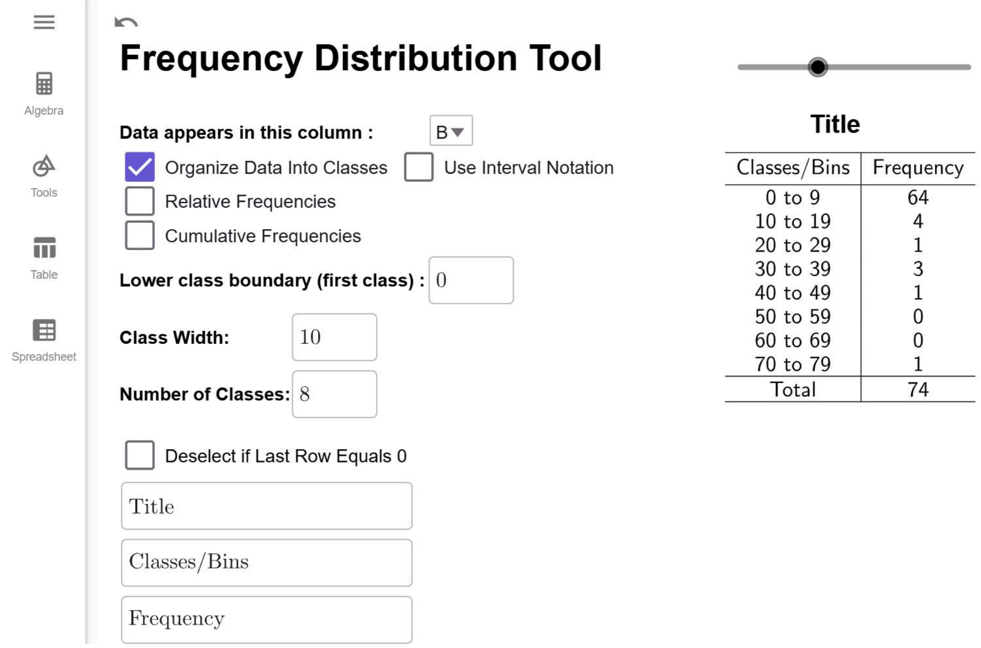
Relative frequencies will always yield a number between 0 and 1. If you get a number larger than 1 or a negative number, double check your work because you made a calculation error!

What is a Relative Frequency Distribution?

A **relative frequency distribution** is a table that lists either the raw data or classes in the first column and the corresponding relative frequencies in the second column.

How Do I Compute a Relative Frequency?

Let's consider our Frequency Distribution from Example 4 in How to Create Frequency Distributions:



Frequency Distribution Tool

Data appears in this column : B▾

☑ Organize Data Into Classes □ Use Interval Notation

□ Relative Frequencies

□ Cumulative Frequencies

Lower class boundary (first class) : 0

Class Width: 10

Number of Classes: 8

□ Deselect if Last Row Equals 0

Title

Classes/Bins

Frequency

Title

Classes/Bins	Frequency
0 to 9	64
10 to 19	4
20 to 29	1
30 to 39	3
40 to 49	1
50 to 59	0
60 to 69	0
70 to 79	1
Total	74

To determine the number of data points, we first choose a class, such as the interval 0 to 9, and note its frequency, which is 64. The relative frequency is found by dividing the class frequency by the total number of year, 74, as follows:

$$\frac{64}{74} \approx 0.8649.$$

Thus, 0.8649 is the relative frequency for 0 to 10, meaning 86.49% of the years from 1950 to 2023 had 0 to 10 deaths (excluding 10). Note we will always round relative frequency to four decimal places, which is a standard practice in statistics.

While it is essential to know how to compute a relative frequency, performing multiple manual calculations to create a relative frequency distribution increases the risk of calculation errors. In our next example, we will see how to use our Frequency Distribution Tool to create a relative frequency distribution.

Example 6

According to National Institutes of Health Cancer Statistics, the rate at which men get Colon-Rectal cancer each year (per 100,000 men, rounded to one decimal place) from 2000 to 2021 is given in the table below. Find the relative frequency distribution for this data if we use a lower class boundary of 35 and a class width of 5.

Copy Data to Clipboard

Colon-Rectal Cancer Diagnosis Rates (Per 100,000 Men)

Year	Rate	Year	Rate
2000	70.8	2011	49.5
2001	69.6	2012	48.0
2002	68.3	2013	46.9
2003	66.3	2014	46.7
2004	64.0	2015	45.5
2005	61.9	2016	44.5
2006	59.1	2017	43.5
2007	58.1	2018	42.9
2008	56.2	2019	42.7
2009	53.1	2020	37.2
2010	51.0	2021	40.8

Solution

Even though the computer does most of the heavy lifting for us, it is still important to know how to perform the calculations by hand to make sure there are no errors in the data or output.

Example 7

Consider the following frequency distribution for average August temperatures in Nashville, TN. (Source: Weather UnderGround) Find the relative frequency for the class **86 to 88.**

Average August Temperature in Nashville, TN (1948-2024)

Degrees Fahrenheit	Number of Years
80 to 82	2
83 to 85	14
86 to 88	33
89 to 91	22
92 to 94	5
95 to 97	1
Total	77

Solution

How to Compute Cumulative Frequency Distributions

Another important concept related to frequency is cumulative frequency, which keeps a running total of frequencies up to a specific value or class. This provides insight into how data accumulates over a specific range of values, and is fundamental in understanding ideas such as median, quartiles, interquartile ranges, computing area under the uniform, normal, and student T curves. By synthesizing raw data into meaningful patterns, cumulative frequency serves as a stepping stone to deeper statistical concepts, emphasizing the relationship between individual data points and their overall context within a dataset.

Cumulative Frequency

What is Cumulative Frequency?

____________________ is the number of data points that are less than or equal to some given number (not necessarily in the data set).

What is a Cumulative Frequency Distribution?

A **cumulative frequency distribution** is a table that lists either the raw data or classes in the first column and the corresponding cumulative frequencies in the second column.

While it is good to have an idea of how to compute a cumulative frequency, the multiple manual calculations needed for a cumulative frequency distribution increases the likelihood of a calculation error. In our next example, we will see how to use our GeoGebra tool to create a cumulative frequency distribution.

Example 8

For the data from Example 1 in How to Create Frequency Distributions, create a cumulative frequency distribution for the raw data.

Copy Data to Clipboard

Solution

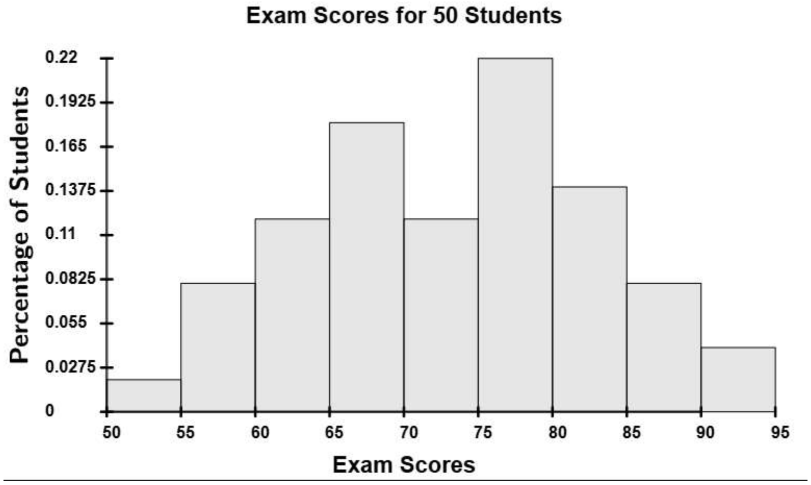
Shapes of Frequency Distributions

The shape of a frequency distribution is an important geometric property that we can determine from its histogram. While there is one shape that is the most important, we will discuss other shapes at various points during the course. Before diving into the different shapes, it is helpful to understand some important concepts related to distributions: the **peak** and **tails** of a distribution.

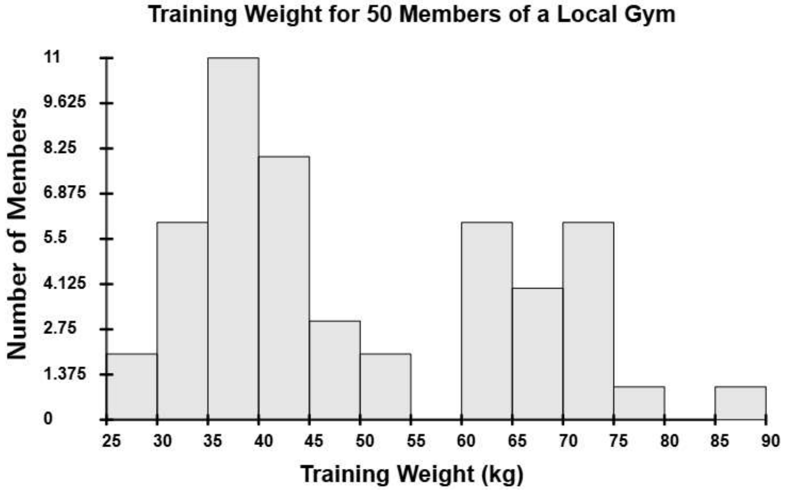
The Peak and Tails

What is the peak of a distribution?

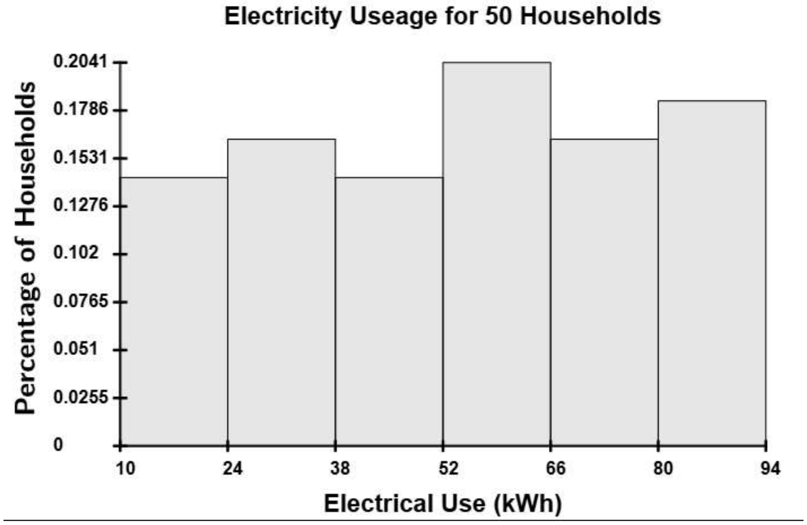
The **peak of a distribution** refers to the tallest part of the histogram, which represents the class(es) with the most data. Distributions can have one peak (**unimodal**)



multiple peaks (**multimodal**)

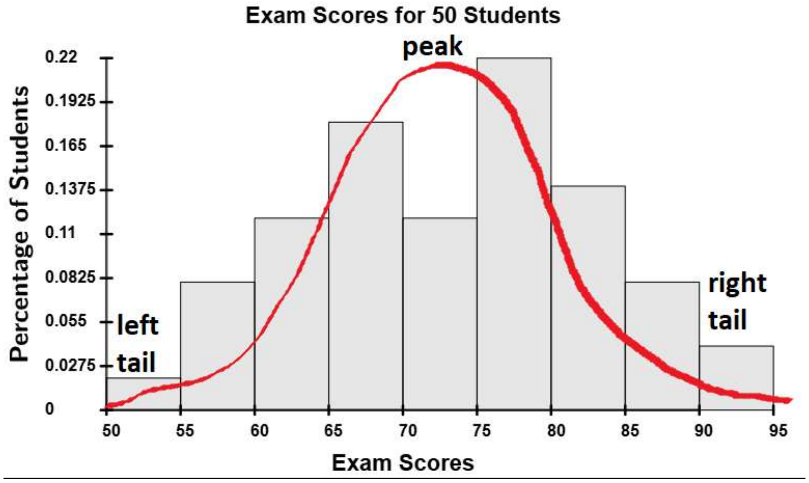


or no distinct peak (**uniform**).



What are the tails of a distribution?

The **tails of a distribution** refer to the far ends of the distribution, which represent the extreme values of the dataset. For example, in a histogram, the left tail corresponds to the smallest values, while the right tail corresponds to the largest values. The shape and behavior of the tails can provide insights into the skewness and outliers of the distribution.



Now that we understand the tails and peaks of frequency distributions, let’s examine the different shapes these distributions can take.

Shapes of Distributions

What are the different shapes of frequency distributions?

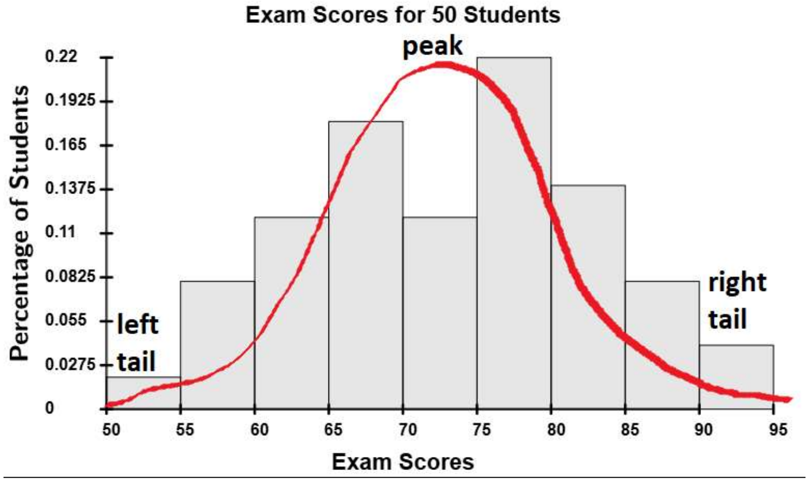
There are three main shapes: **unimodal**, **multimodal**, and **uniform**.

Unimodal Distributions (single-peak)

Unimodal distributions have a single peak in their histogram. These distributions begin with small bars that gradually increase in height from left to right, reach a peak at one class, and then decrease from left to right. There are three subtypes of unimodal distributions: **normal**, **skew-left**, and **skew-right**.

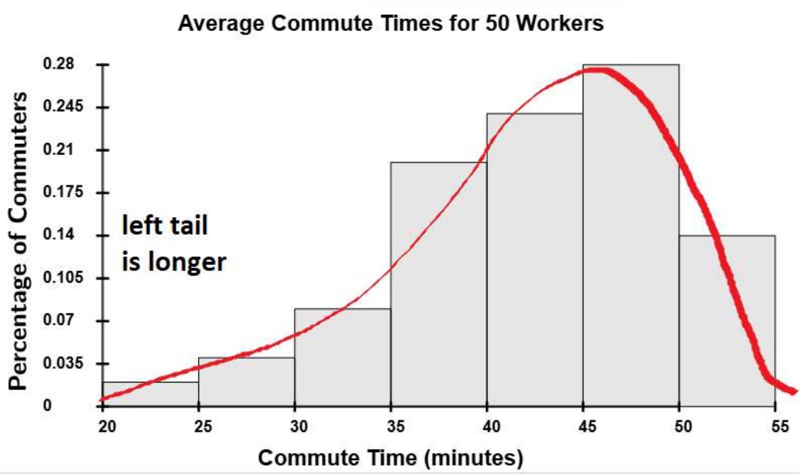
Normal Distributions (aka bell-shaped distribution)

This is the most important distribution in the course. A ____ is symmetric, with its peak located at the center. Its tails extend in both directions but gradually get closer and closer to the horizontal axis as they extend outward. The shape of the histogram for a normal distribution resembles a bell, hence the name "bell-shaped distribution."



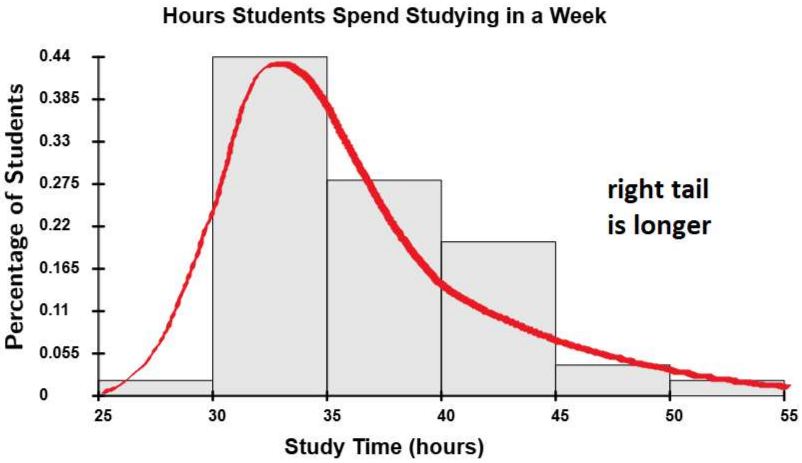
Skew-Left Distributions

A skew-left distribution (also called negatively skewed) is a distribution where the left tail is longer than the right tail. In these distributions, the majority of the data values are concentrated toward the right side of the histogram, with fewer values on the left. This creates a tail that extends toward smaller values.



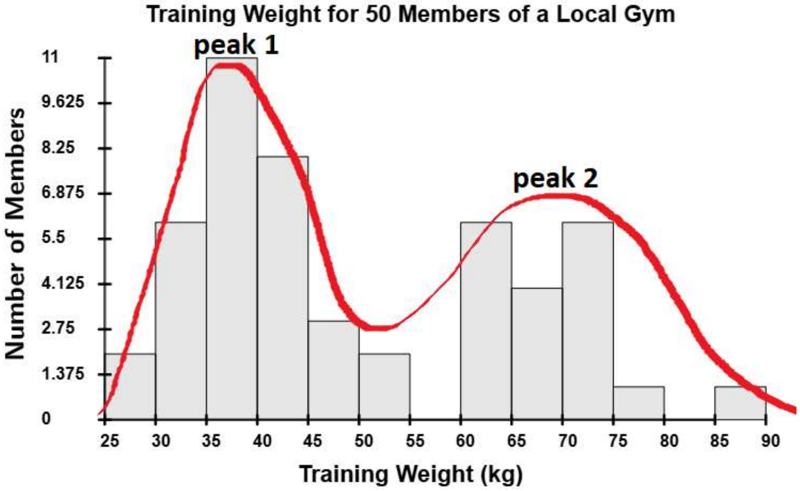
Skew-Right Distributions

A skew-right distribution (also called positively skewed) is a distribution where the right tail is longer than the left tail. In these distributions, the majority of the data values are concentrated toward the left side of the histogram, with fewer values on the right. This creates a tail that extends toward larger values.



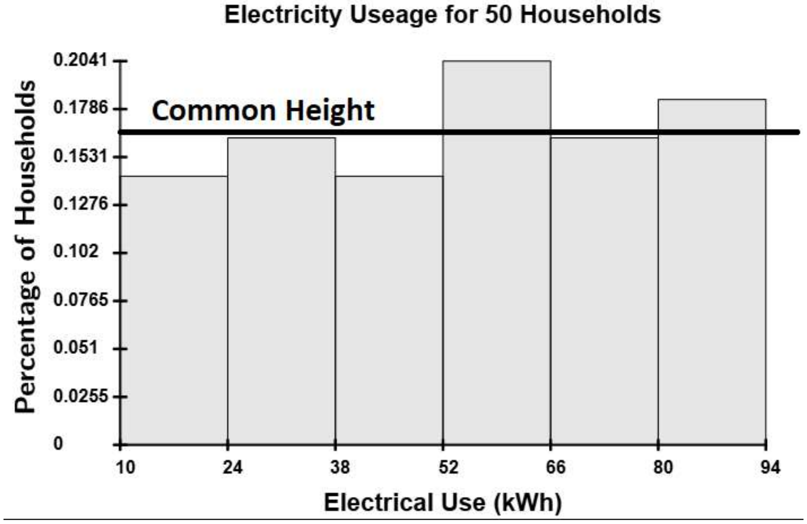
Multimodal Distributions (multiple peaks)

Multimodal distributions have multiple peaks in their histogram. A **bimodal distribution** has two peaks, a **trimodal distribution** has three, and anything with four or more peaks is just called a **multimodal distribution**. While we won't cover multimodal distributions in this course, you may encounter them in advanced statistics courses.



Uniform Distributions

_____________________ have bars that area all approximately the same height (give or take a small margin of error).



Homework Note

Homework problems will ask you to identify the shape of a distribution. The easiest way to determine the shape is by drawing or visualizing a curve through the tops of the bars, as shown in the examples above for unimodal and multimodal distributions. If the distribution is uniform, all the bars should be approximately the same height.