Mini-Lecture Five-point summary and Box-Whisker Plot

Contents

[What is a five-point summary? 1](#_Toc71015245)

[How to find the five-point summary by hand? 2](#_Toc71015246)

How to find the five-point summary by TI 84?.........................................................................................................................2

[How to Find the Five-Point Summary in Excel 3](#_Toc71015248)

[What is Box and Whisker plot? 3](#_Toc71015249)

How to draw the Box and Whisker by hand?......................................................................................................................3

How to draw the Box and Whisker on TI 84 plus?.......................................................................................................................4

[How to draw the Box and Whisker in Excel? 6](#_Toc71015251)

[How to Interpret a Box and Whisker Plot 7](#_Toc71015252)

[What is an Outlier? 9](#_Toc71015253)

[How to identify the outliers using the formula? 9](#_Toc71015254)

[How to identify the outliers from the graphs? 9](#_Toc71015255)

[Why is Finding Outliers Important? 10](#_Toc71015256)

[What are the advantages & disadvantages of the Box-Whisker Plot?........................ 11](#_Toc71015257)

# What is a five-point summary?

One of the main goals in data analysis is to find the center and the variability of data. Using the median to measure a center and the quartiles to show spread is a simpler, more easily visualized method. Also, the median and quartiles are not as easily affected by extreme values compared to other measures of central. It is clear that there are to be five numbers in our summary, but which five? The five-number summary gives you a rough idea about what your data set looks like, center of our data, as well as how spread out the data points are. It is also known as a 5-number summary.

The five-number summary consists of the following:

The minimum – this is the smallest value in our data set.

The first quartile – this number is denoted *Q*1 and 25% of our data falls below the first quartile.

The median – this is the midway point of the data. 50% of all data falls below the median.

The third quartile – this number is denoted *Q*3 and 75% of our data falls below the third quartile.

The maximum – this is the largest value in our data set.

Sometimes, it’s impossible to find a five-number summary. For the five numbers to exist, your data set must meet these two requirements:

Your data must be [univariate](https://www.statisticshowto.com/univariate/). In other words, the data must be a single [variable](https://www.statisticshowto.com/probability-and-statistics/types-of-variables/).

Your data must be [interval](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/nominal-ordinal-interval-ratio/#interval) or [ratio](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/nominal-ordinal-interval-ratio/#ratio).

# How to find the five-point summary by hand?

**Step 1:**Put your numbers in ascending order (from smallest to largest).

**Step 2:**Find the minimum and maximum for your data set. Now that your numbers are in order, this should be easy to spot.

**Step 3:**Find the [median](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median). The median is the middle number.

**Step 4:**Place parentheses around the numbers above and below the median.

(This is not technically necessary, but it makes $Q\_{1 }$and $Q\_{3 }$ easier to find)

**Step 5:**Find $Q\_{1 }$and $Q\_{3 }$. $Q\_{1 }$can be thought of as a median in the lower half of the data, and $Q\_{3 }$ can be thought of as a median for the upper half of data.

**Step 6:**Write down your summary found in the above steps.
minimum =, $Q\_{1 }$ =, median =, $Q\_{3 }$ =, and maximum =.

# How to find the five-point summary on TI 84 plus?

|  |  |
| --- | --- |
| Step 1: STAT] [1] selects the list-edit screen. After entering the data into a List, Press Stat | STAT 1  Type the List |
| Step 2: Move the cursor to CalcChoose #1 | Move the cursor to Calc Choose #1  |
| Step 3: You will be placed on the home screen automatically.  Enter which list you would like to take the five-number summary of:1-Var Stats $L\_{1}$Step 4: Press enter, and you will be given the answer. | Choose the option 1 and press ENTER  |

# ****How to Find the Five Number Summary in Excel****

We can find the five-number summary of a dataset in Excel using the following steps:

|  |  |
| --- | --- |
| **Step 1: Enter the data values in one column.** | Enter the data values in one column |
| **Step 2: Find the five-number summary.**The five values of the five-number summary are shown in column D and the formulas used to find these values are shown in column E: | Five number summary and the formula used to find the five point summary are shown in column E |

# What is the Box and Whisker plot?

Box-Whisker plot is a way to data display created by John Tukey of Princeton in 1977. A box plot is a scaled display in which the data are divided into four equal parts and are used to show the spread of the data relative to the median and the quartiles. We use these box plots or graphical representation to know:

Distribution Shape

Central Values

Variability

When we plot a graph for the box plot, we outline a box from the first quartile to the third quartile. A vertical line that goes through the box is the median. The whiskers (small lines) go from each quartile towards the minimum or maximum value, as shown in the figure below.

One of the easiest ways to visualize a five-number summary is by creating a box-and-whisker plot, which uses a box with a line in the middle along with “whiskers” that extend on each end. The process to abstract a set of data that is estimated using an interval scale is called a box and whisker plot. These are mostly used for data interpretation. It helps us find the outliers of the data.

# How to draw the Box and Whisker by hand?

To draw a box and whisker diagram, we need to find the five-point summary.

Firstly, write the given data in increasing order.

Range = Maximum value – Minimum value

Now, Median = center value of the given data

Now, we need to find the quartiles.

First quartile = Q1 = Median of data values present at the left side of Median

Third quartile = Q3 = Median of data values present at the right side of Median

Therefore, the interquartile range = Q3 – Q1 =

The five-number summary is given by:

Minimum, Q1, Median, Q3, Maximum

Now, we can draw the box and whisker plot, based on the five-number summary.



# How to draw the Box and Whisker on TI 84 plus?

|  |  |
| --- | --- |
| Step 1: [STAT] [1] selects the list-edit screen Enter the numbers in $L\_{1 }$Enter the data points. | STAT 1 Edit the List |
| Step 2: Clear other plots. In this step, you disable any other plots and graphs that could overlay your box-whisker plot. Press [Y=] to open the list of equations and plots.Look at the plots across the top, and look at the column of = signs. If any are enabled (highlighted), disable them.Step 3: Set up the boxplot.Use the arrow keys to get to any highlight, and press [ENTER] to remove the highlight.Clear the grid and enable coordinate display for later use in tracing. | Press Y=Y= with highlights removed |
| Press [2nd ZOOM *makes* FORMAT]. If Grid On is highlighted, press [▼ 3 times] [ENTER].If CoordOff is highlighted, use the [▼] or [▲] key to get to CoordOn, and press [ENTER]. | Format screen with GridOff |
| Select $Plot1$. Press [2nd Y= *makes* STAT PLOT] [ENTER].Select the modified box-whisker plot.Press [▼] [► 3 times] [ENTER] [▼]. The bottom part of the screen may change when you do this. Press [ZOOM] [9], which is ZoomStat or “zoom to statistics”. | boxplot |
|  |  |
| Step 4: Display the boxplot.On the box-whisker diagram, any outliers show as isolated squares. The whiskers are mix and max (disregarding any outliers), and the box is first quartile, median, and third quartile. Press [TRACE]. Use [◄] and [►] to display the numbers in the five-number summary as well as any outliers. | trace showing numerical value of outlier |
|  |  |

# How to draw the Box and Whisker in Excel?

|  |  |
| --- | --- |
| **Step 1: Highlight the data values.** | Boxplot in Excel Highlight the data values |
| **Step 2: In the**Insert **tab in the**Charts **group along with the top ribbon, click the tiny arrow in the bottom left corner to “See All Charts.”** | In the Insert tab in the Charts group along with the top ribbon, click the tiny arrow in the bottom left corner to “See All Charts.” |
| **Step 3: Select “Box & Whisker” and click OK.** | Select “Box & Whisker” and click OK. |
| A box and whisker plot will automatically be displayed. | A box and whisker plot will automatically be displayed. |
| The top whisker represents the max, the top of the box represents the 3rd quartile, the middle line in the box represents the median, the tiny “x” in the box represents the average, the bottom of the box represents the 1st quartile, and the bottom whisker represents the minimum value: | Five number summary boxplot in Excel |

# How to Interpret a Box and Whisker Plot

Box plots allow us to visually see the distribution of the data and divides the data into sections that contain approximately 25% of the data in each section. When looking at a box plot, we can see if there is any skew in the data set or if it has a normal distribution.



SOURCE: By Ever.chae - Own work, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=84824524>

In a normal distribution, the median is located in the middle of the box and the whiskers are spread out evenly on both sides (as seen on the left side of the above diagram). In a normal distribution, the data above and below the median are evenly spaced out. In a positive skew, the median is closer to the bottom of the box and the bottom whisker is shorter. In a positive skew, the data are clustered together below the median but spread out above it. And, in a negative skew, the median is closer to the top of the box and the whisker on top is shorter. Here, the data are clustered together above the median but spread out below it.

Boxplots also help to visualize how scattered the data are. The box shows the middle 50% of the data. The smaller the box, the less scatter in the data. The larger the box, the more scatter in the data.

When you encounter a box plot in an article, there typically is more than one box plot depicted. It is useful to be able to compare multiple boxplots. Parallel box plots are a useful way of comparing groups of data. For example, compare the data from different lab groups or different trials of an experiment. Or compare the data collected from an experiment conducted at different times when a variable is changed, such as time of the day.

**Step 1:** Compare the boxes. If the boxes, which represent the middle half of the data, do *not* overlap, then there *is* a difference between the two groups of data.



SOURCE: https://blog.bioturing.com/2018/05/22/how-to-compare-box-plots/

**Step 2:**Compare the medians. If you compare the median of one boxplot and it lies outside of the box of the boxplot you are comparing it to, then there is likely a difference between the two groups.



SOURCE: <https://blog.bioturing.com/2018/05/22/how-to-compare-box-plots/>

**Step 3:** Compare the length of the whiskers and the boxes. The longer the whiskers or the box, the more spread out the data are. The shorter/smaller the whiskers or box, the more the data are clustered together.



<https://blog.bioturing.com/2018/05/22/how-to-compare-box-plots/>

**Step 4:** Compare the skewness



When the right side of the box-and-whisker plot is longer, it is skewed to the right. The values on this side — the upper end of the scale — are more variable. Most observations concentrate at the low end of the scale.

When a box plot is left-skewed, values gather at the upper end, making a short and tight section there. To the left of that crowd, data points spread out, creating a longer tail.

Box plots are like the base of distribution curves. Skewness suggests that data may not be normally distributed.

# What is an Outlier?

An outlier is a value or point that [differs substantially from the rest of the data](https://en.wikipedia.org/wiki/Outlier). Sometimes outliers might be errors that we want to exclude or an anomaly that we don’t want to include in our analysis. But at other times it can reveal insights into special cases in our data that we may not otherwise notice.

# How to identify the outliers using the formula and from the graphs?

**The formula to find the outliers**: To calculate outliers of a data set, you'll first need to find the five-point median, lower quartile, or Q1, upper quartile Q3. Find the interquartile range by finding the difference between the 2 quartiles. Then use the IQR rule to find the outliers.

[**What is the 1.5 IQR rule?**](https://askinglot.com/goto/616374A9)

Using the Interquartile Rule to Find Outliers Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers). Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier. Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.

# How to identify the outliers from the graphs?

Outliers can look like this:



 

# Why is Finding Outliers Important?

It helps us detect errors, allows us to separate anomalies from the overall trends, and can help us focus our attention on exceptions. While what we do with outliers is defined by the specifics of the situation, by identifying them we give ourselves the tools to more confidently make decisions with our data.

 Some of the other reasons are

* Ensure Data Quality
* Provide Confidence in Analysis
* Contextualize the Findings
	+ Identify High Performers
	+ Identify Low Performers
* Visualization

As we can see, it is imperative to pay attention to outliers because they can bias data analysis. But, in addition to identifying outliers these are some suggested ways to better treat them:

* Exclude the discrepant observations from the data sample: when the discrepant data is the result of an input error of the data, then it needs to be removed from the sample; a separate analysis with only the outliers: this approach is useful when you want to investigate extreme cases, such as students who only get good grades, companies that make a profit even in times of crisis, fraud cases, among others. use clustering methods to find an approximation that corrects and gives a new value to the outlier’s data.
* In cases of data input errors, instead of deleting and losing an entire row of records due to a single outlier observation, one solution is to use clustering algorithms that find the behavior of the observations closest to the given outlier and make inferences of which would be the best approximate value.

Finally, the main conclusion about the outliers can be summarized as follows:

“a given outlier may be what most disturbs the analysis, but may also be exactly what you are looking for.”

# Benefits and drawbacks of Box-Whisker plot

### Benefits

Box and whisker plots are very effective and easy to read, as they can summarize data from multiple sources and display the results in a single graph. Box and whisker plots allow for the comparison of data from different categories for easier, more effective [decision-making](https://asq.org/quality-resources/decision-making-tools).

It displays the range and distribution of data along a number line.

Box plots provide some indication of the data’s symmetry and skew-ness.

Box plots show outliers.

You can use box and whisker plots when you have multiple data sets from independent sources that are related to each other in some way. Examples include:

* Test scores between schools or classrooms
* Data from before and after a process change
* Similar features on one part, such as camshaft lobes
* Data from duplicate machines manufacturing the same products

### Drawbacks

In most cases, the original data is not clearly shown in the box plot. Also, mean and mode cannot be identified in a box plot.

It can be used only for numerical data.

No indication of sample size: Though you can use box plots on non-parametric data, it is best to have a sample size of at least 20 (some might even say 30). For a smaller sample size, consider using individual value plots.

The illusion of bar graphs: Box plots resemble bar graphs in their appearance, yet they present completely different information. Bar graphs compare groups by their absolute counts, while box plots show their distributional ranges. Remember: the size of each section in a box plot shows how widely spread a data range is; it says nothing about the quantity of the group.

The troubles are in the whiskers: Box plots’ whiskers are mistaken as error bars more often than you’d think, especially when there are asterisks representing outliers on top of them. They are not. They show the lowest and highest quartiles of values. They contain half of the data points; the other half is in the box.

The secret box: Box plots sometimes hides important information. When data “morph” but manage to maintain their ranges and medians, their box plots stay the same.

This material is based upon work supported by the National Science Foundation under Grant No.1919613. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



