# Data Analysis: A Practical Introduction for Absolute Beginners

## Lab 6: Data Analysis

### Learning Objectives

- Calculate the mean, median, and mode in Excel.
- Calculate the standard deviation in Excel.
- Calculate the skewness variable in Excel.

### Data Set

Mod4Labs.csv

### What You'll Need

To complete the lab, you will need the online version of Microsoft Excel.

### Overview

This lab is a crash course in the basics of some of the math concepts and summary statistics we'll use for data analysis. We'll tackle the three measures of center (mean, median, mode), standard deviation, and skewness. Buckle up!

### Exercise 1: Mean, Median, Mode

There are three ways to measure the "center" of a data set, and each uses a different type of logic. Here they are:

- **Mean**: The mean is the mathematical average of a data set: add up all the values, then divide that sum by the number of values. The mean is most useful when the data are fairly symmetric (i.e. normally distributed).
- **Median**: The median is the middle number in a data set (when all the values are arranged in order from least to greatest). If there are two values in the middle, the median is the average/mean of those two numbers. The median is usually used as a backup measure instead of the mean when the data are skewed or biased in some way.
- **Mode**: The mode is the most frequent value in the data set. For example, in the list 1, 2, 2, 4, 5, the mode is 2 because 2 shows up more often than the other values. The mode isn't particularly useful with numerical data, but it's very helpful with categorical data because it shows which category was the most popular.

Let's run through all three of these measures in Excel, using the movie runtime data we used back in Module 3.

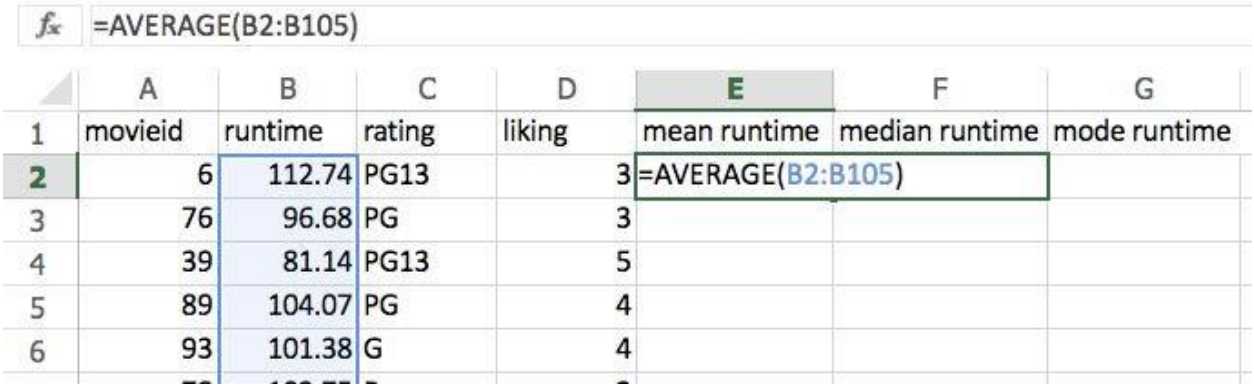
1. Open the data set in Excel, which shows 104 different movies along with their runtime (in minutes), rating, and liking (which gives the audience's average ranking of the movie on a scale of 1–5, with 1 being the lowest). Here's what the first few entries look like:

| | A | B | C | D |
|---|---|---|---|---|
| 1 | movieid | runtime | rating | liking |
| 2 | 6 | 112.74 | PG13 | 3 |
| 3 | 76 | 96.68 | PG | 3 |
| 4 | 39 | 81.14 | PG13 | 5 |
| 5 | 89 | 104.07 | PG | 4 |
| 6 | 93 | 101.38 | G | 4 |
| 7 | 78 | 102.75 | R | 3 |
| 8 | 31 | 92.05 | G | 3 |
| 9 | 47 | 114.65 | R | 3 |
| 10 | 41 | 98.86 | R | 4 |
| 11 | 104 | 85.41 | PG | 5 |
| 12 | 75 | 94.98 | PG | 3 |
| 13 | 60 | 97.85 | PG13 | 3 |
| 14 | 32 | 109.98 | PG13 | 3 |
| 15 | 77 | 110.27 | G | 5 |
| 16 | 71 | 104.81 | PG13 | 4 |

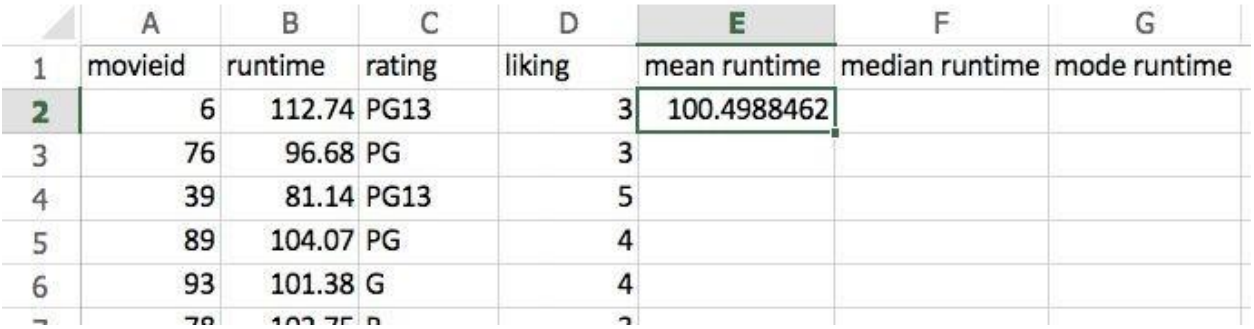
2. Create three new columns for the mean, median, and mode of the "runtime" variable.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | movieid | runtime | rating | liking | mean runtime | median runtime | mode runtime |
| 2 | 6 | 112.74 | PG13 | 3 | | | |
| 3 | 76 | 96.68 | PG | 3 | | | |
| 4 | 39 | 81.14 | PG13 | 5 | | | |
| 5 | 89 | 104.07 | PG | 4 | | | |
| 6 | 93 | 101.38 | G | 4 | | | |

3. In the "mean runtime" column (in cell E2), find the mean of all the values in column B. To do this, use Excel's AVERAGE function. The syntax is **= AVERAGE(first cell:last cell)**. In this case, you want the mean of every single value in the "runtime" column, so the first cell is B2 and the last cell is B105. You can either type B2:B105 directly inside the parentheses of the AVERAGE function, or you can just click inside the parentheses and highlight all the cells from B2 to B105.

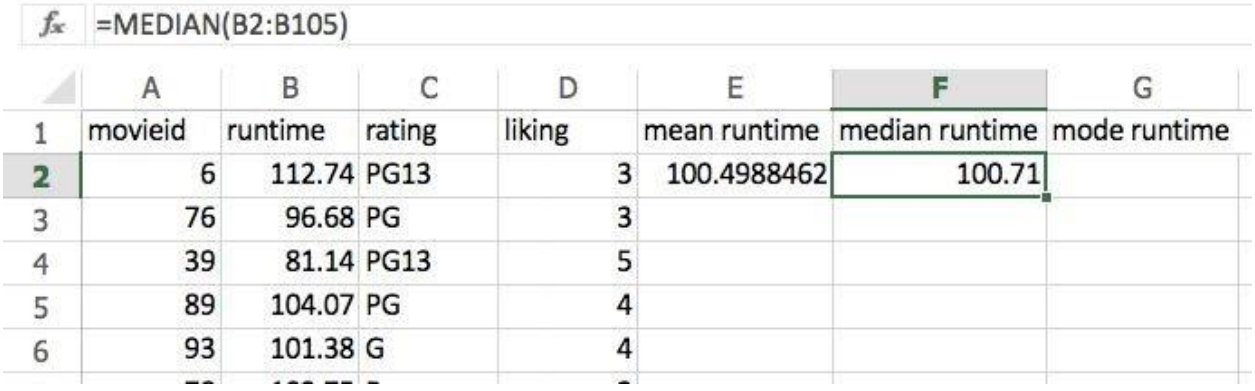



Hit Enter, and Excel will calculate the average for you in cell E2.



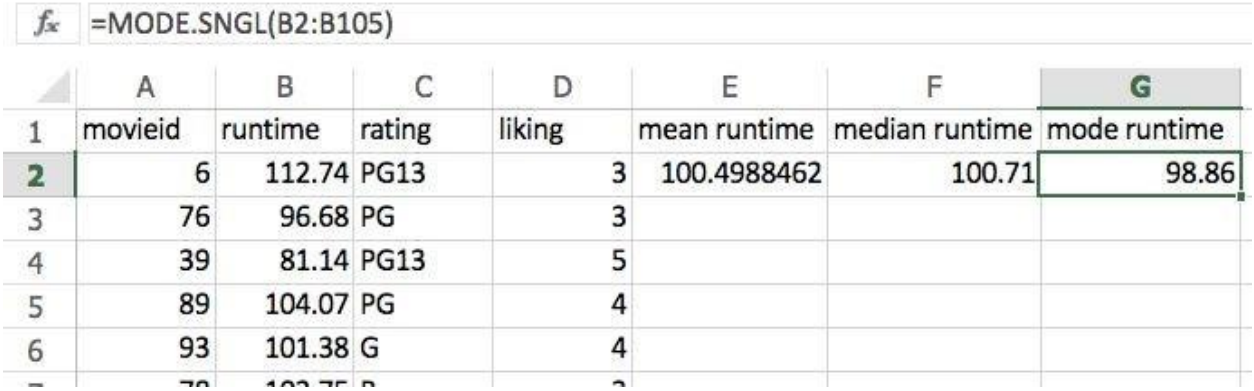

Translation: The average length of these 104 movies was about 100.5 minutes (when we round the value to one decimal place).

4. Now find the median runtime using the MEDIAN function. It works pretty much the same as the mean: The syntax is **= MEDIAN(first cell:last cell)**. You're still looking at the runtime data, so click into cell F2 and type =MEDIAN(B2:B105), then hit Enter.




See how the value of the median is slightly different than the mean? Sometimes they'll be the same, but not this time. What this means is that the middle value in the data set is 100.71 minutes.

5. Find the mode of the runtime data, which is the value that shows up the most often. (There can be more than one mode if multiple values show up with the same frequency.) Excel's MODE.SNGL function gives the single most frequently-occurring number in a data set, so let's use that. The syntax is **= MODE.SNGL(first cell:last cell)**. Once again, you're looking at the runtime values in column B, so the first cell is B2 and the last cell is B105.




| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | movieid | runtime | rating | liking | mean runtime | median runtime | mode runtime |
| 2 | 6 | 112.74 | PG13 | 3 | 100.4988462 | 100.71 | 98.86 |
| 3 | 76 | 96.68 | PG | 3 | | | |
| 4 | 39 | 81.14 | PG13 | 5 | | | |
| 5 | 89 | 104.07 | PG | 4 | | | |
| 6 | 93 | 101.38 | G | 4 | | | |

This means that 98.86 minutes is the most common movie length in our list. For this data set, we had different values for the mean, median, and mode, but you'll often see two or even all three of these values being exactly the same. It just depends on how skewed the data are (more on this later, in Exercise 3).

## Exercise 2: Standard Deviation

The standard deviation of a set of values is a measure of how spread out those values are. The mathematical formula for finding the standard deviation is pretty intense, but thankfully, Excel has a built-in function that'll do the work for you — as long as you're careful to enter the correct range of values.

1. In that same data set from Exercise 1, create a new column for the standard deviation of the "runtime" variable.

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | movieid | runtime | rating | liking | mean runtime | median runtime | mode runtime | std dev runtime |
| 2 | 6 | 112.74 | PG13 | 3 | 100.4988462 | 100.71 | 98.86 | |
| 3 | 76 | 96.68 | PG | 3 | | | | |
| 4 | 39 | 81.14 | PG13 | 5 | | | | |
| 5 | 89 | 104.07 | PG | 4 | | | | |
| 6 | 93 | 101.38 | G | 4 | | | | |

2. Click into cell H2 and use the STDEV formula. The syntax is **= STDEV(first cell:last cell)**. You want the standard deviation of all the values in column B, so once again, your range is B2:B105.

$f_x$ =STDEV(B2:B105)

Hit Enter.

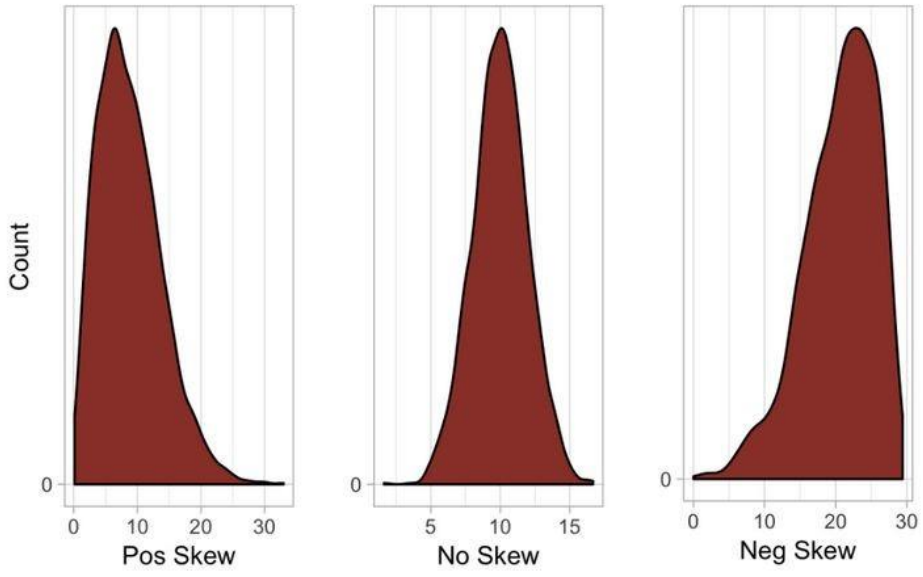
| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | movieid | runtime | rating | liking | mean runtime | median runtime | mode runtime | std dev runtime |
| 2 | 6 | 112.74 | PG13 | 3 | 100.4988462 | 100.71 | 98.86 | 10.6663822 |
| 3 | 76 | 96.68 | PG | 3 | | | | |
| 4 | 39 | 81.14 | PG13 | 5 | | | | |
| 5 | 89 | 104.07 | PG | 4 | | | | |
| 6 | 93 | 101.38 | G | 4 | | | | |

There we go: The standard deviation is about 10.67 when we round it. (Note: The last couple digits of your final answer might be slightly different due to rounding, depending on how wide the cells in column H are.)

What this means, basically, is that most of the movies in our list are within about 10.67 minutes of the *mean* (which we found in column E). In other words, most of these movies are about 100.5 minutes plus or minus 10.67 minutes (100.5 ± 10.67 minutes).

## Exercise 3: Skew

"Skewness" is a measure of how asymmetrical the distribution of a set of data is. A skewness of 0 (zero) means the data are normally distributed, or perfectly symmetrical with no skew. A positive value means the data are **positively skewed**, which means the longer "tail" of the data is on the positive side (to the right). A negative skew value means the data are **negatively skewed**, which means the longer "tail" of the data is on the negative side (to the left). Like this:




In this exercise, we'll find out exactly how skewed the movie runtime data is.

1. Create a new column on the spreadsheet for the skew.

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | movieid | runtime | rating | liking | mean runtime | median runtime | mode runtime | std dev runtime | skew |
| 2 | 6 | 112.74 | PG13 | 3 | 100.4988462 | 100.71 | 98.86 | 10.6663822 | |
| 3 | 76 | 96.68 | PG | 3 | | | | | |
| 4 | 39 | 81.14 | PG13 | 5 | | | | | |
| 5 | 89 | 104.07 | PG | 4 | | | | | |
| 6 | 93 | 101.38 | G | 4 | | | | | |

2. Once again, it's Excel to the rescue: Use the aptly-named SKEW function. The syntax is **=SKEW(first cell:last cell)**. You want the skewness measure of the movie runtime, which is in column B, so again, your first cell is B2 and your last cell is way down at B105.


Hit that Enter key.

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | movieid | runtime | rating | liking | mean runtime | median runtime | mode runtime | std dev runtime | skew |
| 2 | 6 | 112.74 | PG13 | 3 | 100.4988462 | 100.71 | 98.86 | 10.6663822 | -0.16791 |
| 3 | 76 | 96.68 | PG | 3 | | | | | |
| 4 | 39 | 81.14 | PG13 | 5 | | | | | |
| 5 | 89 | 104.07 | PG | 4 | | | | | |
| 6 | 93 | 101.38 | G | 4 | | | | | |

So, the skewness value is -0.16791 (again, your value might show up rounded slightly differently, depending on the width of your column). That means the movie runtimes are *negatively skewed*, but only slightly.