Descriptive Statistics

Table of Contents

INTRODUCTION

Descriptive statistics are used to summarize and organize your data. The product of your descriptive analysis might be enough to meet your needs, or you might choose to do more advanced analyses. This document introduces how to use descriptive statistics to analyze your data.

KEY POINTS IN THIS DOCUMENT

- Start your analysis with frequency and percentile distributions. A few common methods of presenting frequency and percentile distributions are frequency tables, bar charts, pie charts, and histograms. When presenting frequency and percentile distributions with metric data, use categories that do not overlap.
- Mean, median, and modes are ways of finding a measure of the typical score for a variable.
    - The mean is typically considered the best measure of typical score for metric data. However, the mean is sensitive to outliers.
    - The median is said to be robust to outliers because it only looks at the middle value and ignores the extremes.
    - The mode is the only measure of typical score that can be used for nominal data.
    - If your data set has more than one mode, you should be careful about using the mean and median, as mean and median may misrepresent the data by suggesting there is one central point.
- "Measures of dispersion" refers to measuring the amount of variability in a set of metric values. It is important to provide a measure of dispersion when you report on mean, median, or mode for metric data. Three common measures of dispersion are: range, interquartile range, and standard deviation.
    - The range provides limited information because it only accounts for the two most extreme values in the data set.
    - The interquartile range (IQR) is a more informative and reliable measure of dispersion than the range, as it is robust to outliers.
    - The standard deviation is the most commonly used measure of dispersion. The standard deviation is sensitive to outliers. If there are extreme values in the data, the standard deviation can be greatly increased.
- It can be informative to produce descriptive statistics for different sub-populations, such as gender and age.

FREQUENCY AND PERCENTILE DISTRIBUTION

Frequency and percentile distributions should be the first step in your analysis, as they will help you get an understanding of your data set. The tables and charts you create will also be useful for your final reporting.

A **frequency distribution** shows the number of people in each category.

A **percentile distribution** shows the percentage of people in each category (number of people in the category divided by total number of people). Percentile distributions allow you to compare your data to a different population with a different population size.

A few common methods of presenting frequency and percentile distributions are frequency tables, bar charts, pie charts, and histograms. Let's look at each of these methods.

Frequency tables

Frequency tables show the number and/or percentage of people belonging to each category option in a table format. They can be used for categorical and metric (i.e., quantitative) data.

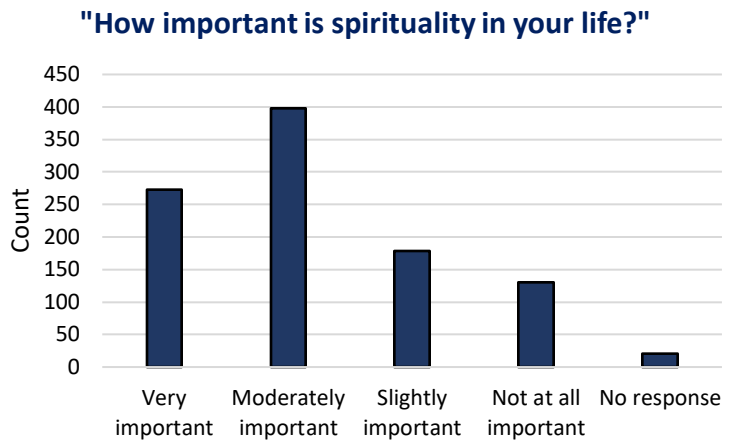
When metric data are presented in a frequency table, you will have to generate categories. Make sure that the categories you create do not overlap. For example, if you collected the variable "age" in your data set, your frequency table might appear as follows:

Age group	Frequency	Percent
10 and under	151	15.1%
11-20	148	14.8%
21-30	145	14.5%
31-40	132	13.2%
41-50	120	12.0%
51-60	114	11.4%
61-70	90	9.0%
71 and over	100	10.0%
Total	**1,000**	**100%**

Bar charts

Bar charts are used to display categorical data. In a bar chart, the height of the bars represents the number or percentage of people in each category.

For example, if one of your survey questions was "How important is spirituality in your life?", the bar chart might look like this:

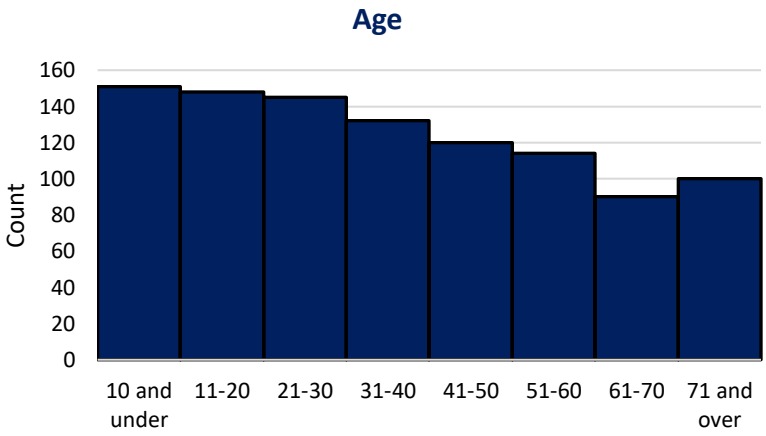


Categorical data: Variables for which there are a set list of categories to select from. There are two types of categorical data:

- **Nominal data:** The category options have no particular order (e.g., gender, religion).
- **Ordinal data:** The category options have a natural order (e.g., a range of response options from "excellent" to "poor").

Histograms

A histogram is used to display metric data. As in a bar chart, the height of the bar represents the number or percentage of people in each category. There is no space between the bars in a histogram, to indicate that the data is continuous. As described in the frequency table, you will have to generate categories that do not overlap. A histogram for the "age" data presented in the frequency table above might look like this:



MEAN, MEDIAN, AND MODE

Mean, median, and modes are ways of finding a measure of the typical score for a variable. This section will explain how to generate these measures manually; however, note that they can be easily calculated using any spreadsheet or analysis software.

Mean (average)

The mean is the average. Mathematically, the average is the sum of all values divided by the total number of values.

Example:

You have a data set that includes 10 people and each reported their age as follows:

Person	Age
Person 1	15
Person 2	37
Person 3	25
Person 4	46
Person 5	59
Person 6	36
Person 7	28
Person 8	85
Person 9	72
Person 10	65

The mean would be calculated as follows: (15+37+25+46+59+36+28+85+72+65)/10 = 46.8. The mean age of the population is 46.8.

Median

The median is the midpoint of a set of values. If you were to line up all the values from smallest to largest, the middle value is the median. If there are an even number of values, the median is the average of the two middle values.

The median can be measured for metric data and ordinal data. The median can be a better measure of typical score than the mean for metric data sets with outliers, as the median is less impacted by outliers than the mean.

Outlier: A value that is substantially different from other values in a data set. Outliers are most easily identified when data sets are visualized in a graph. For example, in the age dataset shown here, the individual with age 85 is an outlier as it is quite a bit higher than the next highest age.

Example:

Using the age data set used for "mean" above, what is the median value?

Person	Age
Person 1	15
Person 2	37
Person 3	25
Person 4	46
Person 5	59
Person 6	36
Person 7	28
Person 8	85
Person 9	72
Person 10	65

Lining these values up in order from smallest to largest, you can see that the middle two values are 37 and 46: 15, 25, 28, 36, 37, 46, 59, 65, 72, 85.

Therefore, the median value is (37+46)/2 = 41.5.

You can see that for this data set, the median (41.5) is slightly lower than the mean (46.8). As discussed above, the mean is heavily influenced by outliers. In this case, there are a couple of outliers with older ages (85, 72). Because of these outliers, median may be a better measure of typical score for this data set.

Mode

The "mode" is the value that occurs most frequently for a variable. There may be more than one mode in a data set. Mode is most useful for categorical data and is the only measure of typical score that can be used for nominal data. In the "marital status" table below, what is the mode?

Marital Status	Number of people
Married	321
Common law	264
Never married	100
Separated	150
Divorced	100
Widowed	65
Total	**1000**

The mode in the table above is "married", since there are more individuals in the "married" category than any other category.

Summary

The mean is typically considered the best measure of typical score for metric data. However, the mean is sensitive to outliers. Outliers are numbers that differ by a large amount from the typical score. This sensitivity is why the mean may not be the best measure of typical score for datasets with outliers. For data sets with extreme outliers, you could remove the outliers and note that you have done this in your reporting, or you could report the median as well as the mean.

The median is said to be robust to outliers because it only looks at the middle value and ignores the extremes.

Other more advanced compromises not described here include the trimmed mean and data transformations.

Note that the mode is the only measure of typical score that can be used for nominal data.

Note also that if your data set has more than one mode, you should be careful about using the mean and median, as mean and median may misrepresent the data by suggesting there is one central point.

MEASURES OF DISPERSION

"Measures of dispersion" refers to measuring the amount of variability in a set of metric values. In other words, measures of dispersion show how spread out or clustered together your data set is.

It is important to provide a measure of dispersion when you report on mean, median, or mode for metric data. The measure of dispersion provides important context about how representative the mean, median, or mode is of the whole population.

Three common measures of dispersion are: range, interquartile range, and standard deviation.

Range

The range is the difference between the minimum and maximum value for a variable. The range is a very simple calculation; however, it provides limited information because it only accounts for the two most extreme values in the data set.

Example:

In the age data set below, what is the range?

Person	Age
Person 1	15
Person 2	37
Person 3	25
Person 4	46
Person 5	59
Person 6	36
Person 7	28
Person 8	85
Person 9	72
Person 10	65

The minimum value is 15 and the maximum value is 85, so the range is 85-15=70.

Interquartile range

The interquartile range, or IQR, is the difference between the 1st quartile and the 3rd quartile of the distribution of values. If you were to line up all the values in the data set from smallest to largest, the 1st quartile would be the value for which 25% of values are smaller and 75% of values are larger. The 3rd quartile is the value for which 25% of values are larger and 75% of values are smaller. For context, the median is also the 2nd quartile.

The interquartile range shows the spread of the centre half of the values. It is a more informative and reliable measure of dispersion than the range, since it is not focused on outliers. In other words, like the median, the IQR is robust to outliers.

The interquartile range can also be thought of as the middle half of the data, because it describes the 50% of the data that is nearest the median.

To calculate the quartiles, sort the data from lowest to highest like you would when calculating the median. Then, take the median of the lower half for the 1st quartile and the median of the upper half of the 3rd quartile.

Example:

Consider the same ten numbers as in the median example, 15, 25, 28, 36, 37, 46, 59, 65, 72, 85.

The lower quartile is the median of {15, 25, 28, 36, 37}, or 28.

The upper quartile is the median of {46, 59, 65, 72, 85}, or 65.

If there is an odd number of values in the set, remove the middle value before splitting the values into upper and lower halves. Some computer programs use a more complicated formula for quartiles, so double check, but do not worry if your by-hand determination disagrees with your computer's determination.

Standard deviation

The standard deviation is the most commonly used measure of dispersion. It is the measure of variation around the mean. The standard deviation is typically presented alongside the mean.

The standard deviation is difficult to calculate manually, but easily calculated using any spreadsheet or data analysis software. Due to the complexity of the calculation, we will not cover it here. If you would like to understand the calculation, see this article from the Khan Academy.

Note that the calculation for the standard deviation of a *sample* is different than the calculation for the standard deviation of the *population*. When you use software to calculate the standard deviation, ensure that you use the right formula.

Like the mean, the standard deviation is sensitive to outliers. If there are extreme values in the data, the standard deviation can be greatly increased.

SUB-POPULATION GROUPINGS

It can be informative to produce descriptive statistics for different sub-populations. For example, you may be interested in looking at responses by gender, by age grouping, or by income. You may also be interested in understanding the differences between your population of interest and comparison populations to provide more context for any differences in outcomes or indicators. Groupings should be intuitive and easy to understand.