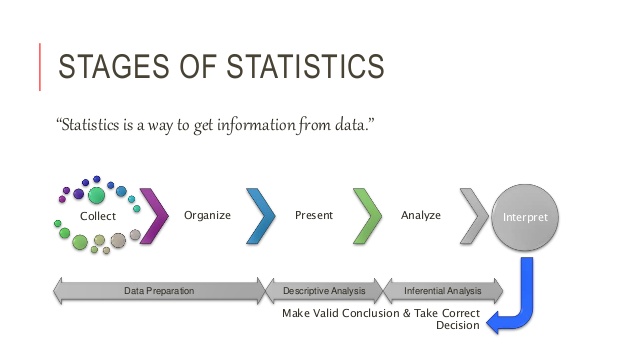
Statistics for Data Science

The science of collecting, describing, interpreting and presenting data is popularly known as Statistical leveraging in Data Science.

Statistics is a way to get information from data. Statistical knowledge helps us to use the proper methods to collect the data, employ the correct analyses, and effectively present the results.



**Two areas of statistics in Data Science:**

Descriptive Statistics – Methods of organizing, summarizing, and presenting data in an informative way

Inferential Statistics – The methods used to determine something about a population on the basis of sample

**Descriptive Statistics**

Descriptive statistics are methods for organizing and summarizing data. For example, tables or graphs are used to organize data, and descriptive values such as average score are used to summarize data.

A descriptive value for a population is called a parameter and a descriptive value for a sample is called statistic.

Collect data : e.g., Survey 

Present data : e.g., Tables and graphs

Summarize data : e.g., Sample mean 

Descriptive statistics is key because it allows us to present large amounts of raw data in a meaningful way. This enables a better interpretation of data. There are usually two types of descriptive statistics:

**(i) Measures Of Central Tendency**

These are statistical measures that describe the central position of a frequency distribution for a large amount of raw data. These measures include many different statistics such as mean, mode and median. Different measures of central tendency are more appropriate to use under different conditions. **A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.**

**Mean:** The mean is the average of all numbers and is sometimes called the arithmetic mean

**Median:** The statistical median is the middle number in a sequence of numbers. To find the median, organize each number in order by size; the number in the middle is the median

**Mode:** The mode is the number that occurs most often within a set of numbers

**When to use mean, median and mode?**

**Mean** – When your data is not skewed i.e normally distributed. In other words, there are no extreme values present in the data set (Outliers).

**Median**– When your data is skewed or you are dealing with ordinal (ordered categories) data (e.g. likert scale 1. Strongly dislike 2. Dislike 3.Neutral   4. Like 5. Strongly like)

**Mode**- When dealing with nominal (unordered categories) data.

**Example**

 In real life, suppose a company is considering expanding into an area and is studying the size of containers that competitors are offering. They would be more interested in the mode because they want to know what size tends to sell most often.

**(i) Measures Of Spread**

|  |
| --- |
| A measure of spread (variability, dispersion, scatter) refers to how the data within the set is "spread out" (or "dispersed", or "scattered") about the mean. |

If the data is clustered around the center value, the "spread" is small.  
The further the distances of the data values from the center value, the greater the "spread".

Measures of spread describe how spread out the distribution is for a particular group of data. Measures of spread give an idea of the range and variation in a given set of data. This helps develop a better understanding of the nature of the data. Measures of spread include things like variance, standard deviation, range, and quartiles.

.

|  |  |
| --- | --- |
| lessspread1 | morespread |
|  |  |
| **Measures of Spread** | **May also be called: Measures of Variability, Measures of Dispersion, or Measures of Scatter** |

**Range :**

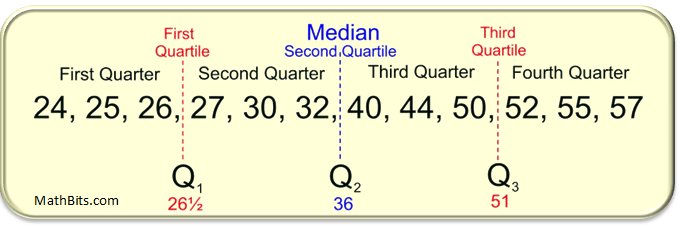
Range is the difference between the largest data value and the smallest data value in the set. While the range is simple to compute, it is often unreliable as a measure of variability. The range is based on only two values within the set, which may tell very little about "how" the remaining values are distributed in the set. For this reason, range is used as a supplement to other measures of spread, instead of being the only measure of spread.

|  |
| --- |
| range This range of 43 tells us very little about how the data in this set is scattered. The range alone cannot tell us, for example, if the data is clustered to one end of the set, or if there is an outlier in the data set. |

**Inter Quartile Range (IQR):**

*It is the measure of variability, based on dividing a data set into quartiles.*

* **Quartile**: Quartiles tell us about the spread of a data set by breaking the data set into quarters, just like the median breaks it in half.  The interquartile range is another form of range which divides the set into four equal parts (or quarters). The three values that form the four divisions are called quartiles: first quartile, Q1; second quartile (median), Q2, and third quartile, Q3. The interquartile range is the difference between the third quartile and the first quartile.
* A median divides a data set into two equal parts. The set can be subdivide further into four equal parts, by values called quartiles. The quartiles divide the data set into quarters, with each quarter containing one-fourth (or 25%) of the data. The quartiles are like additional "medians" of the lower and upper halfs of the data set.



|  |  |  |
| --- | --- | --- |
| **Q1:**The first quartile is the middle (the median) of the lower half of the data set. One-fourth (25%) of the data lies below the first quartile, and three-fourths (75%) lies above. | **Q2:** The second quartile is another name for the median of the entire set. One-half (50%) of the data lies below the second quartile, and one-half (50%) lies above. | **Q3:** The third quartile is the middle (the median) of the upper half of the data set. Three-fourths (75%) of the data lies below the third quartile and one-fourth (25%) lies above. |

* The difference between the third quartile and first quartile is called  
  the **interquartile range (IQR)**.  
  The interquartile range (also called the *midspread* or *middle fifty*), is the distance between the third and first quartiles and is considered a more stable statistic than the "range" of the set.   
  The IQR contains 50% of the data.

|  |  |  |
| --- | --- | --- |
| |  | | --- | | **IQR = Q3 - Q1** | | For the example shown above, the IQR = 51 - 26½ = 24½. |

It may be the case that a data value falls well outside the range of the other values in the set. Such data values are called **outliers** (as they "lie outside" the other values) .

Outliers are defined as those data points that fall more than a specified distance from the first or third quartiles. That specified distance is 1.5 • IQR (one and one-half times the IQR). Data points that fall to the far left, or far right, of an ordered data set should be tested as possible outliers.

|  |  |  |  |
| --- | --- | --- | --- |
| **Outliers** are: greater than Q3 + (1.5 • IQR) (referred to as the upper fence) or less than Q1 - (1.5 • IQR) (referred to as the lower fence) | | | |
| **Box & Whiskers:**  A five statistical summary can be represented graphically as a **box and whisker plot** (or box plot). The first and third quartiles are the ends of the box, the median is indicated with a vertical line in the interior of the box, and the minimum and maximum are the ends of the whiskers (unless an outlier is present). Each of the four "sections" of a box plot represents 25% of the data in the set. | | https://1.bp.blogspot.com/-pAiXmkTmUzs/X0NLpIxW5pI/AAAAAAAAA0A/FIqof_E58EsRd3lzdvxrkkl4gCEvk1Y7gCLcBGAsYHQ/s614/box-plot.png | | |
| Data Set: {1, 30, 40, 44, 44, 44, 45, 46, 47, 51, 54, 54, 55} It certainly looks like the "1" is not in keeping with the rest of these values. Let's test it to see if it is an outlier. First, we need to find the first and third quartiles: | | | |
| outlierpic | | | |
| Now, do the calculations to test for an outlier: Is "1" less than Q1 - (1.5 • IQR)? outliercheck Since "1" is less than 26.25, "1" is definitely an outlier.  The "1" is plotted as a single dot (or asterisk \*), separate from the box's whisker . The whisker then uses 30 as its minimum point. | outlierpicgraph Graph with outlier. | | |

**Variance**

The variance is the average of the squared differences from the mean. A small variance indicates that the data points tend to be very close to the mean and to each other A high variance indicates that the data points are very spread out from the mean and from each other. One problem with the variance is that it does not have the same unit of measure as the original data. For example, original data containing lengths measured in feet has a variance measured in square feet.

*Process:* (1) Find the mean (average) of the set. (2) Subtract each data value from the mean to find its distance from the mean. (3) Square all distances. (4) Add all the squares of the distances. (4) Divide by the number of pieces of data (for population variance).

|  |  |
| --- | --- |
| formula2 | formulakey1 |

*Note:* the notation used to represent "variance" is actually the square of the notation for standard deviation. The notation always reminds us of the relationship betweeen these two quantities.  
  
varSDnotation

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | A1Side | **Measures of Spread** [MathBitsNotebook.com](http://mathbitsnotebook.com/)  [Topical Outline](https://mathbitsnotebook.com/Algebra1/StatisticsData/SToutline.html) | [Algebra 1 Outline](https://mathbitsnotebook.com/Algebra1/Algebra1.html) | [MathBits' Teacher Resources](http://mathbits.com/)  [Terms of Use](https://mathbitsnotebook.com/termsofuseN.html)**Contact Person:** [Donna Roberts](mailto:donrob@twcny.rr.com?subject=Notebook%20-%20Algebra1) divider |   Knowing the mean, or median, of a data set yields a certain amount of information about the typical data within the set. It is possible, however, that many different data sets may have the same mean value (or median value). To determine how these data sets are different requires that we expand our investigation to obtain more information about the set. One additional investigation is the examination of the measure of spread of the data set. How is the data "spread out"?   |  |  | | --- | --- | | definition | A measure of spread (variability, dispersion, scatter) refers to how the data within the set is "spread out" (or "dispersed", or "scattered") about the mean. |   If the data is clustered around the center value, the "spread" is small. The further the distances of the data values from the center value, the greater the "spread".   |  |  | | --- | --- | | lessspread1 | morespread |      |  |  |  | | --- | --- | --- | | |  |  | | --- | --- | | Measures of Spread | May also be called: Measures of Variability, Measures of Dispersion, or Measures of Scatter | |   bullet Range: The first method of measuring "spread" of a data set that you learned was finding the range. Range is the differene between the largest data value and the smallest data value in the set. While the range is simple to compute, it is often unreliable as a measure of variability. The range is based on only two values within the set, which may tell very little about "how" the remaining values are distributed in the set. For this reason, range is used as a supplement to other measures of spread, instead of being the only measure of spread.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | range This range of 43 tells us very little about how the data in this set is scattered. The range alone cannot tell us, for example, if the data is clustered to one end of the set, or if there is an outlier in the data set. | |  |  |  | | --- | --- | --- | | |  |  | | --- | --- | | [ti84c](http://mathbits.com/MathBits/TISection/Statistics2/dispersion.htm) | For calculator help with range  [click here.](http://mathbits.com/MathBits/TISection/Statistics2/dispersion.htm) | | |   bulletInterquartile Range (IQR): ([Read more about IQR](https://mathbitsnotebook.com/Algebra1/StatisticsData/STboxplot.html)) The interquartile range is another form of range which divides the set into four equal parts (or quarters). The three values that form the four divisions are called quartiles: first quartile, Q1; second quartile (median), Q2, and third quartile, Q3. The interquartile range is the difference between the third quartile and the first quartile. You can think of the IQR (also called the midspread or middle fifty) as a "range" between the third and first quartiles. The IQR is considered a more stable statistic than the typical range of a data set, as seen in the first section. The IQR contains 50% of the data, eliminating the influence of outliers.   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | rangequartile   |  |  | | --- | --- | | IQRformula | quartile key | | |  |  |  | | --- | --- | --- | | |  |  | | --- | --- | | [ti84c](http://mathbits.com/MathBits/TISection/Statistics1/FiveNumberSummary.htm) | For calculator help with IQR from 5 number summary [click here.](http://mathbits.com/MathBits/TISection/Statistics1/FiveNumberSummary.htm) | | |  |  | | --- | | For the following methods, you need to understand [*"population" vs "sample" data*](https://mathbitsnotebook.com/Algebra1/StatisticsData/STPopSample.html).Unlike range and interquartile range, these methods utilize ***all***of the values in a data set to produce a measure of spread. |   bulletMean Absolute Deviation (MAD): ([Read more about MAD](https://mathbitsnotebook.com/Algebra1/StatisticsData/STMAD.html)) The mean absolute deviation is the average (mean) of the absolute value of the differences between each piece of data in the data set and the mean of the set. It measures the average distances between each data element and the mean. Process: (1) Find the mean (average) of the set. (2) Subtract each data value from the mean to find its distance from the mean. (3) Turn all distances to positive values (take the absolute value). (4) Add all of the distances. (4) Divide by the number of pieces of data (for population MAD).   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | formula1 | formulakey1 | |  |  |  | | --- | --- | --- | | |  |  | | --- | --- | | [ti84c](http://mathbits.com/MathBits/TISection/Statistics1/MAD.html) | For calculator help with MAD  [click here.](http://mathbits.com/MathBits/TISection/Statistics1/MAD.html) | | |     **bullet** Variance: ([Read more about Variance](https://mathbitsnotebook.com/Algebra1/StatisticsData/STSD.html)) The variance is the average of the squared differences from the mean. A small variance indicates that the data points tend to be very close to the mean and to each other A high variance indicates that the data points are very spread out from the mean and from each other. One problem with the variance is that it does not have the same unit of measure as the original data. For example, original data containing lengths measured in feet has a variance measured in square feet. The process is very similar to finding the MAD. The only difference is the squaring of the distances.Process: (1) Find the mean (average) of the set. (2) Subtract each data value from the mean to find its distance from the mean. (3) Square all distances. (4) Add all the squares of the distances. (4) Divide by the number of pieces of data (for population variance).   |  |  | | --- | --- | | formula2 | formulakey1 |   Note: the notation used to represent "variance" is actually the square of the notation for standard deviation. The notation always reminds us of the relationship betweeen these two quantities.  varSDnotation   |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  |  | | --- | --- | --- | | |  |  | | --- | --- | | [ti84c](http://mathbits.com/MathBits/TISection/Statistics2/dispersion.htm) | For calculator help with variance and standard deviation [click here.](http://mathbits.com/MathBits/TISection/Statistics2/dispersion.htm) | | | |  |  |  | | --- | --- | --- | | |  |  | | --- | --- | | [ti84c](http://mathbits.com/MathBits/TISection/Statistics2/GroupDispersion.htm) | For calculator help with variance with grouped data [click here.](http://mathbits.com/MathBits/TISection/Statistics2/GroupDispersion.htm) | | |   **StandardDeviation** The standard deviation is the average distance of each data point from the mean. The standard deviation is the square root of the variance. Taking the square root will return the same units as expressed in the original data, thus eliminating this problem as found with variance. Now, original data containing lengths measured in feet has a standard deviation measured in feet. A low standard deviation indicates that the data points tend to be very close to the mean. A high standard.deviation indicates that the data points are spread out over a large range of values.  The process is simply square rooting the variance. The process for finding variance is shown above.Process: Find the square root of the variance   |  |  | | --- | --- | | sdformula | formulakey1 | |

**Inferential Statistics**

Inferential statistics are methods for using sample data to make general conclusions (inferences) about populations.

Because a sample is typically only a part of the whole population, sample data provide only limited information about the population. As a result, sample statistics are generally imperfect representatives of the corresponding population parameters.

* Estimation
* e.g., Estimate the population mean weight using the sample mean weight
* Hypothesis testing
* e.g., Test the claim that the population mean weight is 70kg

Inference is the process of drawing conclusions or making decisions about a population based on sample results.

|  |  |
| --- | --- |
| **Descriptive Statistics** | **Inferential Statistics** |
| Concerned with the describing the target population | Make inferences from the sample and generalize them to the population |
| Organize, analyze and present the data in a meaningful manner | Compares, test and predicts future outcomes |
| Final results are shown in form of charts, tables and graphs | Final results is the probability scores |
| Describes the data which is already known | Tries to make conclusions about the population that is beyond the data available |
| Tools- Measures of central tendency (mean/media/mode),Spread of data(range, standard deviation, etc) | Tools- hypothesis tests, Analysis of variance, etc |
| Don’t require any assumptions on part of the users as all the raw data is already there | Some inferential tests require the user to make assumptions or educated guesses before running the test. This can also affect the accuracy and certainty of the result |

**Definition to Basic terms**

**Population :** A Collection, or set, of individuals or objects or events whose properties are to be analyzed. Two kinds of populations: *finite* or *infinite*

**Sample:** A subset of the population

**Variable:** A characteristic about each individual element of a population or sample

**Data(singular):** The value of the variable associated with one element of a population or sample. This value may be a number, a word, or a symbol.

**Data(plural):** The set of values collected for the variable from each of the elements belonging to the sample.

**Experiment:** A planned activity whose results yield a set of data.

**Parameter:** A numerical value summarizing all the data of an entire population

**Statistic:** A numerical value summarizing the sample data

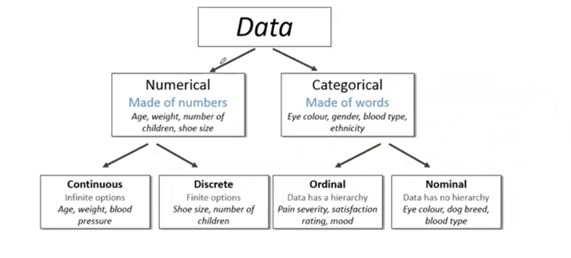
**Types of Data / Variables:**

Data refers to facts and statistics collected together for reference or analysis.

Data can be collected, measured and analyzed. It can also be visualized by using statistical models and graphs. Data can be categorized into two sub-categories:

1.Qualitative Data (Categorical)

2. Quantitative Data(Numerical)



Qualitative

Quantitative

