Thegradientboost
CREATING STRONGER LEARNERS

Descriptive Statistics

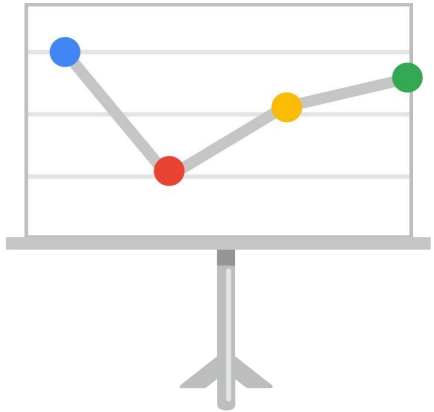
Descriptive statistics refers to the analysis of data that helps describe, show or summarize data in a meaningful way.

If we simply opted to present raw data it would be difficult to understand or explain what the data is showing especially as the size of the data increases.

With descriptive statistics are looking to describe the data without necessarily making any inferences about it.

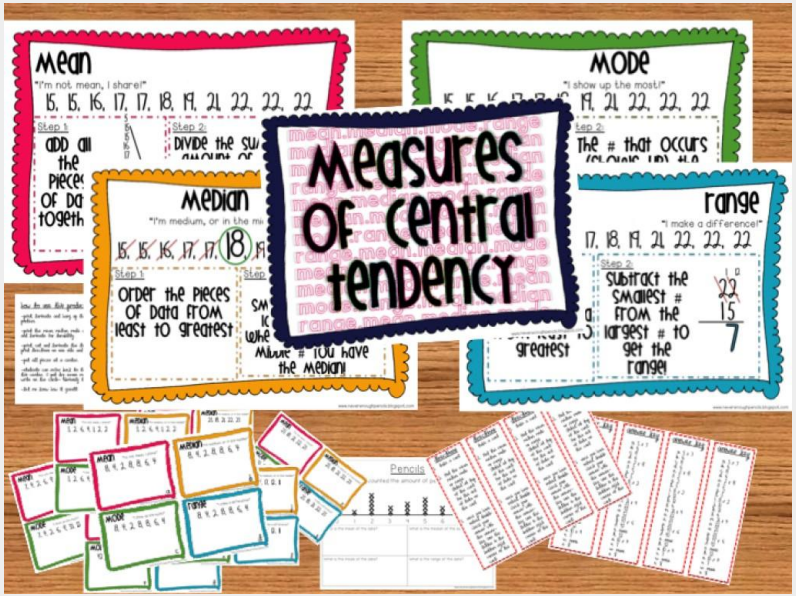
There are two general types of statistics used to describe data:

1. Measures of Central Tendency
2. Measures of Spread/Variability



Source: Descriptive and Inferential Statistics, Laerd Statistics, Available at https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php

Measures of Central Tendency



A measure of central tendency is a value used to describe data by identifying the central position within the data.

There are 3 main measures of central tendency, these are:
i.) the mean
ii.) median
iii.) and mode

Thegradientboost
CREATING STRONGER LEARNERS

Measures of Central Tendency: Mean, Median and Mode

Mean

The mean refers to the sum of all values in a set of values divided by the number of values in the set of values or

$$\text{Mean } (\bar{x}) = \frac{\sum x}{n}$$

Median

The mode refers to the **middlemost number** in a set of values ordered from either lowest to largest or vice versa. If we do not have an exact middle number we would take the average of the two numbers nearest the middle.

Assuming we have a set (1,2,3,4,5) the mode would equal the number in the middle- 3

Mode

The mode is the value(s) that appears the **most frequently** in a set of values or dataset.

Assuming we have a set (1,1,2,3,4,5). The mode would be the number 1 as it appears more frequently than other numbers in the the set

Thegradientboost

CREATING STRONGER LEARNERS

Additional Reading/Videos:

- Measures of Central Tendency: https://www.youtube.com/watch?v=NM_iOLUwZFA
- Data Science Statistics: Measures of Central Tendency [Explanation & Code in R and Python]: https://medium.com/@pratiks.solanki3/data-science-statistics-measures-of-central-tendency-explanation-code-in-r-and-python-12b00b14a4f9
- Chapter 2.1 Descriptive Statistics, Chapter 2.2 Descriptive Statistics: http://greenteapress.com/thinkstats/html/thinkstats003.html
-

Measures of Spread/Variability

A measure of spread, variability or dispersion is a measure of how spread out the data values are from the central value

Together with the measures of central tendency there are values that help us understand the distribution of a set of values.

There are four commonly used measures of spread. These are: range, interquartile range, variance and standard deviation.

Thegradientboost
CREATING STRONGER LEARNERS

Measures of Spread/Variability: Range, IQR and Standard Deviation

Range

The range is defined as a single value that is the difference between the largest and smallest values in a set of values.

Interquartile Range

The interquartile range is a measurement describing where the bulk of the values lie in a set of values. This is calculated by subtracting the first quartile from the third quartile.

NOTE: Quartiles are values that divide a set of values into quarters

Variance

Variance defines how the spread out the values are from the mean

$$s^2 = \frac{\sum (X - \overline{X})^2}{N - 1}$$

To get the variance we subtract the mean from each value in our set, square the differences, add the squared differences and then divide the total by the length of the set of values - 1

Standard Deviation

Standard deviation measures the spread of the data relative to the mean. The standard deviation is calculated by taking the square root of the variance.

The standard deviation represents the typical deviation of observations in a dataset from the mean.

Thegradientboost

CREATING STRONGER LEARNERS

Additional Reading/Videos:

- Measures of Variability: https://www.youtube.com/watch?v=Cx2tGUze60s
- Variability of data with Pandas: https://www.ritchieng.com/pandas-variability/
- Chapter 2.2 Descriptive Statistics, Think Stats, Probability and Statistics for programmers: http://greenteapress.com/thinkstats/html/thinkstats003.html

Descriptive Statistics using graphs

The Big Idea

Graphs provide a visual representation of your data. Understanding each type of graph opens options for you to better understand your dataset

The danger with summary statistics is that while concise, they may obscure the bigger picture.

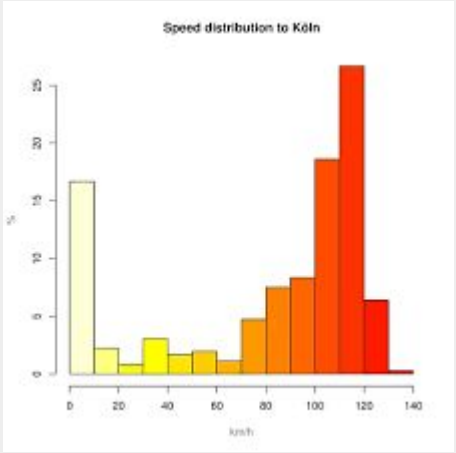
It is often wise to look at both summary statistics and graphs showing the distribution of data to fully understand your data, the skewness of the distribution and other factors that may help you better understand your data.

A distribution describes how often observed values occur in a sample space.

Thegradientboost

CREATING STRONGER LEARNERS

Descriptive Statistics: Histograms



Frequency: how often a value is in a dataset

Probability: how often a frequency is seen out of the total sample size (n)

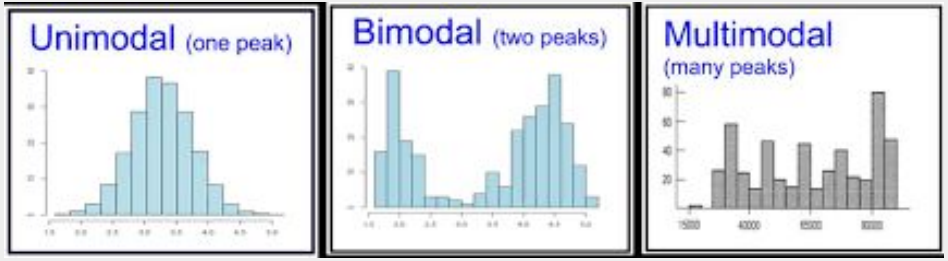
One of the most common ways to represent distribution is a **histogram**: a graph that shows the frequency or probability of each value.

In a histogram, each block represents a bin (or interval). Each value (observation) in the dataset is put into a bin. The number of values in a bin is the frequency. The higher the bar, the more frequent the data is.

For example, if you are creating a histogram of ages in this course you would see bins of 18-24, 25-30, 31-40, 41-50 etc.

Thegradientboost
CREATING STRONGER LEARNERS

Histograms: unimodal, bimodal and multimodal distributions

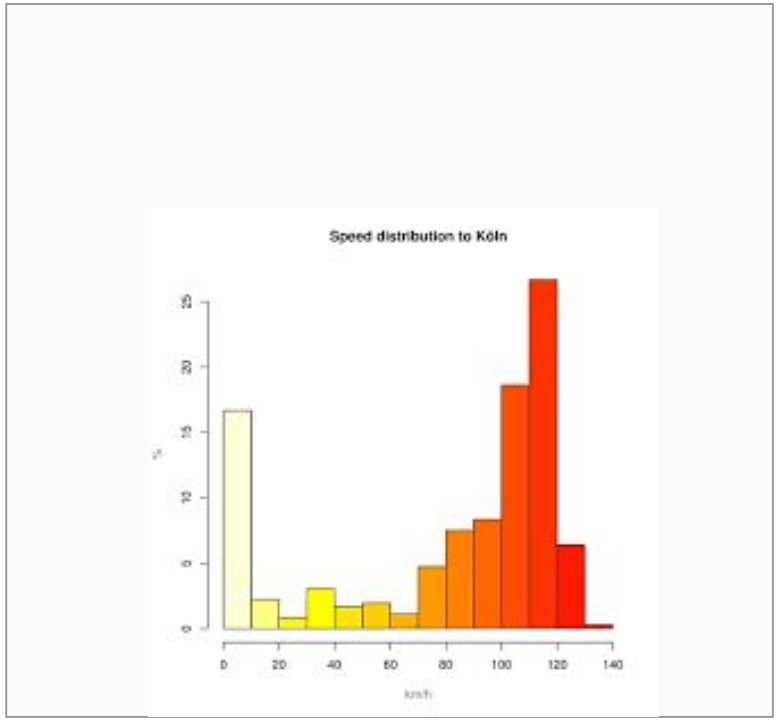


You may notice peaks in your histogram. These represent the value that appears most frequently (the mode). Histograms with a single peak are known as unimodal, bimodal if there are bimodal and histograms with multiple peaks are multimodal

Additional Reading:

Thegradientboost
CREATING STRONGER LEARNERS

Understanding the Histogram: Skewness



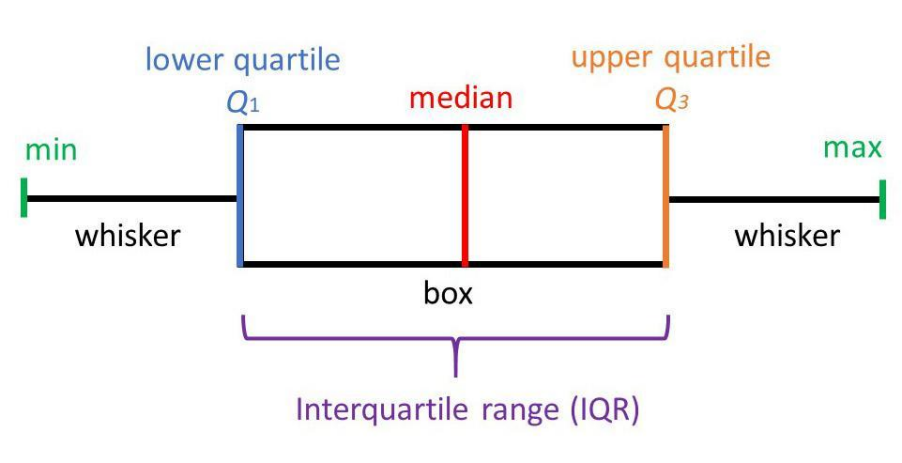
From the histogram we can start to see the skewness of the distribution of our data. This involves looking at whether the bulk of the data lies to the left or right. We call the distribution positively skewed or skewed right if the peak is towards the right and the left tail is longer.

Additional Reading/Videos

- Skewed Histogram: https://www.youtube.com/watch?v=H9ITfdaX2ZQ
- Skewness: Are the Skewness and Kurtosis Useful statistics? https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics#skewness

Thegradientboost
CREATING STRONGER LEARNERS

Descriptive Statistics: Boxplots



We can think of a box plot as a graph that summarises a few statistics. These are the outliers both on the lower and upper end, the first and third quartile (and as a consequence the IQR) and the median.

By outliers we are referring to any observation that appears very different relative to the rest of the data

-

For additional reading please visit these sources:

- Reading and Interpreting boxplots, https://magoosh.com/statistics/reading-interpreting-box-plots/
- Interpreting boxplots, Khan Academy: https://www.khanacademy.org/math/ap-statistics/summarizing-quantitative-data-ap/stats-box-whisker-plots/v/interpreting-box-plots
- Boxplot review, Khan Academy: https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/box-plot-review

Thegradientboost
CREATING STRONGER LEARNERS

Robust Statistics

Not all summary statistics are created equally. Some summary statistics are regarded as more robust. This means that the statistics are less affected by outlier values in a dataset. Summary statistics such as the mean for example, are not robust as a large enough outlier has the ability to skew the mean towards the direction of the outlier whether large or small

For additional reading please visit these sources:

- Chapter 2.1.6: Robust Statistics,OpenIntro Statistics

Thegradientboost

CREATING STRONGER LEARNERS

Understanding different data types

Qualitative data is the data in which the classification of objects is based on attributes and properties.

Quantitative Data is the type of data which can be measured and expressed numerically.

QUALITATIVE

NOMINAL

If there is no natural order between the categories (eg eye colour)

ORDINAL

If an ordering exists (e.g. exam results, survey results on 'how satisfied are you with')

QUANTITATIVE

DISCRETE

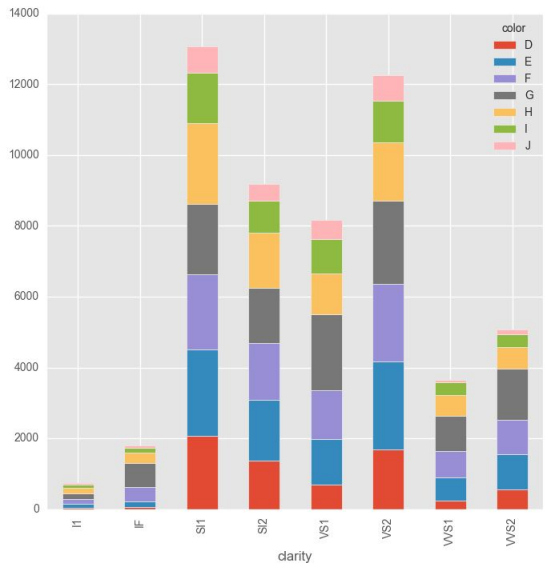
If the measurements are integers (eg number of people in a household)

CONTINUOUS

If the measurements can take on any value, usually within some range (e.g. weight, revenue)

Thegradientboost
CREATING STRONGER LEARNERS

Descriptive Statistics: Categorical data



Supplementary Reading:

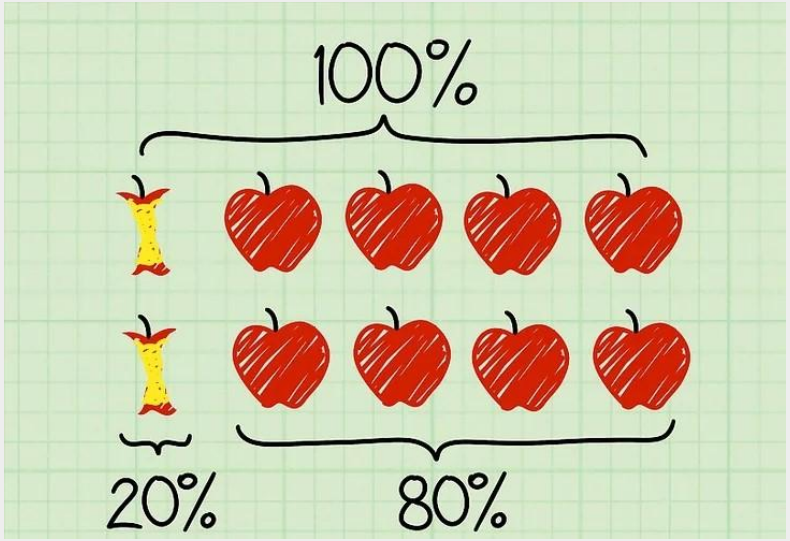
- Plotting with categorical data, https://datascienceplus.com/seaborn-categorical-plots-in-python/

Categorical variables include data that are non quantitative. This could include names or labels. There are multiple plots we can use to visualize categorical data. These plots include; contingency tables, barplots, mosaic plots and pie charts.

You can read more about these plots in Chapter 2.2 Considering categorical data in the book 'Open Intro Statistics' that has been provided to you

Thegradientboost
CREATING STRONGER LEARNERS

Descriptive Statistics: Percentiles



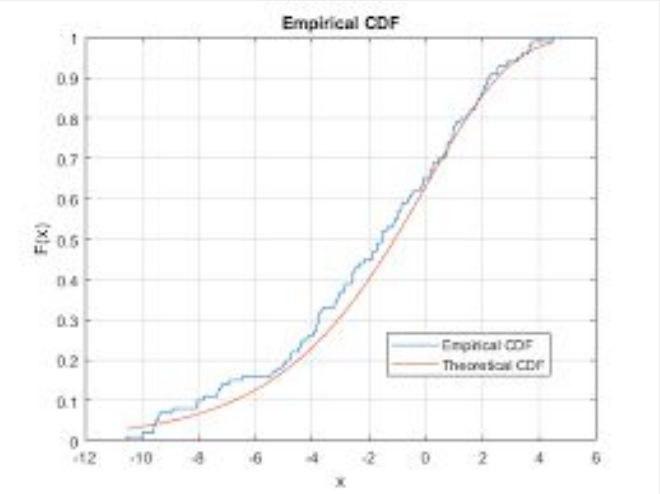
Additional Reading:
- Chapter 3.3 Percentiles ThinkStats

Percentiles can be defined as a number/numbers whereby a certain percent of scores fall below that number. If for example your test score is in the 95th percentile this means that you scored better than 95% of the people who took the same tests.

Percentile rank refers to the proportion of values in a distribution that a specific score is greater than or equal to. Using the above example if you are in the 95th percentile in terms of test scores then your percentile rank would be 95

Thegradientboost
CREATING STRONGER LEARNERS

Descriptive Statistics: CDF



We would typically use a cumulative distribution function to map values to their percentile rank in a distribution.

The CDF is regarded as a measure of how much a variable accumulates.

Additional Reading:

- Chapter 3.4 Cumulative Distribution Functions, ThinkStats
- What is CDF - Cumulative Distribution Function, StackEchange https://math.stackexchange.com/questions/52400/what-is-cdf-cumulative-distribution-function

Thegradientboost
CREATING STRONGER LEARNERS