Practical Statistics

# Descriptive Statistics

Descriptive Statistics is about describing our collected data

Data

Data is essentially information, data can come in any form such as text, tabular form, video and audio and from sensors.

Two types of Data

**Quantitative** data takes on numeric values that allow us to perform mathematical operations (like the number of dogs).

**Categorical** are used to label a group or set of items (like dog breeds - Collies, Labs, Poodles, etc.).

We can divide **categorical data** further into two types: **Ordinal** and **Nominal**.

**Categorical Ordinal** data take on a ranked ordering (like a ranked interaction on a scale from Very Poor to Very Good with the dogs).

**Categorical Nominal** data do not have an order or ranking (like the breeds of the dog).

We can think of **quantitative data** as being either **continuous** or **discrete**.

**Continuous** data can be split into smaller and smaller units, and still a smaller unit exists. An example of this is the age of the dog - we can measure the units of the age in years, months, days, hours, seconds, but there are still smaller units that could be associated with the age.

**Discrete** data only takes on countable values. The number of dogs we interact with is an example of a discrete data type.

Analyzing Quantitative Data

Four aspects to analyzing quantitative data

1. Measures of **Center**
2. Measures of **Spread**
3. The **Shape** of the data.
4. **Outliers**

Measures of Center

There are three measures of center

1. **Mean**
2. **Median**
3. **Mode**

The Mean

The mean is often called the average or the **expected value** in mathematics. We calculate the mean by adding all of our values together, and dividing by the number of values in our dataset.

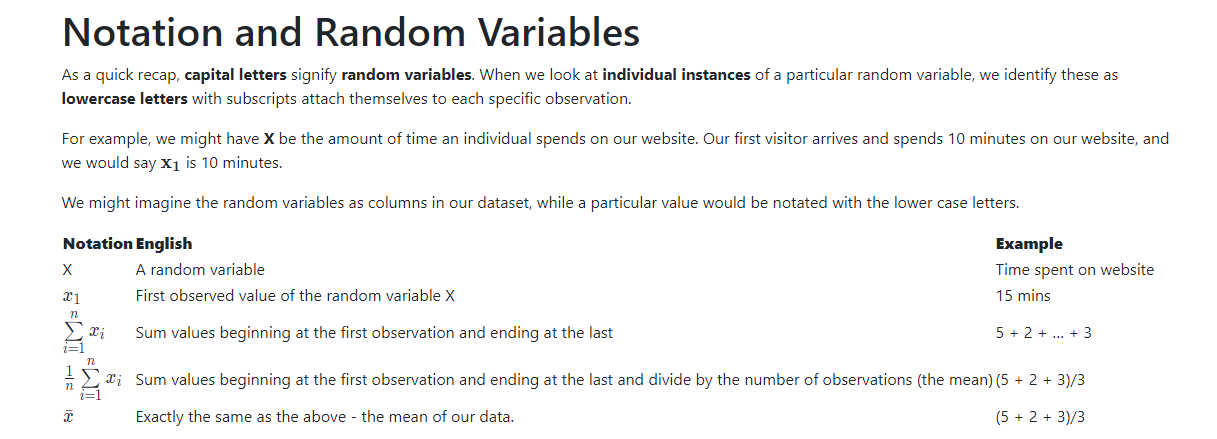
The Median

The **median** splits our data so that 50% of our values are lower and 50% are higher.

Whether we use the mean or median to describe a dataset is largely dependent on the **shape** of our dataset and if there are any **outliers**.

The Mode

The **mode** is the most frequently observed value in our dataset.



Measures of Spread

**Measures of Spread** are used to provide us an idea of how spread out our data are from one another. Common measures of spread include:

1. **Range**
2. **Interquartile Range (IQR)**
3. **Standard Deviation**
4. **Variance**

Histograms

It is the most common visual method used for quantitative data

Histograms are created by grouping data into groups or bins, depending on the specifications of the bins like sizes or different bin edges the histogram may be slightly different

Histograms gives the count of each bins

Histograms are useful to understanding the different aspects of quantitative data like

* center
* spread
* shape
* outliers

## **Calculating the 5 Number Summary**

The five number summary consist of 5 values:

1. **Minimum:** The smallest number in the dataset.
2. **Q1**​: The value such that 25% of the data fall below.
3. **Q2**​: The value such that 50% of the data fall below.
4. **Q3**​: The value such that 75% of the data fall below.
5. **Maximum:** The largest value in the dataset.

Range

The **range** is the difference between the **maximum** and the **minimum**.

IQR

The **interquartile range** is the difference between **Q3**​ and **Q1**​.

Boxplots

Boxplots are useful for quickly comparing the distributions of two datasets (5 Num Summ)

Standard Deviation and Variance

The **standard deviation** is one of the most common measures for talking about the spread of data. It is defined as **the average distance of each observation from the mean**.

When comparing the spread between two datasets, the units of each must be the same.

The standard deviation is associated with risk in finance, assists in determining the significance of drugs in medical studies, and measures the error of our results for predicting anything from the amount of rainfall we can expect tomorrow to your predicted commute time tomorrow.

Shape

The distribution of our data is frequently associated with one of the three **shapes**:

**1. Right-skewed**

**2. Left-skewed**

**3. Symmetric** (frequently normally distributed)

The shape of the distributions can tell us a lot about mean and median of the data

Symmetric (Normal) 🡪 Mean equals Median 🡪 Height, Weight, Errors, Precipitation

Right-skewed 🡪 Mean greater than Median 🡪 Amount of drug remaining in a blood stream, Time between phone calls at a call center, Time until light bulb dies

Left-skewed 🡪 Mean less than Median 🡪 Grades as a percentage in many universities, Age of death, Asset price changes

Outliers

The outliers are the data points that are far away for the rest of the data points

This influences measures like the mean and standard deviation much more than measures associated with the five number summary.

**Working with the outliers**

When outliers are present we should consider the following points.

**1.** Noting they exist and the impact on summary statistics.

**2.** If typo - remove or fix

**3.** Understanding why they exist, and the impact on questions we are trying to answer about our data.

**4.** Reporting the 5 number summary values is often a better indication than measures like the mean and standard deviation when we have outliers.

**5.** Be careful in reporting. Know how to ask the right questions.

# Inferential Statistics

Inferential Statistics **is about using our collected data to draw conclusions to a larger population**. Performing inferential statistics well requires that we take a sample that accurately represents our population of interest.

A common way to collect data is via a survey. However, surveys may be extremely biased depending on the types of questions that are asked, and the way the questions are asked. This is a topic you should think about when tackling the first project.

We looked at specific examples that allowed us to identify the

1. **Population** - our entire group of interest.
2. **Parameter** - numeric summary about a population
3. **Sample** - subset of the population
4. **Statistic** numeric summary about a sample

Many career paths involving **Machine Learning** and **Artificial Intelligence** are aimed at using collected data to draw conclusions about entire populations at an individual level.

# Probability and Statistics

In probability we make prediction about future events based on models or causes that we assume, In statistics we analysis the data from the past events to infer what those models or causes could be. In one(probability) you are predicting data and in other(statistics) you are using data to predict

Python probability practice

# Sampling distributions and the Central Limit Theorem

We will be learning from data to draw conclusions rather than using probabilities to draw our conclusions

Inferential Statistics

Drawing conclusions regarding a parameter using on our statistics is known as **inference**

Sampling distribution is the distribution of statistics

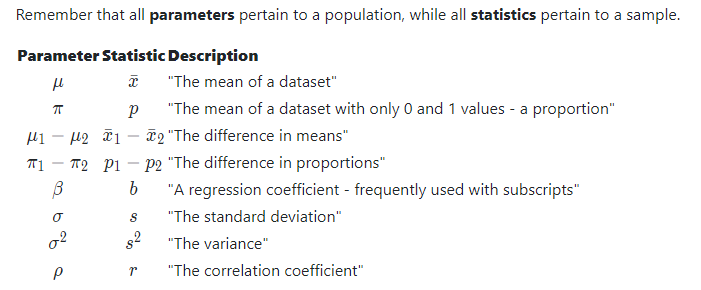
Check out the Jupyter notebook for sampling distributions

We found that for proportions (and also means, as proportions are just the mean of 1 and 0 values), the following characteristics hold.

* The sampling distribution is centered on the original parameter value.
* The sampling distribution decreases its variance depending on the sample size used. Specifically, the variance of the sampling distribution is equal to the variance of the original data divided by the sample size used. This is always true for the variance of a sample mean!

In notation, we say if we have a random variable **X,** with variance of **σ^2 ,** then the distribution X(bar) (the sampling distribution of the sample mean) has a variance of

**σ^2 / n**

****

Two Useful Theorems

Two important mathematical theorems for working with sampling distributions include:

1. **Law of Large Numbers**
2. **Central Limit Theorem**

The **Law of Large Numbers** says that **as our sample size increases, the sample mean gets closer to the population mean**

The **Central Limit Theorem** states that **with a large enough sample size the sampling distribution of the mean will be normally distributed**.

The **Central Limit Theorem** actually applies for these well known statistics:

1. Sample means ( *x*ˉ )
2. Sample proportions (*p*)
3. Difference in sample means ( *x*ˉ1 ​− *x*ˉ2 ​)
4. Difference in sample proportions (*p*1​−*p*2​)

And it applies for additional statistics, **but it doesn't apply for all statistics!**

​