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.

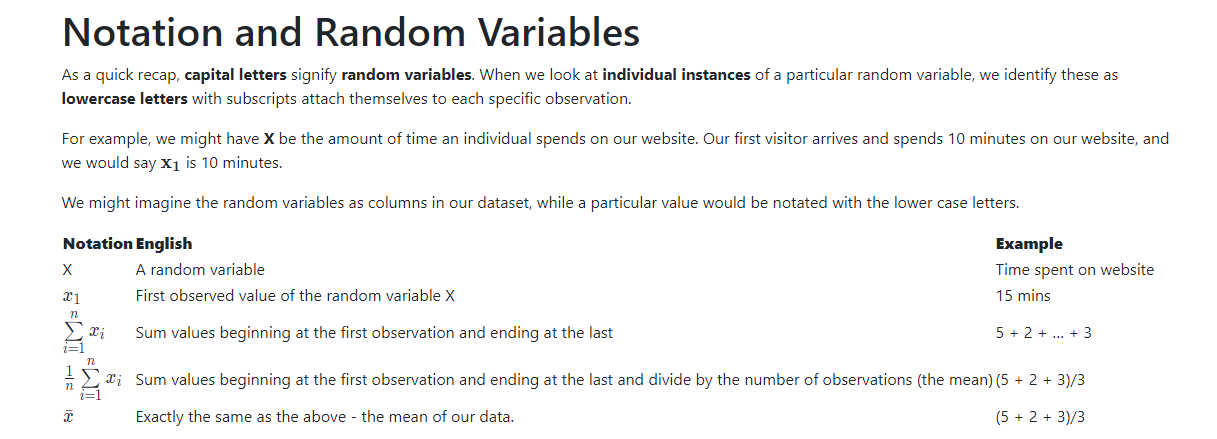
The Median

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, this could be any statistics

**Check out the Jupyter notebook for sampling distributions**

In 1st notebook we have an array of students 1,0,1,1 which represents if they drink coffee or not, We Calculate all summary statistics for population, the stimulate a sample draw using np.random.choice(students, 5) 🡪 This gives me sample of size 5 with replacement

We stimulate it 10000 times and find the summary statistics of bootstrap samples and we find the p\*(1-p) == variance of population

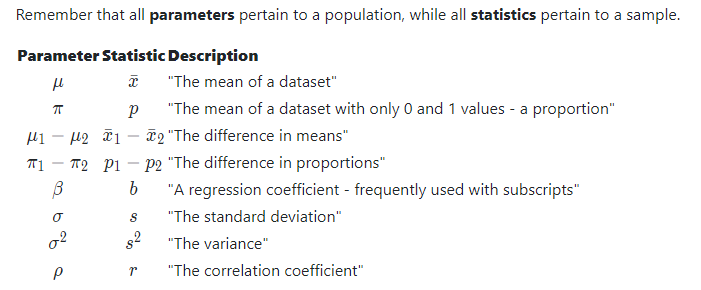
P\*(1-p) / n == Variance of bootstrap samples (where n is the sample size)

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**

How did we determine that the sample mean would estimate a population mean

The most common estimation techniques for finding good statistics are **Maximum Likelihood Estimation, Method of Moments Estimation** and **Bayesian Estimation**

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!**

The **Central Limit Theorem** applies to the sample mean of 100 draws from a right-skewed distribution. However, it did not apply to a sample size of 3 draws from this same distribution. (Jupyter Notebooks)

Bootstrapping

**Bootstrapping** is sampling with replacement.

Using Central limit Theorem we have to have large enough sample size and know which statistics Central limit Theorem applies to, Instead of relying on Theorem we can stimulate the sampling distribution (This introduces a technique known as bootstrapping)

The Application of bootstrap sampling goes beyond use cases here, Bootstrapping techniques have been used for leading machine learning algorithms like Random forest and Stochastic Gradient Boosting

The way we can draw inference about a population parameter by only performing repeated sampling within our existing sample, We actually gain confidence about were parameters likely to exist without having to collect any additional data

Bootstrapping – No more data needed to gain a better understanding of the parameter

When looking at inference techniques online you might find a lot of formulas and built in calculators for computing the final results for this technique, however this built in calculators hide the assumption and potential biases. Knowing about sampling distribution and bootstrapping we can take on not just the built in techniques but you extend this techniques to multitude of other situations

#### **Bootstrapping**

* **Bootstrapping** is a technique where we sample from a group with replacement.
* We can use bootstrapping to simulate the creation of sampling distribution, which you did many times in this lesson.
* By bootstrapping and then calculating repeated values of our statistics, we can gain an understanding of the sampling distribution of our statistics.

# Confidence Intervals

We can use bootstrapping and sampling distributions to build confidence intervals for our parameters of interest.

In the Jupyter notebook we worked to through an example of how to build a confidence interval using the sampling distribution of the statistic that best estimates your parameter of interest. In this case, we used a sample mean height to estimate the population mean height.

Confidence Interval Interpretation: **We are 95% confident, the population mean falls between the bounds that you find**.

In Jupyter notebooks you built a confidence interval for the difference of the average heights for coffee drinkers and non-coffee drinkers. The interval was built at a 95% confidence level, and since the difference did not contain zero, this suggested there was truly a difference in the average heights in the population of coffee drinkers as compared to non-coffee drinkers.

**Statistical vs Practical Difference**

Using confidence intervals and hypothesis testing, you are able to provide **statistical significance** in making decisions.

However, it is also important to take into consideration **practical significance** in making decisions. **Practical significance** takes into consideration other factors of your situation that might not be considered directly in the results of your hypothesis test or confidence interval. Constraints like **space**, **time**, or **money** are important in business decisions. However, they might not be accounted for directly in a statistical test.

**Traditional Confidence Interval Methods** – All this **formulas** have **underlying assumptions** that may or may not be true, Bootstrapping does not have these assumptions but **bootstrapping only assumes that the sample is representative of the population**

With large enough sample sizes this formulas should provide every similar results to that of bootstrapping

There are many traditional **hypothesis tests** (Which are linked to the way we built the confidence Interval) for proportions, Diff between props, Mean and Diff between means. Each of these hypothesis tests is linked to a corresponding confidence interval, but again the bootstrapping approach can be used in place of any of these! Simply by understanding what you would like to estimate, and simulating the sampling distribution for the statistic that best estimates that value.

**Traditional Methods vs Bootstrapping**

With large sample sizes, these end up looking very similar. With smaller sample sizes, using a traditional methods likely has assumptions that are not true of your interval. Small sample sizes are not ideal for bootstrapping methods though either, as they can lead to misleading results simply due to not accurately representing your entire population well.

Assuming you control all other items of your analysis:

1. Increasing your **sample size** will decrease the width of your **confidence interval.**
2. Increasing your confidence level (say 95% to 99%) will increase the width of your confidence interval.

You can compute:

1. The confidence interval **width** as the difference between your upper and lower bounds of your confidence interval.
2. The **margin of error** is half the confidence interval width, and the value that you add and subtract from your sample estimate to achieve your confidence interval final results.

**Confidence intervals** take an aggregate approach towards the conclusions made based on data, as these tests are aimed at **understanding population parameters** (which are aggregate population values).

Alternatively, **machine learning** techniques take an individual approach towards making conclusions, as they attempt to **predict an outcome for each specific data point.**

# Hypothesis Testing

Hypothesis testing and Confidence intervals allows us to use only sample data to draw conclusions about the entire population

A few rules for setting up null and alternative hypotheses:

1. The *H*0​ is true before you collect any data.
2. The *H*0​ usually states there is no effect or that two groups are equal.
3. The *H*0​ and *H*1​ are competing, non-overlapping hypotheses.
4. *H*1​ is what we would like to prove to be true.
5. *H*0​ contains an equal sign of some kind - either =, ≤, or ≥.
6. *H*1​ contains the opposition of the null - either ​=, >, or <.

**"Innocent until proven guilty"** is one that suggests the following hypotheses are true:

***H*0​:** Innocent

***H*1​:** Guilty

Tests of Errors

**Type I errors** have the following features:

1. You should set up your null and alternative hypotheses, so that the worse of your errors is the type I error.
2. They are denoted by the symbol *α*.
3. The definition of a type I error is: **Deciding the alternative (*H*1​) is true, when actually (*H*0​) is true.**
4. Type I errors are often called **false positives**.

### **Type II Errors**

1. They are denoted by the symbol *β*.
2. The definition of a type II error is: **Deciding the null (*H*0​) is true, when actually (*H*1​) is true.**
3. Type II errors are often called **false negatives**.

Common Types of Hypothesis Tests

Common hypothesis tests include:

1. Testing a population mean [(One sample t-test)](http://sites.utexas.edu/sos/guided/inferential/numeric/claim/one-sample-t/).
2. Testing the difference in means [(Two sample t-test)](https://www.isixsigma.com/tools-templates/hypothesis-testing/making-sense-two-sample-t-test/)
3. Testing the difference before and after some treatment on the same individual [(Paired t-test)](http://www.statstutor.ac.uk/resources/uploaded/paired-t-test.pdf)
4. Testing a population proportion [(One sample z-test)](http://stattrek.com/statistics/dictionary.aspx?definition=one-sample%20z-test)
5. Testing the difference between population proportions [(Two sample z-test)](https://onlinecourses.science.psu.edu/stat414/node/268)

You are always performing **hypothesis tests on population parameters,** never on **statistics.** Statistics are values that you already have from the data, so it does not make sense to perform hypothesis tests on these values.

**There are literally 100s of different hypothesis tests!** However, instead of memorizing how to perform all of these tests, you can find the statistic(s) that best estimates the parameter(s) you want to estimate, you can **bootstrap to simulate the sampling distribution.** Then you can use your sampling distribution to assist in choosing the appropriate hypothesis.

How do we choose between Hypotheses (Jupyter Notebooks)

1. Using your confidence interval, you can simply look at if the interval falls in the null hypothesis space or in the alternative hypothesis space to choose which hypothesis you believe to be true.
2. We Assume that the Null is true and stimulate what the sampling distribution would look like if it came from the Null Hypothesis. We std of samples as Standard deviation of the null distributions.

P-value

P value is **the probability of observing your statistic (or one more extreme in favor of the alternative) if the null hypothesis is true**.

1. Simulate the values of your statistic that are possible from the null.
2. Calculate the value of the statistic you actually obtained in your data.
3. Compare your statistic to the values from the null.
4. Calculate the proportion of null values that are considered **extreme** based on your alternative.

***H*0​:*μ* <= 5** p = (null\_vals > sample\_mean).mean()

***H*1​:*μ* > 5**

***H*0​:*μ* >= 5** p = (null\_vals < sample\_mean).mean()

***H*1​:*μ* < 5**

***H*0​:*μ* = 5**

*H*1​:*μ* != 5

P = (null\_vals < sample\_mean).mean() + (null\_vals > null\_mean + (null\_mean – sample\_mean)).mean()

The p-value is the probability of getting our statistic or a more extreme value if the null is true.

Therefore, small p-values suggest our null is not true. Rather, our statistic is likely to have come from a different distribution than the null.

When the p-value is large, we have evidence that our statistic was likely to come from the null hypothesis. Therefore, we do not have evidence to reject the null.

By comparing our p-value to our type I error threshold (*α*), we can make our decision about which hypothesis we will choose.

**pval ≤ *α* ⇒ Reject *H*0​**

**pval > *α* ⇒ Fail to Reject *H*0​**

The word **accept** is one that **is avoided** when making statements regarding the **null and alternative.** You are not stating that one of the hypotheses is true. Rather, you are making a decision based on the **likelihood** of your data coming from the null hypothesis with regard to your type I error threshold.

Therefore, the wording used in conclusions of hypothesis testing includes: **We reject the null hypothesis** or **We fail to reject the null hypothesis**. This lends itself to the idea that you start with the null hypothesis true by default, and "choosing" the null at the end of the test would have been the choice even if no data were collected.

Things to consider when Hypothesis testing

One of the **most important aspects** of **interpreting any statistical results** (and one that is frequently overlooked) is assuring that your **sample** is **truly representative** of your **population of interest.**

Particularly in the way that data is collected today in the age of computers, **response bias** is so important to keep in mind.

**With large sample sizes,** hypothesis testing leads to even the smallest of findings as **statistically significant**. However, these findings might not be practically significant at all.

With larger sample sizes we can do better than estimating the parameter of the population

**Hypothesis testing** takes an **aggregate approach** towards the **conclusions** made based on data, as these tests are aimed at understanding population parameters (which are aggregate population values).

Alternatively, **machine learning** techniques take **an individual approach** towards making **conclusions,** as they attempt to predict an outcome for each specific data point.

Multiple Tests

When we perform **multiple tests** on same data, the chances that our result came from **null hypothesis increases** and depends on **alpha**

When performing more than one hypothesis test, your type I error compounds. In order to correct for this, a common technique is called the **Bonferroni**correction. This correction is **very conservative**, but says that your new type I error rate should be the error rate you actually want divided by the number of tests you are performing.

Therefore, if you would like to hold a type I error rate of 1% for each of 20 hypothesis tests, the **Bonferroni** corrected rate would be 0.01/20 = 0.0005. This would be the new rate you should use as your comparison to the p-value for each of the 20 tests to make your decision.

A two-sided hypothesis test (that is a test involving a != in the alternative) is the same in terms of the conclusions made as a confidence interval as long as:

1−*CI*= *α*

For example, a 95% confidence interval will draw the same conclusions as a hypothesis test with a type I error rate of 0.05 in terms of which hypothesis to choose, because:

1−0.95= 0.05

assuming that the alternative hypothesis is a two sided test.

A/B Testing in Jupyter notebooks

# Regression

Machine Learning

**Machine Learning** is frequently split into **supervised** and **unsupervised** learning. Regression, which you will be learning about in this lesson (and its extensions in later lessons), is an example of supervised machine learning.

**In supervised machine learning,** you are interested in predicting a label for your data. Commonly, you might want to predict fraud, customers that will buy a product, or home values in an area.

**In unsupervised machine learning,** you are interested in clustering data together that isn't already labeled.

Intro to Linear Regression

In simple linear regression, we compare two quantitative variables to one another.

The **response** variable is what you want to predict, while the **explanatory** variable is the variable you use to predict the response. A common way to visualize the relationship between two variables in linear regression is using a scatterplot.

The **response or dependent** variable is on **y axis** and **explanatory or independent** variable is on **x axis**

In simple linear regression we compare two variables of quantitative types

Scatter plots

Scatter plots are a common visual for comparing two quantitative variables. A common summary statistic that relates to a scatter plot is the **correlation coefficient** commonly denoted by **r**.

Though there are a [few different ways](http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/) to measure correlation between two variables, the most common way is with [Pearson's correlation coefficient](https://en.wikipedia.org/wiki/Pearson_correlation_coefficient). Pearson's correlation coefficient provides the:

1. Strength
2. Direction

of a **linear relationship**. [Spearman's Correlation Coefficient](https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient) does not measure linear relationships specifically, and it might be more appropriate for certain cases of associating two variables.

Correlation coefficients provide a measure of the **strength** and **direction** of a **linear** relationship.

0.7≤∣*r*∣≤1.0 🡪 Strong     0.3≤∣*r*∣<0.7 🡪 Moderate           0≤∣*r*∣<0.3 🡪 Weak

What Defines a Line

A line is commonly identified by an **intercept** and a **slope**.

The **intercept** is defined as **the predicted value of the response when the x-variable is zero**.

The **slope** is defined as **the predicted change in the response for every one unit increase in the x-variable**.

Linear regression line 🡪 *y*^​=*b*0​+*b*1​*x*1​

Where y^ is the predicted value

bo is intercept, b1 is slope

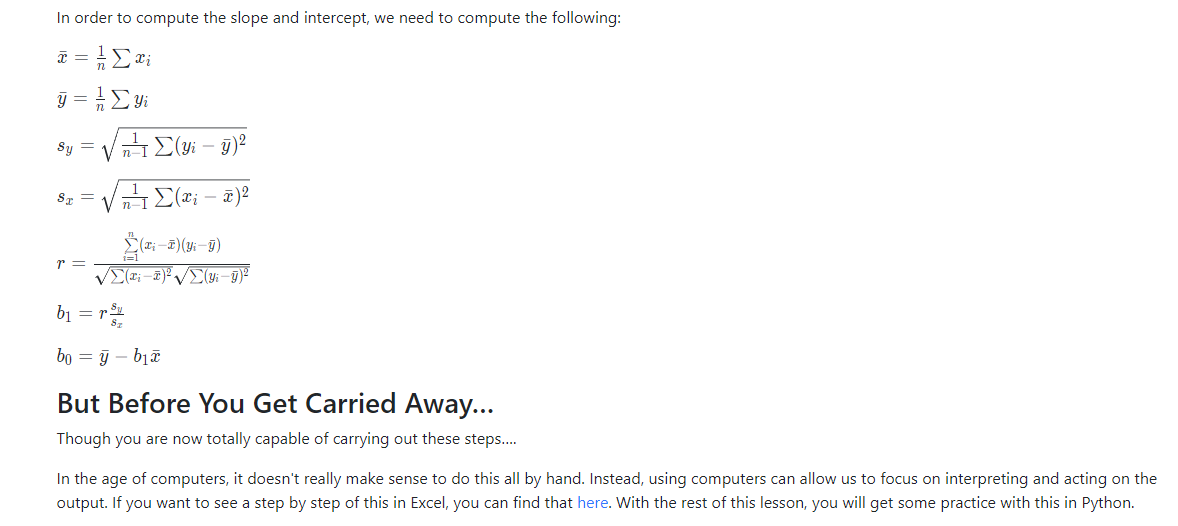
x1 is the explanatory variable

y is an actual response value for a data point in our dataset

Fitting A Regression Line

The main algorithm used to find the best fit line is called the **least squares algorithm,** which finds the line that minimizes

There are other ways we might choose a "best" line, but this algorithm tends to do a good job in many scenarios.



Jupyter notebooks

​