Intro to Statistics with Matplotlib

The Data Bootcamp | January 11, 2018

Central Tendency

- The **MODE** of a data set is the most frequently occurring element.
    a. For example, in a list like `[1, 1, 2]`, 1 would be the mode.
- The **MEDIAN** of a data set is the middle element
    a. To find the median, we first sort the data, and select the middle element. In `[1, 2, 3]`, 2 is the median.
    b. For even-length datasets, we have *two* elements in the middle of the list.
        i. We generally return the average of the two elements as the median of such a list.
- The **MEAN** of a data set is what is commonly called the *average* of a data set.
    a. To calculate the mean, we sum all of the numbers in the data set, and divide by the length of the data set.

Variance

Variance of a data set is a single number that describes how "far apart" its values are.

Two types of variance:

Population:

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Sample:

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - X_{avg})^2}{n-1}$$

The difference between the two disappears for large populations…

Variance

Which to use?

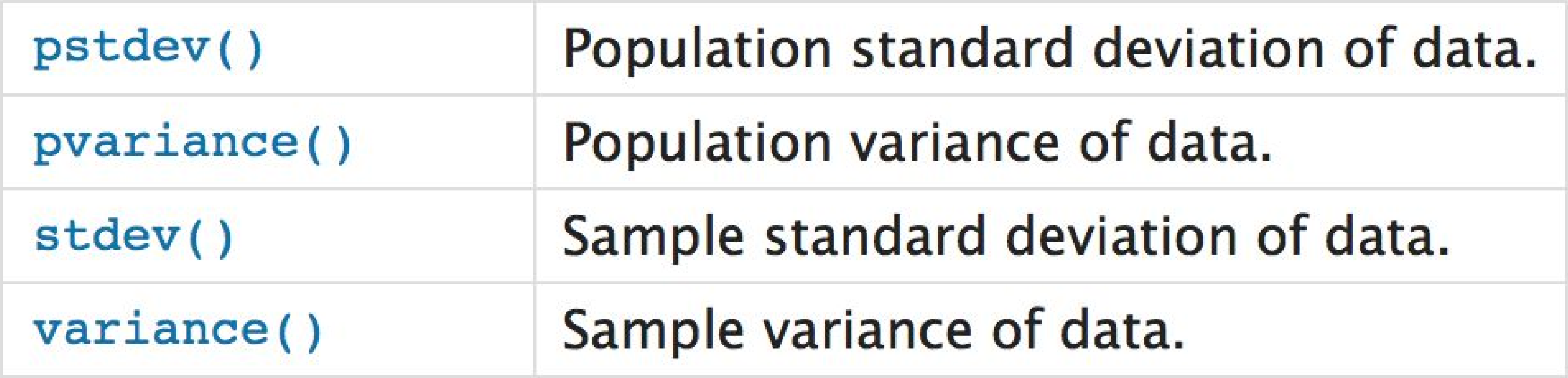
Probably Sample Variance...

When we are estimating the variance of a population from a sample, though, we encounter the problem that the deviations of the sample values from the mean of the sample are, on average, a little less than the deviations of those sample values from the (unknown) true population mean.

That results in a variance calculated from the sample being a little less than the true population variance. Using an n-1 divisor instead of n corrects for that underestimation.

Variance

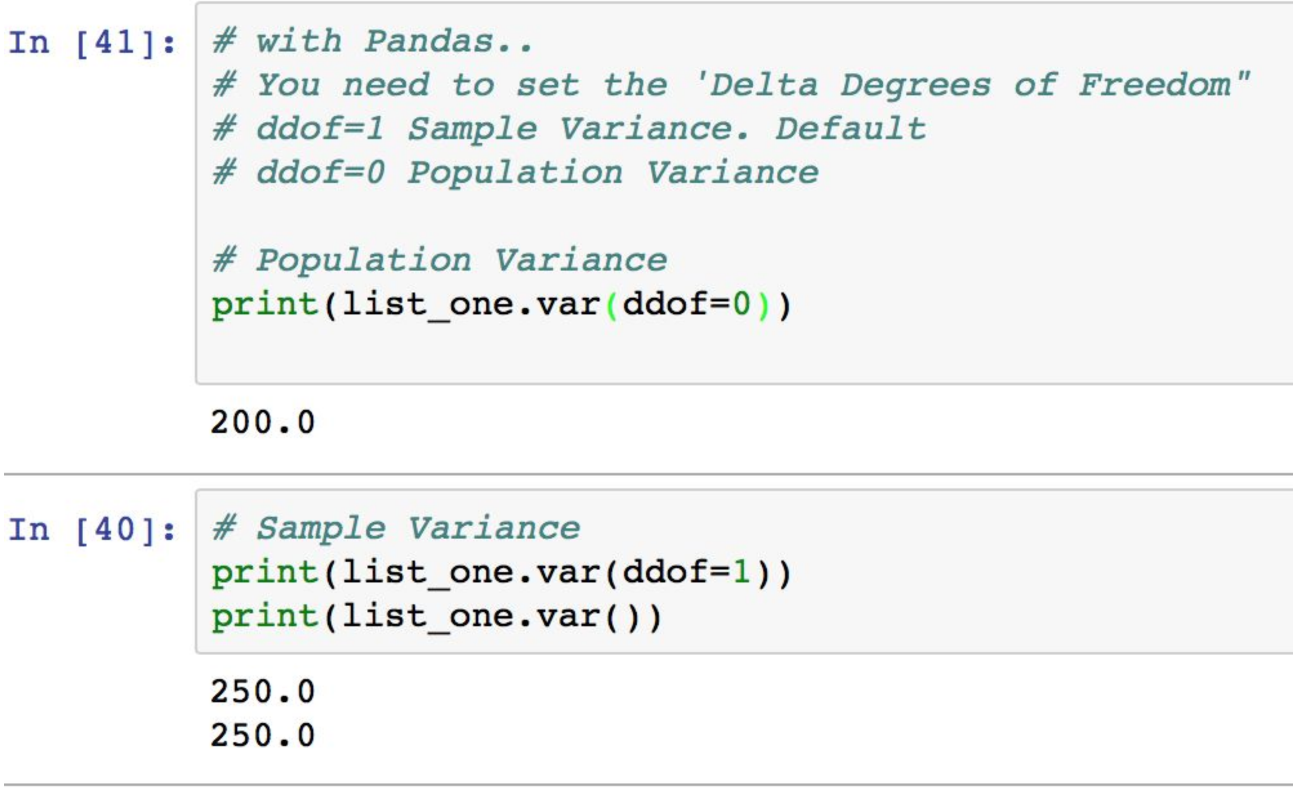
Import statistics



`pstdev()`	Population standard deviation of data.
`pvariance()`	Population variance of data.
`stdev()`	Sample standard deviation of data.
`variance()`	Sample variance of data.

Variance

Pandas



```
In [41]: # with Pandas..
         # You need to set the 'Delta Degrees of Freedom"
         # ddof=1 Sample Variance. Default
         # ddof=0 Population Variance

         # Population Variance
         print(list_one.var(ddof=0))

         200.0
```

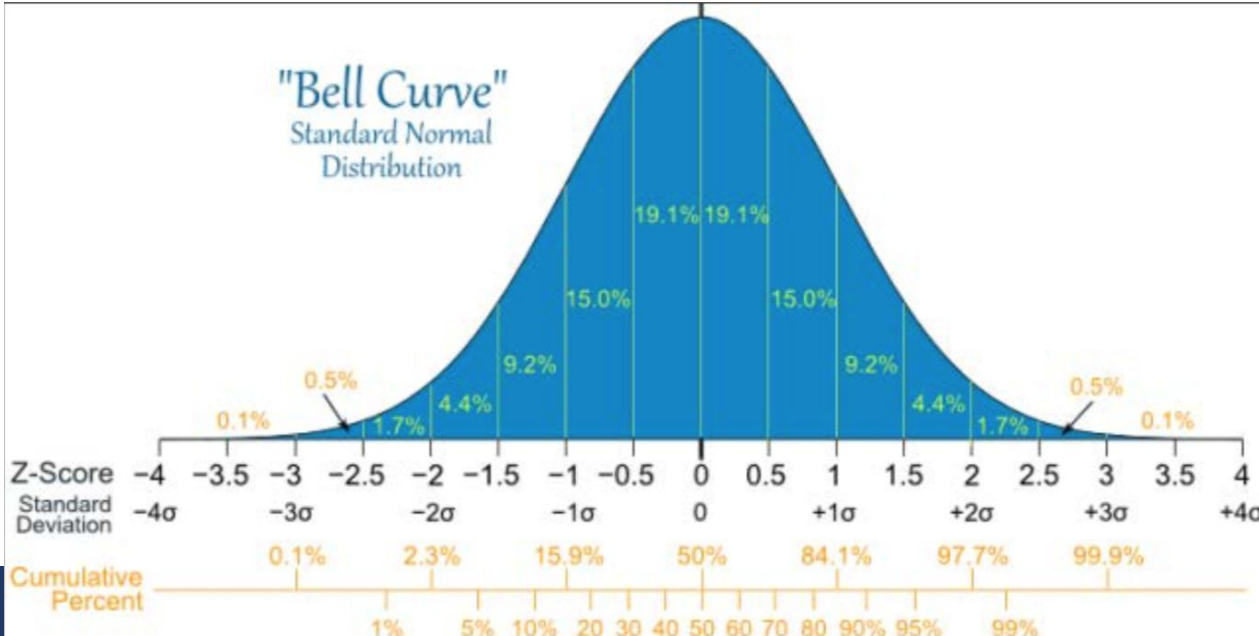
```
In [40]: # Sample Variance
         print(list_one.var(ddof=1))
         print(list_one.var())

         250.0
         250.0
```

Z-Score

- Variance and standard deviation are *single* numbers that describe the *whole* data set
- The **z-score** is a statistic that describes how far away from the mean any *single* number in the data set is.
- The z-score for a number in a data set tells us how many standard deviations away from the mean that number is.
- No need to memorize! The libraries we will work with — namely SciPy — have them built-in.

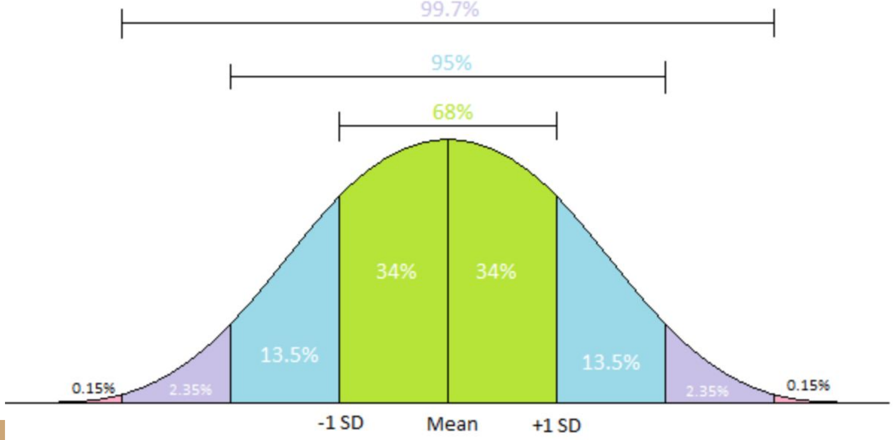
$$z\ score = \frac{(x - \mu)}{\sigma}$$



Standard Deviation

- The way we often use the standard deviation as a unit to describe how far individual numbers in a data set are away from the mean.
  a. Intuitively, the standard deviation tells us how far away from average any number in the data set is. For example, consider price data of: [30, 31, 31, 32, 32, 40, 41, 41, 1000].
  b. If you run the numbers, 1000 would have a z-score of 2.83. That is, 2.83 standard deviations away from average, meaning it is *far* above the mean

For comparison, 41 is only -0.3 standard deviations away from average, meaning it is *just a little below* the mean.

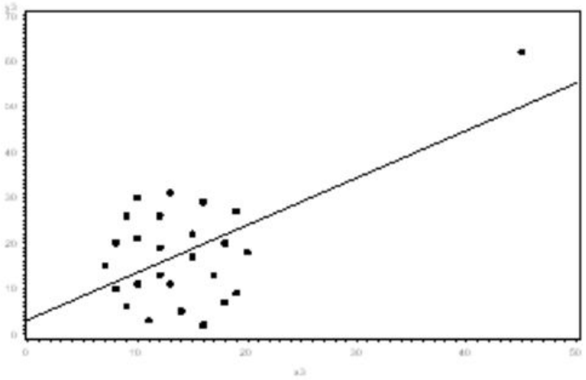


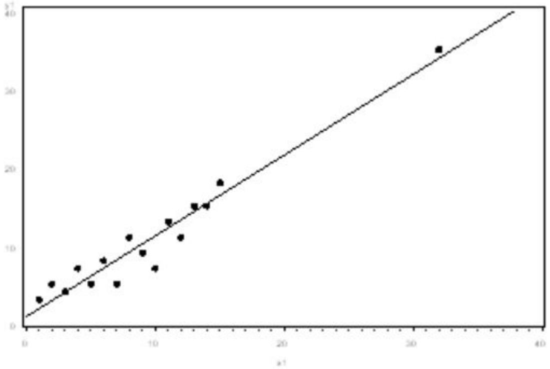
Examples

Outliers

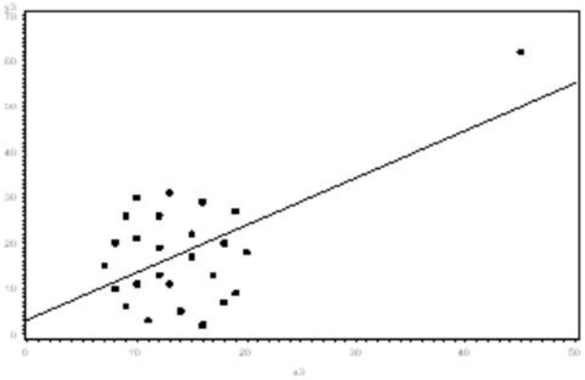
- Extreme values often do not describe the data…
- It is okay to remove outliers if any of the following are true:
  - The data is due to bad measurements.
  - If the outliers *create* trends that wouldn't exist without them, you *should* drop them.

Outliers - To remove or not to remove?

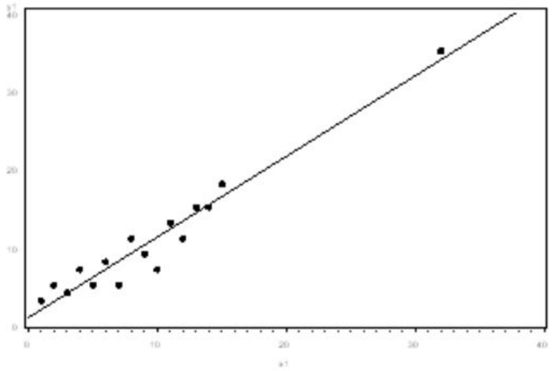




Outliers- To remove or not to remove?



If the outliers *create* trends that wouldn't exist without them, you *should* drop them



The outliers do *not* change your results. In this case, it is okay to drop them, but it is best to make a note of having done so.

http://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/

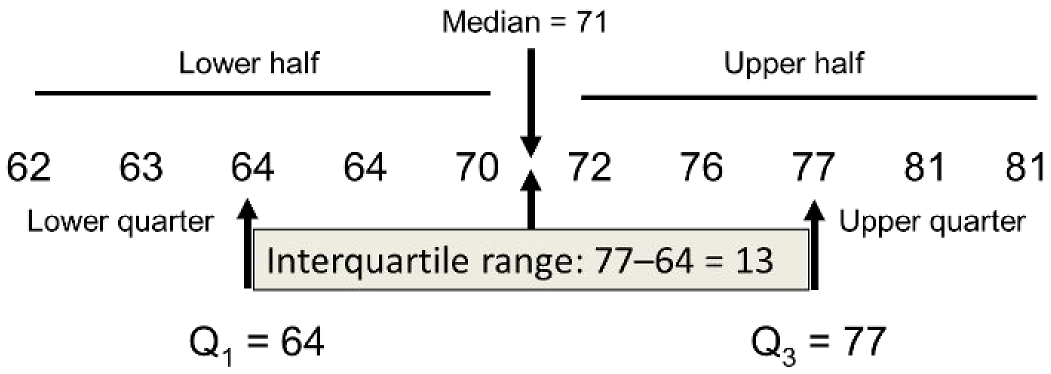
Tukey Method For Outlier Detection

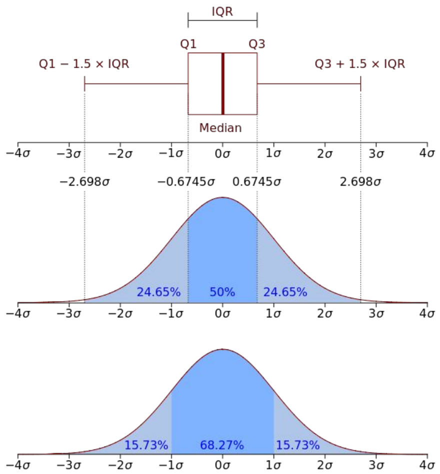
Rather than visually inspecting charts, we want the computer to detect outliers for us.

The Tukey Method was one of the first methods developed and still used today.

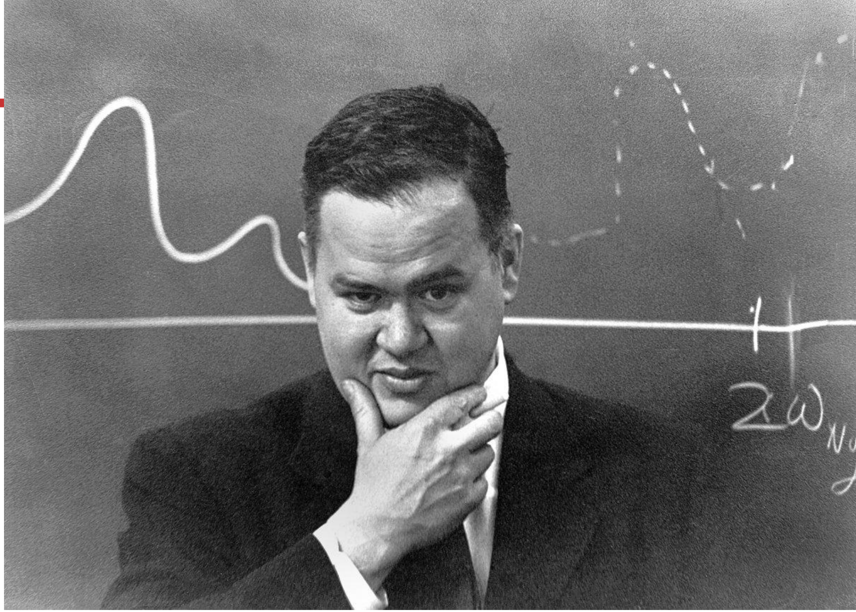
This method is a great general method, but is by no means the only way of detecting outliers. With time series data for example, spikes are considered outliers and are detected with different methods. The Tukey method works great with salary, housing data.

IQR - The interquartile range (IQR) is the length of the middle 50% of that interval of space.





Outliers



John Tukey defined :

q1-(1.5*iqr) and q3+(1.5*iqr) as “inner fences”
q1-(3*iqr) and q3+(3*iqr) as “outer fences”

the observations between an inner fence and its nearby outer fence as “outside”, and anything beyond outer fences as “far out”. The “outside” and “far out” observations can also be called possible outliers and probable outliers, respectively. This is a good starting point.

Sometimes… Outliers are the Focus

Candy Crush Saga publisher King brought in $2 billion in revenue in 2015 across all of its games, and all of that money came from just **2 percent its total players**. In Silicon Valley and Las Vegas parlance, these are called: ‘whales’

Approximately 738,000 people spent an average of $1,400 each.

What's even more startling is that 1% of players spent about $1.65 billion

Whole teams are focused on how to get a small number of users to spend big.

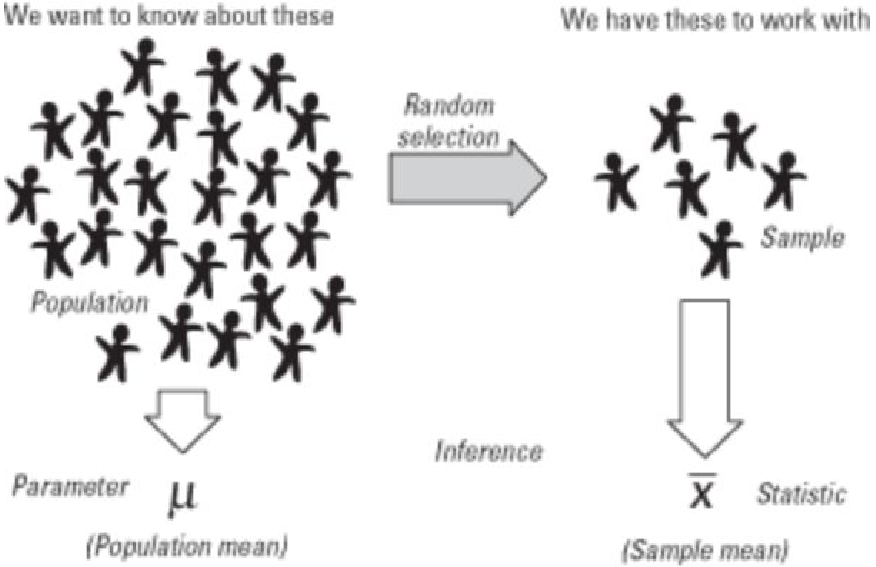
Focusing on outliers can be a valid approach in machine learning.

https://www.buzzfeed.com/blakemontgomery/heres-who-drops-the-most-cash-on-candy-crush-and-clash-of-cl?utm_term=.ypqn22n1m#.pyaallaGw

Examples

Standard Error

We will focus on the notions of standard error and how well measurements of a section of a population (e.g., voting habits of Americans in cities) represent that population as a whole (e.g., all Americans).



Standard Error

This process can lead to imperfect predictions….why?

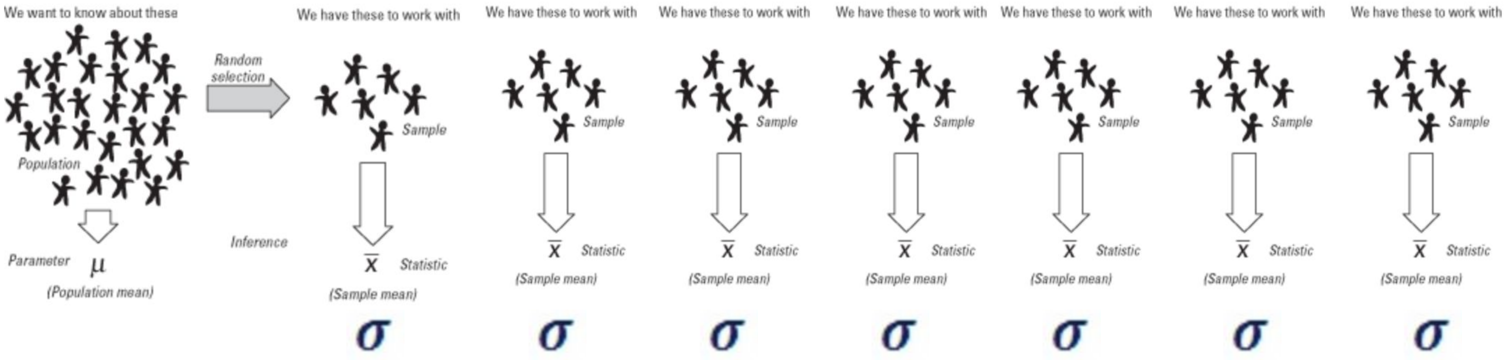
Standard Error

This process can lead to imperfect predictions….why?

- A small sample might not realistically represent a large population, and that even a large sample, if chosen poorly, might not describe the population at large.
- For example, if we poll only college students and people who live in cities, we will almost certainly make bad predictions, because this sample does not represent the entire population.

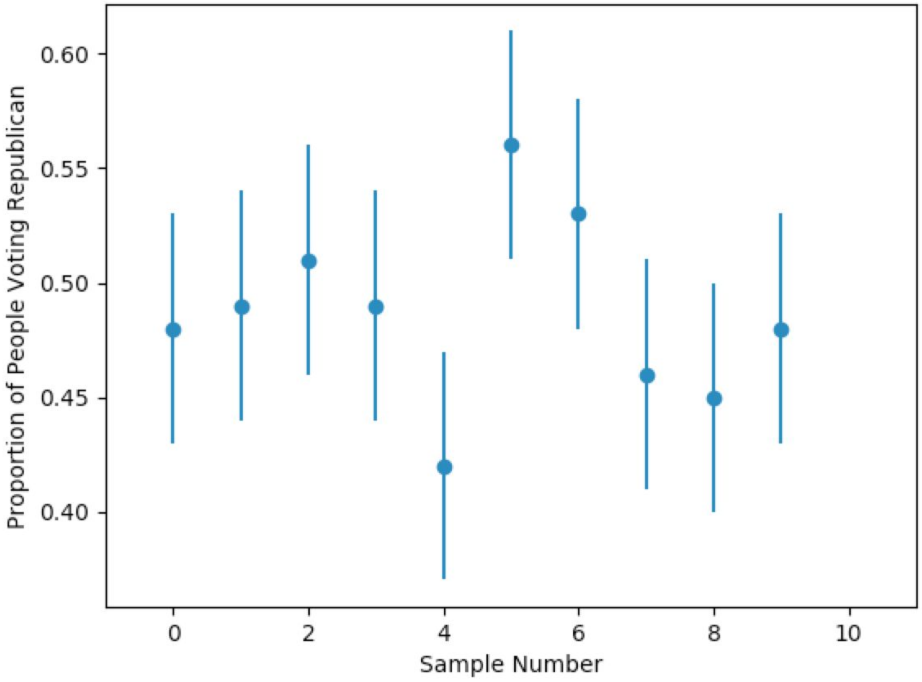
Do we trust our data?

- If we take multiple samples, the collection of samples *themselves* is a data set.
- If we have 100 samples, we create a list of the samples' standard deviations.
- We can use these numbers to calculate something called standard error, which is an estimate how well the samples represent the population.
- Each sample *standard error* describes how far its mean is from the **population's** "true" mean.
- There is a function in SciPy that does this for us.



SEM and Error Bars - Example

- **SEM:** Standard Error of the Mean.
- **Summary**: errorbars to provide a visual indicator as to how confident we were in the proximity of our sample means to the "true" population mean.



Error Bars

- If we plotted the errorbars of our data and found that most of them did not overlap, it would raise questions about the data.
- Suppose, for instance, that about half of the means had error bars that overlapped one another, and the other half had error bars that overlapped one another, but neither cluster's error bars overlapped the other's.
- In this case, we might expect that the two clusters *were not* randomly selected from the same population

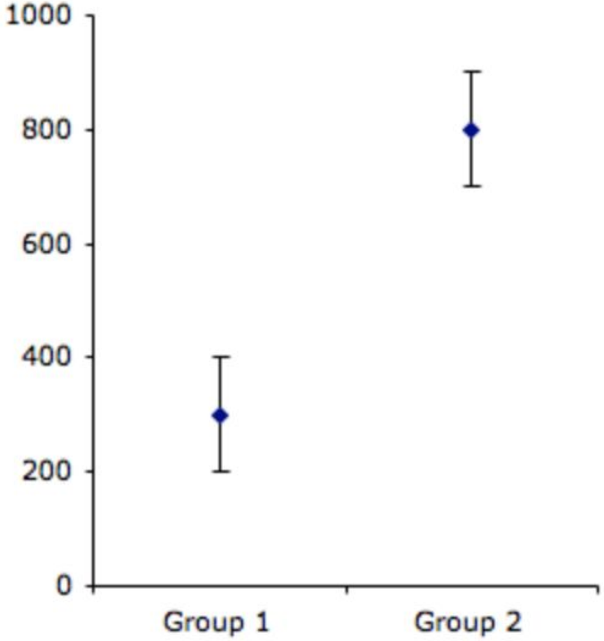
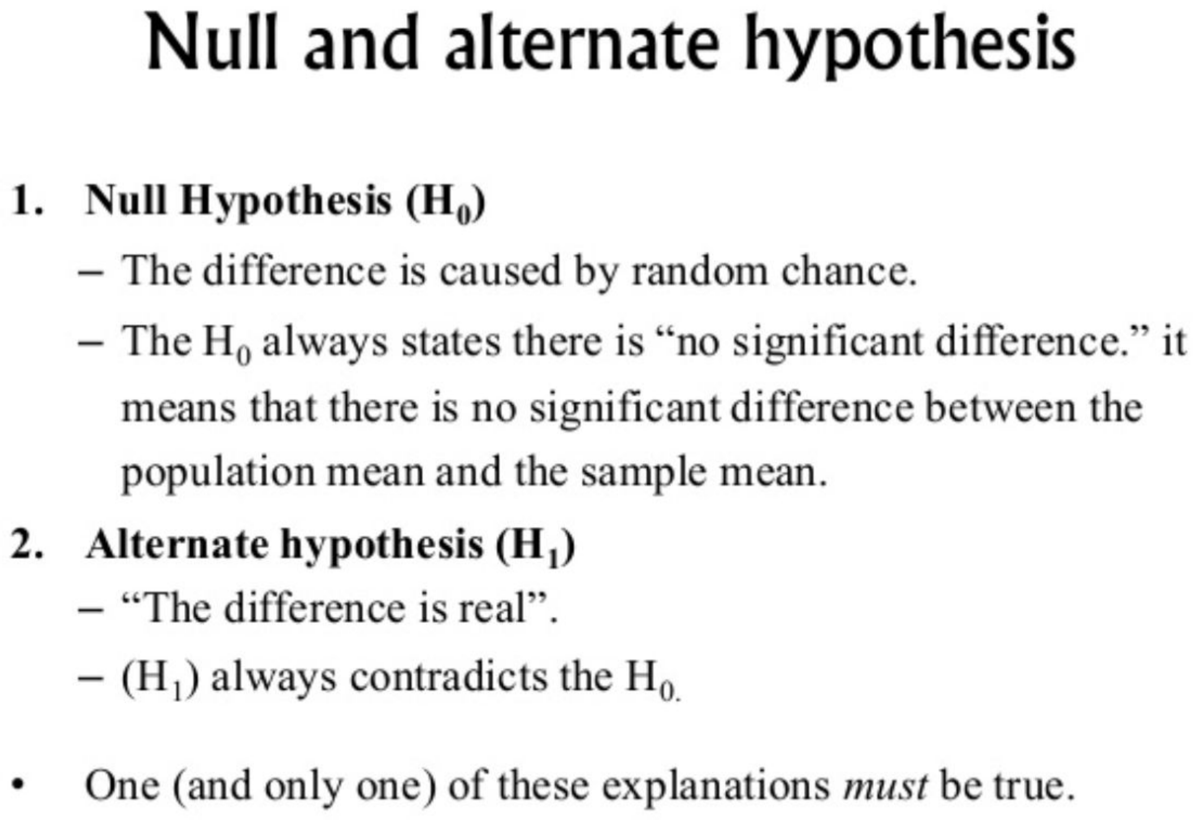


Figure 2: Mean reaction time (ms) and standard error for Group 1 (n=36) and Group 2 (n=34).

Examples

T-tests



Null and alternate hypothesis

1. **Null Hypothesis (H_0)**
   - The difference is caused by random chance.
   - The H_0 always states there is “no significant difference.” it means that there is no significant difference between the population mean and the sample mean.
2. **Alternate hypothesis (H_1)**
   - “The difference is real”.
   - (H_1) always contradicts the $H_{0.}$

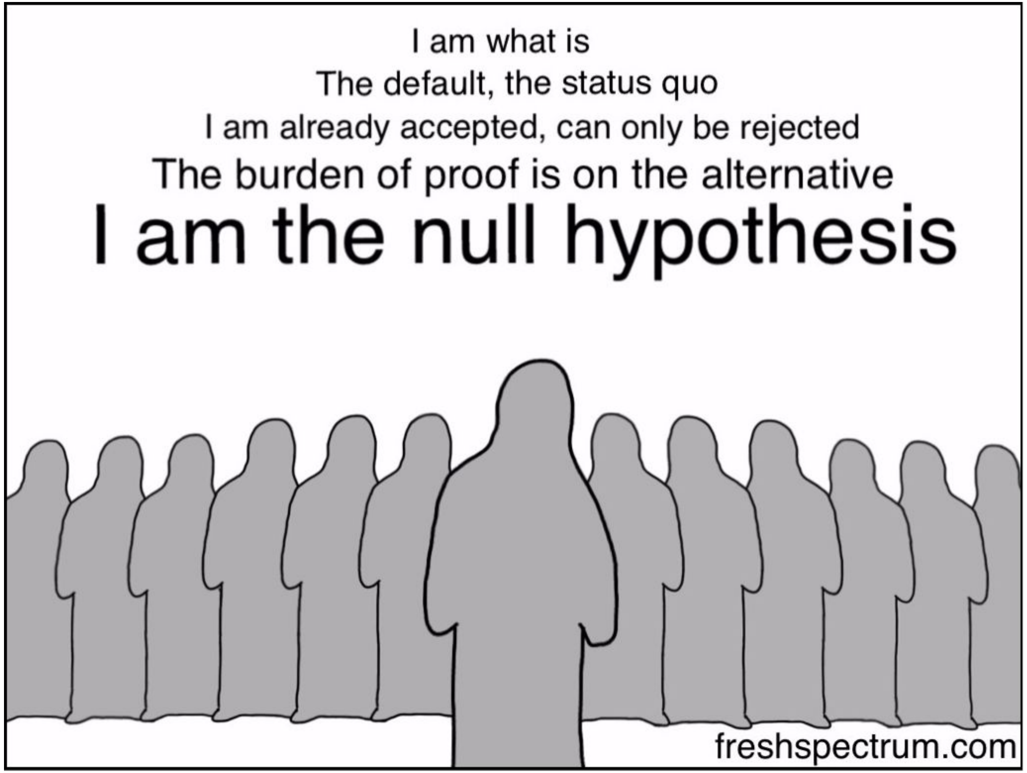
- One (and only one) of these explanations *must* be true.

Null Hypothesis

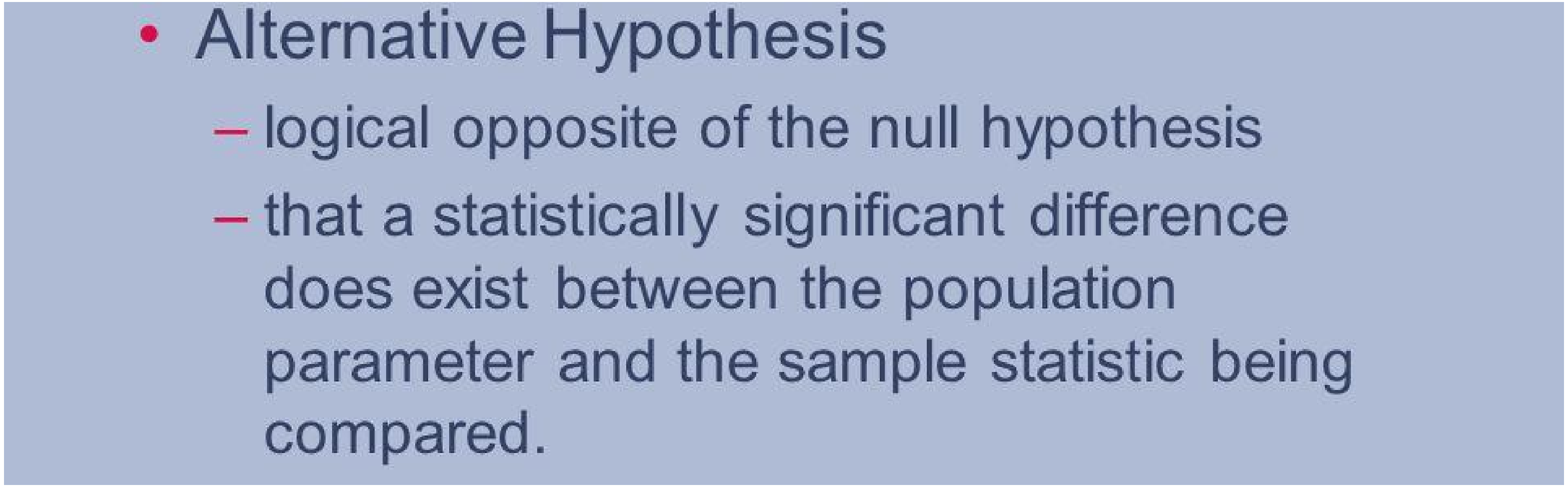
The exact form of the null hypothesis varies from one type test to another:

If you are testing whether groups differ, the null hypothesis states that the groups are the same.

For instance, if you wanted to test whether the average age of voters in your home state differs from the national average, the null hypothesis would be that there is no difference between the average ages.


I am what is
The default, the status quo
I am already accepted, can only be rejected
The burden of proof is on the alternative
I am the null hypothesis
freshspectrum.com

Alternative Hypothesis



- Alternative Hypothesis
  - logical opposite of the null hypothesis
  - that a statistically significant difference does exist between the population parameter and the sample statistic being compared.

The exact form of the alternative hypothesis will depend on the specific test you are carrying out. Continuing with the previous example, the alternative hypothesis would be that the average age of voters in your state does in fact differ from the national average.

Null and alternative hypothesis

- Memory score for 1,000 subjects. They eat 40 grams of spinach daily for 90 days.
- Null hypothesis (H0) is that the mean memory score for this sample of 1,000 subjects will not differ from the control group of 1,000 subjects who consumed no spinach during the same time period.

The null hypothesis is -nullifying- what we’re trying to prove with our experiment.

We focus on trying to prove or disprove the null hypothesis because we can calculate the probability that our results are due to chance.

There is no easy way to calculate the probability of the alternative hypothesis (H1) since any improvement in long-term memory could be due to other factors besides eating spinach.

Alpha and p-values

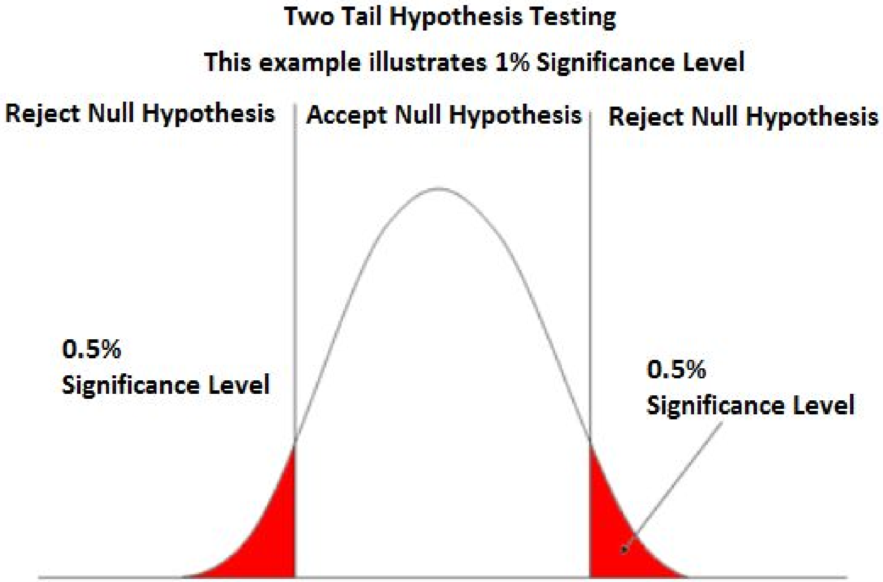
The p-value is the probability of the result we observe by assuming that the null hypothesis is true.

The p-value can also be thought of as the probability of obtaining a test statistic as extreme as or more extreme than the actual obtained test statistic, given that the null hypothesis is true.

The significance level, or alpha value is the threshold value against which we compare p-values. This gives us a cut-off point in order to accept or reject the null hypothesis. It is a measure of how extreme the results we observe must be in order to reject the null hypothesis of our experiment. The most commonly used values of alpha are 0.05 or 0.01.

Significance Level

Once you have the null and alternative hypothesis in hand, you choose an alpha, the significance level. The significance level is a probability threshold that determines when you reject the null hypothesis.



After carrying out a test, if the probability of getting a result as extreme as the one you observe due to chance is lower than the significance level, you **‘reject the null hypothesis.’**

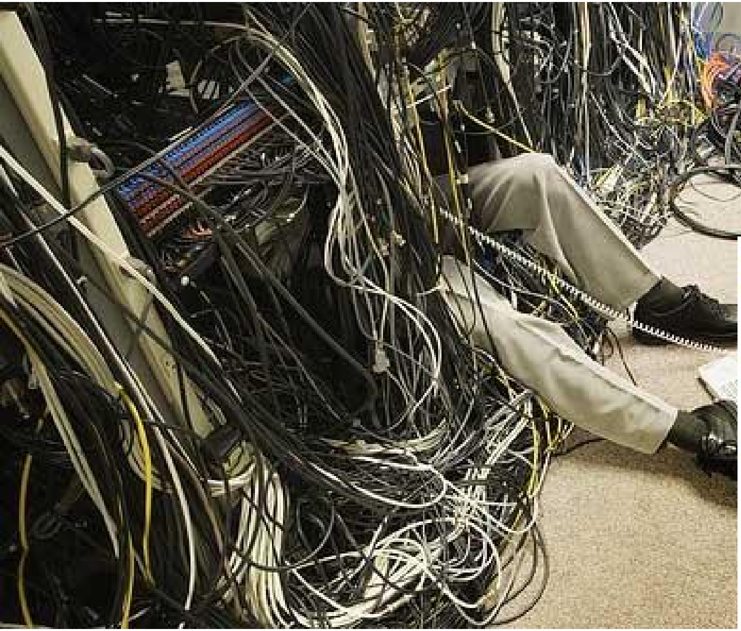
Dangers with p-values

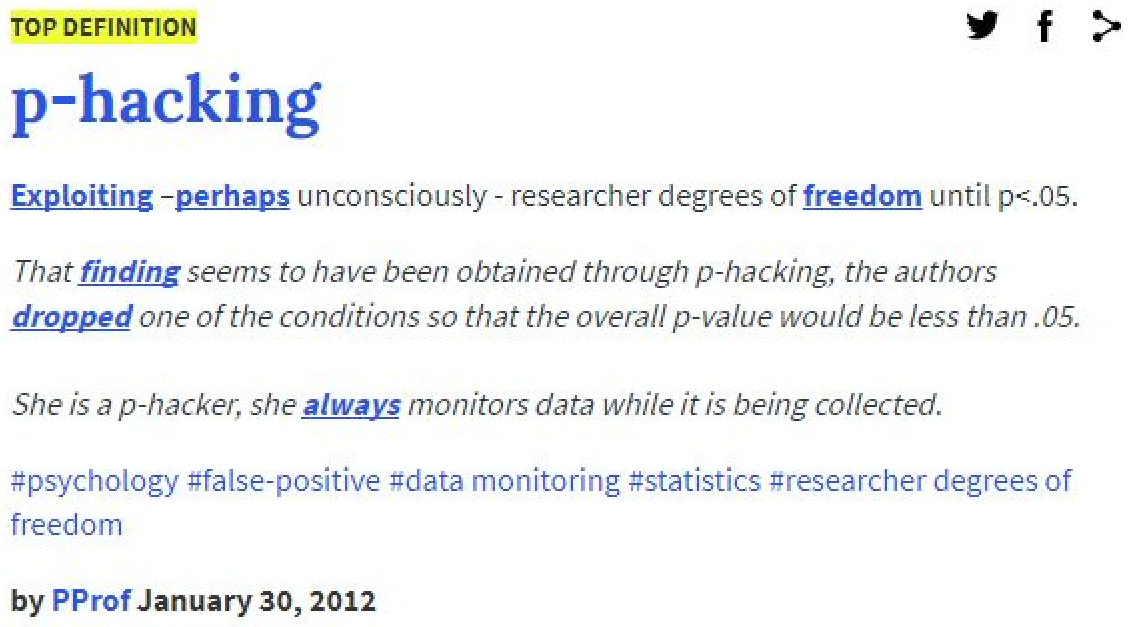
The seemingly arbitrary values of alpha in usage are one of the shortcomings of this methodology, and there are many questions concerning this approach.

The following article in the Nature journal highlights some of the problems: http://www.nature.com/news/scientific-method-statistical-errors-1.14700.

“Science progresses one funeral at a time. The future depends on some graduate student who is deeply suspicious of everything I have said.” - Geoff Hinton, grandfather of deep learning September 15, 2017

p-hacking





TOP DEFINITION

p-hacking

Exploiting –**perhaps** unconsciously - researcher degrees of **freedom** until p<.05.

*That **finding** seems to have been obtained through p-hacking, the authors **dropped** one of the conditions so that the overall p-value would be less than .05.*

*She is a p-hacker, she **always** monitors data while it is being collected.*

#psychology #false-positive #data monitoring #statistics #researcher degrees of freedom

by PProf January 30, 2012

T-Test

The T-test is a statistical test used to determine whether a numeric data sample differs significantly from the population or whether two samples differ from one another.

- T-Test devised by William Gosset to monitor the quality of Guinness Stout. He found that 80% of the time, the measurement from just two observations was accurate enough.

Different variants for different questions...

- **One Sample T-Test:** I know one mean, is the second mean the same as the first?
- **Two Sample T-Test:** Are the means of two normally distributed populations equal?

One Sample T-Test Example

Is known population mean, μ, different from the mean of a sample population?

Example:

We know μ=0.293 is the error rate the decision tree algorithm ID3 has on categorizing a given data set.

I trained neural nets to categorize the same data set and the mean error rate was 0.227

Are neural nets better on this data set? Or was that a happy accident?

I’d use a one-sample t-test to find out.

One Sample T-Test Example

Example:

We know μ=0.293 is the error rate the decision tree algorithm ID3 has on categorizing a given data set.

I trained neural nets to categorize the same data set and the mean error rate was 0.227

Are neural nets better on this data set? Or was that a happy accident?

What is the Null hypothesis?

One Sample T-Test Example

Example:

We know μ=0.293 is the error rate the decision tree algorithm ID3 has on categorizing a given data set.

I trained neural nets to categorize the same data set and the mean error rate was 0.227

Are neural nets better on this data set? Or was that a happy accident?

Null hypothesis: There is no significant difference between the sample mean and the population mean.

Neural nets perform no better than ID3 on this data

One Sample T-Test Example

Example:

We know μ=0.293 is the error rate the decision tree algorithm ID3 has on categorizing a given data set.

I trained neural nets to categorize the same data set and the mean error rate was 0.227

Are neural nets better on this data set? Or was that a happy accident?

What is the Alternate hypothesis?

One Sample T-Test Example

Example:

We know μ=0.293 is the error rate the decision tree algorithm ID3 has on categorizing a given data set.

I trained neural nets to categorize the same data set and the mean error rate was 0.227

Are neural nets better on this data set? Or was that a happy accident?

Alternate hypothesis: There is a significant difference between the sample mean and the population mean.

Neural nets DO perform than ID3 on this data

Two Sample T-Test

A two-sample t-test investigates whether the means of two independent data samples differ from one another. In a two-sample test, the null hypothesis is that the means of both groups are the same. Unlike the one sample-test where we test against a known population parameter, the two sample test only involves sample means.

Example Two-Sample T-Test

Example:

To motivate citizens to conserve gasoline, a non-profit is considering a national conservation campaign. Before doing so, they are going to conduct a local experiment. 12 families are randomly selected, and the amount of gasoline monitored one month before and after the advertising campaign.

What is the Null hypothesis?

Example Two-Sample T-Test

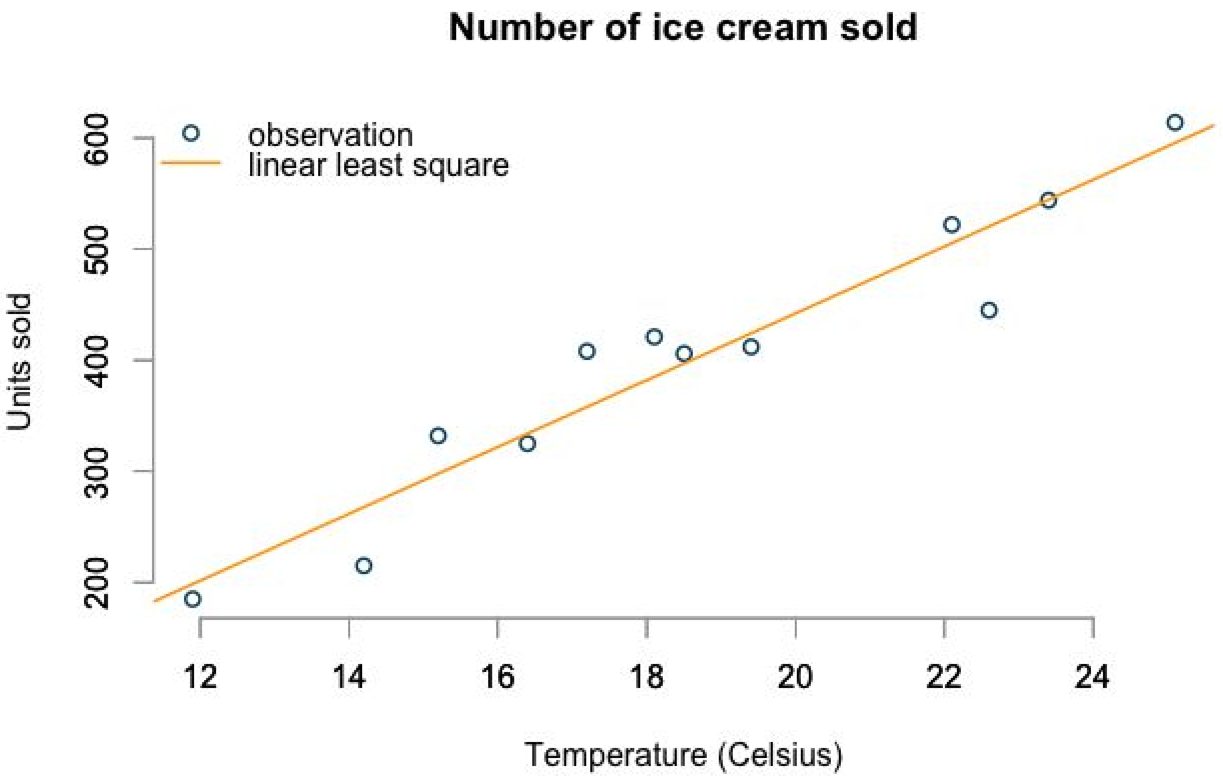
Example:

To motivate citizens to conserve gasoline, a non-profit is considering a national conservation campaign. Before doing so, they are going to conduct a local experiment. 12 families are randomly selected, and the amount of gasoline monitored one month before and after the advertising campaign.

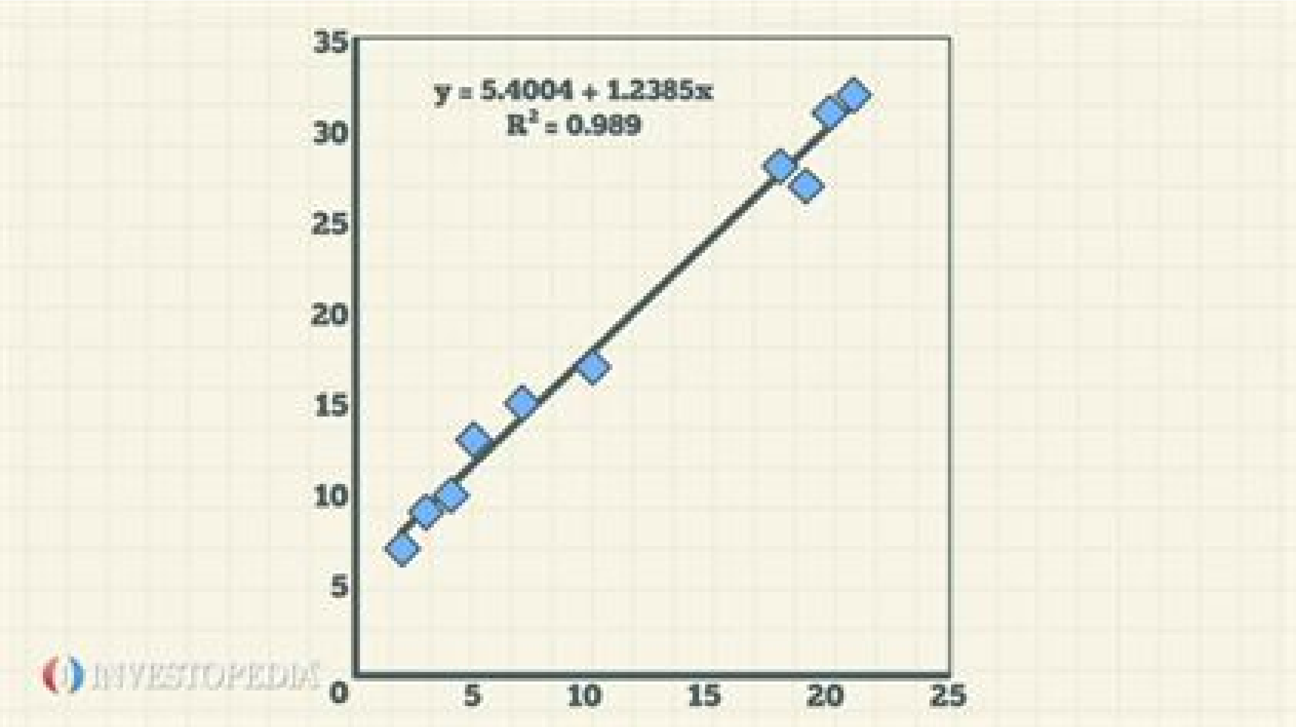
Null hypothesis. Conservation campaign has no effect.
Alternative hypothesis: Conservation campaign affects the amount of gasoline learned.

Examples

Linear Regression



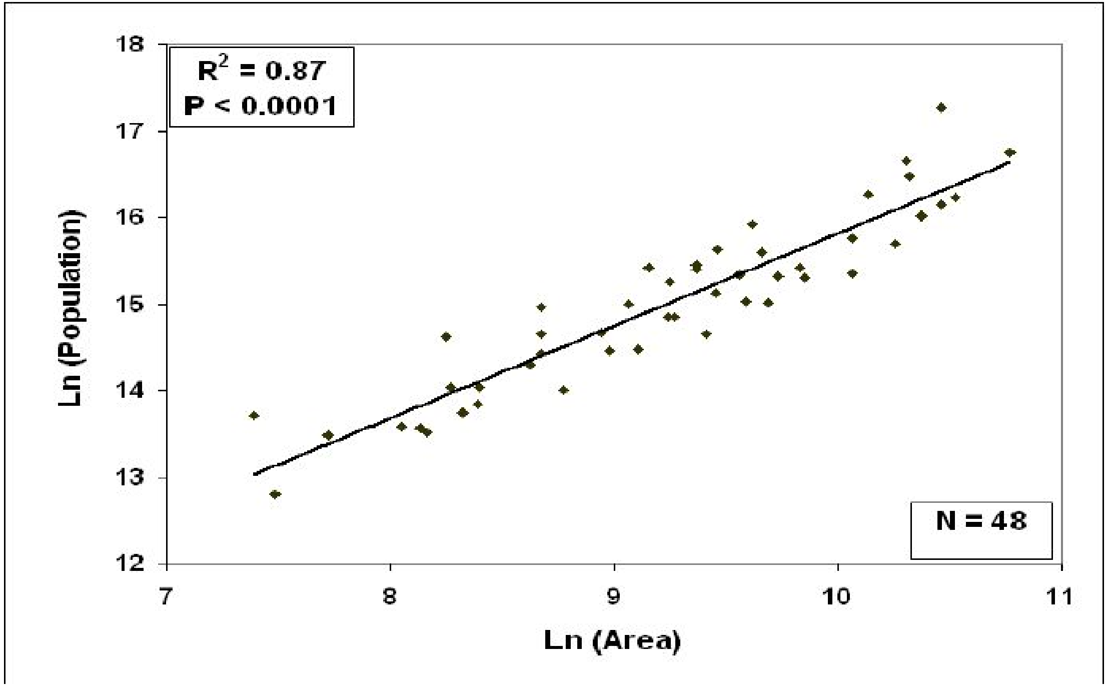
R-Squared



What Is R-squared?

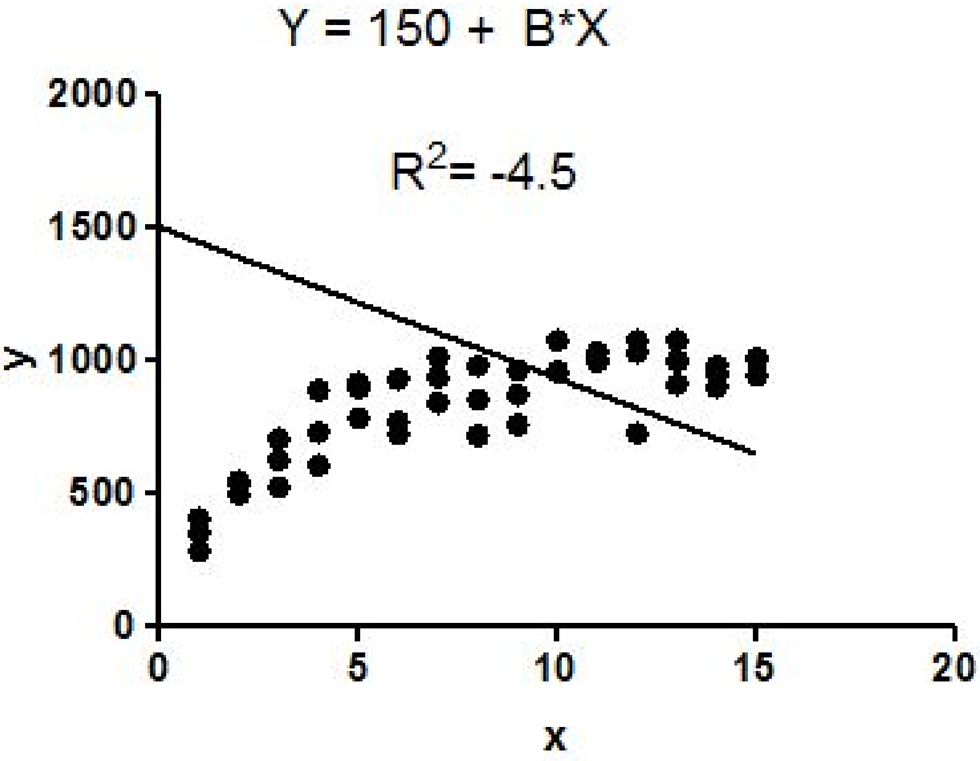
R-squared is a statistical measure of how close the data are to the fitted regression line.

R-Squared

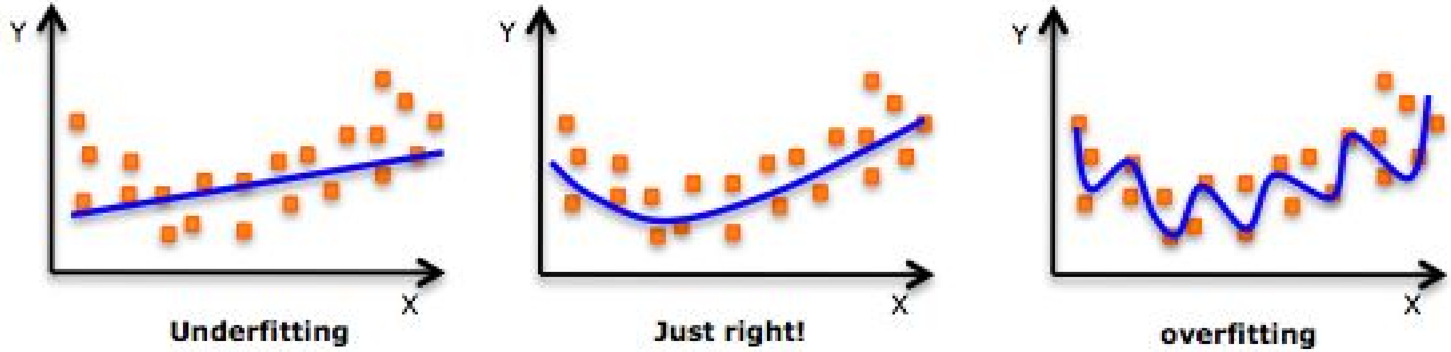


$R^2 = 0.87$
$P < 0.0001$
N = 48
Ln (Population)
18
17
16
15
14
13
12
7
8
9
10
11
Ln (Area)

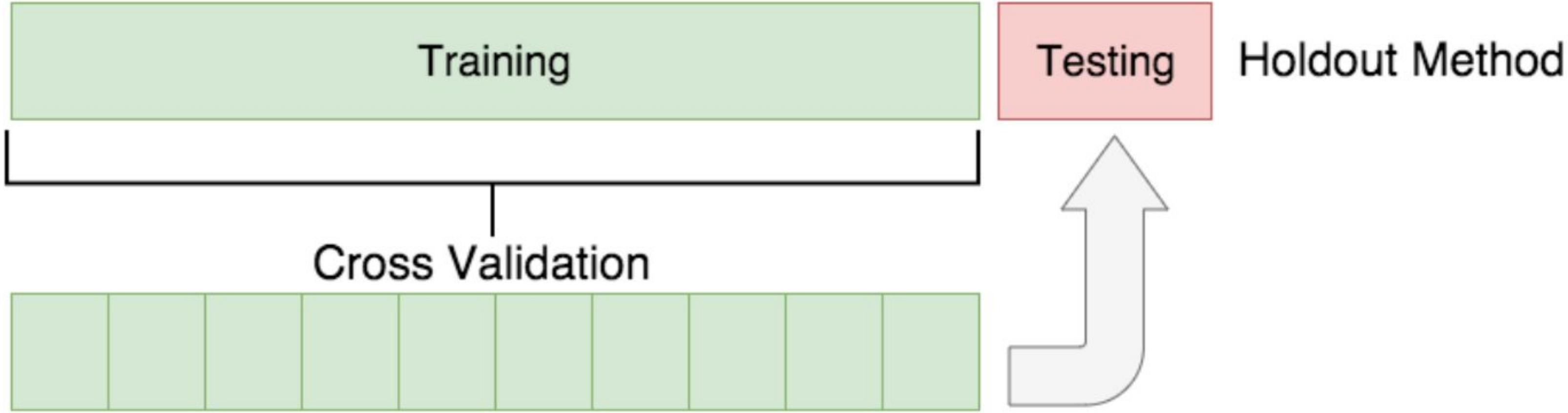
R-Squared



Overfitting



Test / Train Split. The Holdout Method



Many ways of doing this. We will focus on the Holdout Method today. K-fold cross validation is also very popular.