# Statistics for data analysts (part 1)


YAY STATISTICS

## Useful Statistics for data analysts with data questions

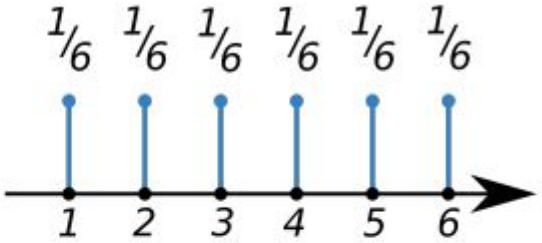
1. **Descriptive statistics -** ways to describe data
2. **Probability -** common misconceptions | conditional probability
3. **Hypothesis formulation -** starting with logical assumptions

\+ Using the p value in correlation analysis

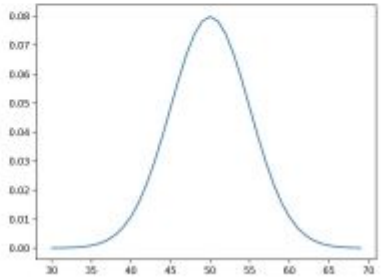
4. **Addressing uncertainty** - why confidence intervals are useful

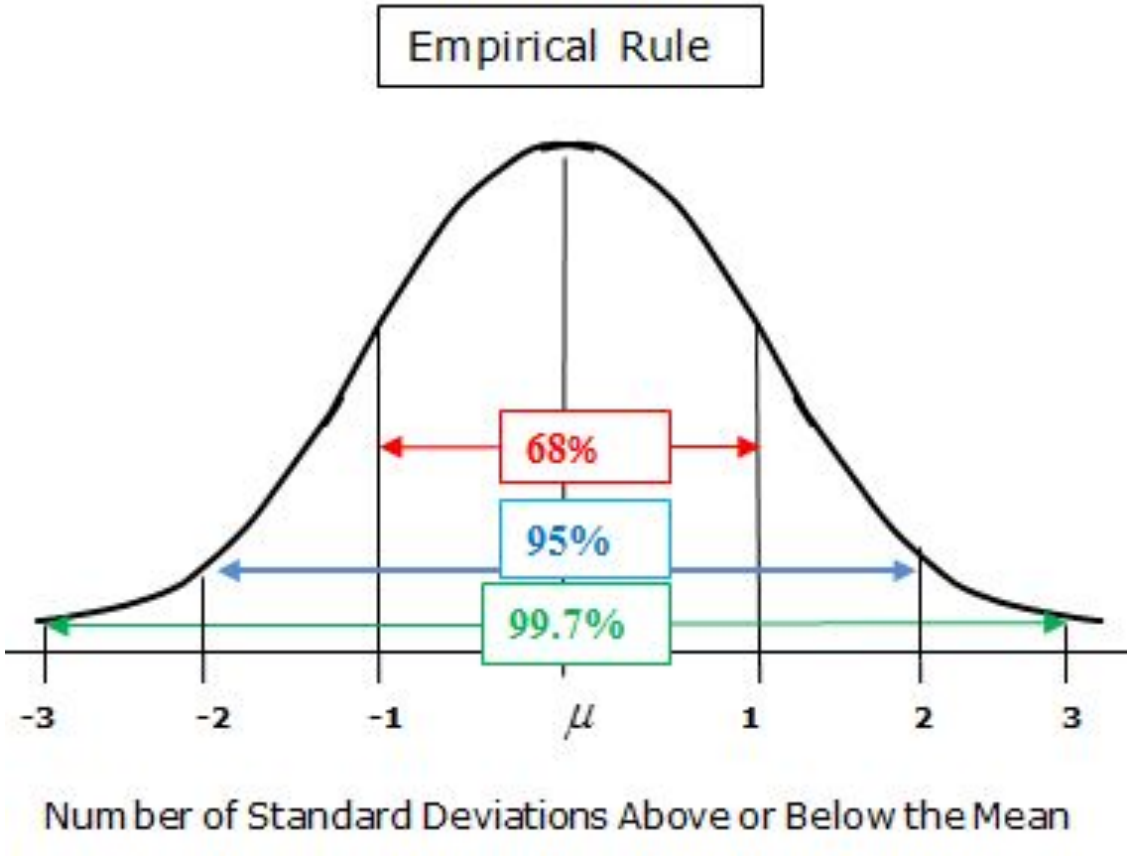
# Distributions by data type

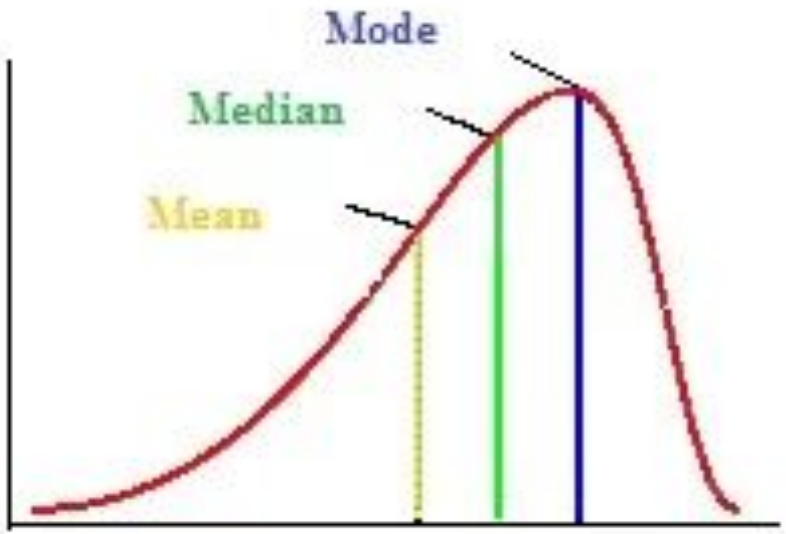
Discrete →



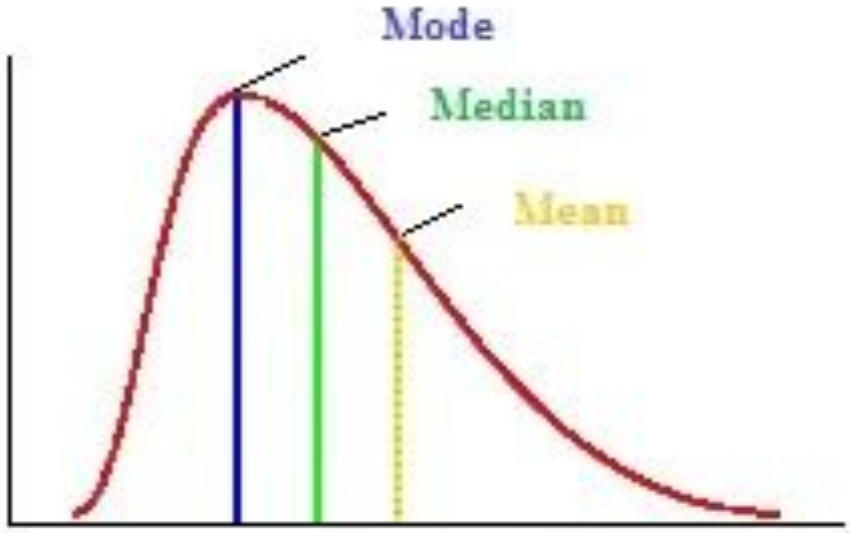

Continuous →



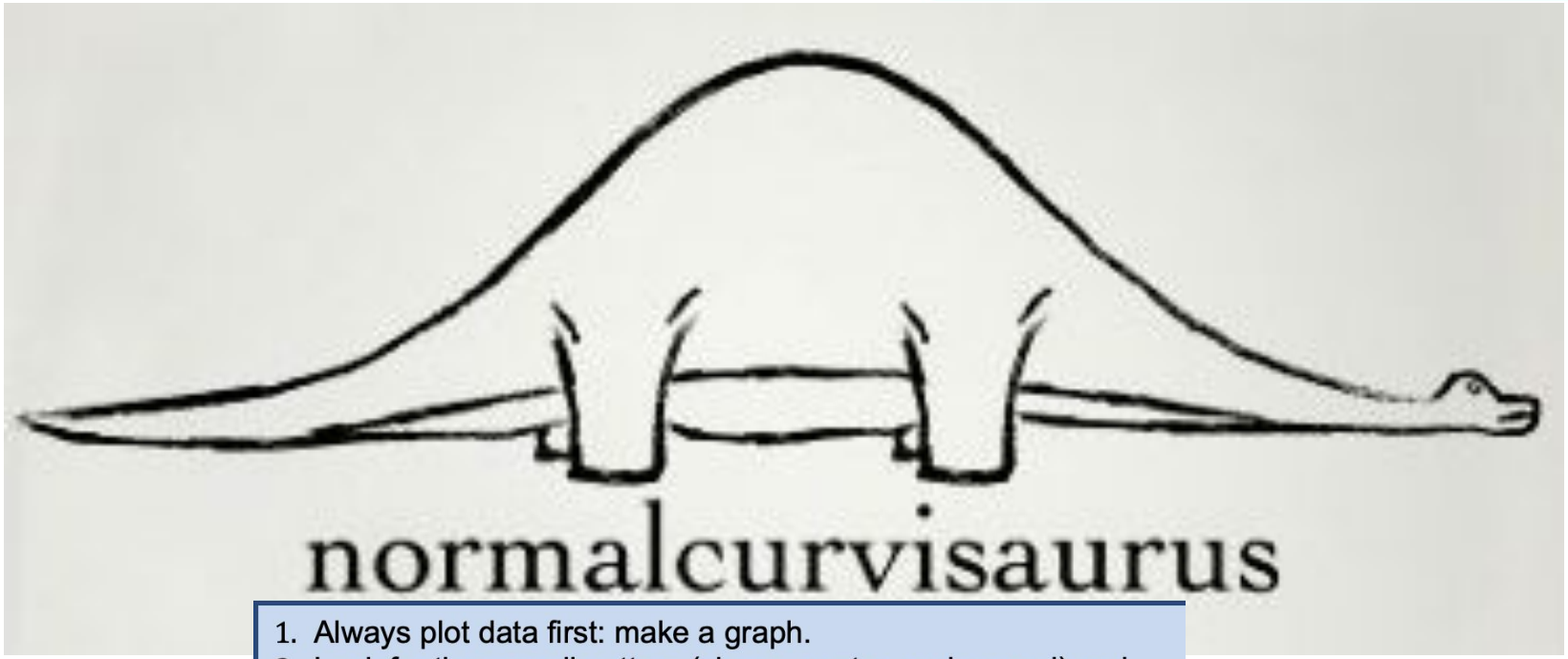

Empirical Rule
68%
95%
99.7%
-3
-2
-1
μ
1
2
3
Number of Standard Deviations Above or Below the Mean



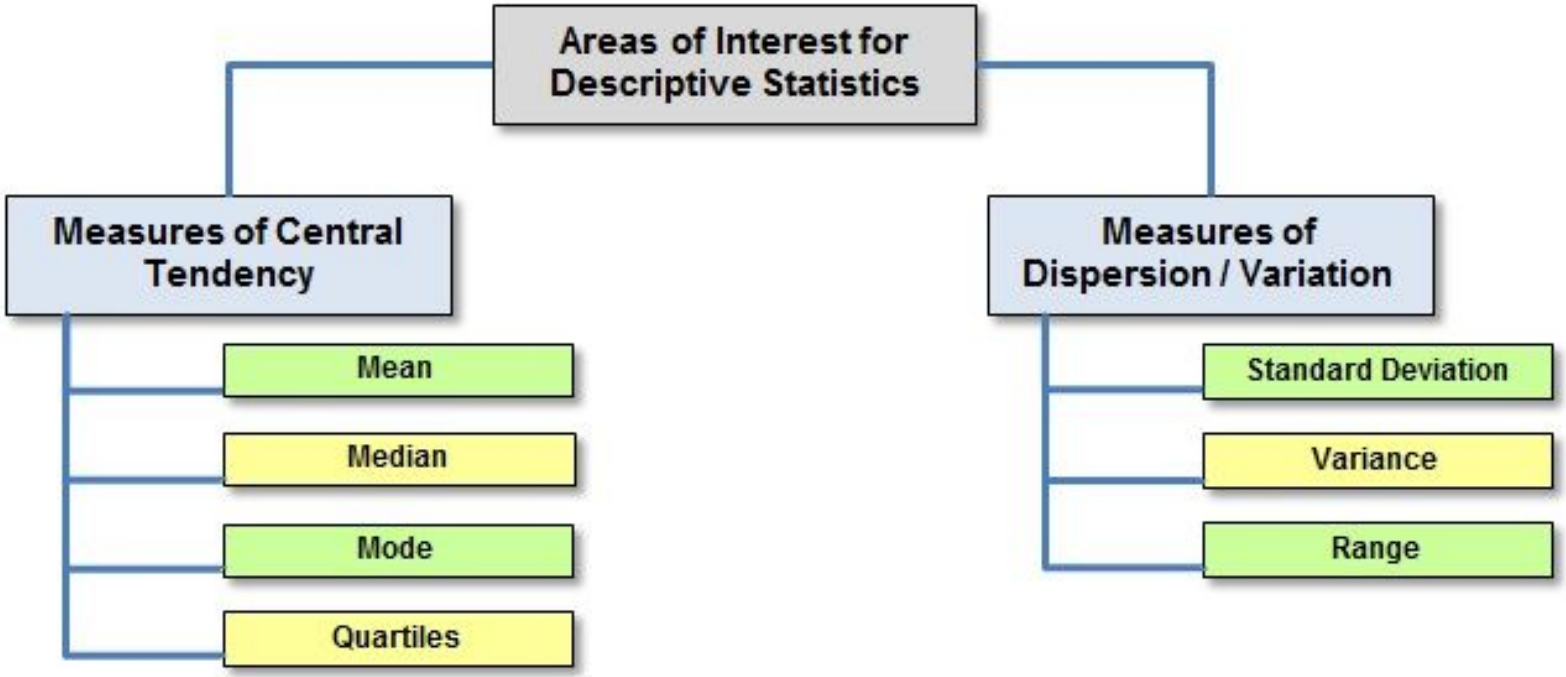

Left-Skewed (Negative Skewness)




Right-Skewed (Positive Skewness)




1. Always plot data first: make a graph.
2. Look for the overall pattern (shape, center, and spread) and for striking departures such as outliers.
3. Calculate a numerical summary to briefly describe center and spread.
4. Sometimes the overall pattern of a large number of observations is so regular that it can be described by a smooth curve.




Summarising the characteristics of a data set helps us understand the data retrieved

## Mean, median, mode

**[1, 2, 2, 4, 6, 8, 50]**

Mean = sum(1, 2, 2, 4, 6, 8, 50) / 7 = 10,43

Median [1, 2, 2, 4, 6, 8, 50] = 4

Mode [1, 2, 2, 4, 6, 8, 50] = 2

## Useful terminology

**Deviations (errors, residuals):** Difference between observed values and an estimate.

**Variance:** The sum of squared deviations from the mean divided by n - 1.

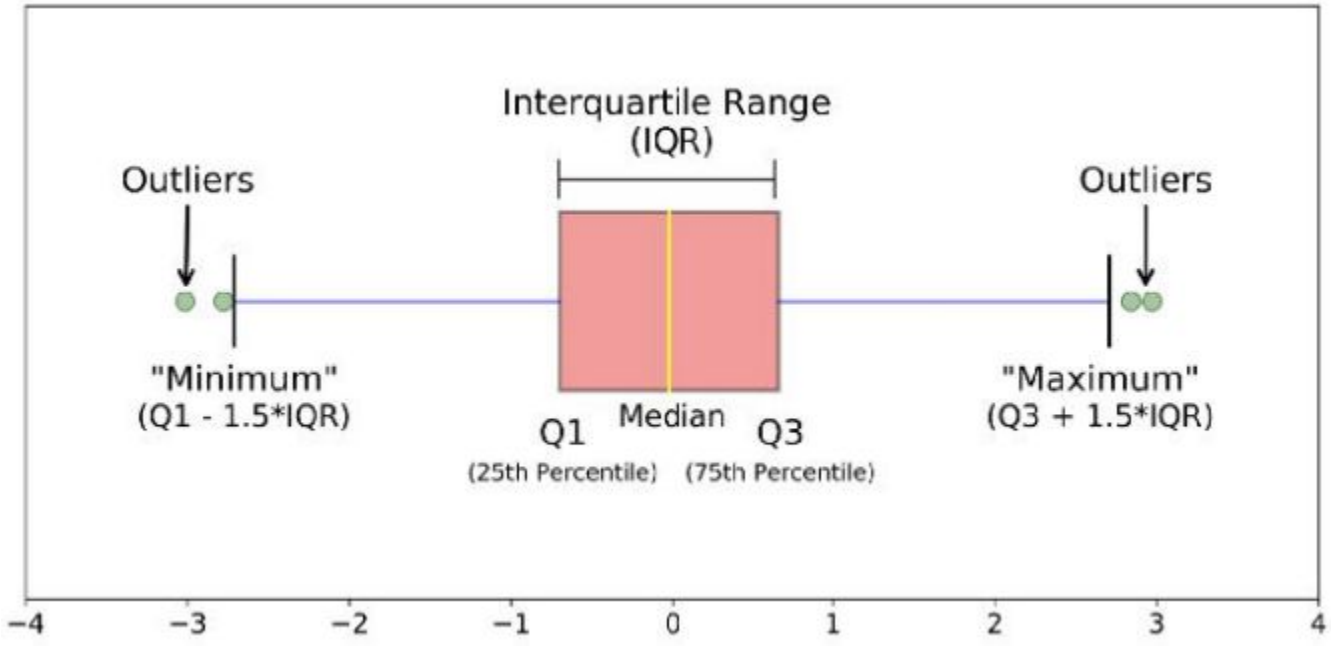
**Standard deviation:** Square root of the variance.

**Range:** Difference between the largest and the smallest value in a data set.

**Percentile:** The value such that P percent of the values take on this value or less.

**Interquartile range:** The difference between the 75th percentile and the 25th percentile

# Descriptive statistics on a box and whisker plot



## Use the correct notation

| | |
|---|---|
| $N$ | Population size |
| $\mu$ | Population mean |
| $\sigma$ | Population standard deviatic |

| | |
|---|---|
| $n$ | Sample size |
| $\bar{x}$ | Sample mean |
| $s$ | Sample standard deviation |

# probability

## Probability - True or False?

**1.**

**I've spun an *unbiased* coin 3 times and got 3 heads. It is more likely to be tails than heads if I spin it again.**

**16.**

**I have thrown an unbiased dice 12 times and not yet got a six. The probability of getting a six on my next throw is more than 1/6.**

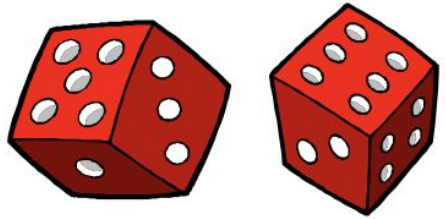
## Probability - True or False?

**15.**

**It is not worth buying a national lottery card with numbers 1, 2, 3, 4, 5, 6, on it as this is less likely to occur than other combinations.**

**8.**

**If six fair dice are thrown at the same time, I am less likely to get 1, 1, 1, 1, 1, 1 than 1, 2, 3, 4, 5, 6.**



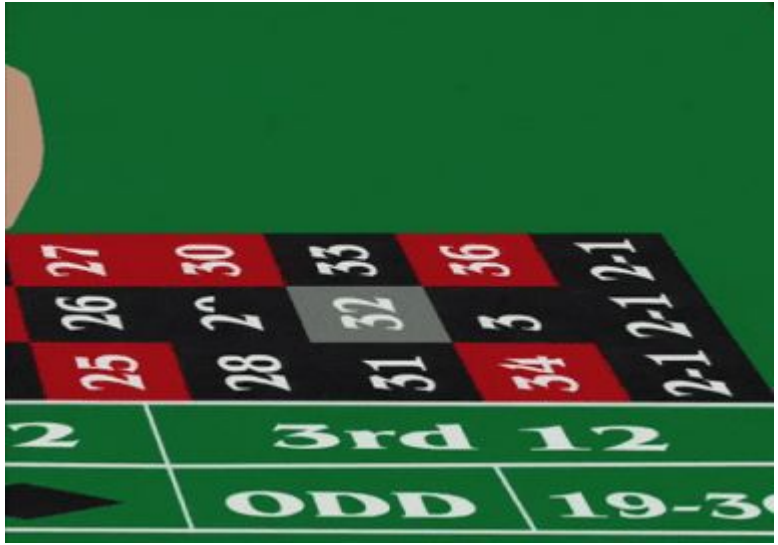
## Probability - True or False?

**14.**

**My Granpa smoked 20 cigarettes a day for 60 years and lived to be 90, so smoking can't be bad for you.**

**7.**

**Mr. Brown has to have a major operation. 90% of the people who have this operation make a complete recovery. There is a 90% chance that Mr. Brown will make a complete recovery if he has this operation.**




## Gambler's Fallacy

The gambler's fallacy implies that when we come across a local imbalance, we expect that the future events will smoothen it out. We will act as if every segment of the random sequence must reflect the true proportion and, if the sequence has deviated from the population proportion, we expect the imbalance to soon be corrected.

In fact this is unreasonable – coins, unlike people, have no sense of equality and proportion

## Law of small numbers

We intuitively believe that inferences drawn from small sample sizes are highly representative of the populations from which they are drawn.

- We have preconceived notions of what randomness looks like.
- We also have a tendency to believe in the self-correcting process in a random sample
- This generates expectations about sample characteristics and representativeness, which are not necessarily true.

# 1. The Prosecutor's Fallacy

"The chance of a coincidental match with an innocent man is 1 in 40,000."

What the Expert Says

"The chance that the accused in innocent is 1 in 40,000, so the odds that he is guilty must be 39,999 to 1."

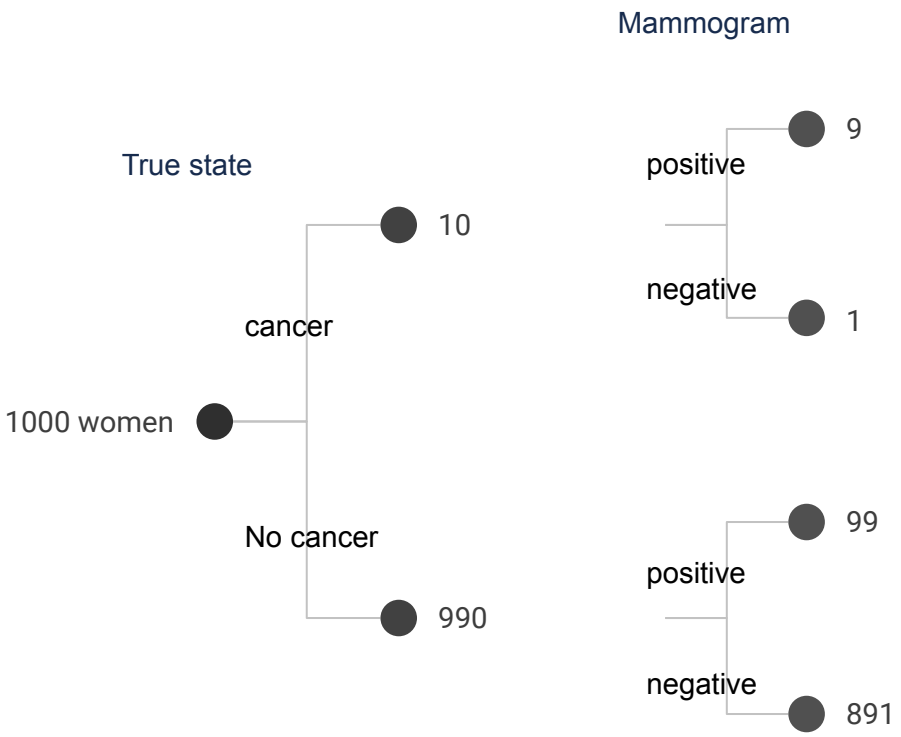
What the Jurors Think

## Conditional Probability

When screening for breast cancer, mammography is roughly 90% accurate ( 90% of the women with cancer and 90% of the women without cancer will be correctly classified.)

Suppose 1% of women have breast cancer.

How do we work out the probability that a woman who has a positive mammogram test really has breast cancer?




Given that there 10/1000 women have cancer, there's a 8% chance it will be correctly screened with a mammogram

- Hypothetical sample
- Not everyone who has cancer goes for a mammogram
- Based on estimates - across ? - where is it 90% ?
- Under what conditions

Assumes all things being equal


Let's talk about hypotheses


1LIVE
SUIC

# NULL HYPOTHESIS EXAMPLES

THE NULL HYPOTHESIS ASSUMES THERE IS NO RELATIONSHIP BETWEEN TWO VARIABLES AND THAT CONTROLLING ONE VARIABLE HAS NO EFFECT ON THE OTHER.




ThoughtCo.

# Null & Alternative Hypotheses

1. Company XYZ manufactures calculators with an average mass of 450g. An engineer believes that average weight to be different and decides to calculate the average mass of 50 calculators. State the null and alternative hypotheses.

2. The teachers in a school believes that at least 80% of students will complete high school. A student disagrees with this value and decides to conduct a test. State the null and alternative hypotheses.

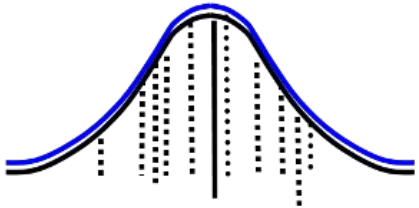
3. A teacher wishes to test if the average GPA of students in the high school is different from 2.7. State the null and alternative hypotheses.

4. The percentage of residents who own a vehicle in town XYZ is no more than 75%. A researcher disagrees with the value and decides to survey 100 residents asking them if they own a vehicle. State the null and alternative hypotheses.

Do smokers weigh the same as non-smokers?

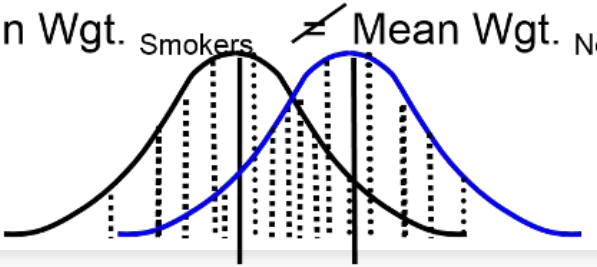
Null Hypothesis ($H_0$): the average weight does *not* differ

$H_0$: Mean Wgt. $_{smokers}$ = Mean Wgt. $_{Non\text{-}smokers}$



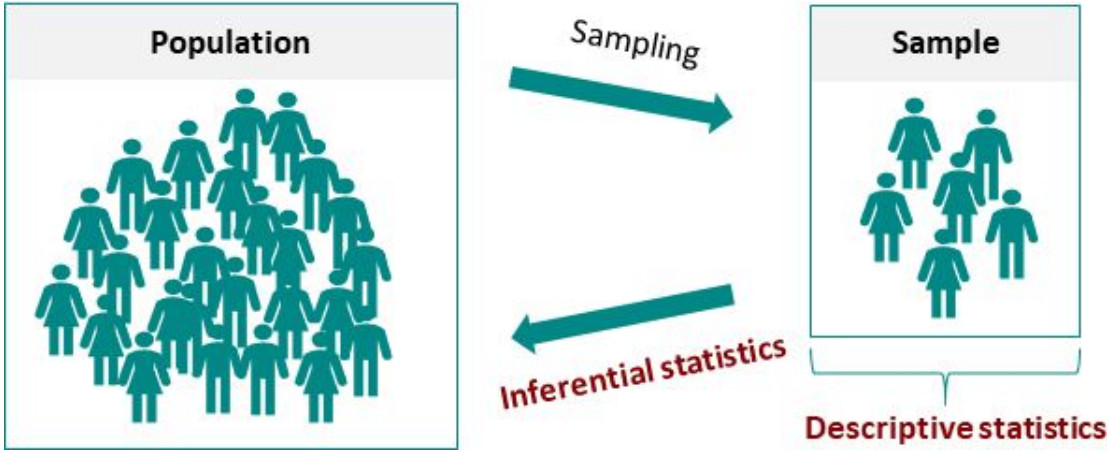
Alternative Hypothesis ($H_A$): the average weights differ

$H_A$: Mean Wgt. $_{Smokers}$ $\neq$ Mean Wgt. $_{Non\text{-}smokers}$



**Descriptive Statistics =** describes / summarizes the data through some statistics which include mean, median, mode, frequency, standard deviation, and variance.

**Inferential Statistics** = helps us to draw inferences about the population, from the given sample data. The goal is to draw conclusions from smaller number of subjects/samples and generalize them on a larger population (where the complete data is unavailable or unknown)

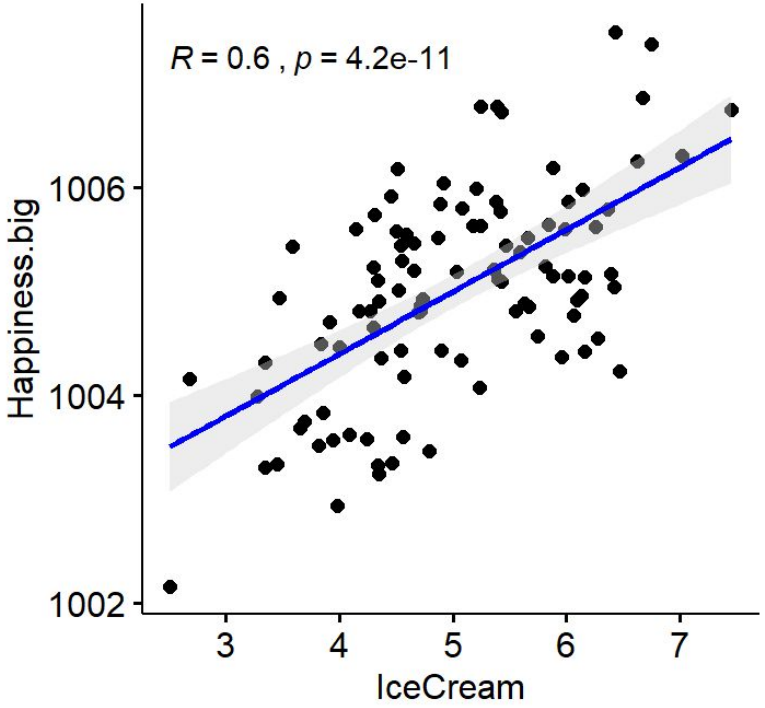


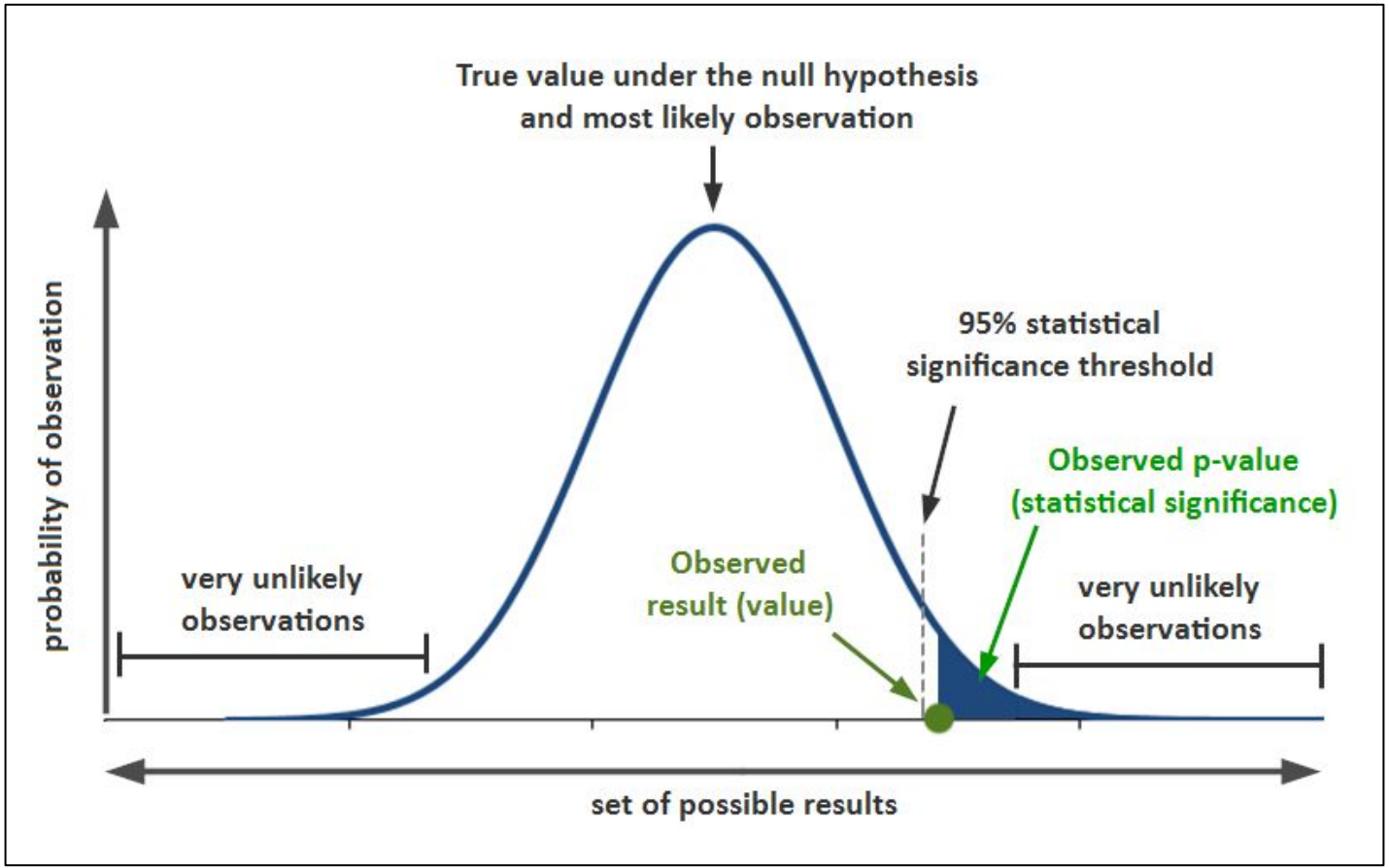
# Using the P value to support a logical starting point

P value can be used to help reject a H0

In a regression model the P-Value for an independent variable tests the Null Hypothesis that there is “No Correlation” (in your sample)

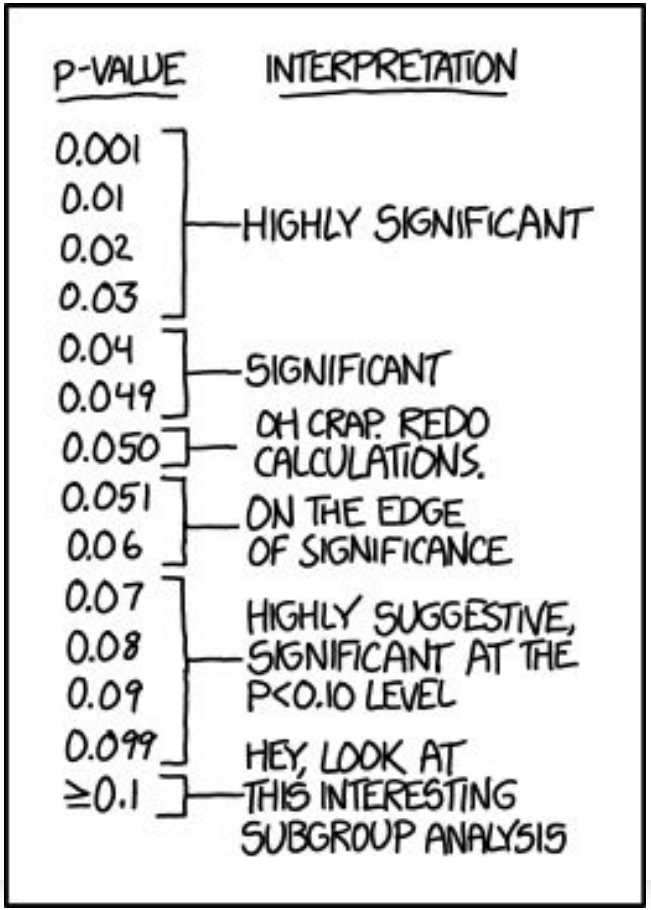
So, if the P-Value is less than the significance level - usually 0.05- then your regression line fits the data (in your sample) well.



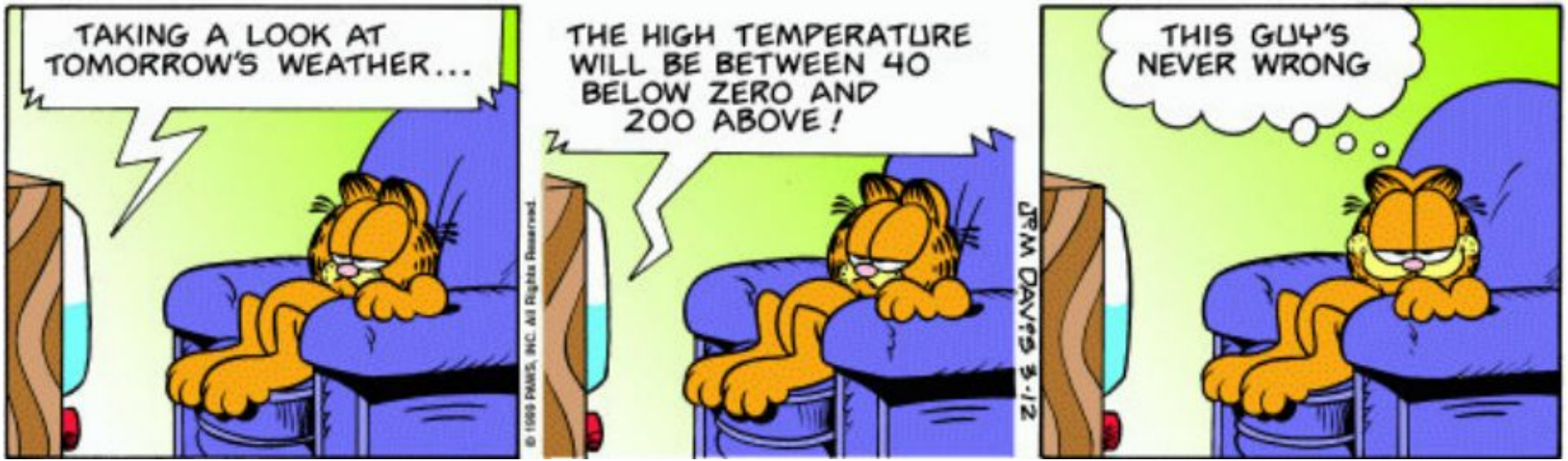

True value under the null hypothesis
and most likely observation
95% statistical
significance threshold
Observed p-value
(statistical significance)
Observed
result (value)
very unlikely
observations
very unlikely
observations
probability of observation
set of possible results

# The P Value is NOT

- The p-value is not the probability the null hypothesis is false
- The p-value is not evidence we have a good linear model
- A high p-value does not necessarily mean there is no relationship between two variables
- P value + R squared do not make a business decision



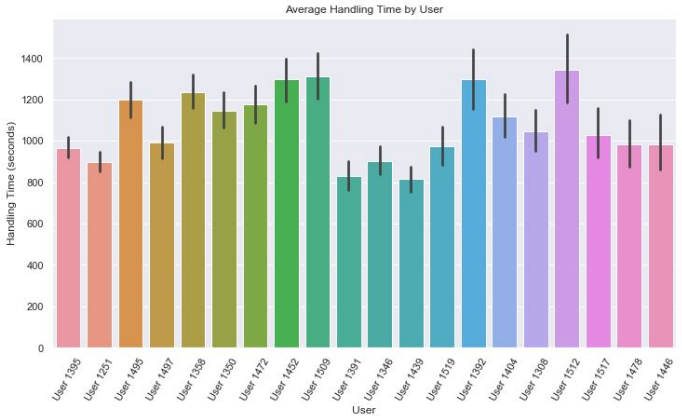
# Confidence (or lack of)

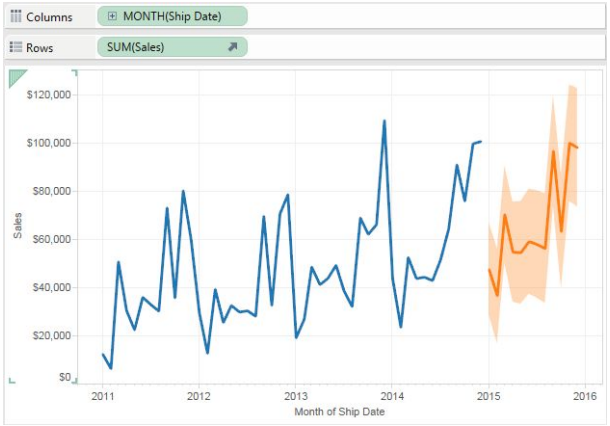

TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
JIM DAVIS 3-12

# Dealing with Uncertainty

- For any given research question, are you easily able to define the population?
- What methods can be undertaken to collect data which accurately reflects the target population?
- Margin of error should be reflected in reported results
- Display charts with confidence interval (normally 95%)

(Does not include systematic error)



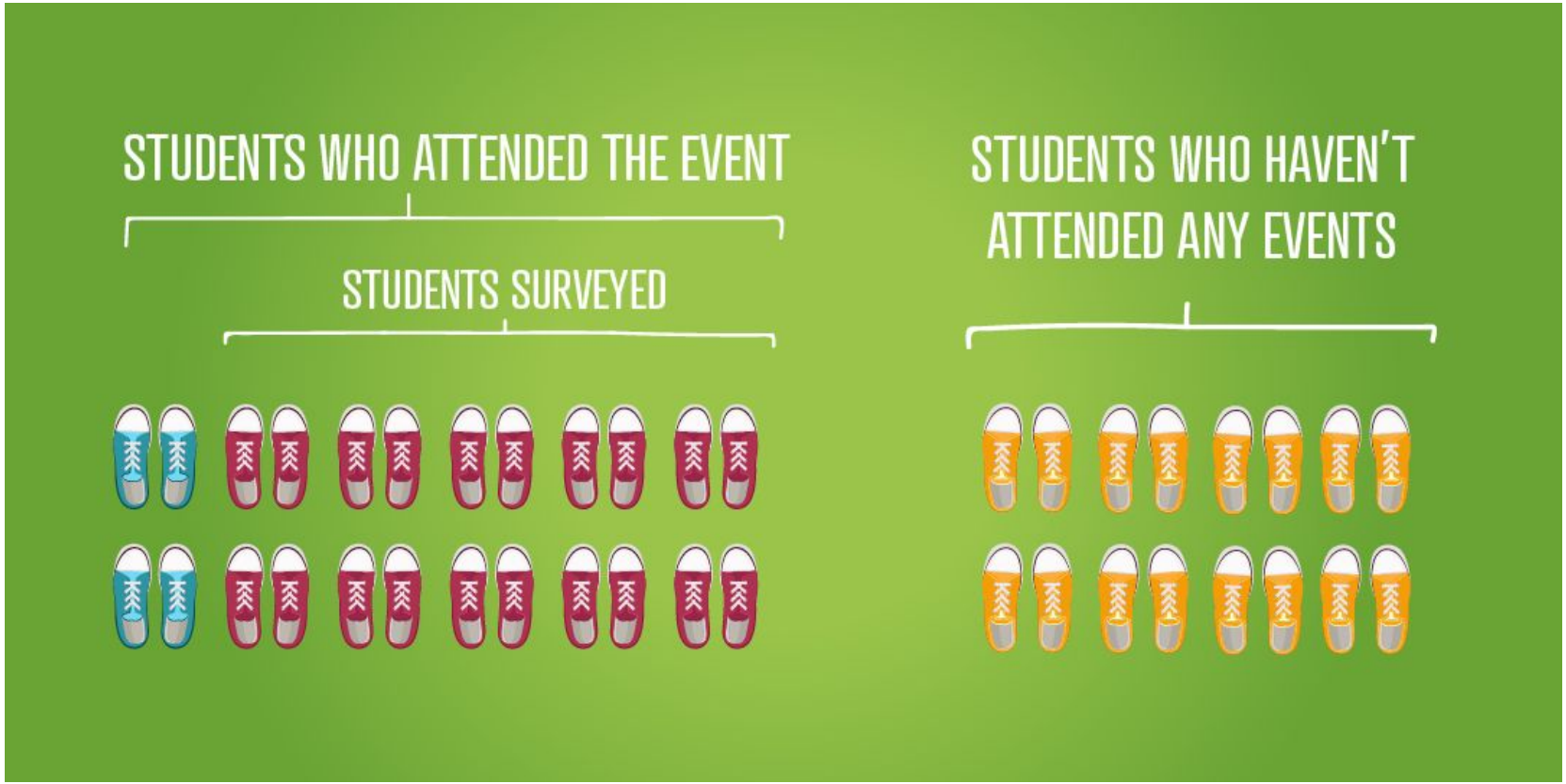
# Sampling and bias

## Bias

Systematic error: one or more parts of the population are favored over others to become part of the sample


## Unbiased sample

A good sample is the one that is representative of the entire population


STUDENTS WHO ATTENDED THE EVENT
STUDENTS SURVEYED
STUDENTS WHO HAVEN'T ATTENDED ANY EVENTS

# What is a confidence interval

**A confidence interval is a range of values that encloses a parameter with a given likelihood.**

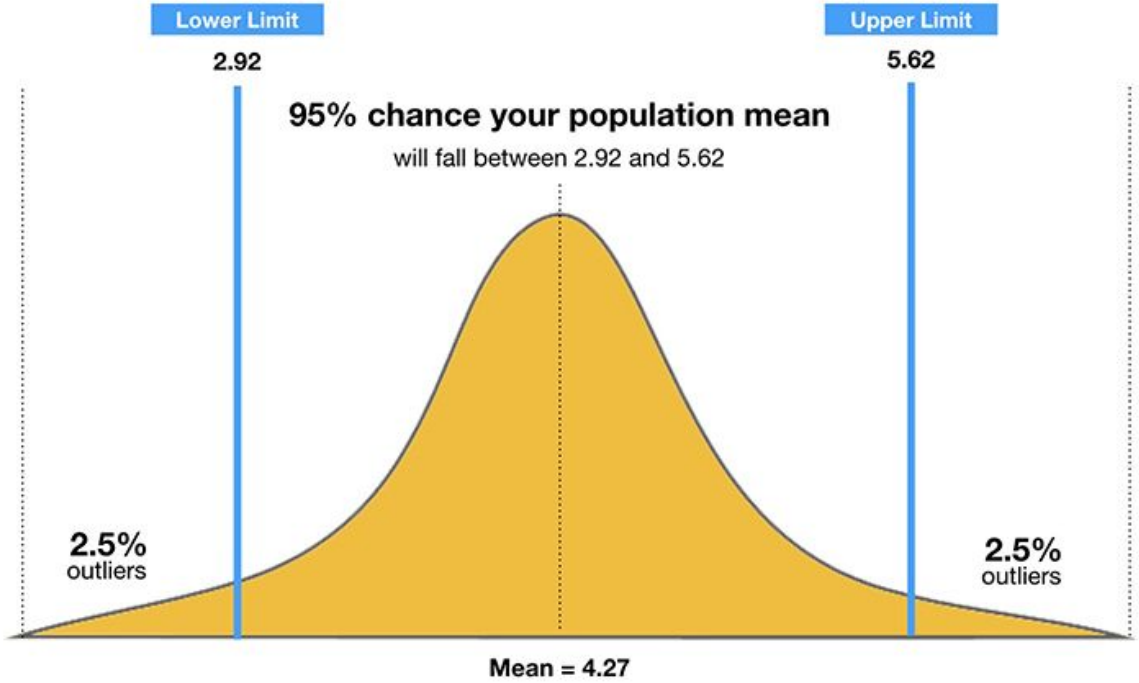
So let's say we've a sample of 200 people from a population of 100,000. Our sample data come up with a correlation of 0.41 and indicate that **the 95% confidence interval for this correlation runs from 0.29 to 0.52.**

**This means :**

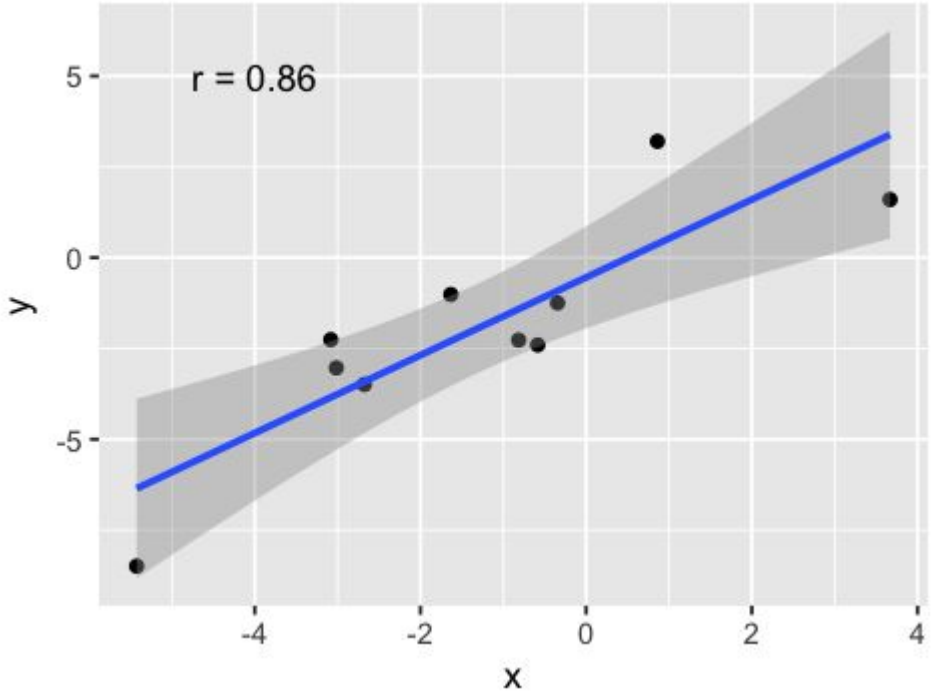
- the *range of values* -0.29 through 0.52
- has a *95% likelihood* of enclosing the *parameter* (the correlation for the entire population) that we'd like to know.

a confidence interval tells us how much our *sample* correlation is likely to differ from the *population* correlation we're after.

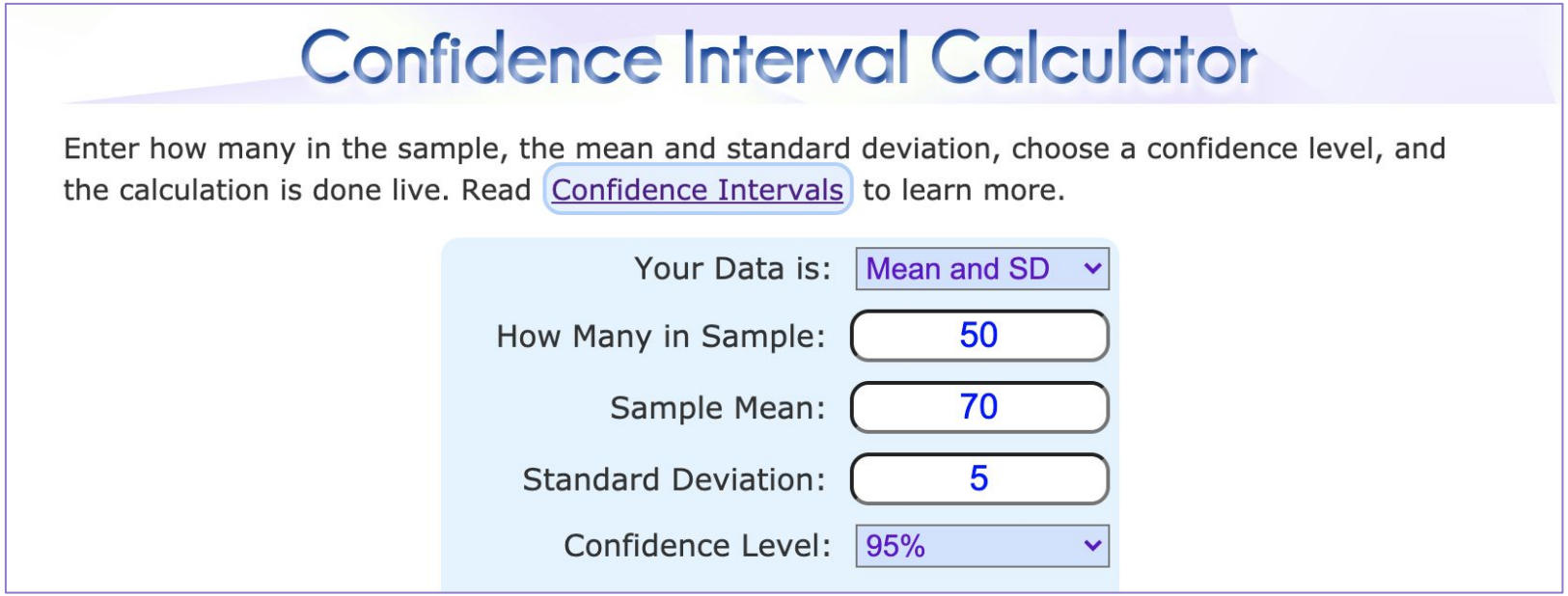
# Visualising a confidence interval



# Visualising a confidence interval



# Calculate the confidence interval



# Key takeaways for data analysts

# How do we apply these principles?

1 - Propose an explanation for a phenomenon

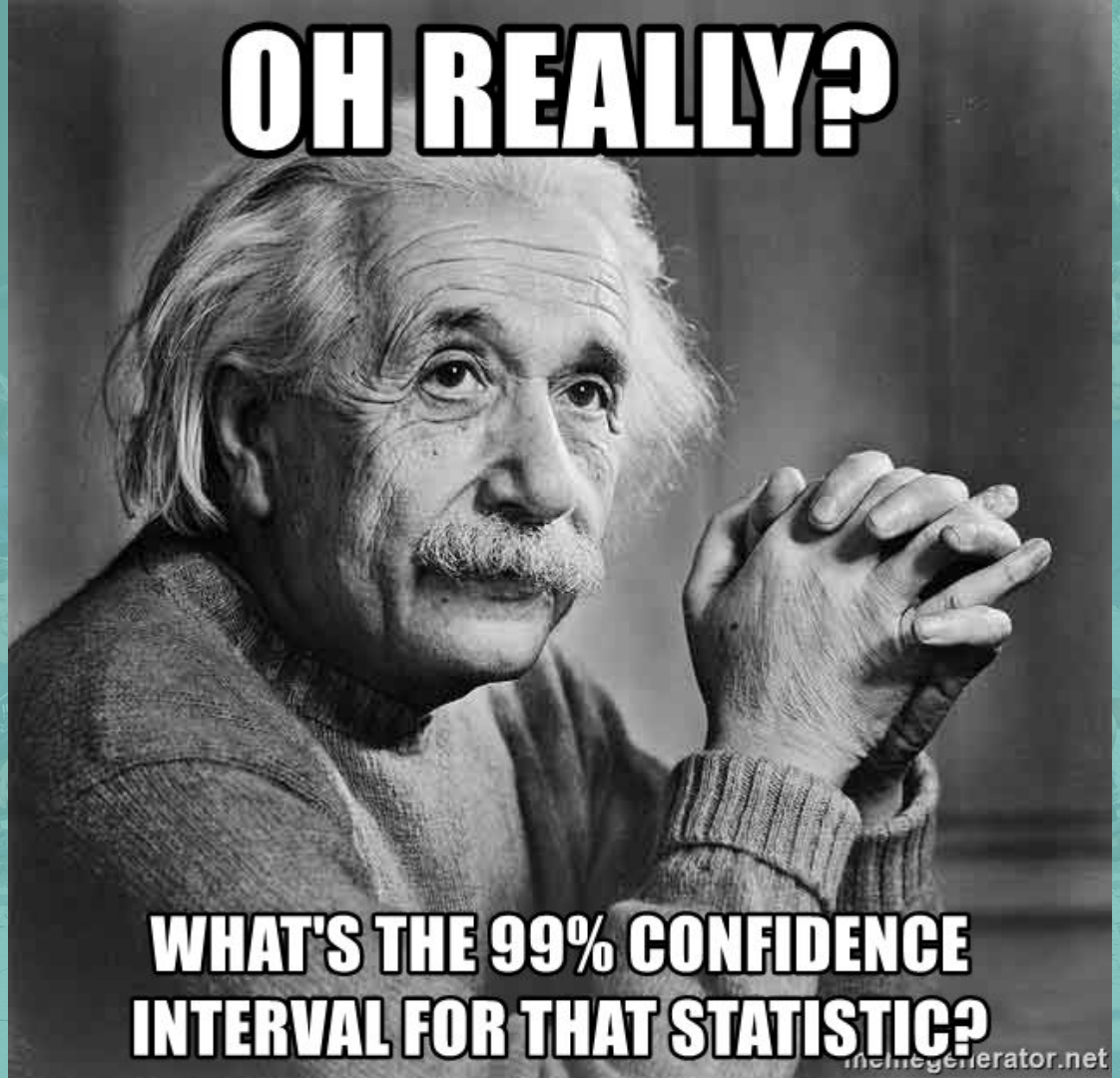
- this is the working assumption, not the absolute truth

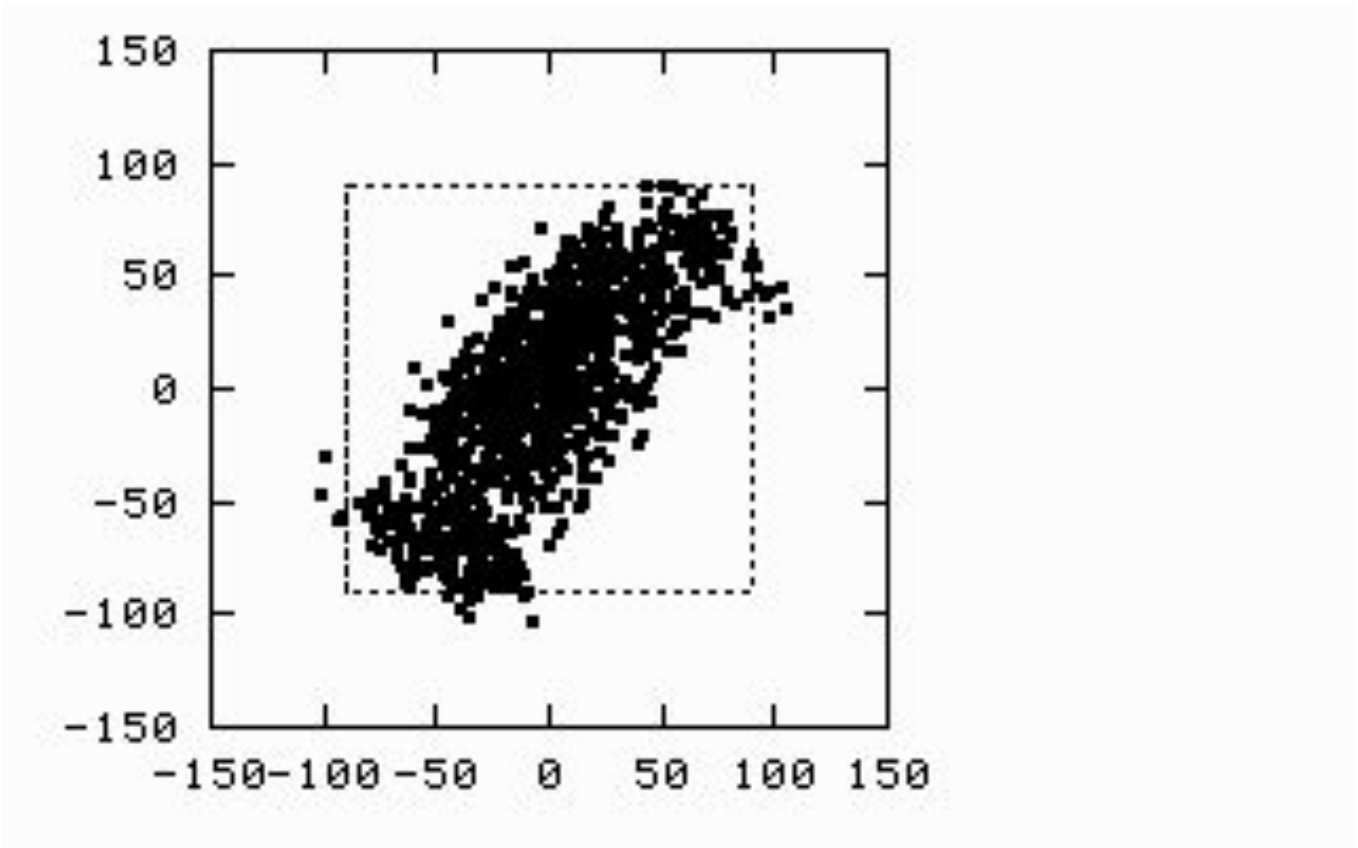
**remember - We all have an innate need to ‘discover’ things**

2 - Assume null until proven otherwise - through experimentation, analysis

3 - Consider natural variation in a population (+ display confidence interval?)

4 - Assume bias, or that observations are not independent


OH REALLY?
WHAT'S THE 99% CONFIDENCE INTERVAL FOR THAT STATISTIC?
memegenerator.net


150
100
50
0
-50
-100
-150
-150-100-50 0 50 100 150