Mathematics and Problem Solving

Lecture 11

Descriptive Statistics

> “An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.”
>
> John Tukey

Overview

- Statistics
- Measures of Central Tendancy
- Measures of Spread
- Data Visualisation
- Correlation
- Probability Distributions

Statistics

- A statistic is a number that provides a summary of a set of other numbers
  - Statistics are key to modern science
  - Machine Learning is essentially sophisticated statistics

A simple statistic

- Lets say you have some text that has been encrypted using a simple cipher. It looks something like this:
  - `rhy iuksq plcaz bcd nuwfo cxyl rhy tmjg vce`
- We can **count** each letter
  - In English, ‘e’ is more common than ‘q’.
- Count is a number that summarises our data. It’s a statistic.

Types of Data

- Data comes in 3 types
  - Numeric
  - Ordinal
  - Nominal

Numeric Data

- Often our data takes the form of numbers
  - 1, 4, 2.2, 5.3, 8.2, 2.1, … etc.
- i.e. numeric or ratio or scalar data
  - Temperature measurements
  - Game scores
- Numeric data **can be ordered**, and the **differences between values** is meaningful

Ordinal Data

- Ordinal data can be ordered, but the ratio between the values is not meaningful
- e.g. we ask people to rate how much they enjoyed using our app, on a scale of 1-5
  - 1, 2, 5, 2, 3, 4, 1, 2
  - In this case 2 does not necessarily mean they enjoyed it twice as much as 1

Nominal Data

- Nominal data has no natural ordering
- e.g. we ask 6 people for their favourite animal
  - Cat, Dog, Squirrel, Dog, Dog, Rabbit

Exercise 1:

For each of the types of data, say whether they are nominal, ordinal, or numeric:

1. Idle temperature of a CPU
2. Gender
3. Age
4. Height
5. Paper size (A3, Fullscap, Letter, A4, etc.)
6. Dog breed ordered by average height

Statistics

- There are two types of statistics
  - Descriptive Statistics
    - Describe properties of your data (e.g. counting occurances)
  - Inferrential Statistics
    - Used to infer things about the world from data. (e.g. hypothesis testing)

Applying Descriptive Statistics

- Lets say you're designing a game that involves probability (e.g. rolling dice)
- There are three ways to analyse it
  - Playtesting
  - Analytically
  - Statistically – through simulation
- You could use descriptive statistics to understand, e.g.
  - How long does the game last, usually? How much does this vary?
  - What is the most likely outcome in x situation? What is the range of likely outcomes?
  - What is the most common move?

Applying Inferrential Statistics

- A/B testing is widespread online
  - Deliver a new version of a website/game to a subset of users
  - Measure something (click throughs, conversions, etc.)
  - Compare the test population against a control group to determine whether the change improved engagement (hypothesis testing)

Summary

- Statistics are useful
  - Describe what things are like
  - Test links between things
- What type of data do you have?
  - Nominal
  - Ordinal
  - Numeric

Measures of Central Tendancy

Measures of Central Tendancy

- Tells you roughly where the middle of your dataset is. I.e. what's at the centre of your sample
- Why?
  - Want a single value that can be used for comparison
    - e.g. 'does this cost more or less than average?'
  - A single value that represents all other values
    - How quickly am I downloading data?

Arithmetic Mean

- The most common meaning of “average”
- Sum a set of numbers and divide by the number of numbers
  - Sometimes denoted using an overline
- Can’t use for nominal or ordinal data

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Exercise 2:

Calculate the arithmetic mean of the following set of numbers:

1. { 2, 2, 3, 4 }

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

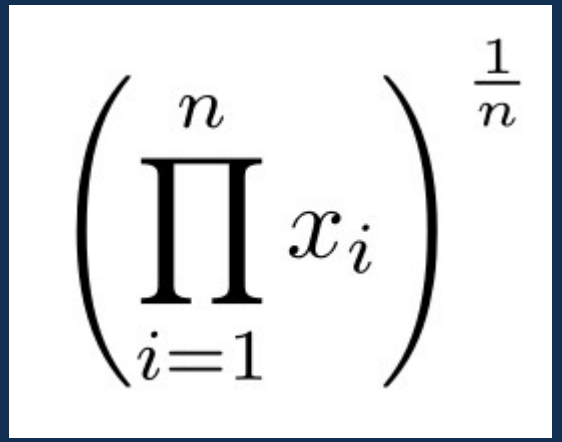
Weighted Mean

- The weighted mean
  1) multiplies each value by a weight
  2) divides by the sum of weights

$$\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Geometric Mean

- Similar to arithmetic mean, but
  1) values are multiplied
  2) the nth root is found



Median

- Median is the element in the middle
  - Sort all elements
  - Pick the element in the middle
    - If there is no element in the middle, take the mean of the two closest values to the middle
- Median can only be calculated for sortable values
  - (numeric or ordinal)

Exercise 3:

Find the median of the following sets of numbers:

1. 3, 1, 2
2. 4, 1, 2, 3, 4, 1

Mode

- The mode is the most common value
  - Count occurances of each value in your dataset. The mode is the value with most occurances
- Works for all types of data

Exercise 4:

Find the mode of the following set of numbers:

1. 1, 2, 3, 4, 1

Exercise 5:

For each of the follow, what is the type of data? and give an appropriate measure of central tendancy

1. Download speed
2. Character choices in Mario Kart
3. Colour

Outliers

- Outliers are data points that are far away from other data points
  - May be due to underlying variability in construct being measured
    - (e.g. sometimes people are just really tall)
  - May be due to experiment error
    - (e.g. misrecorded value)
- Outliers can cause significant problems in statistical analyses

Outlier Detection

- There is no single standard for detecting outliers
- One method is Tukey's fences, proposed by statistician John Tukey
  - An outlier is an observation outside the range

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)]$$

  - Where Q_1, Q_2, and Q_3 are the quartiles
  - And k is a constant. A standard value of k for outlier detection is 1.5

Exercise 6:

For the following data, calculate the maximal value of k of any outlier

- 4, 8, 12, 15, 15, 19, 21, 40

Report the following before and after outliers (using k = 1.5) have been excluded

- Mean
- Median
- Mode

What difference do the outlier(s) make?

- Hint:
  - Tukey's Fences

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)]$$

  - Here
    - $Q_1 = 10$
    - $Q_2 = 15$
    - $Q_3 = 20$

- Mean is usually extremely sensitive to outliers
  - i.e. outliers have a big effect
  - Can lead to odd conclusions
- Median is more robust in the presence of outliers
- Mode is very robust

Question:

Why is this the case?

Measures of Spread

Measures of spread

- How spread out is our data?
  - Measures of Spread describe the variability in a sample or population
- Used with a Measure of Central Tendancy to describe a set of data
- Examples
  - Range
  - IQR
  - Standard deviation

AV Example

Exercise 7:

Say I've written an algorithm

- Takes as input a snapshot from a dashboard camera
- As output, tells an AV to perform an emergency stop as it's about to cause a collision

In order to be safe, this algorithm has to be guaranteed to execute within 0.25 seconds. A sample of timing data, in seconds, is given below.

- 0.1, 0.1, 0.1, 0.2, 0.3, 0.3, 0.3

1. What is the average execution time? Is the algorithm safe to use?

- Above 0.25 almost half the time!
- In order to work out whether our data has certain properties that fit the real world we often need to not just look at where the ‘middle’ of the data is but also look at the spread of the data

Range

- Range is the crudest and easiest measure of spread to calculate
- Range is the difference between the highest and lowest scores in a data set.

range = largest - smallest

Range

- Our data from earlier:
  - 0.1, 0.1, 0.1, 0.2, 0.3, 0.3, 0.3
- Calculate range and mean:
  - Range = 0.2
  - Mean = 0.2
- Values for algorithm a range from 0.1 to 0.3
  - If there is a safety threshold of 0.25, we know we can’t use this algorithm!

Range

- Range is very sensitive to outliers
  - (extreme high or low values)
- For example
  - Range({-2,1,0,1,2,-1,5,2,0,2,1000}) = 1002
- We (usually) don’t want a couple of extreme values to have such a big effect, otherwise such values render our statistics effectively meaningless

Inter-Quartile Range (IQR)

- Similar to range, but resilient to outliers
- Calculate the range of the middle 50% of the data.
  - Between the lower quartile (<25%) and the upper quartile (>75%)
- As outliers are found in the lower and upper quartiles, so they have very little effect.

Inter-Quartile Range (IQR)

- Order the data and find the median
- Split the data upper half (> median) and lower half (< median)
  - Median of lower half = Q1
  - Median of upper half = Q3
- IQR = Q3-Q1

Inter-Quartile Range

- Given the data D = {0,1,2,2,2,3,4,4,5,6,7,8}
  - D_{lower} = {0,1,2,2,2,3}
  - D_{upper} = {4,4,5,6,7,8}
- Median(D_{lower}) = mean(2,2) = 2
- Median(D_{upper}) = mean(5,6) = 5.5
- IQR = 5.5-2 = 3.5

Inter-Quartile Range

- Given the data D = {0,1,2,3,6,7,8}
  - D_{lower} = {0,1,2}
  - D_{upper} = {6,7,8}
- Median(D_{lower}) = 1
- Median(D_{upper}) = 7
- IQR = 7-1 = 6

Exercise 8:

Calculate the range, upper quartile, lower quartile and IQR for the following data:

- 1, 2, 3, 4, 5, 6, 7, 8, 9
- 1, 2, 3, 4, 5, 6, 7, 8

Exercise 9:

Identify outliers (if any) in the following dataset using Tukey's Fences.

1. 4, 5, 7, 7, 7, 8, 8, 8, 9, 10, 12

What are the potential problems of detecting outliers in this way?

When are they more likely to occour?

- Earlier we saw Tukey's Fences
  - An outlier is an observation outside the range

    $[Q_1 - k(Q_3 - Q_1), k(Q_3 - Q_1)]$
  - Where Q_1, Q_2, and Q_3 are the quartiles
  - And k is a constant. A standard value of k for outlier detection is 1.5

Variance

- Average of the squared differences from the mean
- There are two ways to calculate variance depending on whether your dataset is a population or a sample
  - Population Variance (σ^2)
  - Sample Variance (s^2)
- Are you calculating the variance of *this data*, or are you *approximating* the variance of a wider population?

Population Variance

- Population Variance (σ^2) is the square of the average difference between a value and the mean of our data, divided by the size of our dataset

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_1 - \overline{x})^2}{n}$$

- n is the size of our dataset
- x_i is the ith data point of our dataset
- $\overline{x}$ is the mean of all the data points.

Sample Variance

- Sample Variance (s^2) is the same as Population variance, but you divide by the size of the dataset - 1

$$s^2 = \frac{\sum_{i=1}^{n}(x_1 - \overline{x})^2}{n-1}$$

- n is the size of our dataset
- x_i is the ith data point of our dataset
- $\overline{x}$ is the mean of all the data points.

Standard Deviation

- Square-root of variance
- Like variance, there are two ways to calculate standard deviation
  - Population Standard Deviation (σ)
  - Sample Standard Deviation (s)
- Are you calculating the standard deviation of *this data*, or are you *approximating* the standard deviation of a wider population?

Population Standard Deviation

- Often denoted by σ
- Square root of population variance (σ^2)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_1 - \overline{x})^2}{n}}$$

- n is number of items
- $\bar{x}$ is to the overall mean of items

Sample Standard Deviation

- Often denoted by *s*
- Square root of population variance (s^2)

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_1 - \overline{x})^2}{n-1}}$$

- *n* is number of items
- $\bar{x}$ is to the overall mean of items

Exercise 10:

You measure the ages of 5 children in a nursery. The values are given below. Approximate the standard deviation of age of all the children in the nursery.

- 2, 3, 2, 4, 4

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_1 - \overline{x})^2}{n-1}}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_1 - \overline{x})^2}{n}}$$

Summary

- Measures of Spread include
  - Range
  - Inter Quartile Range (IQR)
  - Variance
    - Population
    - Sample
  - Standard Deviation
    - Population
    - Sample

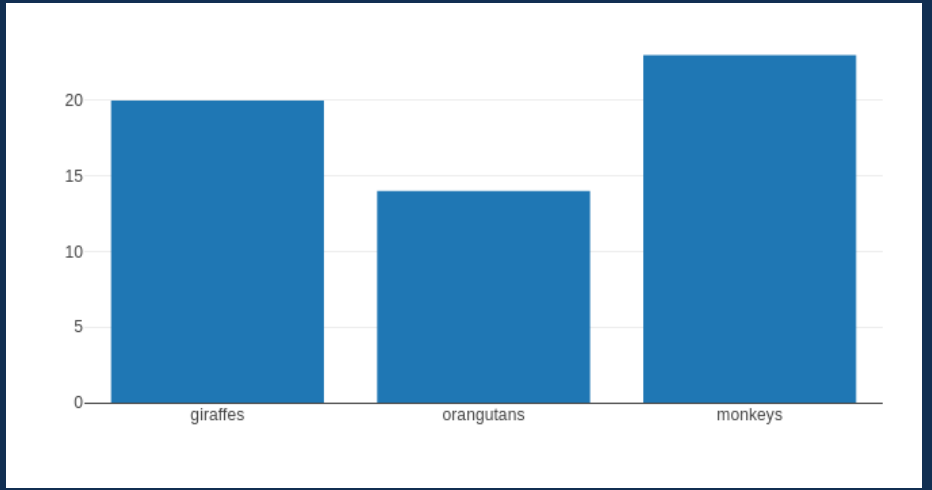
Data Visualisation

Visualising Data

- Why?
  - Look at your data
  - Understand your data
  - Communicate your findings

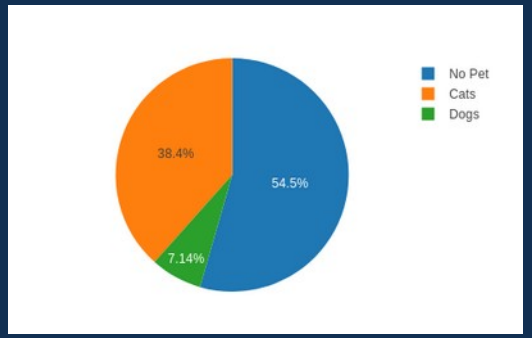
Bar Charts

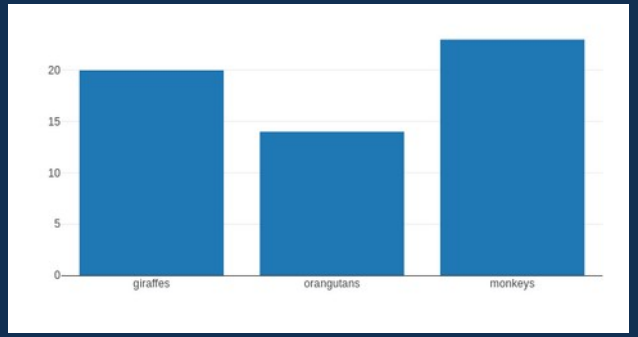
- Show the count of each variable
- Appropriate for nominal data



Proportion

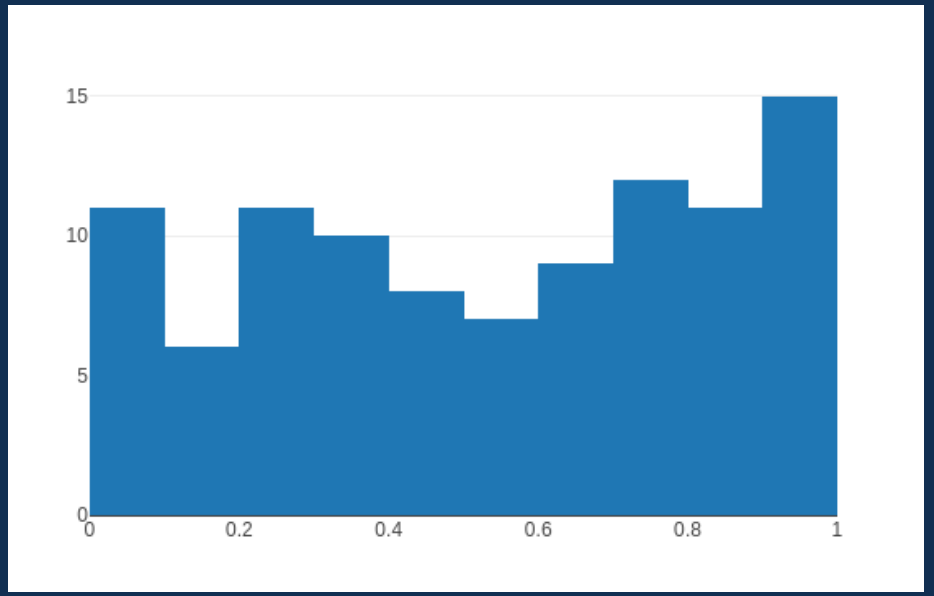
- Pie charts
  - Conveys the idea of part of a whole
  - Harder to read
    - Similar size slices are hard to compare
  - Avoided in research literature





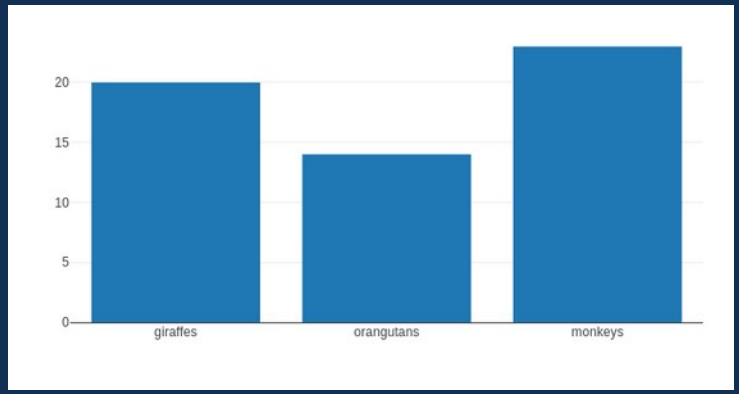
Histogram

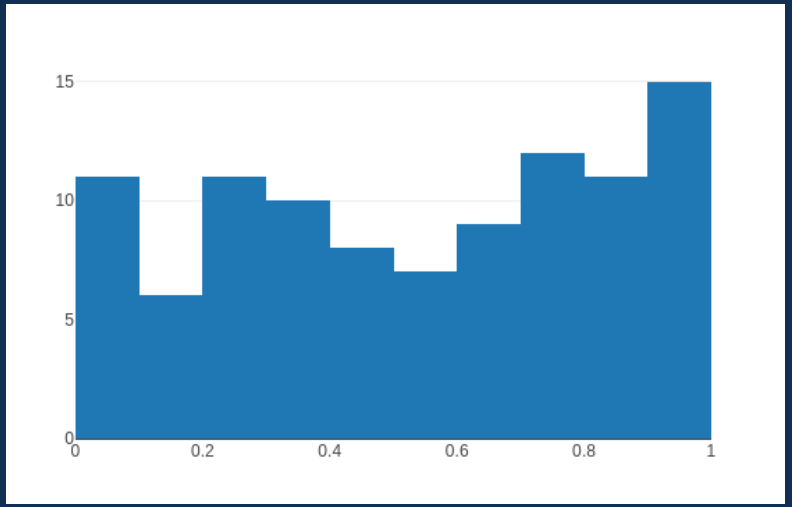
- Group into bins
  - Can vary in size
- For numeric data



Ordinal Data

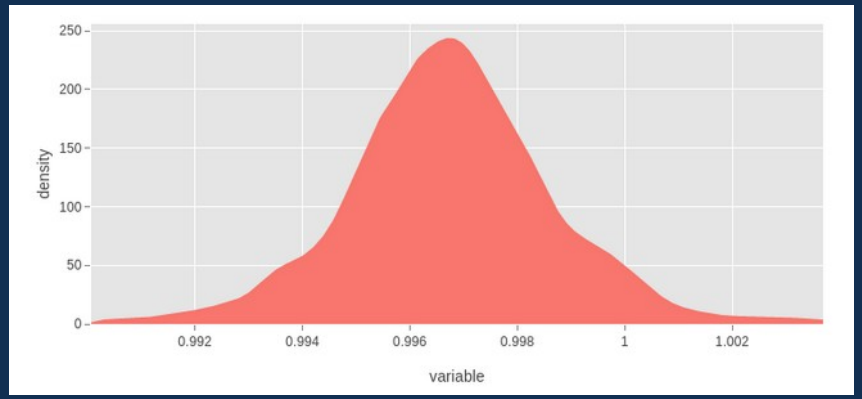
- Few values → bar chart
- Many values → histogram

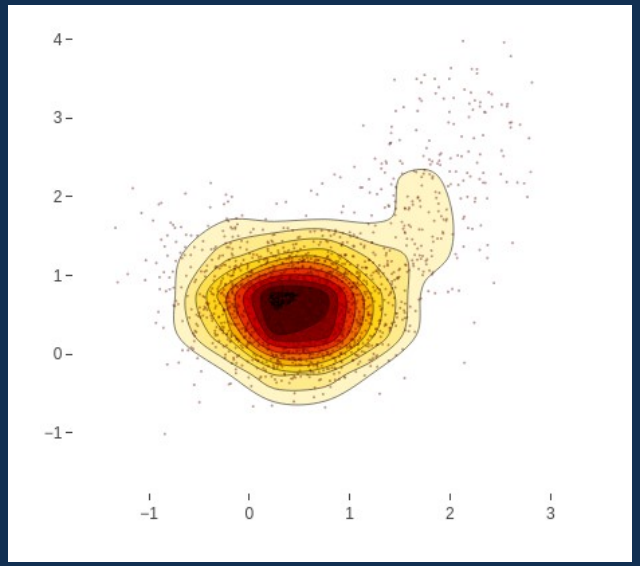




Density Plot

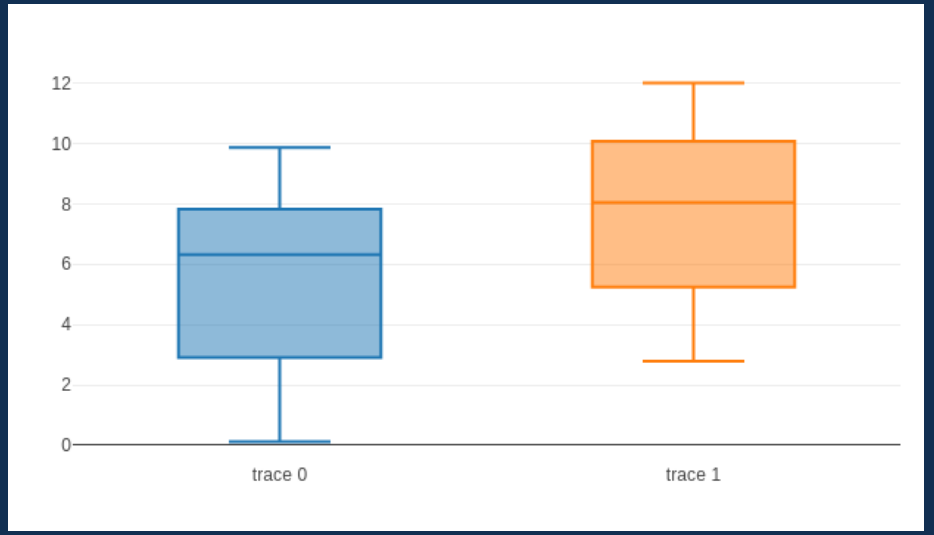
- Show distribution of data





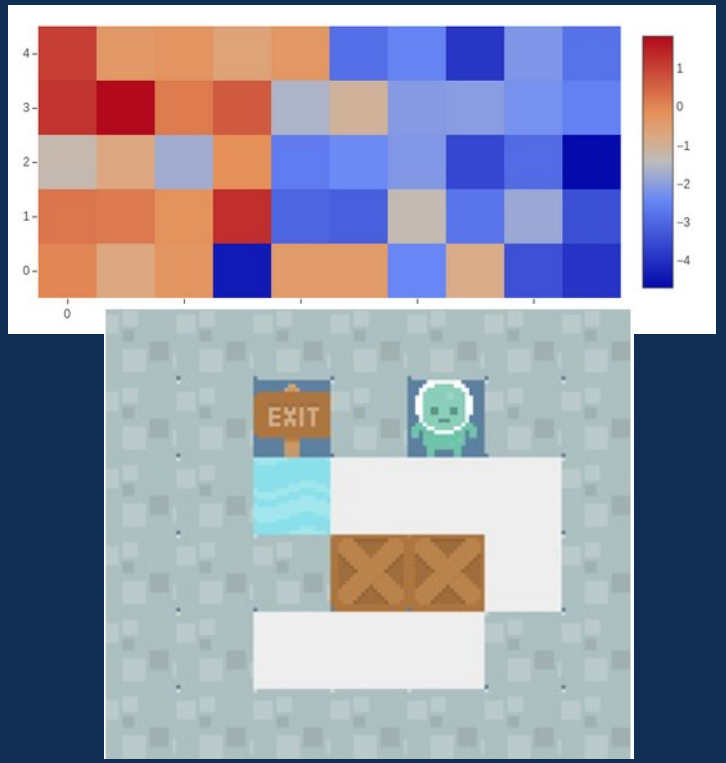
Boxplot

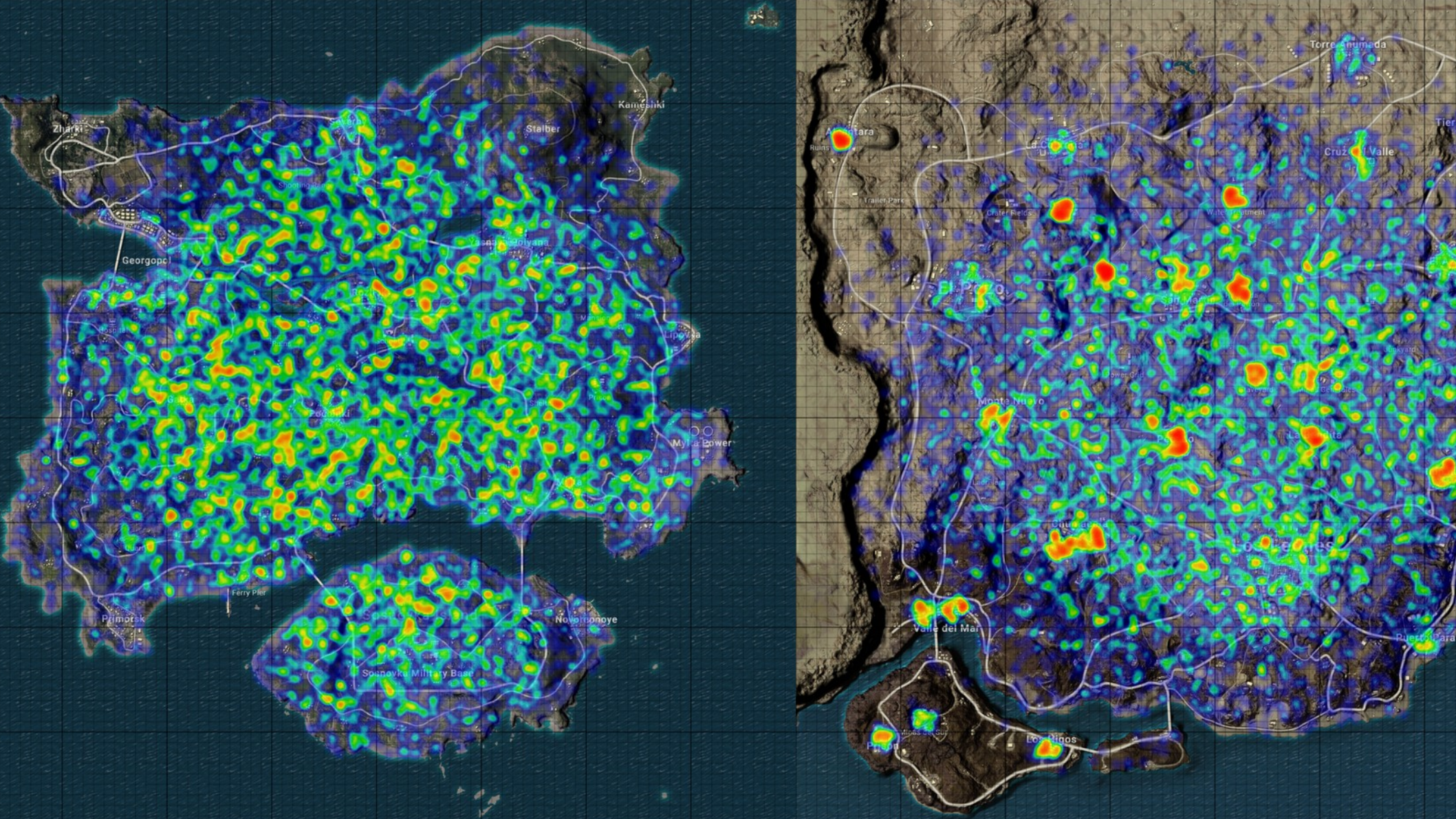
- Boxplot
  - Line = mean
  - Box = IQR
  - Whiskers = range (minus outliers)
  - Outliers as dots
- Standard way of illustrating a hypothesis test comparing multiple groups



Heatmap

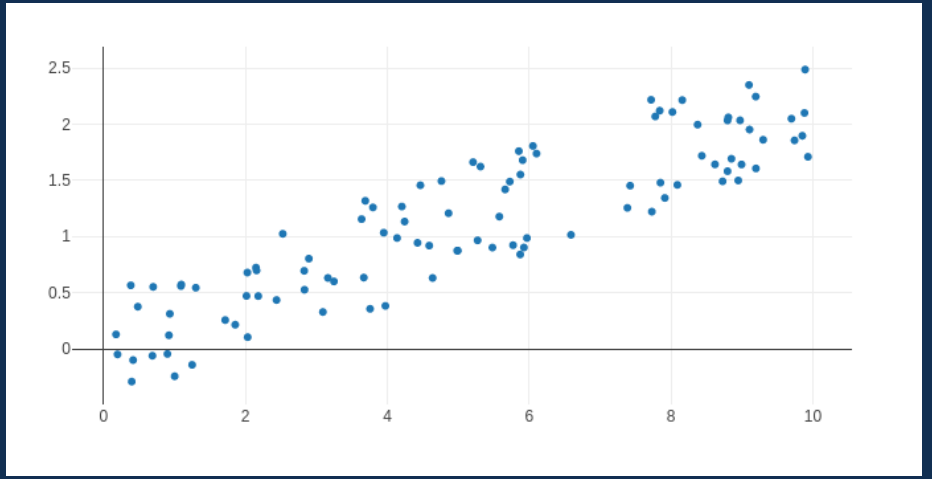
- Shows magnitude in 2D data
  - Easy to see clusters in data
  - Correspondance to a 2D space




Zharki
Stalber
Kameshki
Georgopol
Yasnaya Polyana
Lipovka
Mylka Power
Ferry Pier
Primorsk
Novoreonoye
Sosnovka Military Base
Torre Anumada
Cruz del Valle
Trailer Park
Crater Fields
Water Treatment
El Pozo
San Martin
Monte Nuevo
Valle del Mar
Los Pinos
Prison

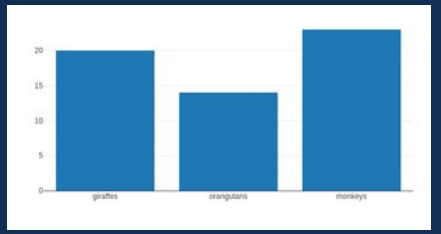
Correlations

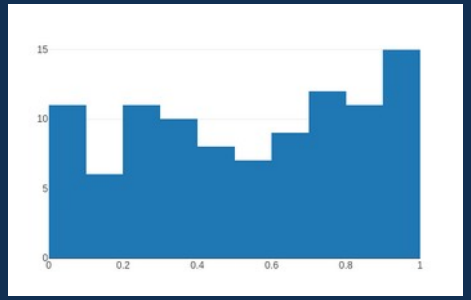
- Correlation?
  - Linear relationship between variables
  - Correlations between -1 → 1
- Scatterplot

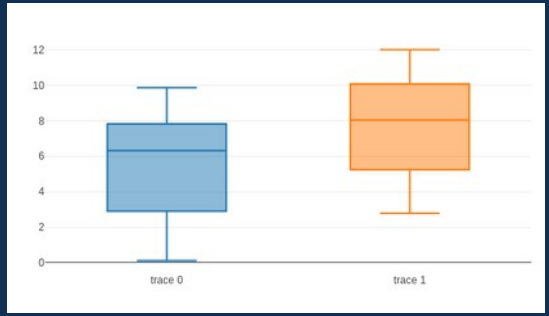


Summary

- Nominal
  - Bar Chart
- Numeric
  - Histogram
- Hypothesis test
  - Box plot





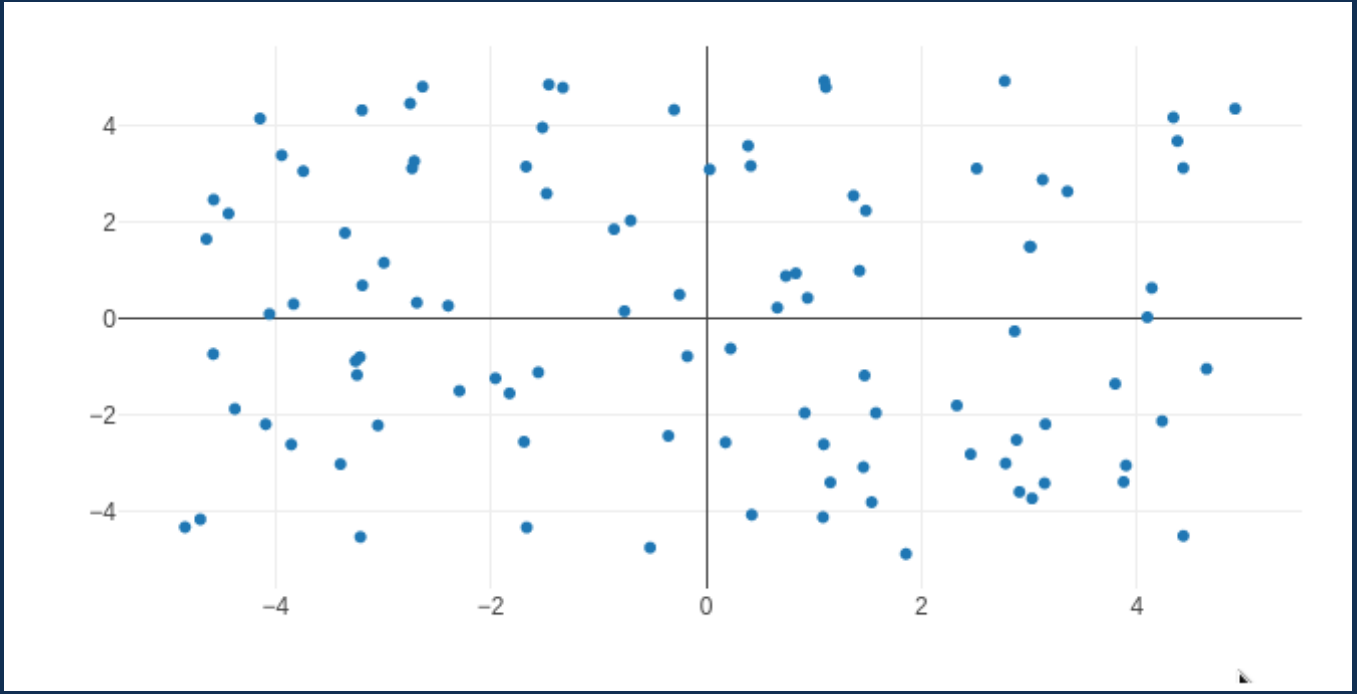


Correlations

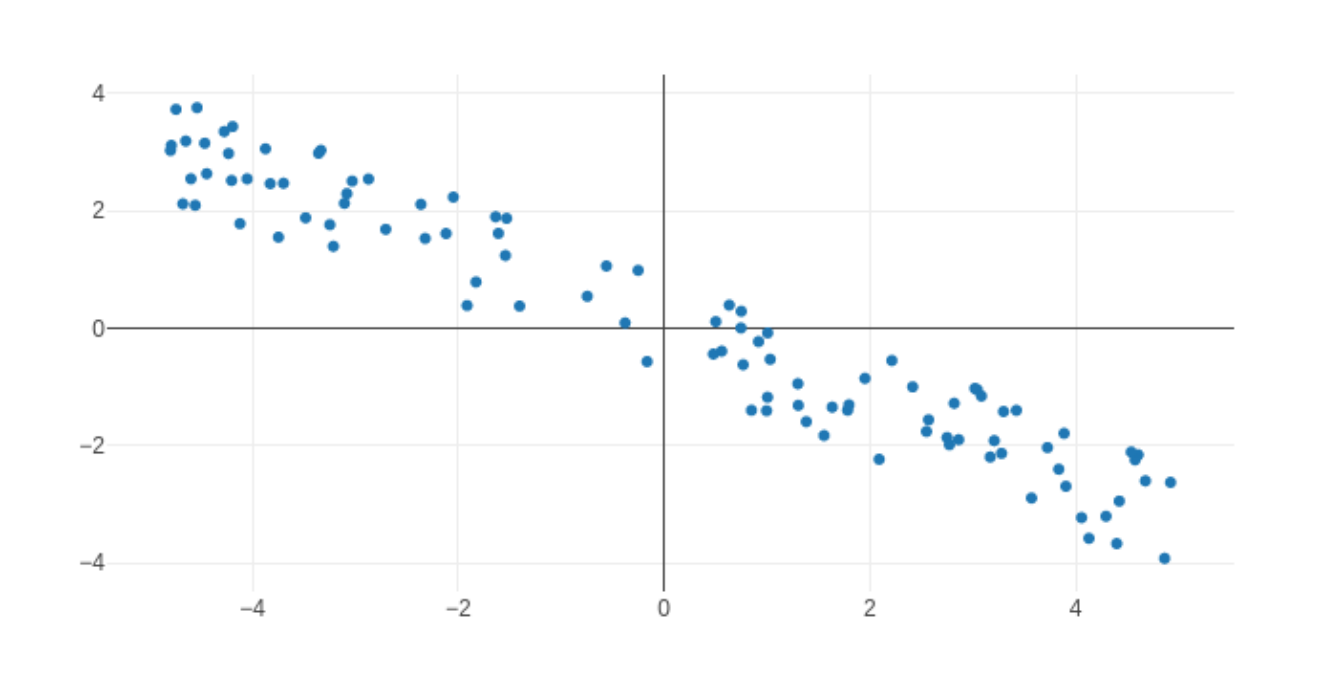
Correlation

- Linear relationship between variables
- Multiple measures of correlation
  - Expressed between -1 → 1
  - -1 (perfect negative correlation)
  - 0 (no correlation)
  - 1 (perfect positive correlation)

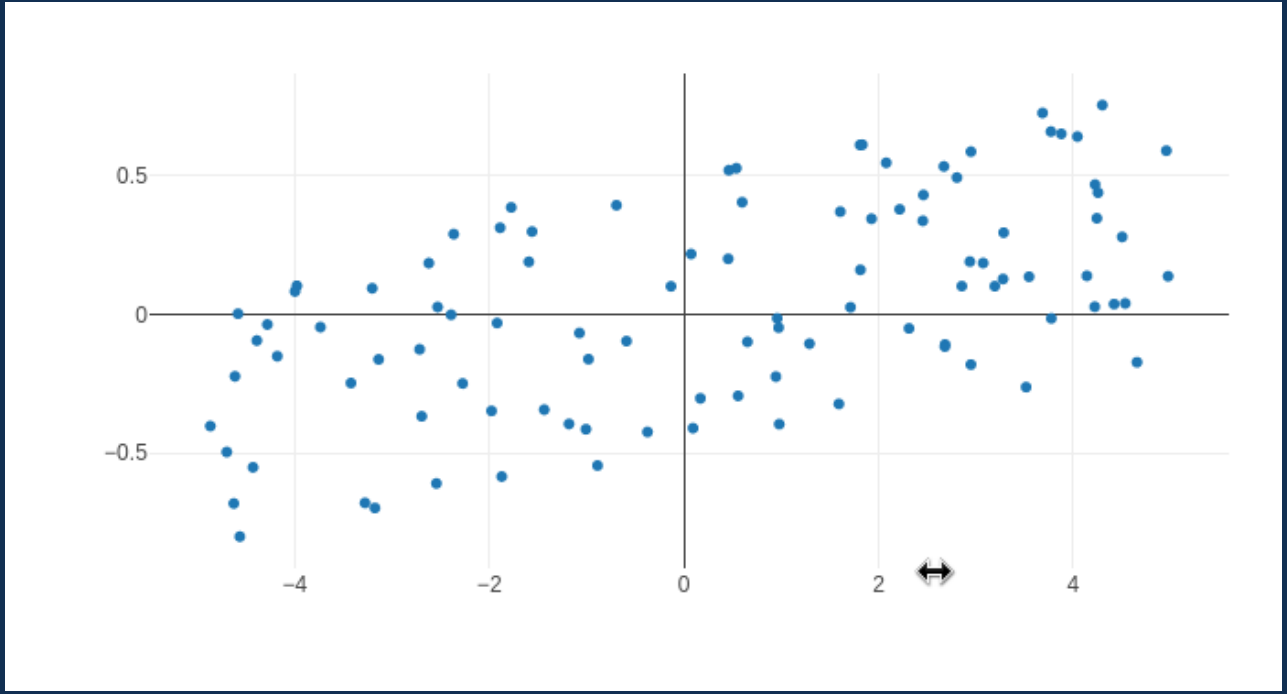
Is there a correlation?



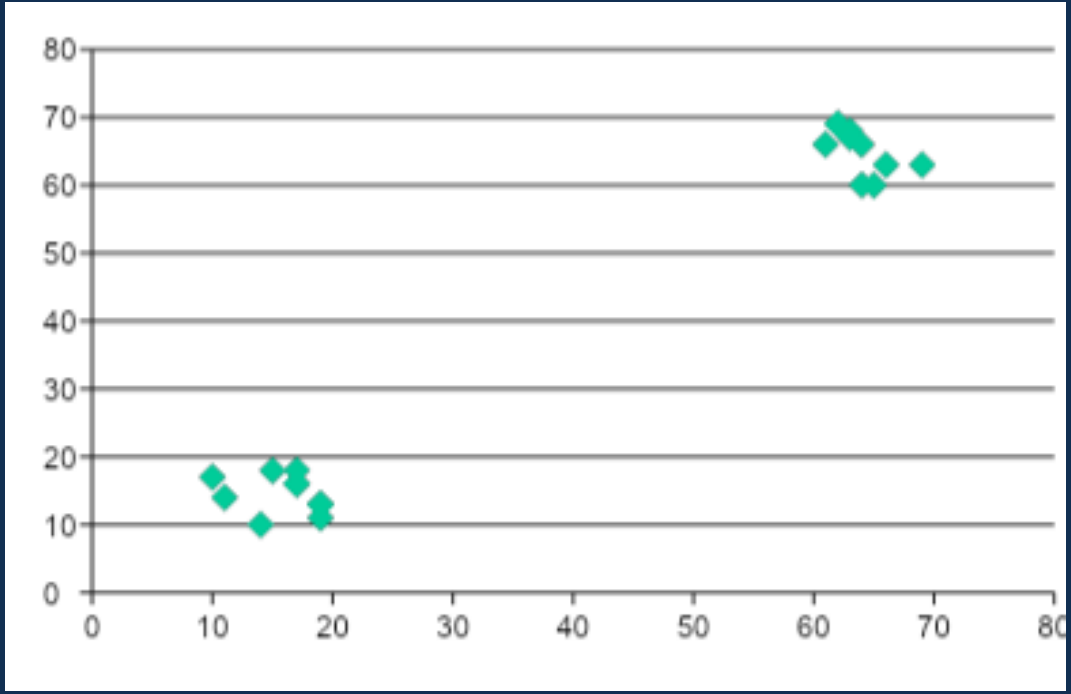
Is there a correlation?



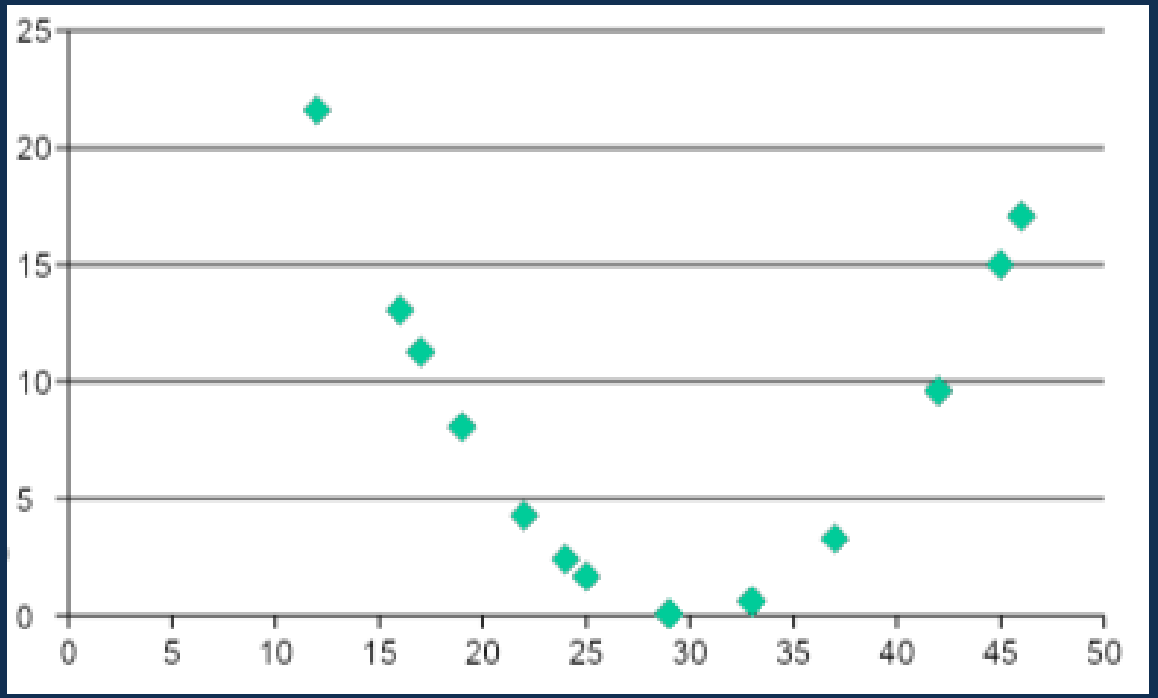
Is there a correlation?



Is there a correlation?



Is there a correlation?



Pearson's r

- Product-moment correlation coefficient
- Checks for linear relationship
- How to calculate
  - Divide into quadrants
  - If lots in 1+3, then +ve, if most in 2+4, -ve
  - Formula expresses this in maths

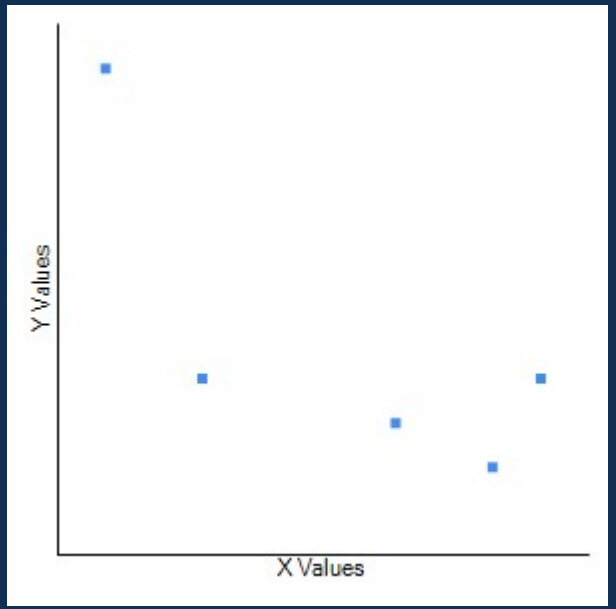
$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}}$$

Pearson's r

Example:

Calculate Pearson's r for the following data:

x	y
2	3
4	-4
8	-5
10	-6
11	-3



Pearson's r

Example:

Calculate Pearson's r

1) Find formula
2) Calculate means

x	y
2	3
4	-4
8	-5
10	-6
11	-3

$\bar{x} = 7 \quad \bar{y} = -3$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Pearson's r

Example:

Calculate Pearson's r

1) Find formula
2) Calculate means
3) Calculate values

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
	2	3	-5	6	-30	-25	36
	4	-4	-3	-1	3	-9	1
	8	-5	1	-2	-2	-1	4
	10	-6	3	-3	-9	-9	9
	11	-3	4	0	0	16	0
Σ					-38	60	50

$\bar{x} = 7 \quad \bar{y} = -3$

Pearson's r

Example:

Calculate Pearson's r

1) Find formula
2) Calculate means
3) Calculate values
4) Calculate r

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$\Sigma(x-\bar{x})(y-\bar{y})$	-38
$\Sigma(x-\bar{x})^2$	60
$\Sigma(y-\bar{y})^2$	50
$\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2$	3000
sqrt($\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2$)	54.77
r	-0.69

Pearson's r

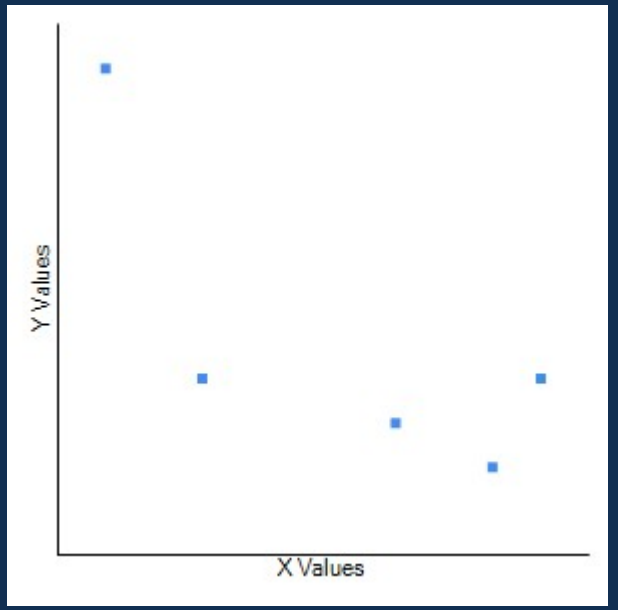
Example:

Calculate Pearson's r for the following data:

Answer:

r = -0.69

x	y
2	3
4	-4
8	-5
10	-6
11	-3



Spearman's Rho

- Measure of correlation for non-parametric data
- Checks for **monotonic relationship** between variables
- Calculate Pearson's r on ranks
  - (i.e. you only care about order, not exact values)

Linear Regression

- Model the relationship between independent and dependant variables
- Try and find a straight line function $y = a + bx$
- Model each point as $y_i = a + b_i + \varepsilon_i$
  - ε_i is error for point *i*
  - Try to minimise error

Least Squares Method

- A formula for finding values of a and b such that

$$y = a + bx$$

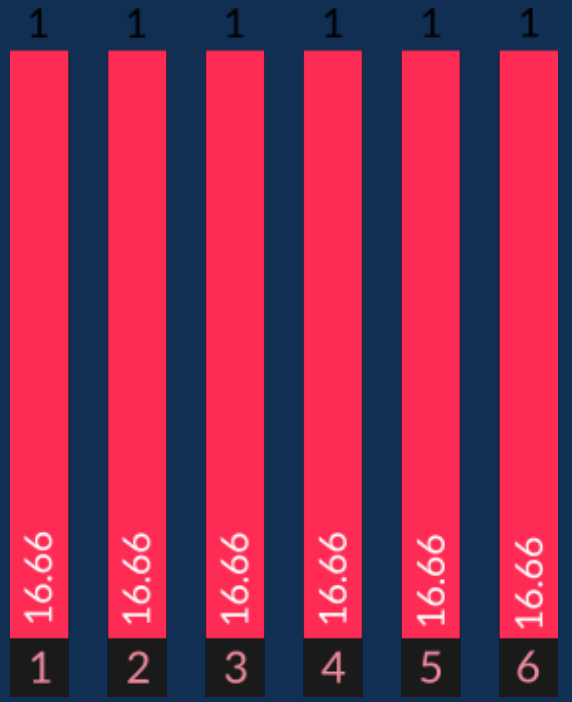
$$b = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

$$a = \overline{y} - b\overline{x}$$

Probability Distributions

Probability Distributions

- Roll one dice
  - The **probability** of getting each number **is the same**



Probability again

Exercise 12:

Imagine you roll two 6-sided dice. What's the probability of getting:

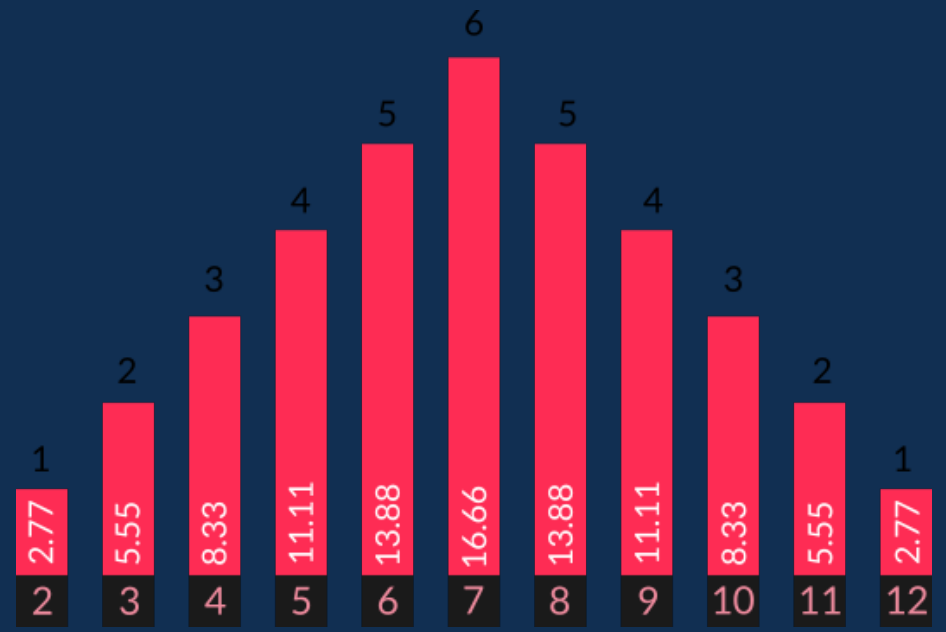
1. 7
2. 12

- What’s the probability of rolling 7?

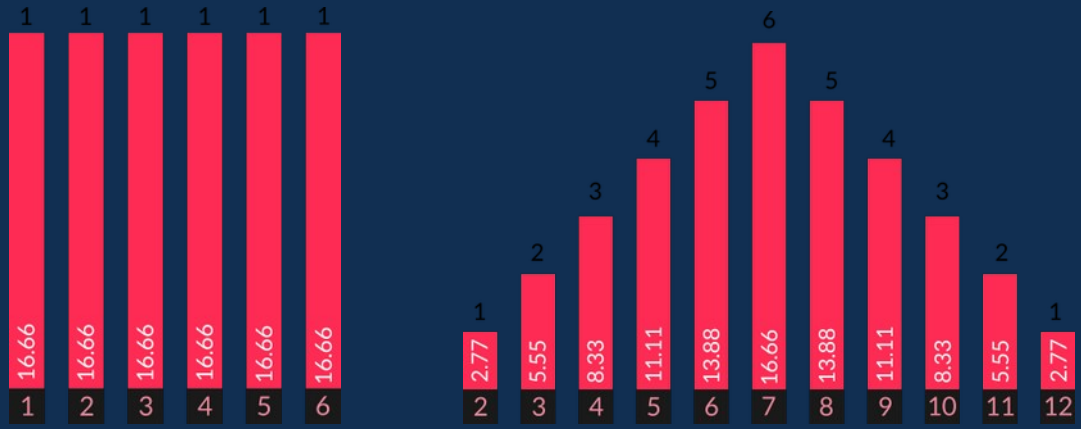
  6/36 = 1/6 = 16.66

P(1) * P(6)
+ P(2) * P(5)
+ P(3) * P(4)
+ P(4) * P(3)
+ P(5) * P(2)
+ P(6) * P(1)

- Roll two dice and add them together
  - The **probability** differs for **each number**



- These are two different **probability distributions**

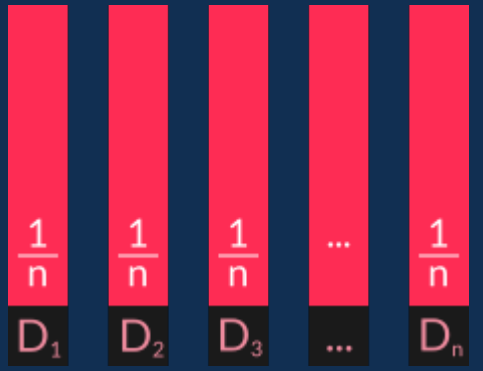


Discrete Distribution

- Discrete
  - **Countable** $D_1 \ldots D_n$
- **Probability mass function** = list of values and their probabilities
  - Sum to 1

Discrete **Uniform** Distribution

- Discrete
  - **Countable** $D_1 \dots D_n$
- Uniform
  - All **same probability**



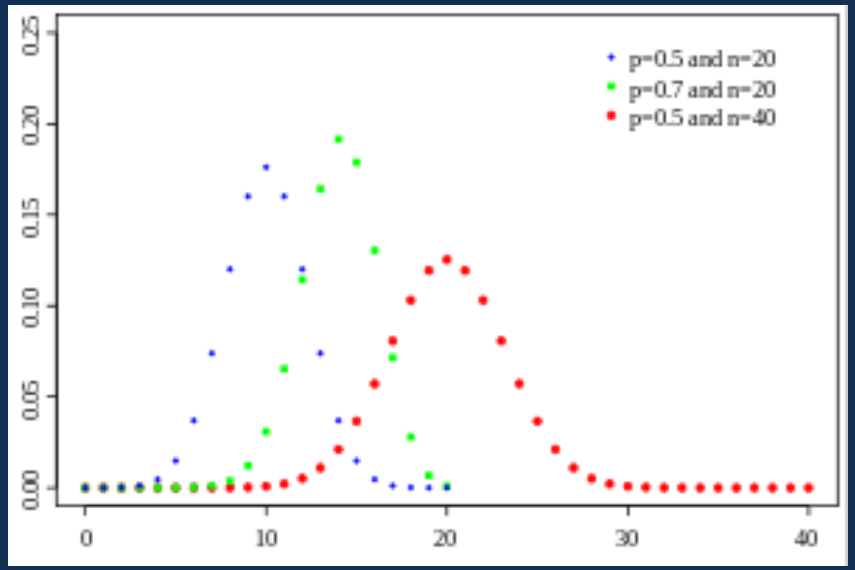
Binomial Distribution

- Flip a coin, twice
  - Probability of success constant = 0.5
  - Independant trials
  - Variable = number of heads

Number of Heads	Probability
0	0.25
1	0.5
2	0.25

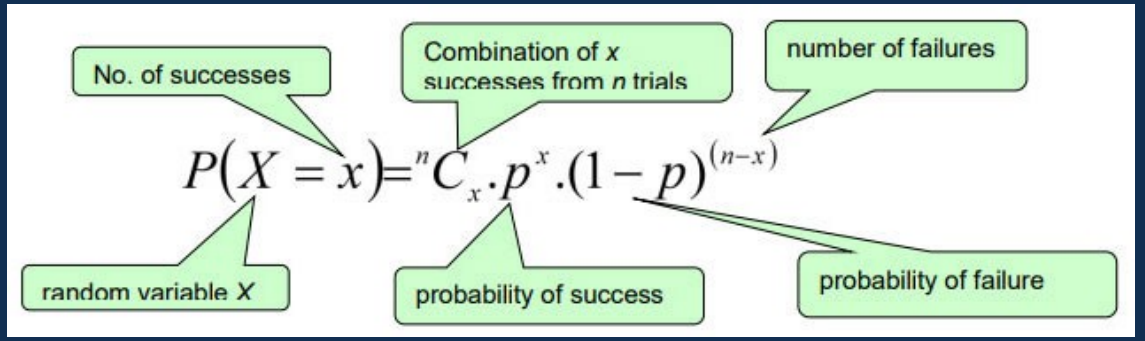
Binomial Distribution

- Number of successes in a sequence of yes/no questions
  - e.g. flip a coin 10 times, count the number of heads
- Defined by
  - **n** (number of questions) and
  - **p** (probability of a “yes”)



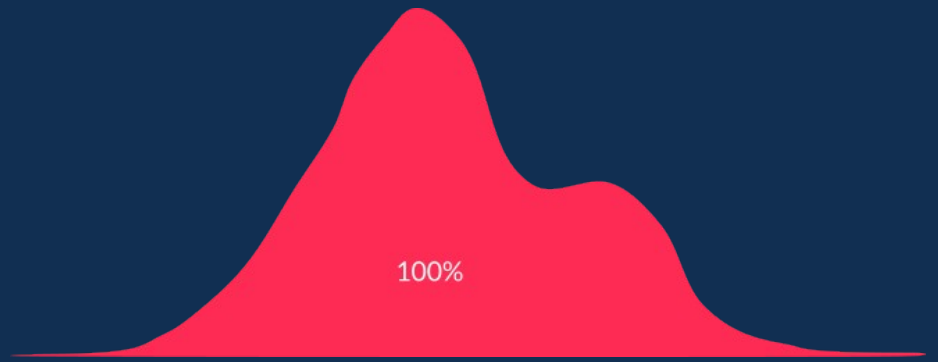
Binomial Distribution

- Binomial formula
  - Probability of x successes



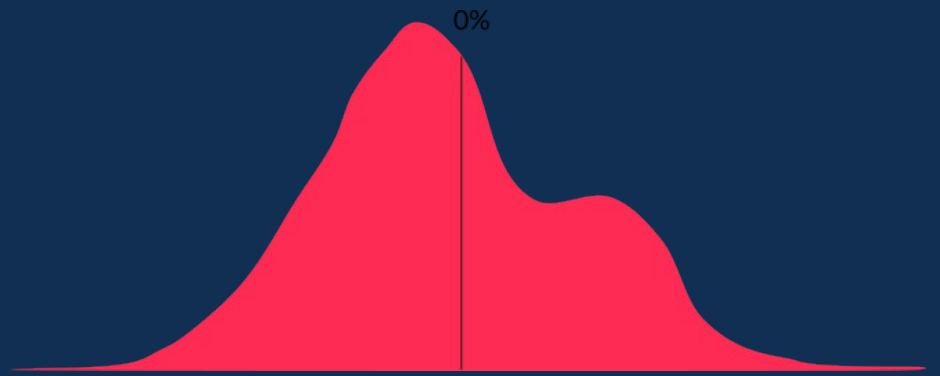
Continuous Distribution

- Probability density function represents a curve
  - Area under curve **adds up to 1**



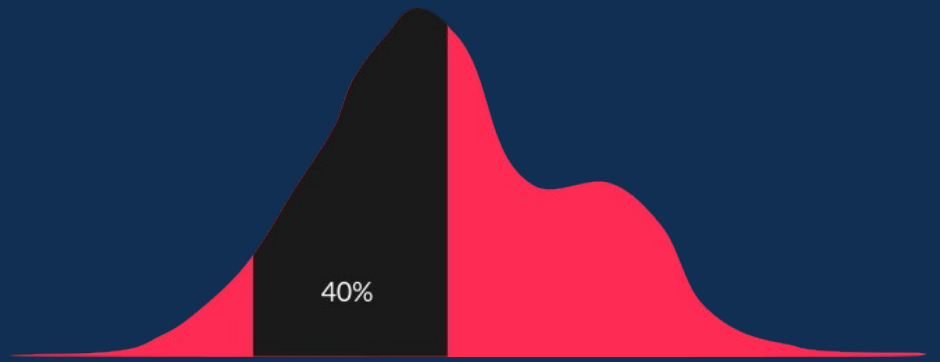
Continuous Distribution

- Probability density function represents a curve
  - Probability of **single value** = 0



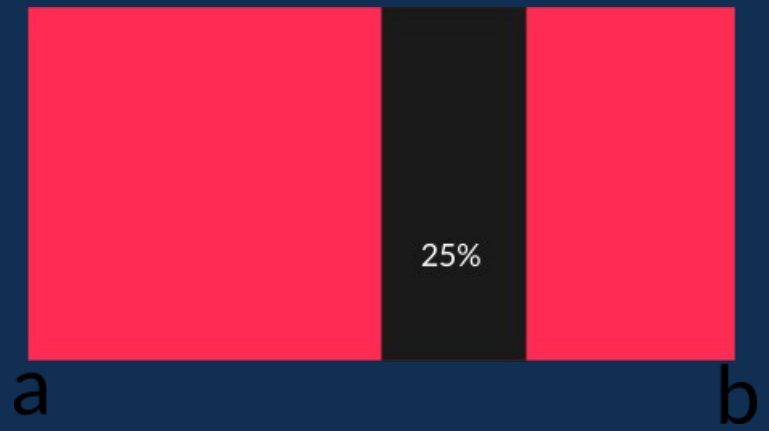
Continuous Distribution

- Probability density function represents a curve
  - **Area under the curve** is probability of **range** of values



Continuous **Uniform** Distribution

- Continuous
  - **Range** a → b
- Uniform
  - All **same probability**

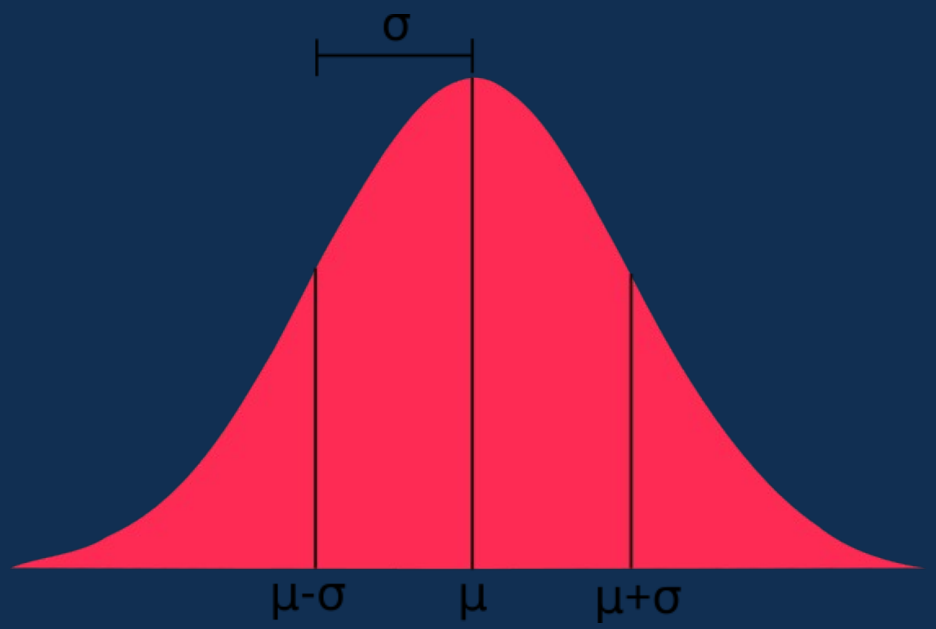


Normal Distribution

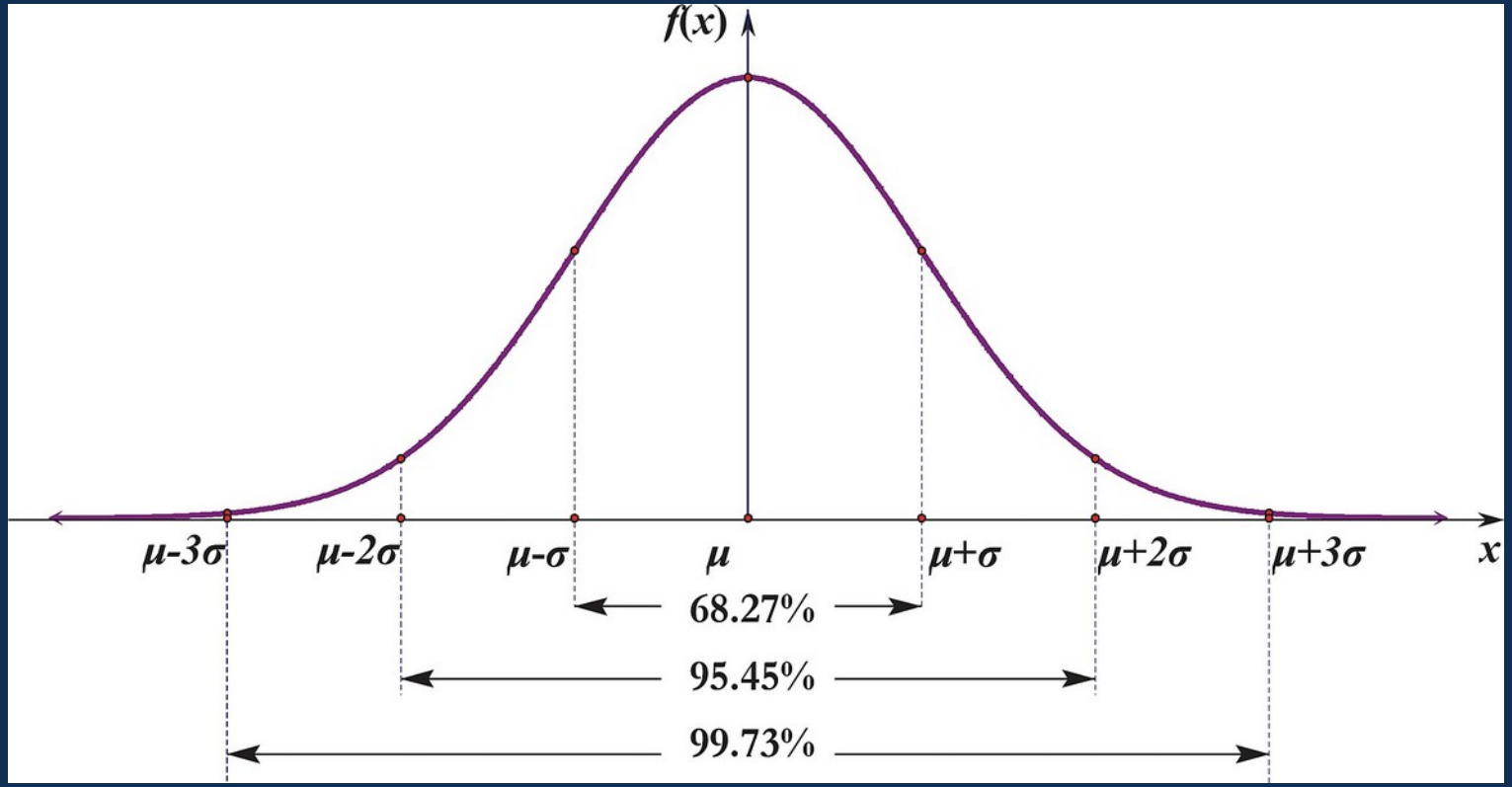
- Measure height of everyone in a class
  - 1.56 1.37 1.14 1.38
    1.09 1.18 1.68 1.22
    1.15 1.43 1.58 1.59
    1.02 ….

Normal Distribution

- Defined by:
  - mean (μ)
  - standard deviation (σ)
- Used a lot in inferential statistics
- Also called a Gaussian Distribution



Normal Distribution



Exercise 13:

Which of the following are discrete, which are continuous?

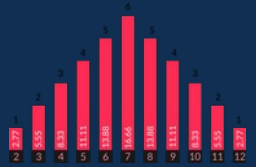
1. Rolling a biased dice
2. Java's Math.random()
3. A random integer 0-100
4. A binomial distribution
5. A gausian (normal) distribution

Exercise 14:

1. What is the liklihood of getting 0.5 on a continuous uniform distribution of numbers from 0-1?
2. What do the probabilities in a probability mass function sum to?
3. What do you get if you integrate a probability density function?

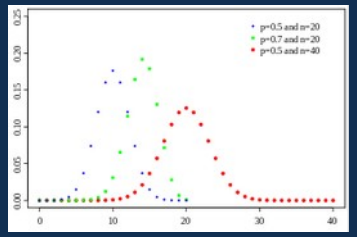
Summary

- Discrete vs. Continuous



- Uniform

- Binomial



- Normal

Summary

Summary

- Statistics
- Measures of Central Tendancy
- Measures of Spread
- Data Visualisation
- Correlation
- Probability Distributions