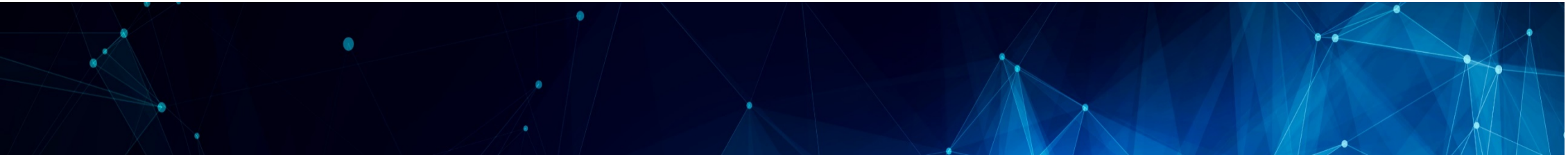
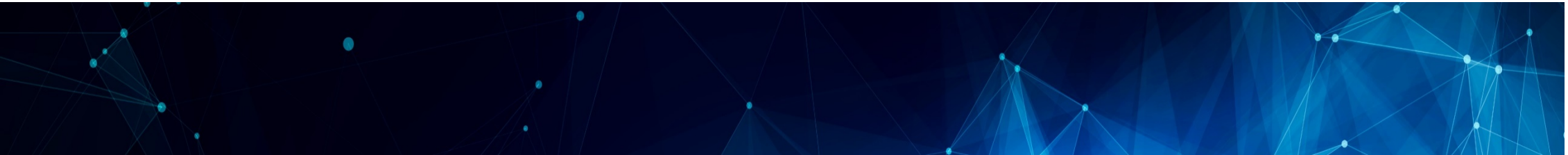
Fundamentals of Data Science

Statistics for Data Analysis



Southampton Data Science Academy

Statistics & Machine Learning

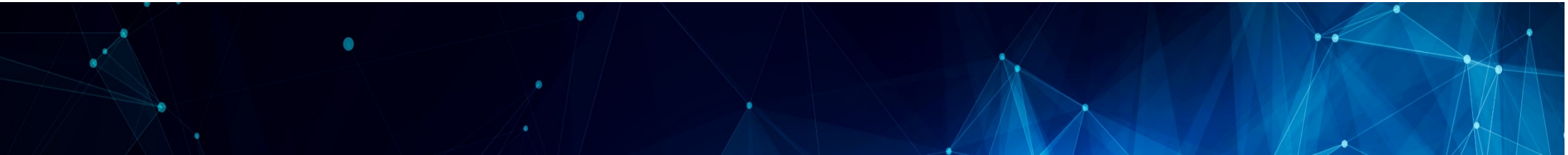


Southampton Data Science Academy

Outline

This week we will:

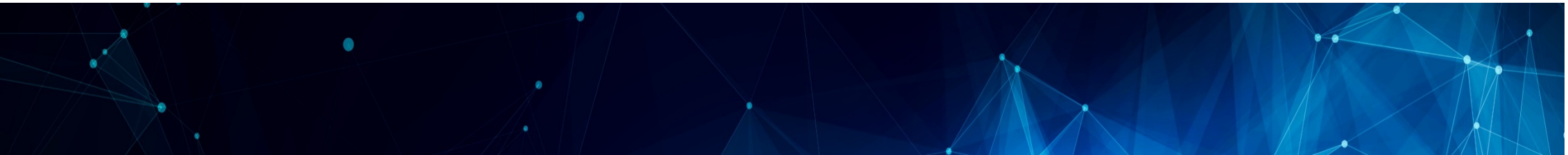
- Describe the role of statistics in data science
- Describe essential techniques to solve statistical problems in data science
- Describe the purpose of machine learning
- Compare three basic machine learning algorithms



Southampton Data Science Academy

Terminology

Data	Collected observations (documents, polls, recordings)
Statistics	Collect, analyse, summarise, interpret and draw conclusions from data
Population	Complete set of elements we would like to study (set of messages from a particular user, all web documents, etc.)
Sample	Smaller subset of population. Good sample is collected uniformly at random



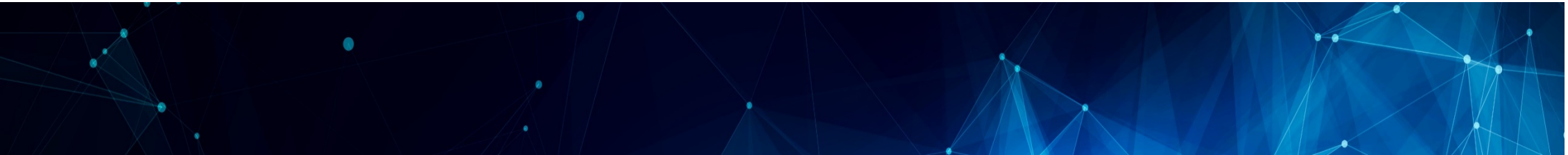
Southampton Data Science Academy

Why statistics?

Describing properties of a given data set (Distributions, typical element -mean/median-, variance, co-dependencies

Identifying what is special about given data (comparison to reference models)

Comparing datasets



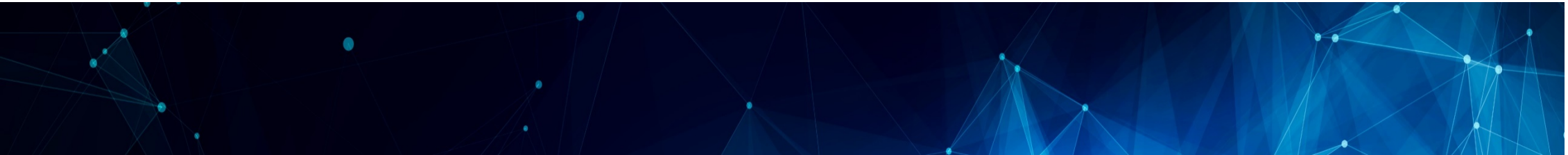
Southampton Data Science Academy

The nature of statistics

Related to, but distinct from, probability

- Probability: “I’ve got a data generating process (e.g. throwing two dice), now tell me what outcomes I can expect.”
- Statistics is the inverse: “I have some outcomes (i.e. data); what can I infer about the process that generated them?”

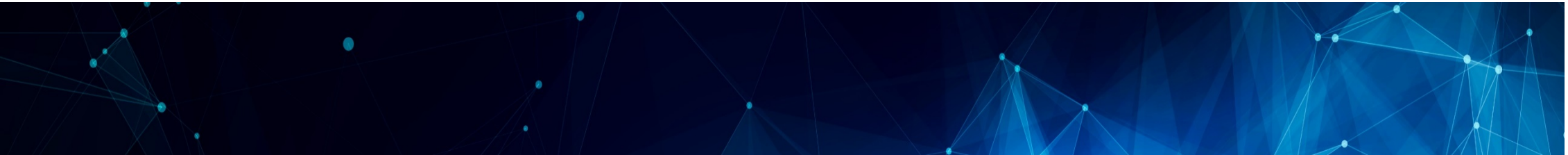
Central tendencies, variance and distributions.



Southampton Data Science Academy

Major properties of a dataset

Centre	The middle, or typical element within the dataset
Outliers	The untypical elements within the dataset
Variation	A measure of variety of a property value along all dataset elements
Distribution	Distribution of values within the data set



Southampton Data Science Academy

Mini Cooper example dataset

Description	Year	Mileage	Price
2003 Mini One 1.6 in Bright Red 92K Full Mot immaculate condition great 1st car	2003	92000	2999
STUNNING MINI CONVERTIBLE LOW MILES SERVICE HISTORY LONG MOT	2004	59000	3495
Mini Countryman Cooper 1.6 D 5dr DIESEL MANUAL 2013/62	2013	61000	8975
12 MONTH WARRANTY! (2009) MINI ONE 1.4 Pepper Pack New Model WHITE One Owner- MINI History- Top Spec	2009	56000	5995
12 MONTH WARRANTY! (55) MINI COOPER Chilli Pack WHITE Lady Owned- Immaculate- MINI History- Top Spec	2006	60000	5495
2010 MINI COUNTRYMAN COUNTRYMAN 1.6 COOPER S ALL4 5DR SAT NAV! LEATHER!	2010	26400	10995
MINI Cooper 1.6 low mileage	2007	42300	4500
2004 MINI Convertible 1.6 Cooper S 2dr	2004	85000	4490
Mini Cooper 1.6 red full service mot ?1575	2003	130000	1575
2005 Mini 1.6 (Pepper) One Convertible Cabriolet Low Insurance Power Hood FSH	2005	75000	3999
2008 mini	2008	90000	3395
2015 MINI HATCH 1.5 COOPER D AUTO HATCHBACK DIESEL	2015	12000	277
2008 Mini Cooper	2008	65000	4195
Mini Mini 1.6 Cooper	2011	59000	6995
2014 MINI Hatch 1.5 Cooper 3dr (start/stop)	2014	5065	9450
2013 Mini Mini Countryman 1.6 One salt 5 door FINANCE AVAILABLE PX SWAP	2013	21700	9995
2015 MINI HATCH 1.5 COOPER D AUTO HATCHBACK DIESEL	2015	12486	14450
2005 Mini 1,6 cc	2005	95000	2695
2004 54 Mini Mini 1.6 Cooper S WITH FULL JCW BODY KIT FINANCE AVAILABLE PX SWAP	2004	74000	3995
2004 Mini Cooper 1.6 Petrol 12m MOT - Low Miles!	2004	68000	2495
2003 MINI HATCH COOPER 1.6 COOPER FULL LEATHER, STOCK CLEARANCE	2003	120000	2290
Mini Cooper S (supercharged) - Panoramic roof	2002	134000	1550
Mini Cooper S SuperCharge	2003	95000	2300

Table 1: Example dataset, 25 most recent adverts for Mini Cooper cars.
https://www.gumtree.com/cars/uk/mini (accessed: 21.08.2016)

Central Tendency in the Mini Cooper data set

Central Tendency examples (typical value)

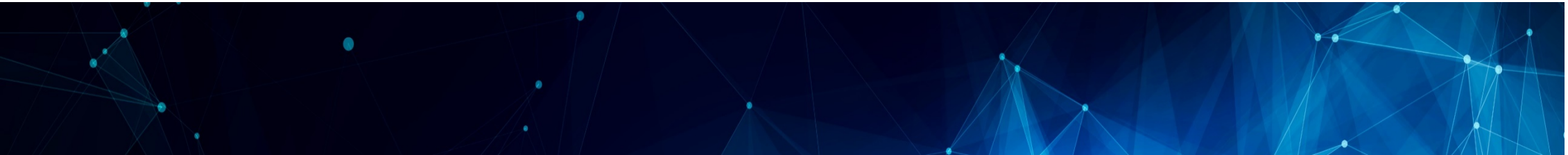
- **Arithmetic Mean:** the average score (x) within the data set

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

- **Median:** the middle score
  - Order by value, take the centre
  - 1,2,2,3,**5**,6,7,8,10 = **5** (odd number of elements)
  - 1,2,2,**3,5**,6,7,12 = (3+5)/2=**4** (even number of elements)
- **Mode:** the most commonly occurring score

	Price	Mileage
Mean	4644.18	71589.1
Median	3695	73000
Mode (1)	5995	99000

Table 2: Central Tendency in the Mini Cooper data set



Southampton Data Science Academy

Variation in the Mini Cooper dataset

Examples of the score spread measures within a dataset

- **Range**: Distance between the maximum and the minimum scores

$$Max(score) - Min(score)$$

- **Variance**: Average squared distance between the mean (μ) and every score(x) in the data set.

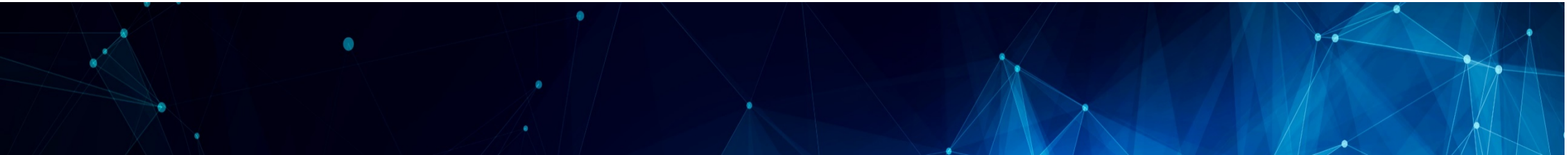
$$Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

- **Standard Deviation**:
  The average distance of the values to the mean (square root of the variance)

$$\sqrt{Var(x)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

	Price	Mileage
Range	(800-9000) 8200	(18500-124500) 106000
Variance	3260611.96	638549084
Standard deviation	1769.23	24758.98

Table 3: Variation within the Mini Cooper dataset



Southampton Data Science Academy

Frequency Distributions

A distribution can be a set of measurements, numbers, or data points.

Frequency Distribution: A histogram that displays the frequency of various outcomes in a sample.

- E.g The number of heads and tails from tossing a coin.
- E.g The term frequency in a document set

Distribution visualisation example

Distribute values into equidistant bins

Round price to nearest 500, mileage to nearest 10000m

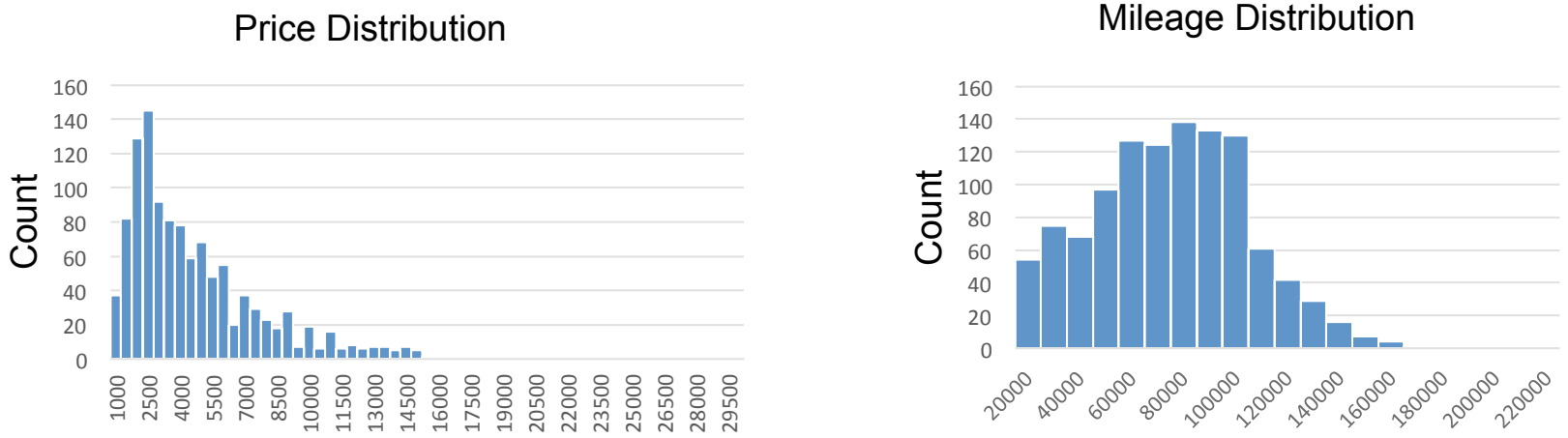


Image 1: A histogram of observations from the price and mileage distribution from adverts for 1155 mini Cooper

Skewness of the distribution

The direction and degree of ‘lopsidedness’ in the distribution. A *left-* or *negative-skewed* distribution is “right-walled”, and vice versa.

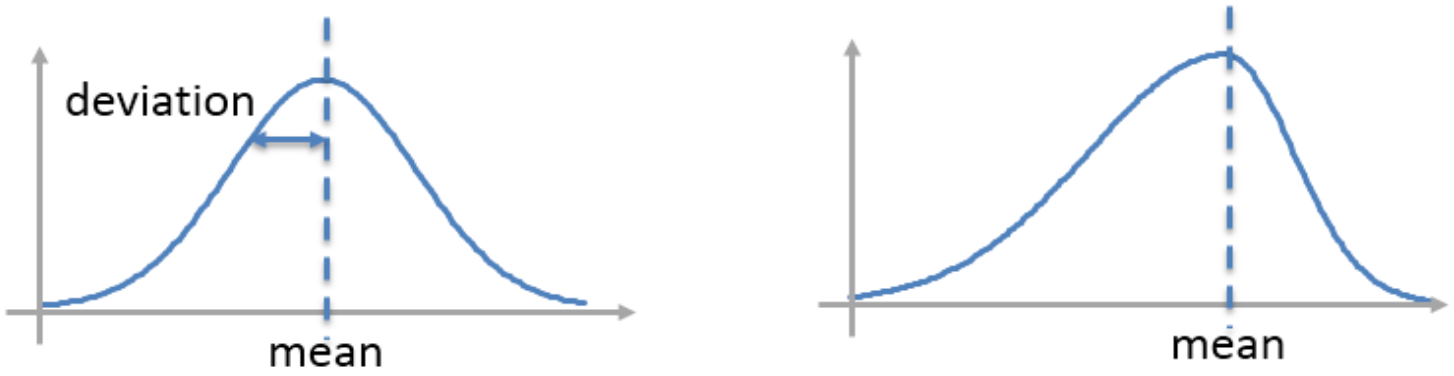
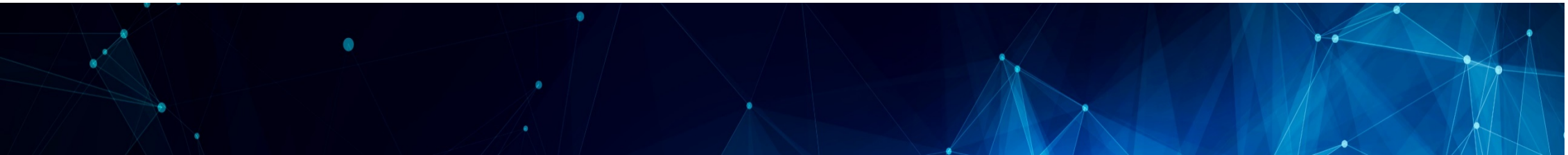


Image 2: Examples of a symmetric (a) and a left skewed (b) distribution

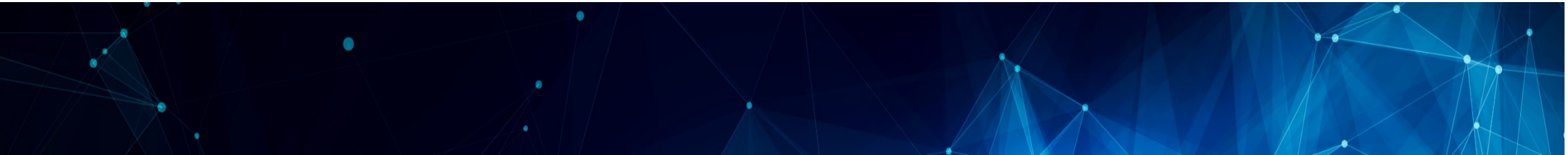
Sampling and central limit theorem



Southampton Data Science Academy

Description of a Population

As noticed earlier, often the population we would like to examine can be infinite. In the next chapter we will discuss, how the characteristics of such population can be estimated using sampling.



Southampton Data Science Academy

Populations and samples

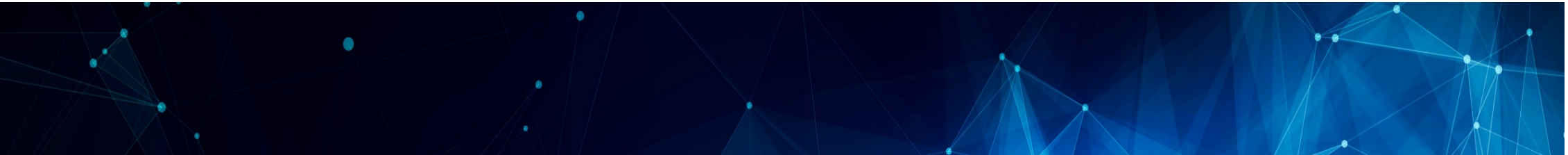
A Population includes all items or events relevant for the measurement. It represents the big picture.

A Population can have a small size, be large, or even infinite. For example:

- A set of messages from a particular user.
- All web documents.
- All documents generated using English language

Sampling

- We often don’t know the shape of a population’s distribution (all fish in the ocean or all documents on the web).
- In order to discover it, we need to sample the population
- Sample can provide information about the population but needs to be selected methodically at random and be of sufficient size
  - **Empirical sampling** – in order to learn about the distribution of term frequencies select randomly 1000 web documents
  - **Repeated runs** of a process simulation



Southampton Data Science Academy

Data generation processes

A process may have an infinite number of outcomes and an infinite extent

- We would like to look at how the predictions are distributed across an infinite number of runs. For example: Throwing two dice and estimating the probability for each score (total of two dice)

Probability density functions

Illustrates what the long-run distribution of a process looks like

Histogram that shows the probability of an outcome, rather than a count of occurrences

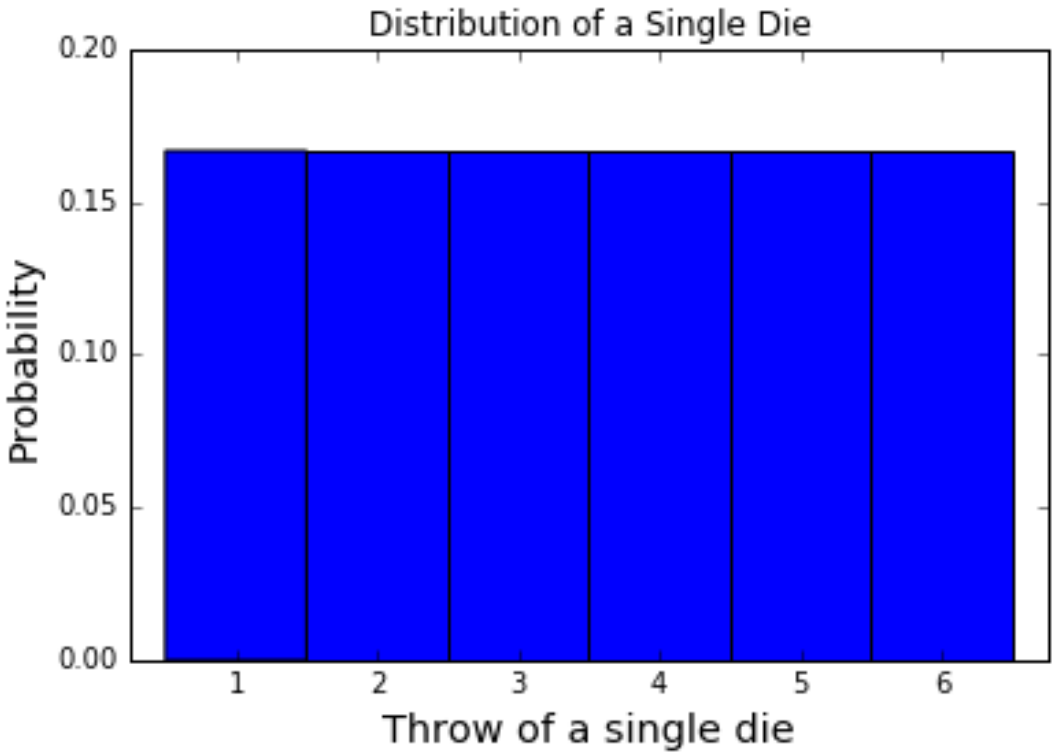


Image 3: Probability density function for the outcome of a die throw

Sampling from known distributions

Consider the results from throwing a single die:

- the uniform distribution across all digits 1-6
- this has a mean of 3.5, variance of 2.917 and standard deviation of 1.708

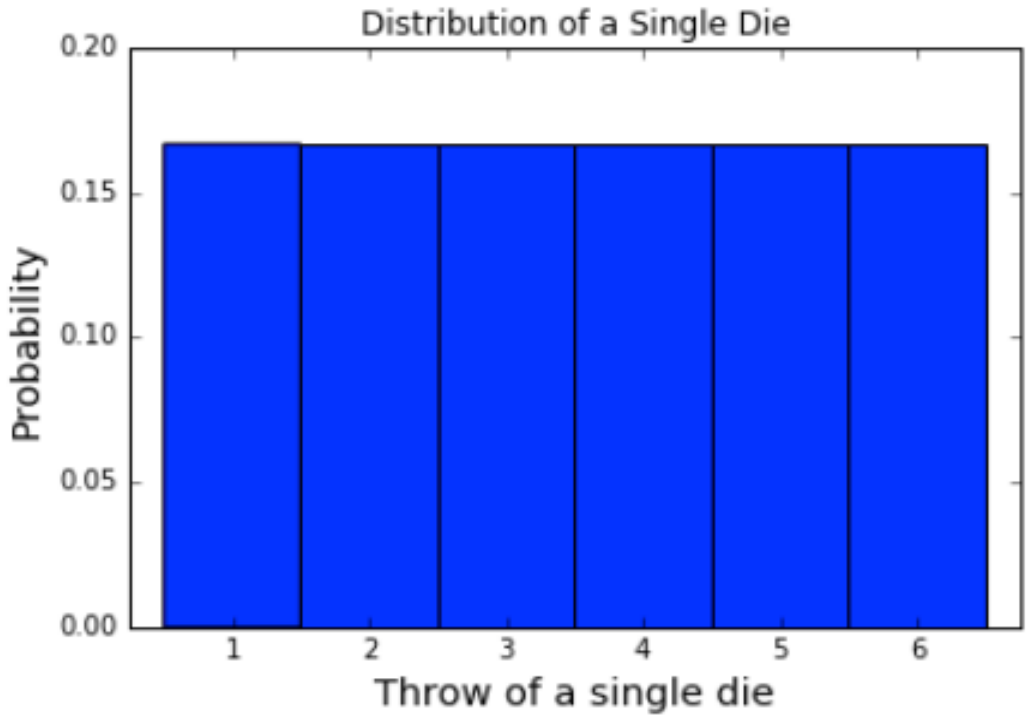


Image 4: Expected probability distribution for a fair dice

Sampling from known distributions (2)

We can simulate drawing some samples to see how the sample size affects our attempts to draw conclusions.

- For example, a sample of size n=1 would be a single variate from the population (one throw of the die).

Varying the size of the sample

In both cases, the shapes did not represent that of the true distribution. The bigger sample returned a more accurate estimate of the population mean.

How much variation should we expect if we keep drawing samples of a particular size? E.g. say we drew a sample of size n=3, 10,000 times

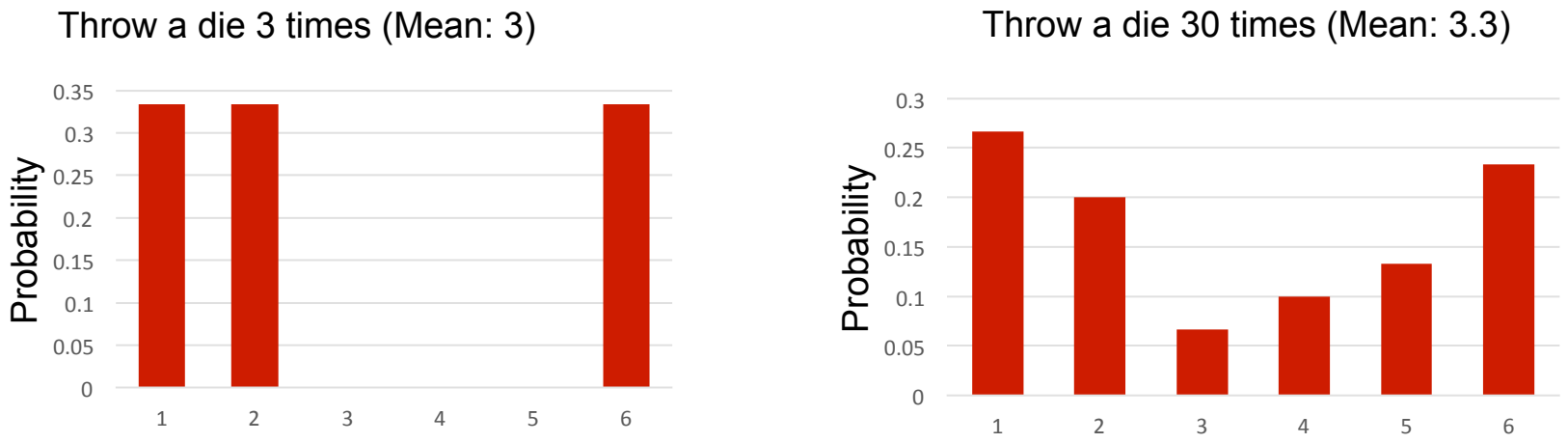


Image 5: Example score distributions in samples of different sizes

Sample distribution of the mean

It looks like the mean of the sample means centres in towards the true mean of 3.5.

- there is a high variation.
- with a small sample size, extreme results (sample mean of 1 or 6) occur quite often

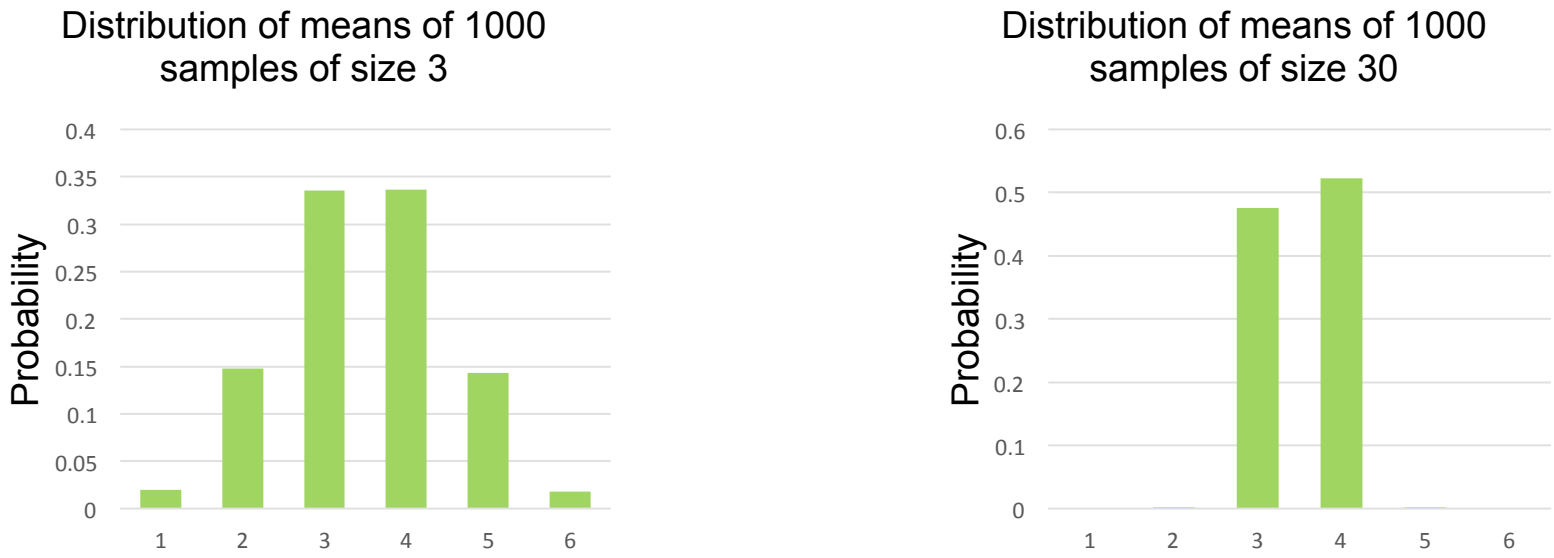


Image 6: Example distribution of means for different sample sizes

Sample distribution of the mean (2)

- The distributions of sample means are shaped like a bell curve
- The width of the bell curve gets smaller as the size of the samples increases
  - Bigger samples give more accurate estimates
- Even with really small sample sizes, the distribution appears centred around the true mean though any particular sample could be a big outlier.

Central Limit Theorem

States that if the number and the size of the samples is sufficiently large, then the distribution of the mean will be approximately normal, converging on the **true population mean** (regardless of the shape of the original distribution).

Standard error is the standard deviation of the sample mean:

- If sample size > 30 the distribution will be close to normal

Mini Cooper price example

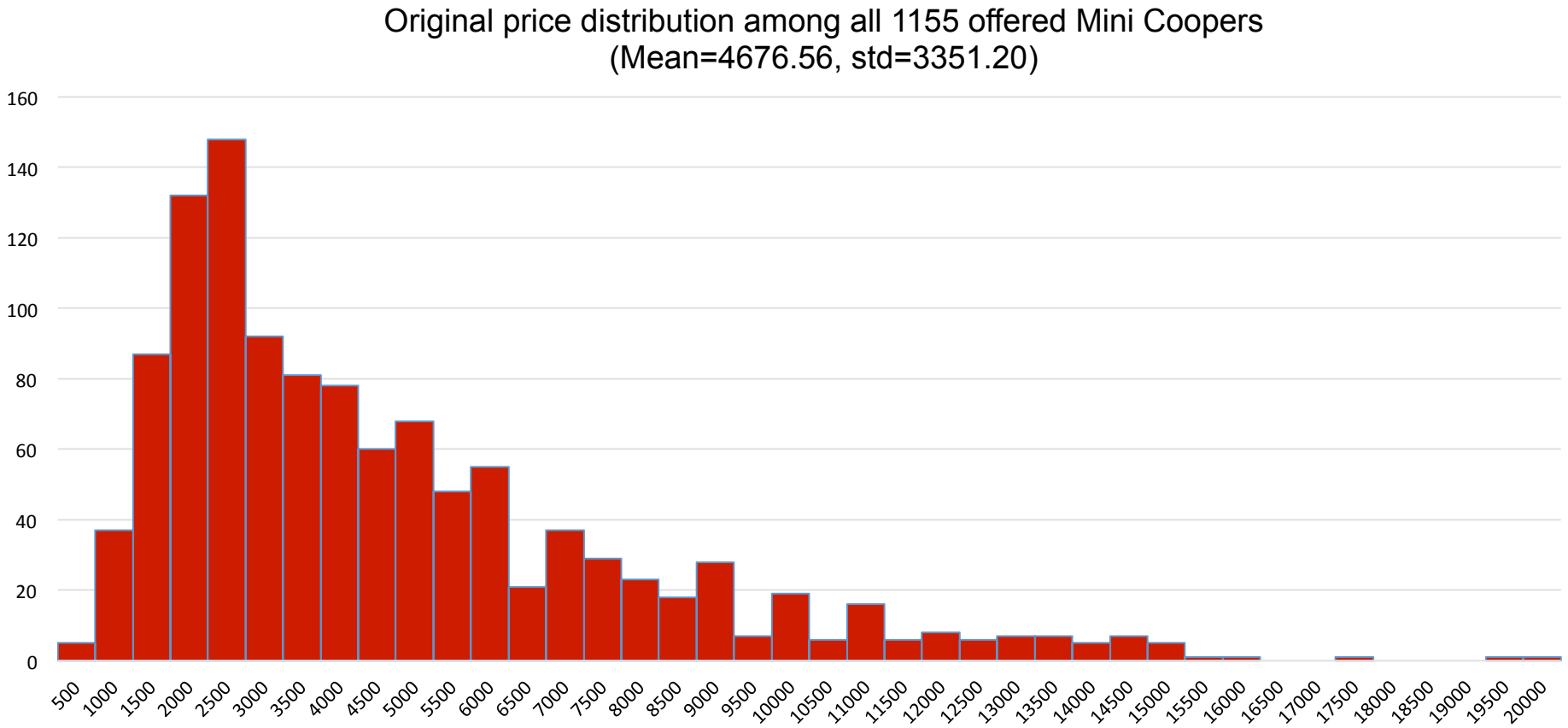
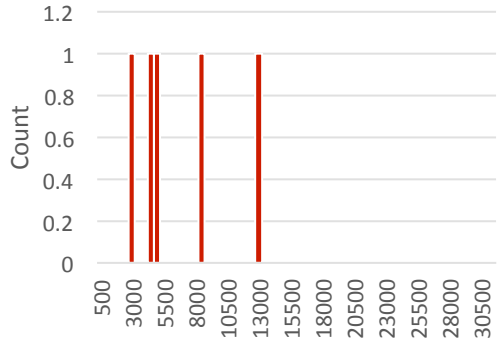


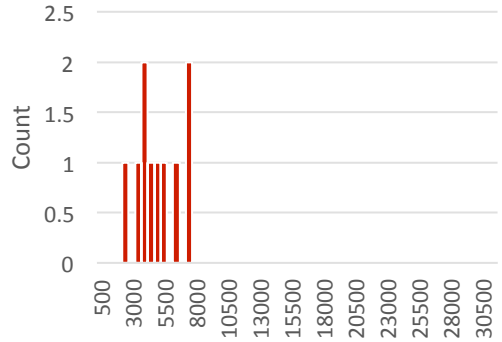
Image 7: Price distribution of Mini Cooper cars. Data source: https://www.gumtree.com/cars/uk/mini (accessed: 21.08.2016)

Sampling procedure

Sample of size 5 (Mean: 5786.0)



Sample of size 10 (mean: 3546.5)



Sample of size 30 (Mean 3733.27)

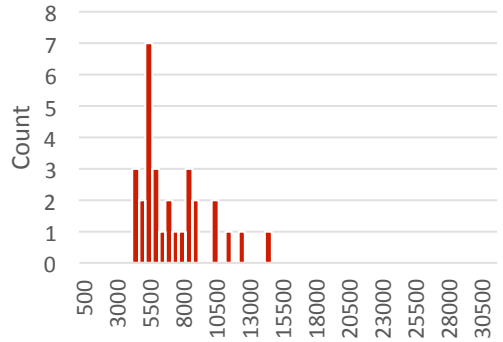
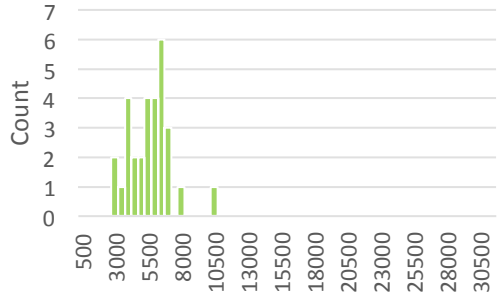
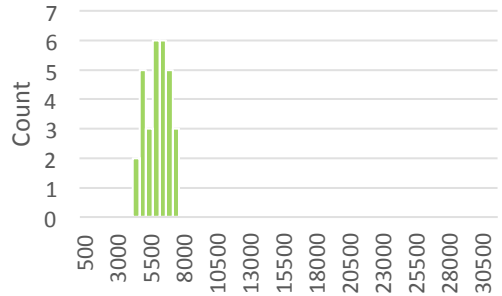


Image 8: Sample mean tends to be normally distributed regardless of the initial distribution shape

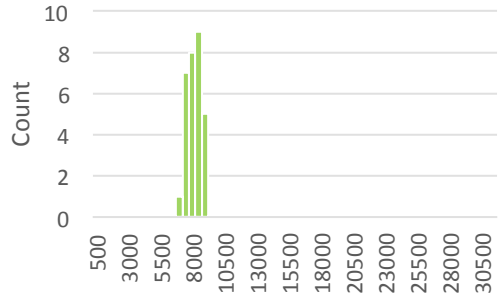
30 variates of sample size 5 (Mean: 4638.7, standard error: 1589.52)



30 variates of sample size 10 (Mean: 4615.5 , standard error: 908.11)



30 variates of sample size 10 (Mean: 4689.37, standarederror: 570.13)

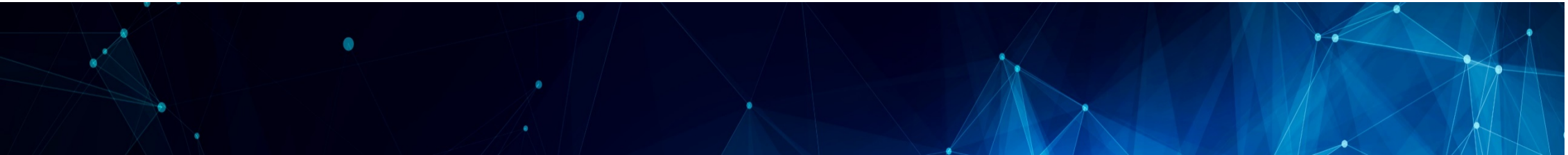


Southampton Data Science Academy

Sample mean summary

If we repeatedly sample from a population, with samples of size N:

- The mean of sample means will converge on the true mean of the population.
- The sample error will tighten up in proportion to: $\frac{1}{\sqrt{N}}$
- So accuracy improves with bigger samples, but with diminishing returns.

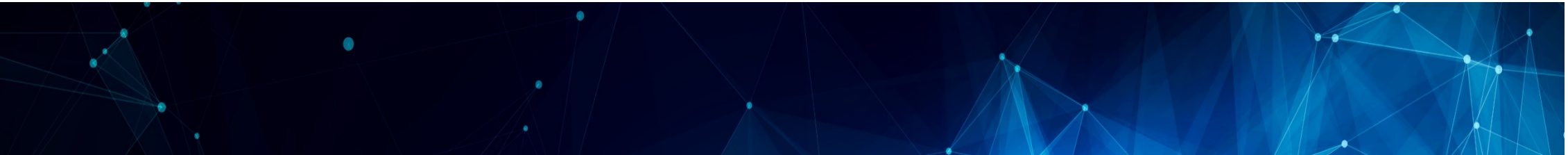


Southampton Data Science Academy

Summary

In this Chapter we learned how statistics can describe a dataset and how sampling can estimate properties of a Population

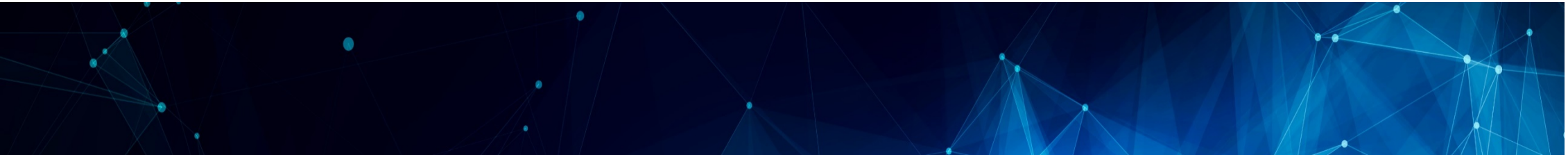
In the next chapter we will show how machines can use statistics to analyse given data to learn from it.



Southampton Data Science Academy

Machine learning

Taking statistics further for data science

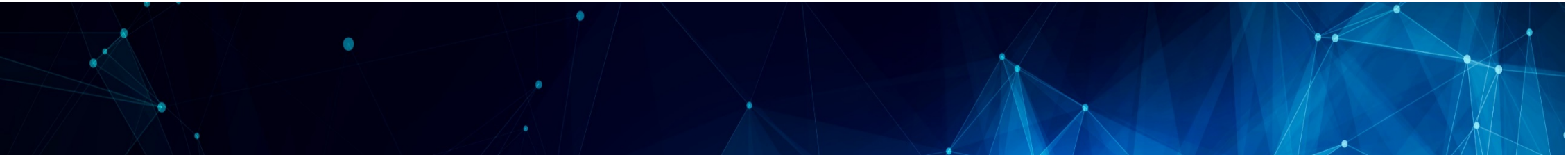


Southampton Data Science Academy

Introduction to Machine Learning

“The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience”.

Tom Mitchell: *Machine Learning* 1st Edition (1997)



Southampton Data Science Academy

Types of machine learning

There are two main types of problems that machine learning targets

- Regression analysis: can we predict a ‘response’ variable from a particular ‘predictor’ variable?
    - E.g. predicting a house price from the square meters or car price from its mileage
- Classification: can we determine where a particular instance belongs to?
    - E.g. is a particular spam email or not?

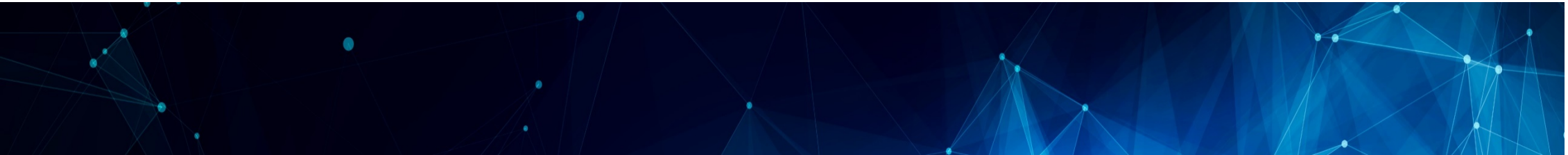
Unsupervised vs supervised learning

Unsupervised

- Dataset does not have a labelled response variable
- Common for clustering analysis (identifying similar groups or patterns)

Supervised

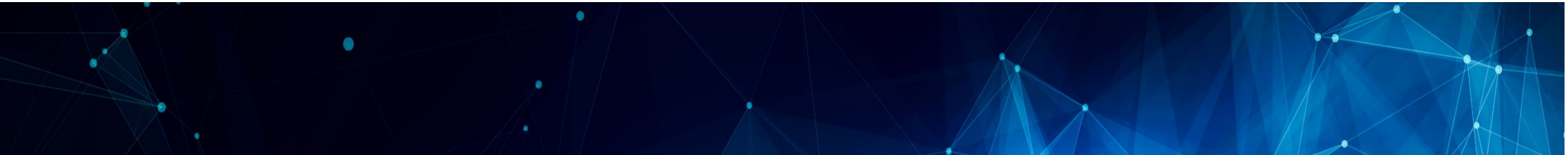
- Uses a known dataset to make future predictions or classifications
- The dataset contains both predictor and response labels



Southampton Data Science Academy

3 basic Machine Learning algorithms

- Linear regression
- Support vector machines (SVM)
- Naïve Bayes



Southampton Data Science Academy

Linear regression

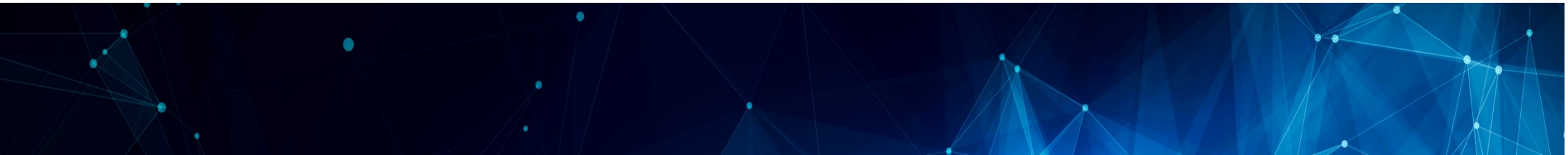
- Predict Y based on X
- Can we predict a price of the car based on its mileage using a training dataset

Example dataset

(25 most recent adverts of Mini Cooper cars made between 1990 and 2000)

Description	Year	Mileage	Price
1991 Classic Mini Mayfair Silver All Original Genuine 18500 miles	1991	18500	5245
Mini Mayfair 998 56k miles 12 months mot	1990	56000	3500
Rover mini Sprite 1275 classic mini	1994	78000	4450
1998 Rover Mini Cooper MPi, MOT til October, Modified	1998	96000	2450
Mini 30	1989	80000	3000
CLASSIC ROVER MINI TAHITI LIMITED EDITION	1994	72000	5495
1992 R1 engined mini Mayfair	1991	30000	8000
Rover Mini Rio 1275cc	1993	78950	3000
MINI COOPER SPORTS PACK LOW MILEAGE	1999	39000	2500
Classic Mini Mayfair	1993	67000	4300
Rover mini 35	1994	124500	3000
Rover Mini Sidewalk 1996	1996	38000	3500
1990 ROVER MINI 1.3 CHECKMATE	1990	75000	3999
Rover Mini 1.3 i Sprite	1996	81000	2895
Classic mini rio 1275	1994	65000	3995
John cooper converted mini Mayfair	1990	55865	5500
Classic 1989 Rover Mini Mayfair	1989	110560	2995
classic mini cooper 91 plate	1991	88788	3500
Classic Austin rover mini	1989	100000	3100
1996 (N) ROVER MINI 1.3i	1996	64000	9000
1991 Classic Mini City E	1991	68000	800
Rover Mini 40	1999	54800	5500
Rover mini	1995	84000	1800
Classic Mini Cooper sport. 1275cc	1996	43600	3000
Classic 1991 Mini Mayfair 998cc	1992	93700	2800

https://www.gumtree.com/cars/mini



Southampton Data Science Academy

Linear Regression

- consider a scatter plot showing ‘Price’ vs ‘Mileage’
- £5245 for a car with 18500m, £3000 for 124500m
- can we learn a model to predict price based on mileage?
- regression uses a predictor variable to adjust our estimates

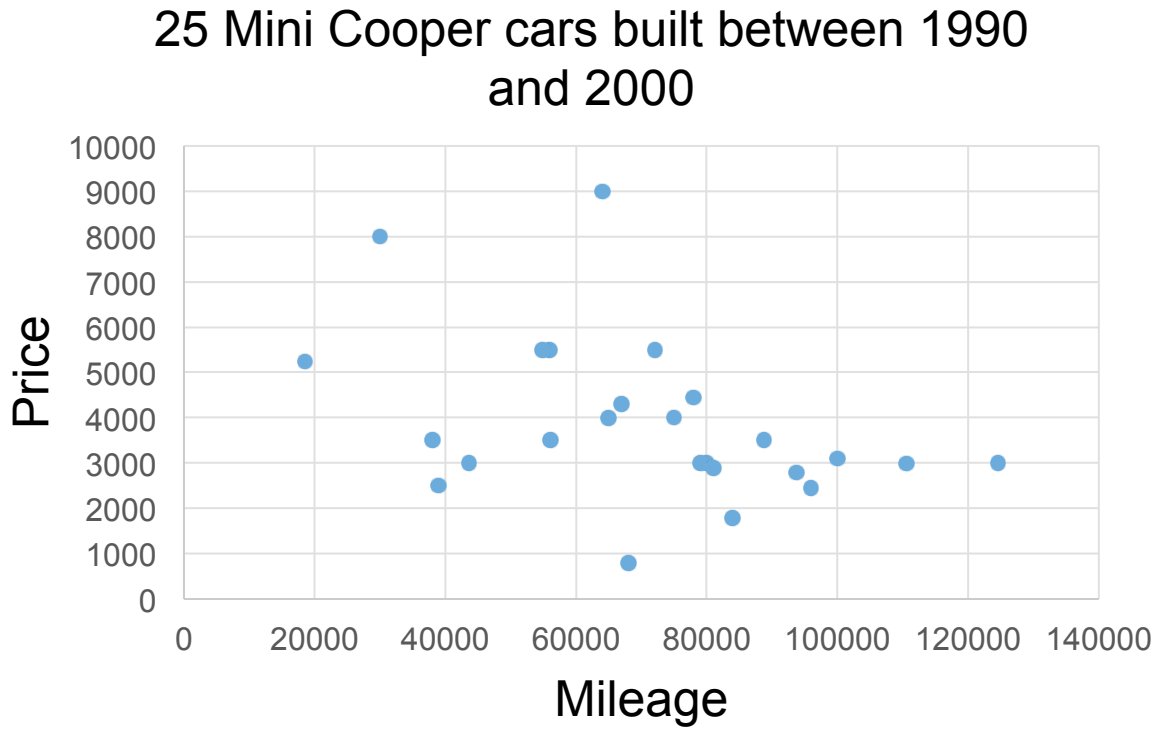


Image 9: Observed Price vs Mileage for mini cars built between 1990 and 2000

Southampton Data Science Academy

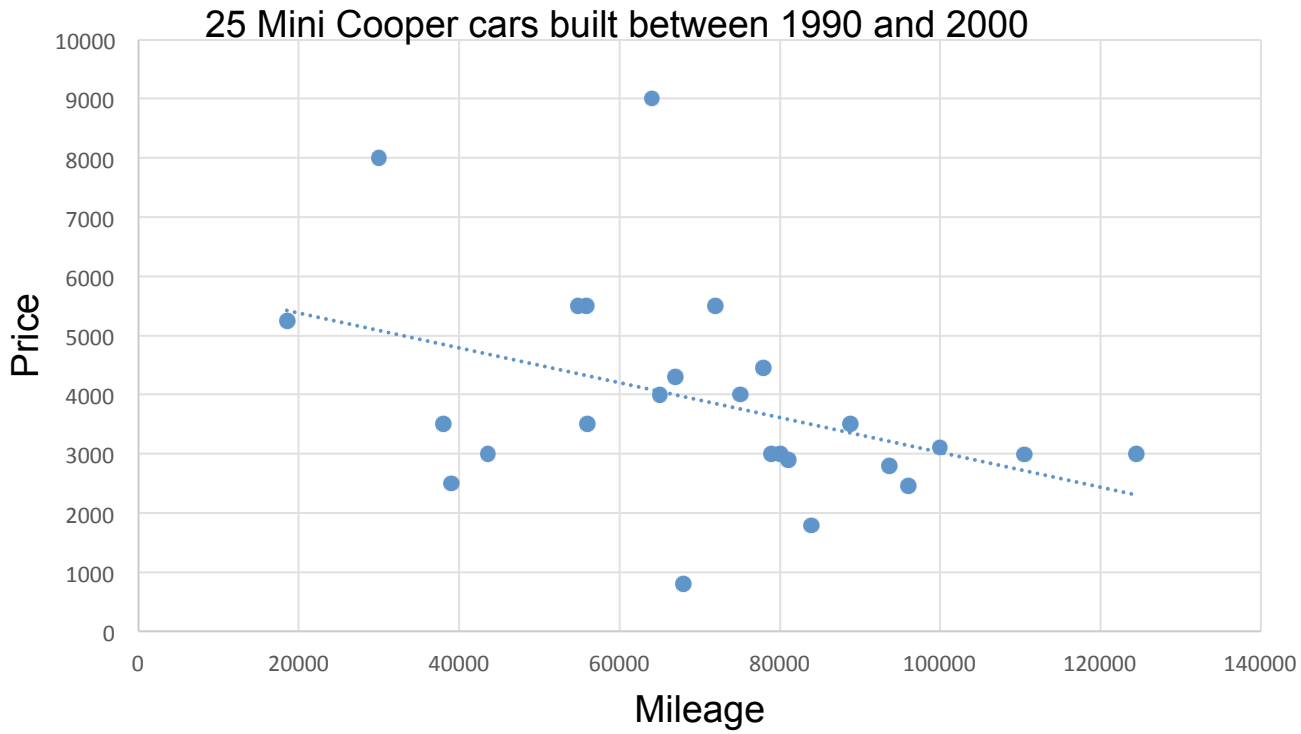
Linear Regression (2)

Generating a ‘line of best fit’

Computational problem:

- Long process involving minimising the least squares difference between predicted line and actual outcome

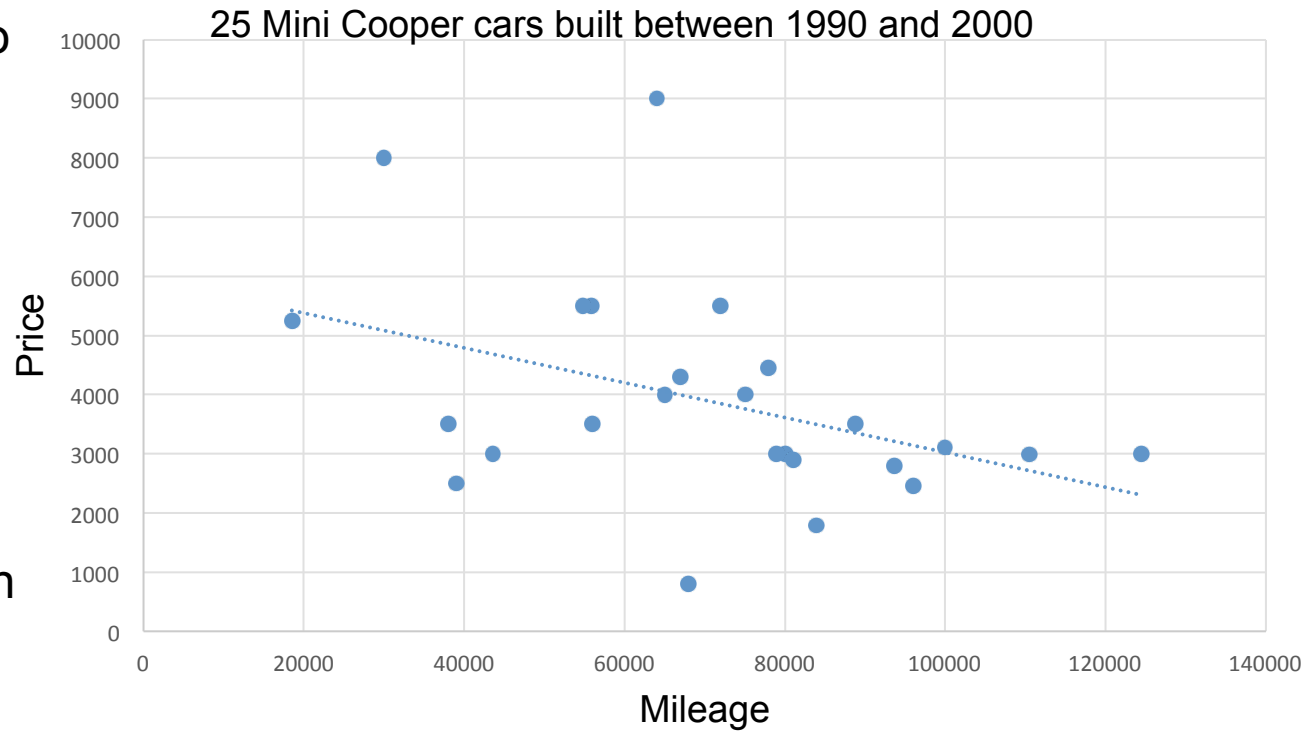
- Provides an equation of the form $Y↓i = (b↓0 + b↓1 X↓i)$



The regression process

- Any new values can be used in the equation to gain a new output y=mx+b
- The graph has an intercept (of the Y axis) of £5,967.9, the slope is = -0.0294 £/m
- y = -0.0294x + 5,967.9
- So for a mini built between 1990 and 2000 with mileage of 50,000, the value would be:

  Price = -0.0294*50,000 + 5967.9 = **£4497.9**



Southampton Data Science Academy

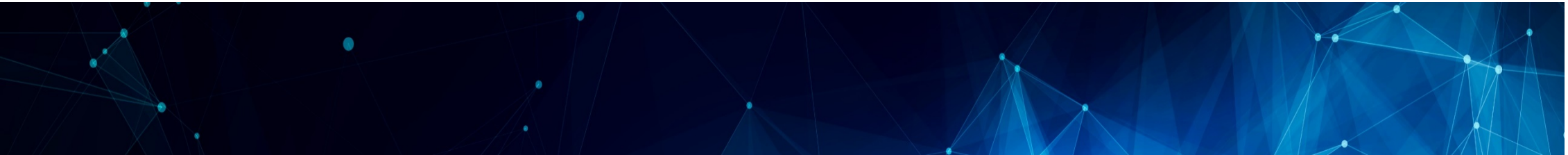
Regression Pitfalls

A model is only a model!

- There will always be error, so this is not the exact price you would get for the car.
- Confidence intervals are usually added to the equation to represent the accuracy.

Extrapolations are dangerous

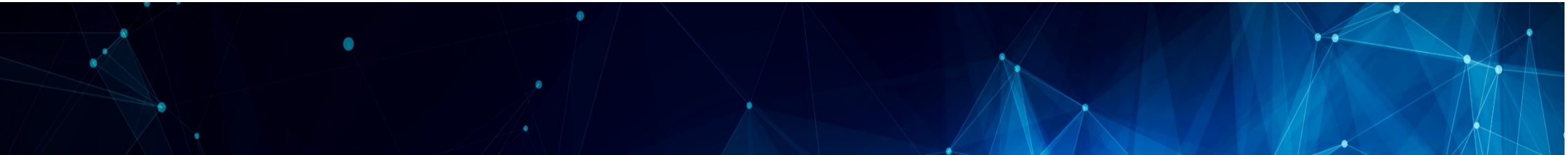
- It is defined only for the observed range on X and Y



Southampton Data Science Academy

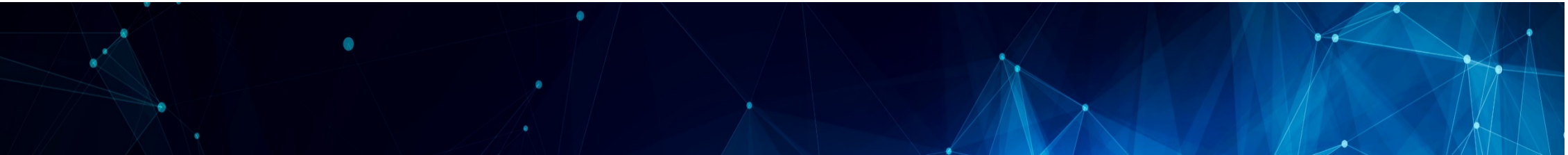
Summary

- Regression can help predicting the outcome value based on an observed value.
- In the next chapter we will see how a machine can learn to categorize objects.



Southampton Data Science Academy

Classification problems



Southampton Data Science Academy

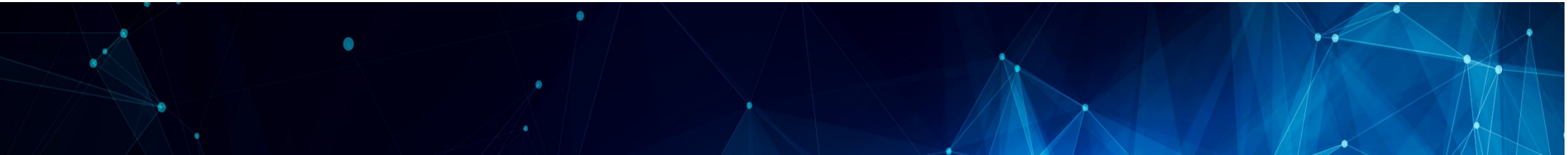
Classification problems

Classification problems are set up so that:

- an object has certain features (colour, size, number of terms etc.)
- a model has certain categories (cats/dogs or spam emails/ham emails)
- classification predicts which category the object should go in, based on its features.

For example

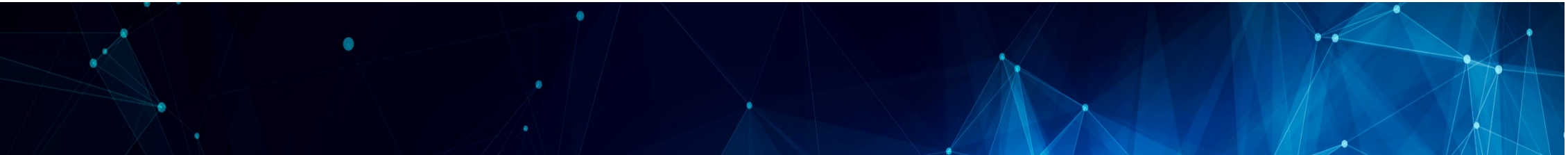
- language of a document
- sentiment analysis of text
- spam filters



Southampton Data Science Academy

Classifying spam email

- Classifying email is a common example
- Two techniques will be covered
  - Support Vector Machines (SVM)
  - Naïve Bayes
- Both use a set of labelled data where the correct category is known
  - This lets us train and test our classifier



Southampton Data Science Academy

Support Vector Machines (SVM)

- Define features, usable for classification
- Create a feature vector space
- Labelled data are placed according to their value
- A ‘hyperplane’ is used to delineate categories in the vector space
- The data closest to the separation hyperplane are called “support vectors”

SVM for email

- each word in an email is assigned a score, using a metric such as frequency
- “Profit” and “Guarantee” are both words that could legitimately be used in emails

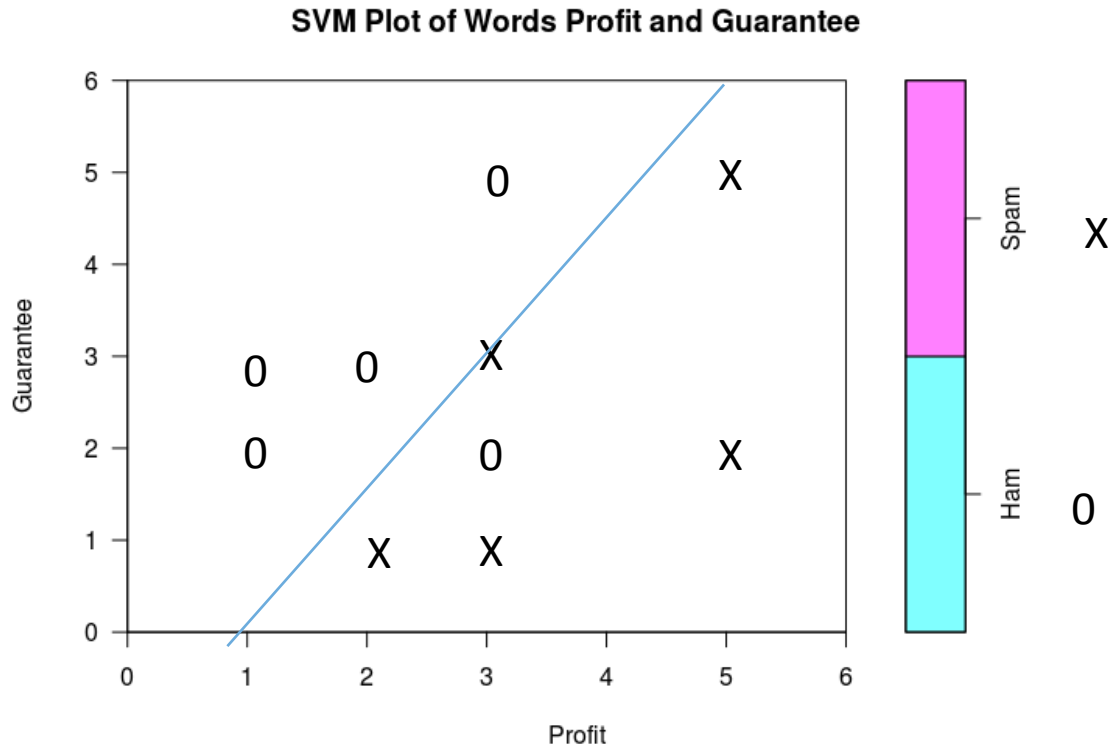


Image 11: Vector space model for spam classification

SVM for email

The plot shows data which are linearly separable

- A single straight line can cut the categories so that all data are on the correct side
- The hyperplane maximises the distance between the support vectors (the points closest to it)

This is repeated for all words in the email, producing a multidimensional plane called (hyperplane)

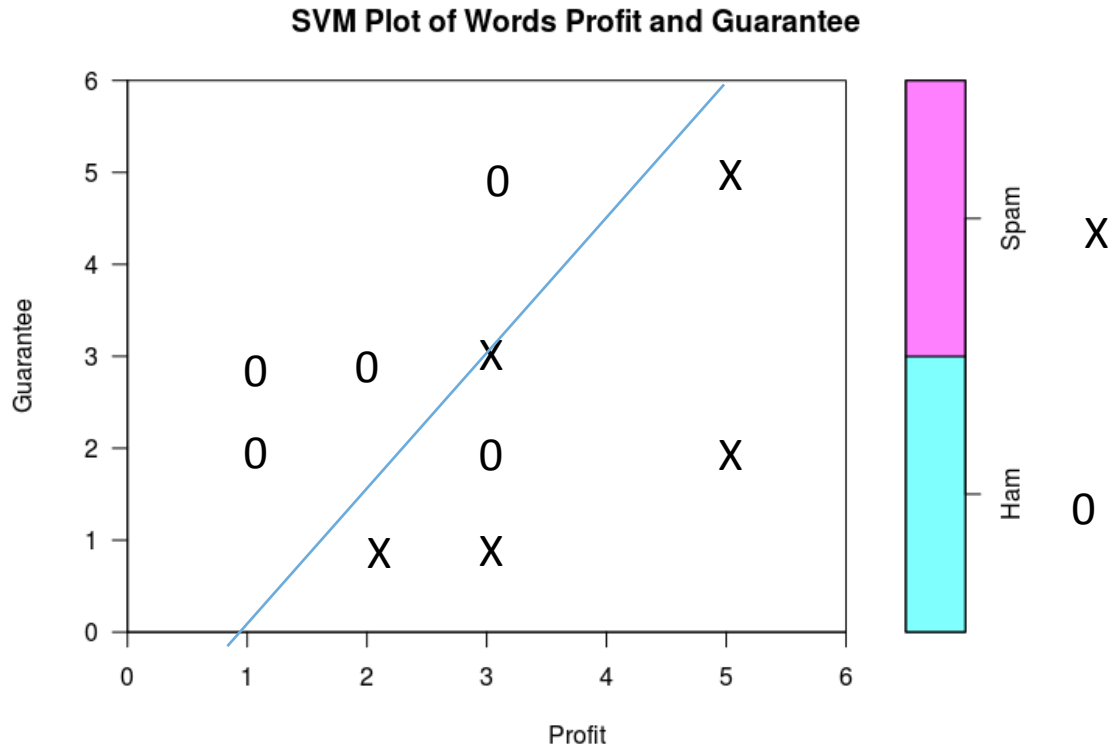


Image 11: Vector space model for spam classification

The main properties of SVM (summary):

- Performs well in high dimensional feature spaces
- Fast classification process (calculate, which side of the hyperplane a given object is placed)
- Relatively complex training process (iterative approximation)

Naïve Bayes

Naïve Bayes represents another family of classification algorithms and is based on probability theory

- Estimation of the probability for the a given instance to belong to a given class
- Estimation is based on the features distribution over the classes in the training set

“Naïve” assumes independence between the features

Spam example

- 50% of spam, 50% of ham emails in our sample
- the terms are: “Guarantee”, “Profit”

	Spam	Ham	
“Guarantee”	9	13	
“Profit”	18	3	
Sum	**27**	**16**	

Table 3: Term statistics

Training process

- First, for each class C we calculate the prior probability of an instance to belong to this class

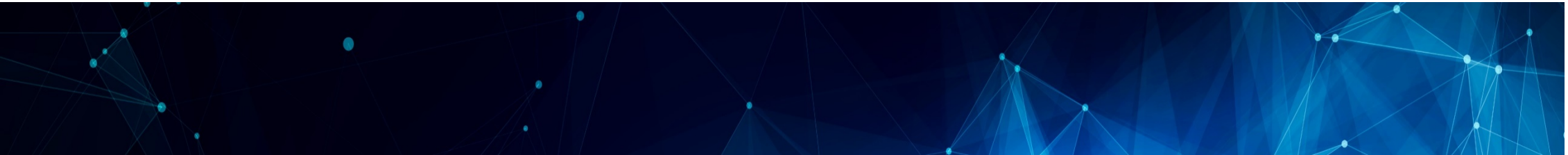
$$\mathrm{P(C)}\,\frac{Nc}{N}$$

- For each feature and class we compute the probability to for the feature to belong to the class

$$P(C|term) = \frac{count(term)}{\sum_{t \in terms\ in\ C} count(t)}$$

Training process example

1. P(spam)=P(ham)=0.5
2. P(spam|Guarantee)=9/(9+18)=0.33
3. P(spam|Profit)=18/(9+18)=0.66
4. P(ham|Guarantee)=13/(13+3)=0.81
5. P(ham|Profit)=3/(13+3)=0.19



Classification process

1. Given a set of features for an instance, multiply the probabilities to obtain the probability of the instance to belong to the class

$$P(C|instance) = P(C) * \prod_{t \in terms\ in\ C} P(C|t)$$

   be assigned to each term and class, even if absent in the training set
2. Select class with the highest probability

Classification example

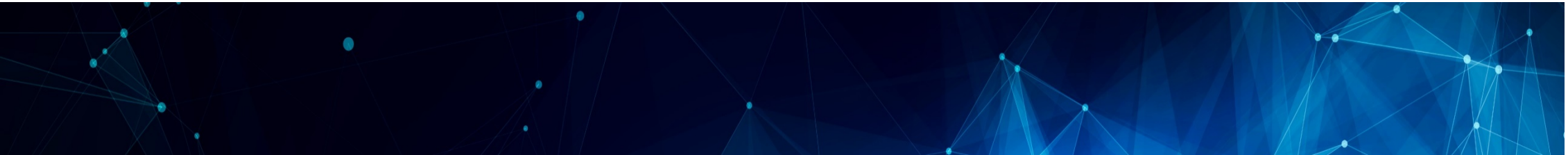
- an email contains only the term “Guarantee”
- P(Guarantee)=0.5*0.33 + 0.5*0.81 = 0.57
- P(spam|document)=P(spam|Guarantee)/P(Guarantee) = 0.5*0.33/0.57 =0.289
- P(ham|document)=P(ham|Guarantee)/P(Guarantee) = 0.5*0.81/0.57 =0.71

Machine learning: summary

- For a pair (X,Y) of highly correlated properties, linear regression helps to estimate the value of Y, given a particular value of X within the observed range
- The choice of SVM vs. Naïve Bayes should depend on the properties of the data.
- Often, SVM performs more accurate and efficient in classification of high dimensional data and Naïve Bayes is preferable when the size of example data is rather small or training times are critical.

Module summary

- In this lecture series we first concentrated on the role of statistics in data since and introduced techniques to describe a dataset in terms of the central element and score variation.
- Additionally, we learned how main characteristics of a large population can be estimated using sampling.
- Finally, we described the purpose of machine learning algorithms that can improve with experience and compared three popular machine learning algorithms



Southampton Data Science Academy