University of Lancashire

Data Analysis using Statistics

Karen O'Shea

Where opportunity creates success

Learning Objectives

University of Lancashire

Evaluate basic statistics for data analysis

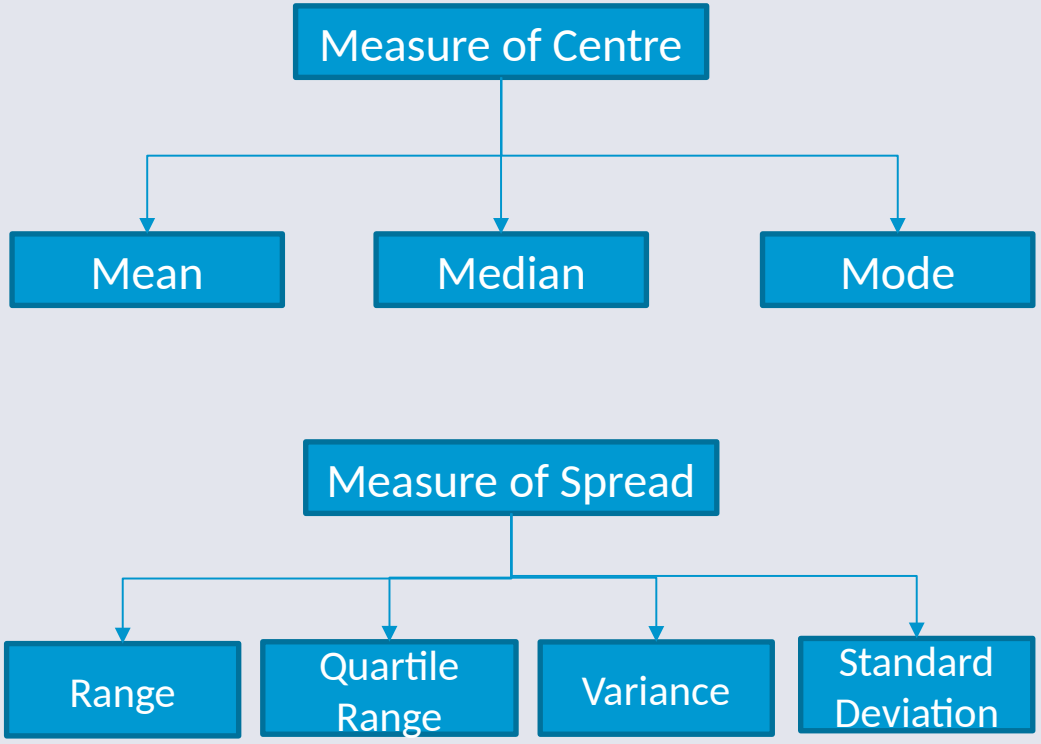
Examine descriptive and probability (inferential) statistics

Types of Statistics

- **Descriptive statistics** is a method used to describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. Descriptive statistics is mainly focused upon the main characteristics of data. It provides a summary of the data.

- **Inferential statistics** makes inferences and predictions about a population based on a sample of data taken from the population in question. Inferential statistics generalises a large dataset and applies probability to draw a conclusion. It allows us to infer data parameters based on a statistical model using sample data.

Descriptive Statistics Categories



Measures of Centre

University of Lancashire

There are three main measures of centre:

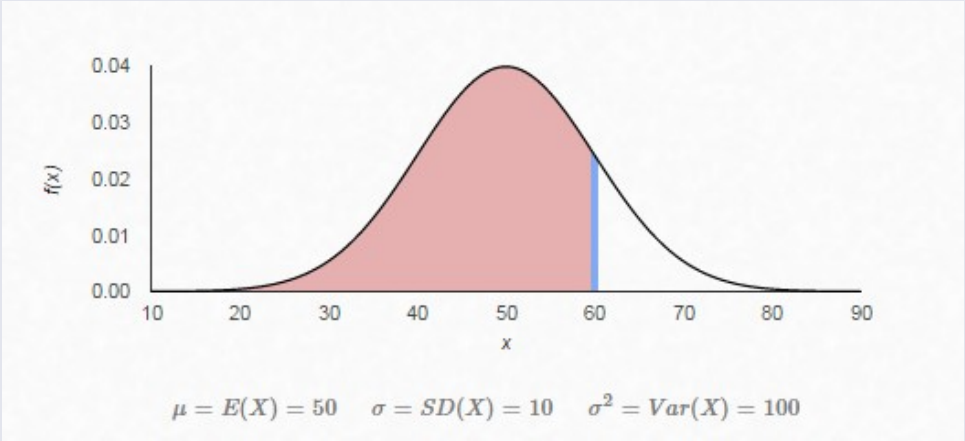
- **Mean**: Measure of the average of all the values in a sample is called Mean.
- **Median**: Measure of the central value of the sample set is called Median.
- **Mode**: The value most recurrent in the sample set is known as Mode.

Measures of Spread

- **Range**: It is the given measure of how spread apart the values in a data set are.
- **Quartile**: Quartiles tell us about the spread of a data set by breaking the data set into quarters, just like the median breaks it in half.
- **Variance**: It describes how much a random variable differs from its expected value. It entails computing squares of deviations.
- **Standard Deviation**: It is the measure of the dispersion of a set of data from its mean.

Standard Deviation – Z-Score (Variance)

- By definition, z-score simply means how many standard deviation a given value is away from the distribution mean. A z-score can be positive or negative.
- We can calculate the z-score as (x-Mean)/Standard Deviation



In Summary: Descriptive Statistics

- Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

- With descriptive statistics you are simply describing what is or what the data shows.

- Descriptive Statistics are used to present quantitative descriptions in a manageable form

Inferential Statistics (Probability)

Probability is the measure of how likely an event will occur. To be more precise probability is the ratio of desired outcomes to total outcomes:

- (desired outcomes) / (total outcomes)

- The probabilities of all outcomes always sums up to 1

Consider the famous rolling dice example:

- On rolling a dice, you get 6 possible outcomes
  - Each possibility only has one outcome, so each has a probability of 1/6
  - For example, the probability of getting a number ‘2’ on the dice is 1/6

In Summary: Probability

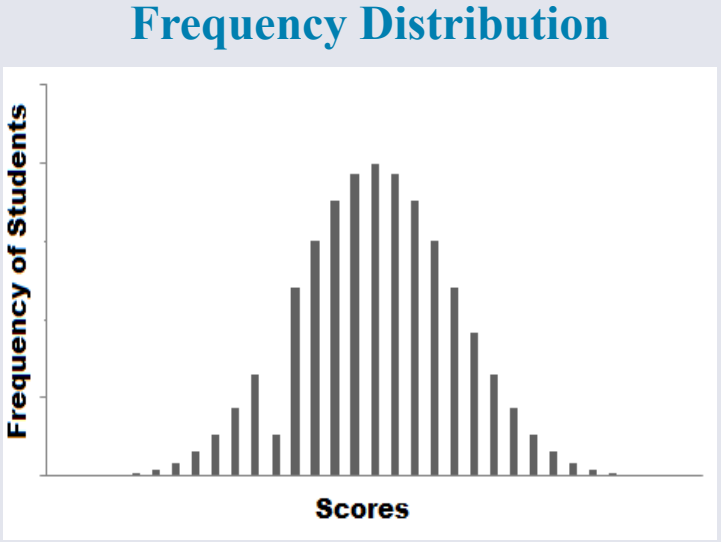
University of Lancashire

- Probability tells how likely an event is going to happen.
- How are we able to find out probability of any given event from a population, even if the event is not in the sample data.
- This is where probability distribution models come into play.
- If we know the probability distribution model for a population, we are able to tell probabilities of all possible events in that population.

Normal Distribution

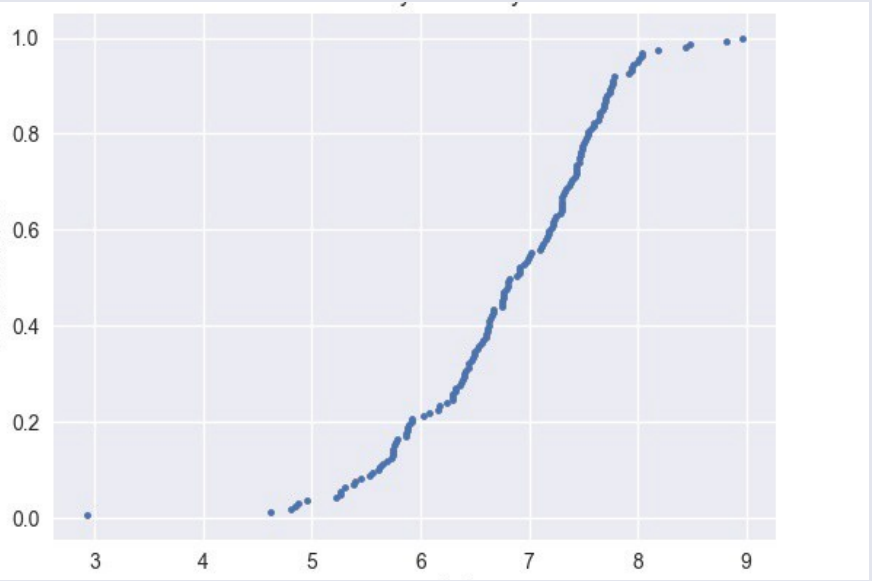
To find out what possible missing values are, we could plot a chart like the following using all available scores.

S. No.	Scores
1	25
2	27
3	38
4	42
5	
6	16
7	35
8	46
9	48
10	31



Cumulative Distribution

- The cumulative distribution function (cdf) provides an integral picture of the probability distribution. As the name cumulative suggests, it is simply the probability that a variable will take a value less than or equal to a particular value. In the example given x=6, the cdf tells us the sum probability of all random variables from 1 to 6.



- What percentage of variables have a value less than 6? About 20%
- What is an approximate percentage of variables that have a value less than 8? About 97%-98%

Uniform Distribution

A Uniform Distribution is a type of distribution of probabilities where all outcomes are equally likely; each variable has the same probability that it will be the outcome.

- A deck of cards has a uniform distribution because the probability that a heart, club, diamond, or spade is pulled is the same.
- The coin also has a uniform distribution because the probability of either the head or the tail in the coin toss is the same.
- The dice tossing has a uniform distribution too, because each side of a dice has the same probability to be drawn.

Binomial Distribution

If we want to know the probability for multiples trials, for instance:

Given 10 flips of a fair coin, what is the probability of getting 6 heads?

The answer to this question is to use Binomial Distribution. The Binomial Distribution is the probability distribution of a sequence of experiments where:

- Each experiment produces a binary outcome
- Each of the outcomes is independent of all the others.

Binomial Example – Coin toss

Experiment: 3 Heads, 2 Heads, 1 head, None

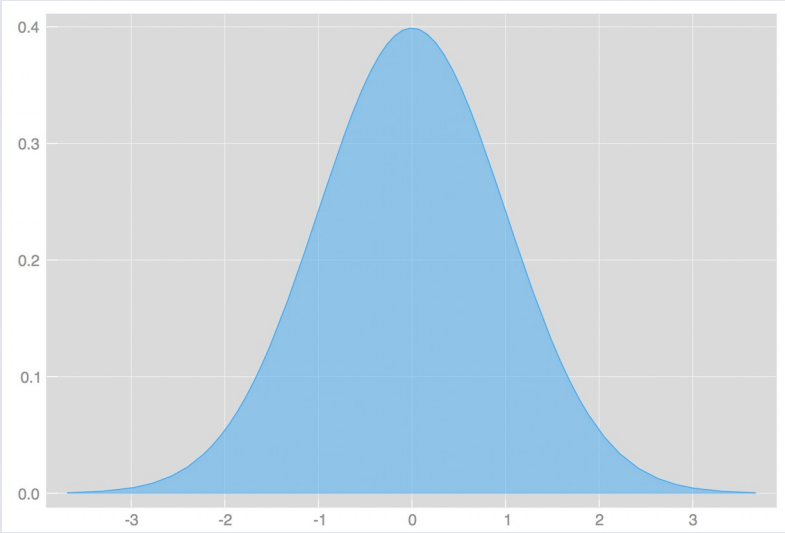
- P(Three Heads) = P(HHH) = 1/8
- P(Two Heads) = P(HHT) + P(HTH) + P(THH) = 1/8 + 1/8 + 1/8 = 3/8
- P(One Head) = P(HTT) + P(THT) + P(TTH) = 1/8 + 1/8 + 1/8 = 3/8
- P(Zero Heads) = P(TTT) = 1/8

Distribution: Symmetric or skewed

Normal/Gaussian Distribution (1)

- Normal distribution, also known as the Gaussian distribution, is a continuous probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

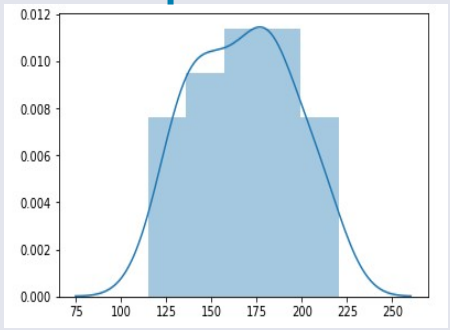
- In graph form, normal distribution will appear as a bell curve.



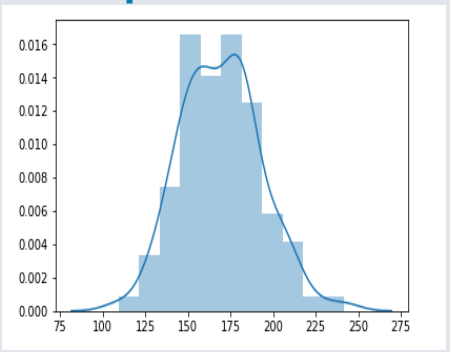
Normal/Gaussian Distribution (2)

Suppose we have a random variable , X, and we use this number to measure an adult's height. With common sense, we know that the probabilities of X's values are not equal. (You don't see 200cm tall people as often as 175cm tall ones in the street).

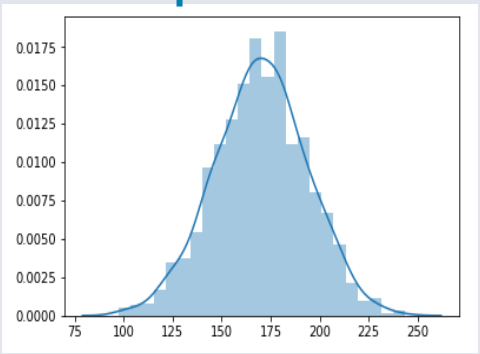
Sample size: 50

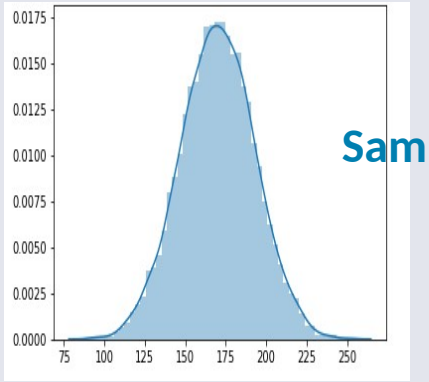


Sample size: 100



Sample size: 1000





Sample size: 10000

Normal/Gaussian Distribution (3)

One of the main reasons for the popularity of the Normal Distribution is that it occurs very commonly in most of the things we see in nature around us. For example:

- Finance, like the salary distribution in an office;
- Healthcare;
- Height/weight distributions;
- Grading distribution.

Reflect on Different Distributions

University of Lancashire

- The common distributions are common because they occur again and again in different and sometimes unexpected domains.
- These distributions occur so often in many applications from our life. Consider them as 3rd party libraries that have been verified and tested, so we can use them directly in our problems.
- It is useful to know the probability density function for a sample of data in order to know whether a given observation is unlikely an outlier or anomaly and whether it should be removed.
- Get familiar with the common probability distributions as it will help you to identify a given distribution from a histogram.

Hypothesis

A statement or claim!

- Examples…?

- All customers over the age of 50 will be accepted for a loan
- Loan acceptance requires higher levels of incomes
- Previous history
- Loan amount

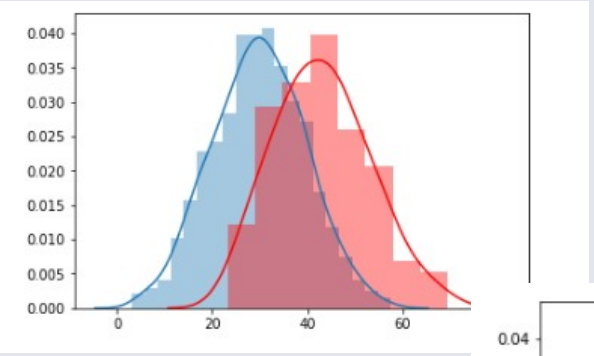
Problem Statement – Loan Predication

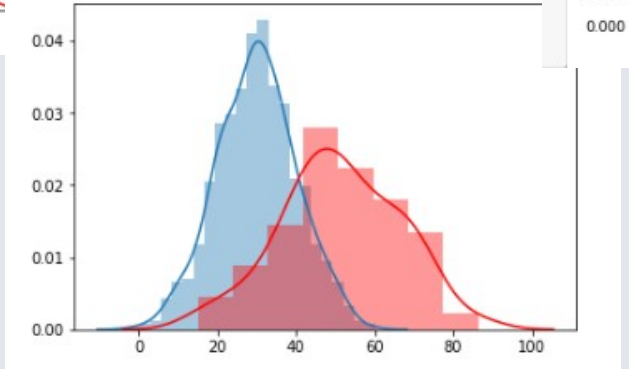
Finance Associates deal in home loans. Customers apply for a loan which is evaluated and validated for eligibility. The company wants to automate the process in 'real-time', which will include details such as: gender, marital status, education, number of dependants, income, credit history and loan amount. Which customer features need identifying to target those that are eligible?

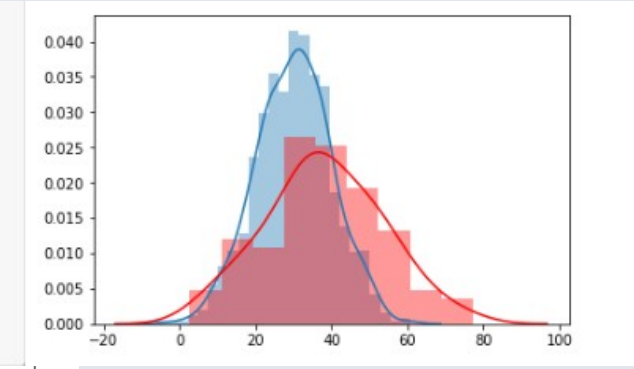
https://www.kaggle.com/datasets/altruistdelhite04/loan-prediction-problem-dataset

Can you prove your Hypothesis?

University of Lancashire







Feature Selection (1)

University of Lancashire

- Data overload – which columns are likely to contribute to your results?
- Is some data irrelevant? GIGO
- What is relevant? Feature selection will help us here.

Task: Example – Loan Predication:

- gender, age, marital status, education, number of dependants, income, credit history and loan amount.

Feature Selection (2)

- How do I know what is important?
  - Domain knowledge, expertise, experience

- Retail data – focus maybe features that influence the purchases a customer makes
- Wine quality – chemical constituents and how they affect preferences
  - Changing constituents could affect the levels of other substances

- Large datasets have many features. Data scientists may look to try different combinations to see what gives the best results.

In Short…

- Domain knowledge is just as important than data analysis skills
- Asking the right questions is more important than elaborate algorithms
- It’s about the ‘right’ data

Learning Objectives

University of Lancashire

Evaluate basic statistics for data analysis

Examine descriptive and probability (inferential) statistics