Mathematics 2

Lecture 1: Introduction to Statistics
Describing Data

Danica Solina

Acknowledgements: Stephen Bush

Teaching Materials (Statistics)

- Lecture notes
  - Available on CANVAS.
  - Worked examples available after the lecture.
- Tutorial exercises
- End of chapter exercises
- Past examination papers
- Recommended Textbook
  - Devore, J.L., **Probability and Statistics for Engineering and the Sciences**, 9th ed. Thompson, 2015. (earlier editions can be used)

What is Statistics?

- Reference: Devore §1.1, 1.2 & 1.3

- What is Statistics?
  - Statistics is a body of principles for designing the process of data collection, developing techniques for data analysis and making inferences about the population from the information in the sample.
  - Statistics helps to understand and quantify the variation in the data and to identify sources that contribute to this variation.
  - Statistics provides answers for making decisions and creating policy.

The role of Statistics in the Workplace

- There are many problems that use Statistics as a problem solving tool:
  - Quality Control in manufacturing processes
  - Calibration models
    - Eliminate systematic measurement biases in field
  - Modelling the number of bit transmission errors made on a digital transmission channel
  - Testing whether there are significant differences in the diameter of steel rods coming from different machines
  - Testing the likelihood that large passenger vehicles will roll over
    - And the extent of the damage if they do.
  - Developing a strategy to maximise the yield in a semiconductor fabrication plant
  - Modelling the number of flaws on a magnetic disc

What else can Statistics do?

- Statistics is not only useful in the workplace. There are many other questions that Statisticians can (try to) answer:
  - Who is the better batsman – Ricky Ponting, or Sachin Tendulkar?
  - What things do people find important when choosing a pizza?
  - Is this new medical treatment better than the current one? Is it worth the extra cost?
  - Where should the new hospitals be built?
  - Who is going to win the next election?

Some Types of Models in Statistics

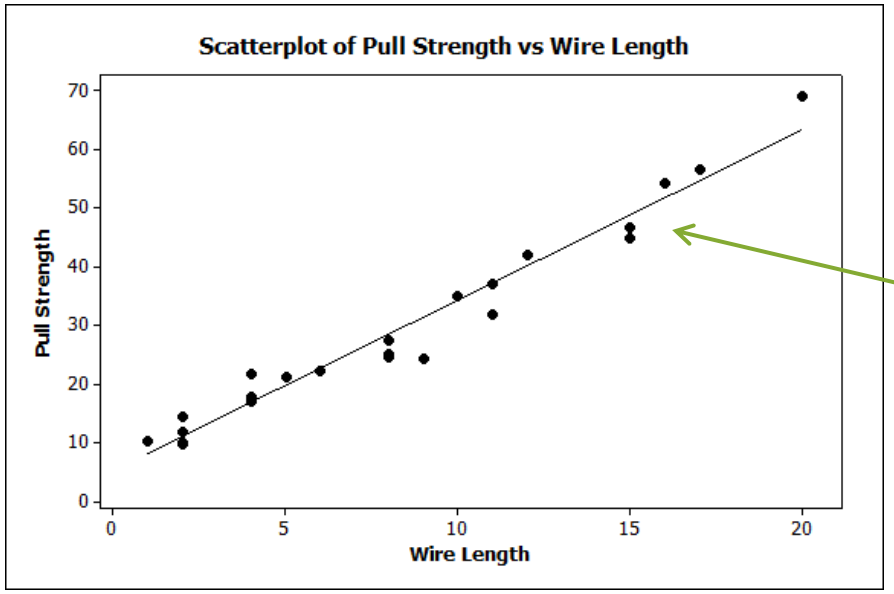
- A **mechanistic model** is built from our underlying knowledge of the basic physical mechanism that relates variables.

- An **empirical model** uses our scientific knowledge of the phenomenon, but is not directly developed from our theoretical or first-principles understanding of the underlying mechanism.
  - This is where the area of Statistics is useful.

Some Types of Models in Statistics

Error in models

- Y = the amount of force requires to bond in a semiconductor.
- X = the length of the wire used to construct the semiconductor.



$$y = a + bx + \varepsilon$$

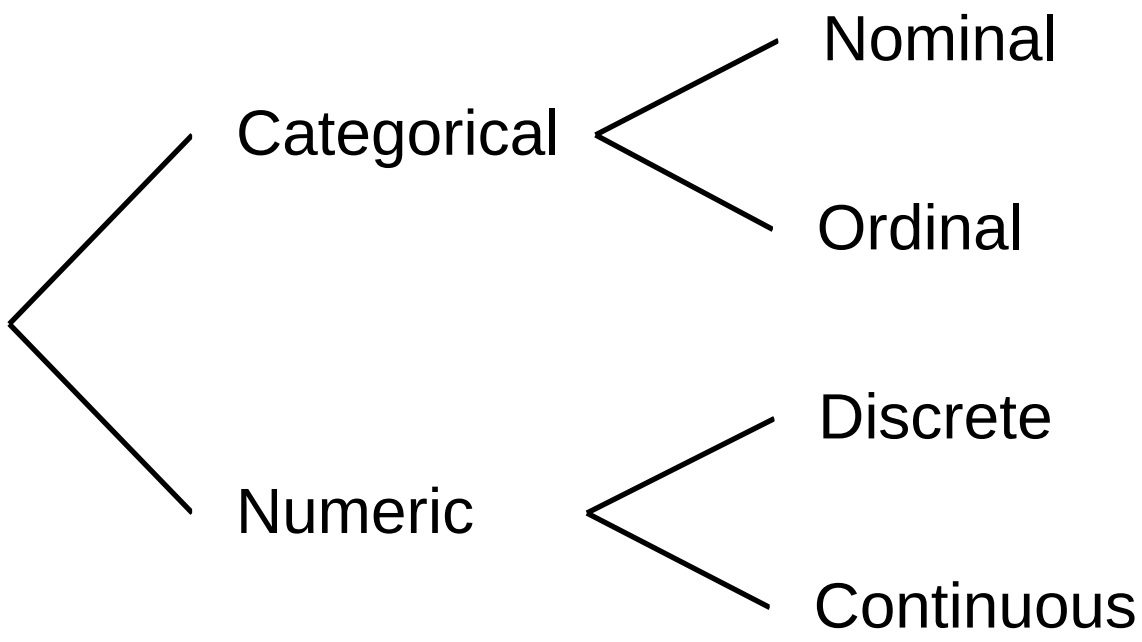
Error – as in variation (not a mistake)

We need data

- How can we collect data?
  - **Observational study**
    - We collect data by observing the objects of interest and measure variables of interest, but we do not attempt to influence the responses.
  - **Perform an experiment**
    - We collect data by imposing treatments on the objects of interest in order to observe their responses

Types of Data

- The type of data that we collect will determine how the data is analysed



Types of Data

- A data set is **numerical** if the individual observations are numerical responses where numerical operations generally have meaning.

- We can further classify numerical data into:
  - **Discrete**: arises from a counting process and can assume only a countable number of values.

  - **Continuous**: arises from a measurement process and can assume infinitely many values.

Types of Data

- A variable is **categorical** if it classifies the objects of interest in a way that numerical operations do not have meaning.

- We can further classify categorical data into:
  - **Nominal**: distinct categories for which there is no single, sensible, way of ordering them.

  - **Ordinal:** distinct categories for which there is an obvious way of ordering them.

Collecting data

- A **population** is the set of all individuals of interest in a statistical study.
  - Example – All students at UTS
- A **sample** is a randomly chosen subset of the population.
  - Example – a randomly chosen group of UTS students
- If all of the population is given the survey/ experiment, then we call this a **census**.
  - Since it is very expensive (or impossible) to conduct a census, we usually use a well chosen sample. The results from the sample will vary from sample to sample.

Collecting Data

- Identify the populations and samples in the following situations:
  - A company wishes to determine the average lifetime of a light bulb. 12 light bulbs were selected, and the length of time to failure measured.
  - The NRMA wishes to test the fuel economy of the latest Holden Commodore. Five cars are selected, and their fuel consumption in city traffic and on the highway are measured.
  - A mining company selects 15 candidate sites from a particular region to take drilling samples to determine the viability of further exploration.

Major Branches of Statistics

- **Exploratory Data Analysis**
  - Consists of procedures that we use to summarise and present the information contained in a set of data.

- **Inferential Statistics**
  - Inferential Statistics consists of procedures that are used to make inferences (draw conclusions and/or make decisions) about population characteristics based on information contained in a sample.
  - This means that we can use a sample to make conclusions about a population (with some level of uncertainty).

Statistical Software

- Statistics involves a large amount of computational work and statistical software packages are used to do the ‘grunt’ work as finding the mean of 200 observations by hand is very boring! Examples of packages are:
  - Minitab, SPSS, ‘R’**, and Excel

- In MM2, we will be using **EXCEL**
  - You will need to use EXCEL for labs, and be able to interpret EXCEL output for exams.

Describing Data

- When we collect a sample, we usually want to be able to describe each variable individually, and also describe the relationships between variables.

- When describing a single variable, there are four features that we are interested in:
  - Where is the centre of the distribution?
  - How spread out is the distribution?
  - Is the distribution symmetric?
  - Are there any unusual observations?

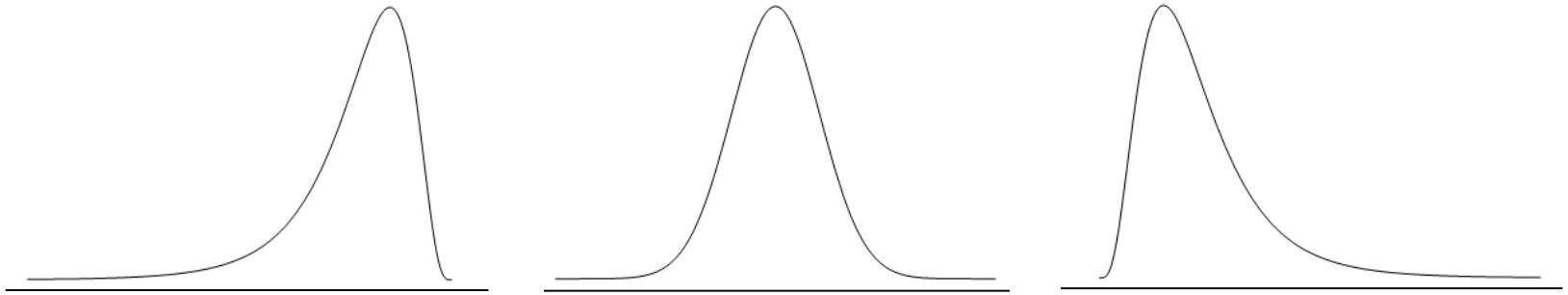
- We can summarise data **numerically** and **graphically**.

Numerical Descriptions of Data

- A **Parameter** is a numerical summary measure that is computed from the entire **population**.
- A **Statistic** is a numerical summary measure that is computed from only the **sample**.
- In order to distinguish between parameters and statistics, we use different notations:
  - Greek letter for parameter (e.g. μ and σ)
  - Roman letter for statistic (e.g. $\bar{X}$ and *s*)
- Parameters are fixed and statistics are random

Symmetry

- We measure the symmetry of a distribution by calculating the skewness of the distribution.
  - If the skewness is negative, then the distribution is **left skewed**.
  - If the skewness is positive, then the distribution is **right skewed**.
  - If the skewness is close to 0, then the distribution is reasonably **symmetric**.
- In general, we will only state that a distribution is skewed if the skew is strong and obvious.



Left-Skewed **Symmetric** **Right-Skewed**

Measures of Central Tendency

- One important feature of the data that we collect is the location of the data, that is, what are typical values for my observations?
- One way to describe the location of the data is to calculate a measure of central tendency. That is, where is the middle of the data?
- There are three common measures of central tendency
  - Mean
  - Median
  - Mode

Measures of Central Tendency - Mean

- Sample mean
  - With sample size n

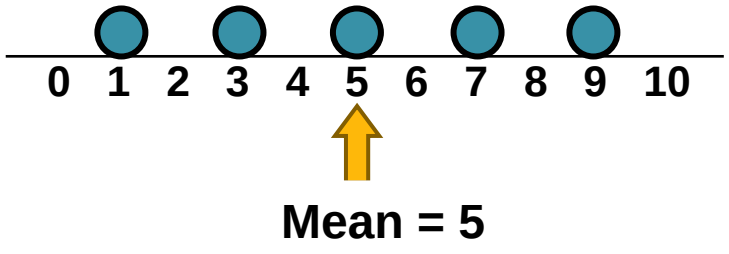
$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

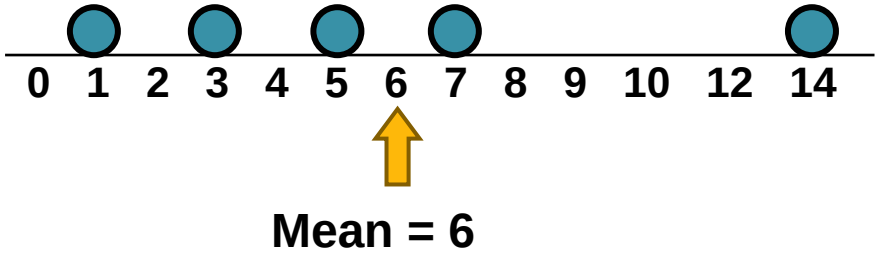
- Population mean
  - With population size N

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Measures of Central Tendency - Mean

- Problems with using the mean:
  - Can be sensitive to skewness.
  - The mean can be very sensitive to a few extreme values (Outliers).
- So we use the mean when the distribution is not strongly skewed, and has no outliers.





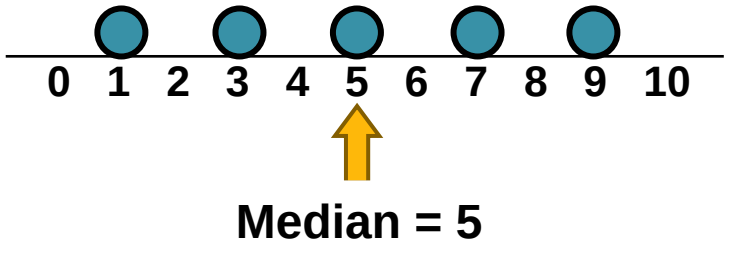
Measures of Central Tendency - Median

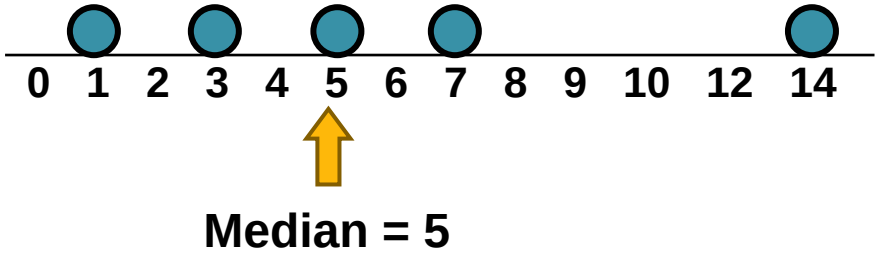
- The **median** is obtained by first ordering the *n* observations from smallest to largest (with any repeated values included, so that every observation appears in the ordered list).

- Then
  - If *n* is odd, the median is the middle observation.
    - That is the $(n + 1)/2^{th}$ observation
  - If *n* is even, the median is the mean of the 2 middle numbers
    - That is the mean of the $n/2^{th}$ and $(n/2 + 1)^{th}$ observations

Measures of Central Tendency - Median

- The median is a good measure of the centre if outliers exist or if skewness is present.
- Extreme values do not affect the median as much.
- We use the median for strongly skewed numerical data, or for numerical data with outliers.





Measures of Central Tendency - Median

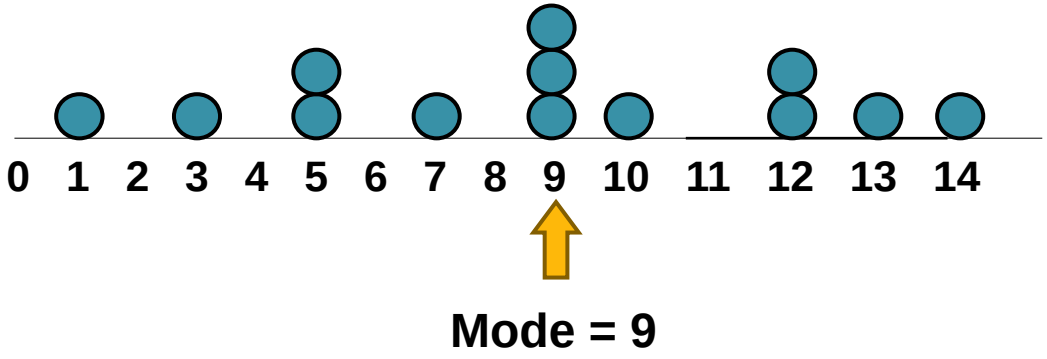
Examples

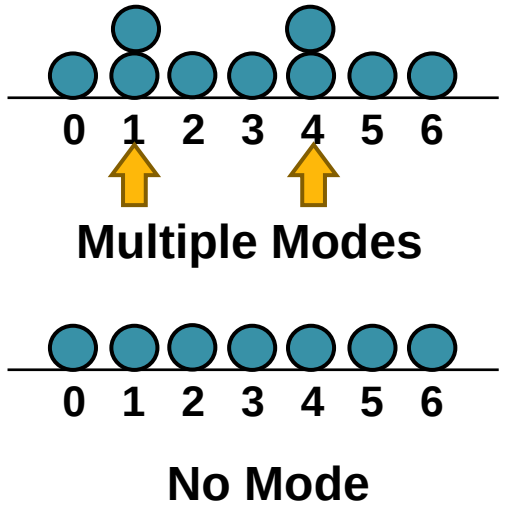
- Data: 14, 9, 12, 10

- Data: 10, 9, 12, 15, 14

Measures of Central Tendency - Mode

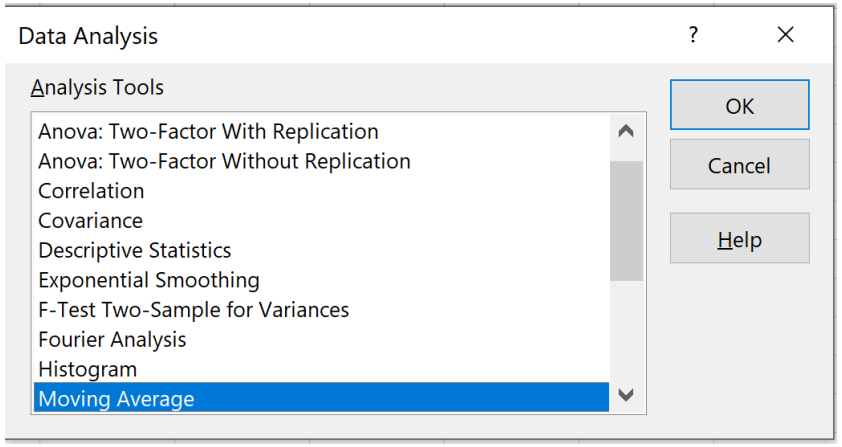
- The **mode** is the value that occurs most often.
  - The mode is not affected by extreme values.
  - There may be several modes.
  - There may not be a mode at all.
  - Used for either numerical or categorical data.

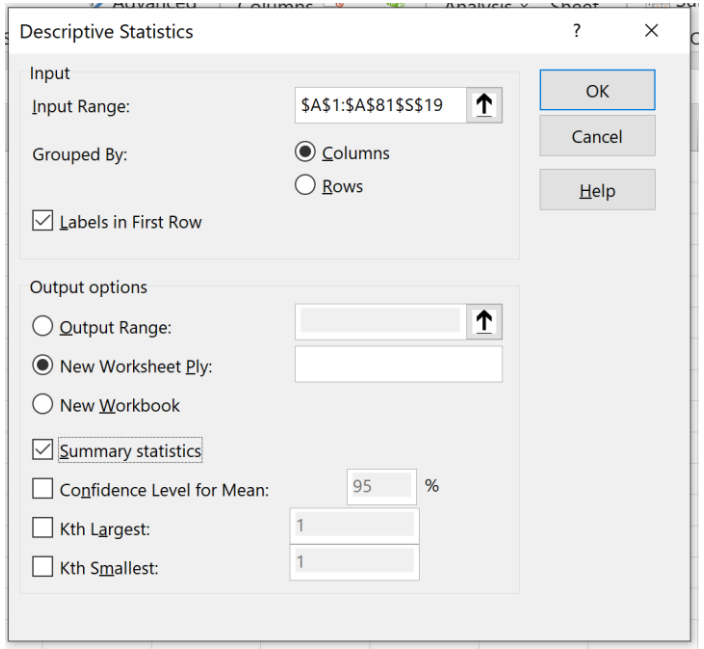




Using Excel to find Descriptive Statistics

- First add the Analysis Toolpak to EXCEL
  - File>>Options>>Add-ins>>Analysis Toolpak
- There is an EXCEL file on CANVAS with the data on the following slide. Open the file in EXCEL.
- Click on ‘Data’ menu and then ‘Data Analysis’
- From the pop-up menu select ‘Descriptive Statistics’
- Add your data range. I have a label for my column so have selected “Labels in the First Row” so that it ignores my label. Select ‘Summary Statistics’-in a new sheet you will obtain the information on the following sheet





Strength Data Set

This table contains the strengths in pounds per square inches (psi) for a new Aluminium-Lithium alloy that could potentially be used to make structural elements in aircraft.

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	258	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Strength Data Set

Descriptive Statistics: Strength

Mean	163.9125
Standard Error	3.958895
Median	163
Mode	160
Standard Deviation	35.40943
Sample Variance	1253.828
Kurtosis	0.352724
Skewness	0.123048
Range	182
Minimum	76
Maximum	258
Sum	13113
Count	80

EXCEL- limitation for multiple Modes

- Note there is an issue. The Descriptive Statistics output gives one mode but there are two in the data. In the EXCEL file on CANVAS I have done the statistics using each of the individual commands including the multi.mode command, which will give you all modes present.

Lecture 1 Revision Exercises

Use a software package to calculate summary statistics

- Devore §1.1 Q1, 2, 5, 9a
- Devore §1.3 Q33a-b, 37, 39, 41

Questions with selected answers are available on CANVAS for the first couple of weeks….until you buy your own copy of the textbook.

Lecture 1 Objectives

- Broadly understand the role that statistics plays in analysis.
- Understand the difference between Mechanistic and Empirical models.
- Be able to classify data into either categorical or numeric, and further classify into nominal, ordinal, discrete or continuous.
- Be able to differentiate between a sample and a population.
- Be aware of the two (broad) branched of statistics (Descriptive and Inferential)
- Be able to select an appropriate measure of central tendency, and be able to calculate it using a software package or by hand.