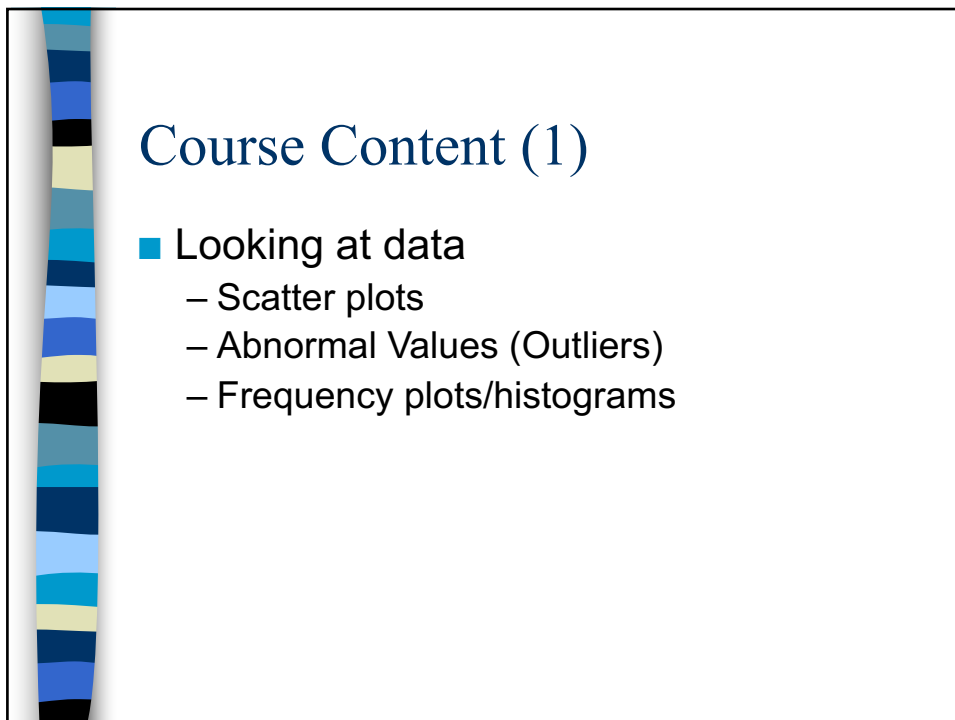




1



2



Course Content (2)

- Descriptive Statistics
 - Measures of central tendency
 - Mean
 - Mode
 - Median
 - Measures of dispersion
 - Standard Deviation
 - Percentiles
 - Inter-quartile range
 - Skewness

3



Course Content (3)

- Data visualisation
 - Creating clean and clear visuals for your report
 - Focusing on the key data and messages
 - “Telling the story”

4



Specific techniques that we might will cover as we go

- Time series
- Correlations
- Statistical Process Control (SPC) Charts

5



Principles

- The course is designed to be quite fluid, allowing for different students to work at a different pace to others, depending on previous skills and experience
- Everyone will have access to two training data sets (Births data and AE dataset), and by the end of the three lectures, hopefully everyone will be able to produce a report, which is written in a way that would be understandable to a (non-expert) lay audience
- You can use whatever statistical software that you are most comfortable with
 - e.g. Excel, R, Python, PowerBI, SAS, Minitab, Tableau, Qlik
- The examples shown will be excel-based, as everyone should have access to it, but the other packages may have better data visualisations built into them, so feel free to experiment and have fun
- Use Google
 - Be self sufficient
 - Statistics is a very broad church, so if you don't know something, look it up...I have to!
 - That includes the syntax and functions required for the different stats packages
- Treat these sessions as “clinics” in which you can ask me anything, using the Teams chat. I may not know the answers immediately, but I'll try to find out

6



Looking at the data

- Graphical representation is best
- Gives you a feel for data before you start doing any analysis. (How is the data distributed?)
- Helps with detection of data anomalies
- Allows us to select the appropriate measures when summarising the data

7

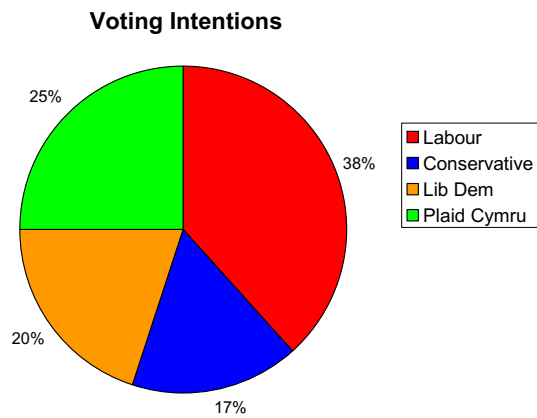


Types of Data

- | | |
|-----------------------|---|
| ■ Numerical: | <u>Examples</u> |
| – <i>Continuous</i> | age, weight, time |
| – <i>Discrete</i> | usually counts, no. of admissions |
| ■ Categorical: | |
| – <i>Nominal</i> | religious groups, occupation |
| – <i>Ordinal</i> | agree / neutral / disagree,
deprivation quintiles,
cancer tumour stage I, II, III ,IV |

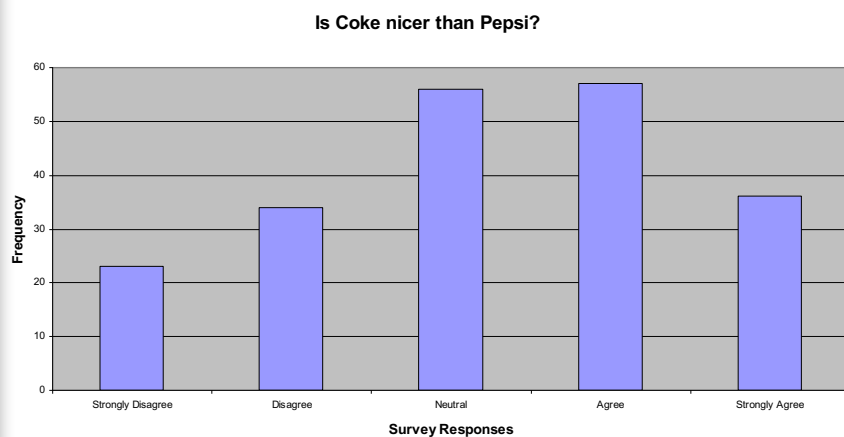
8

Nominal Categorical Data

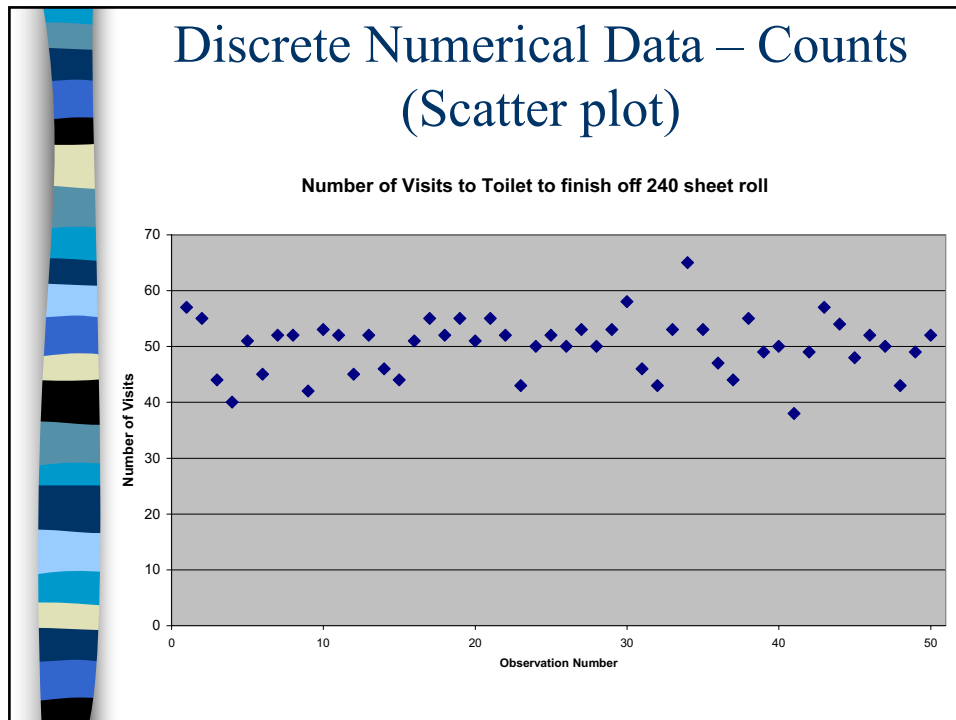


9

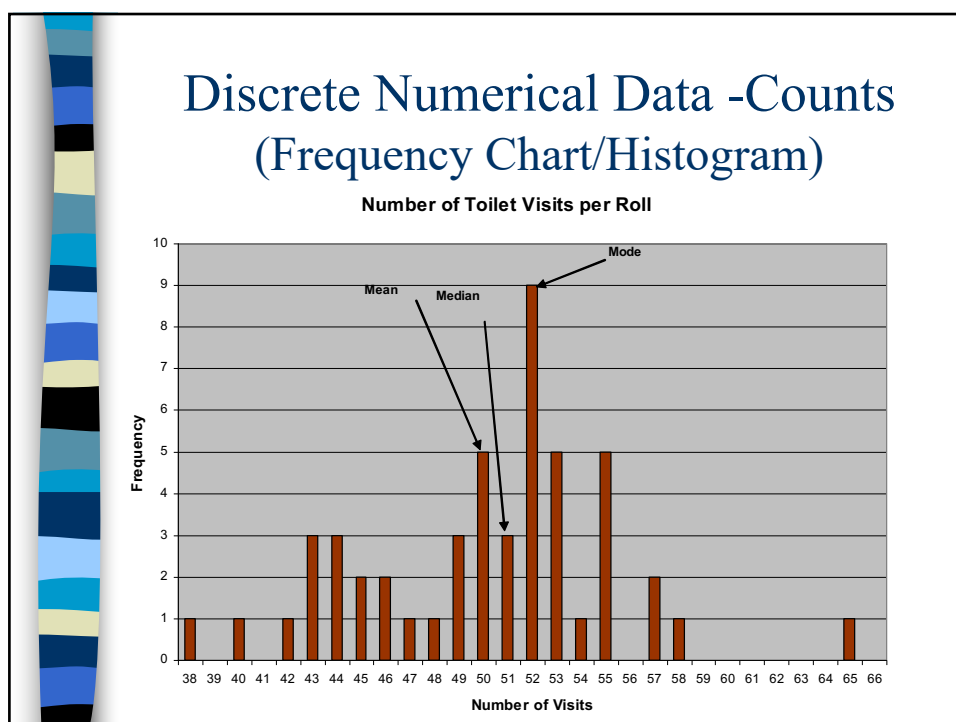
Ordinal Categorical Data



10

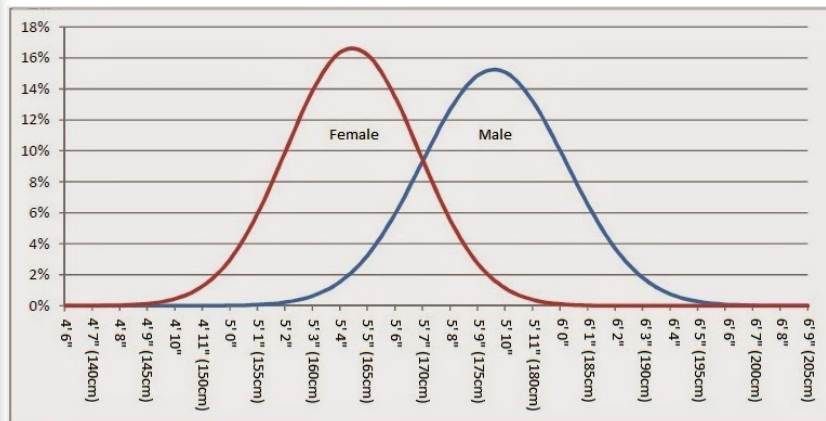


11



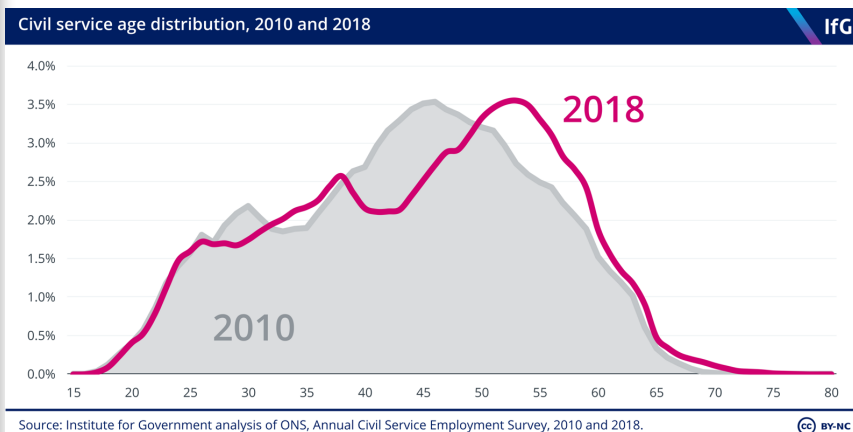
12

Continuous data – Heights (Frequency/Density plot)



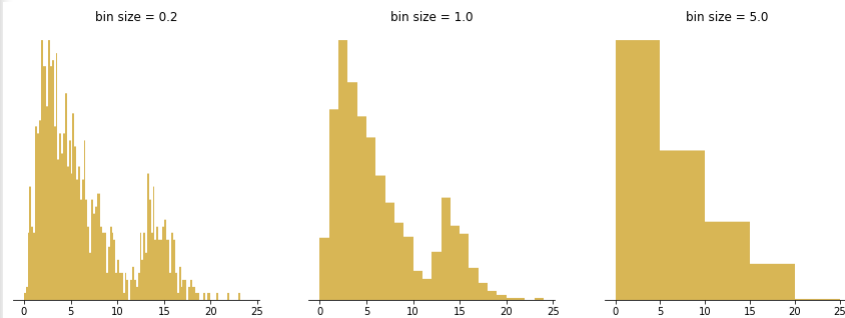
13

(Discretized) Continuous data – Age (Frequency/Density plot)



14

Grouping data into “bins” to create frequency/histogram plots where data is continuous



15

Exercise 1

- Using the data file “Births Data.xls”
- Using a tool of your choice, draw scatter plots of values for the following variables, making a note of your initial findings
 - Gestational Age (Weeks)
 - Maternal Age
 - Weight at Birth (KG)
- Create rules for the identification of abnormal/outlier values for each of these three variables, adding derived columns to your dataset with “flags” denoting whether that record contains an outlier value or not
- Produce frequency/histogram plots for each of the three variables; one set with the outlier values included and another with them excluded. (3x2=6 charts in all)
(Consider the use of appropriate bins for the Weight values)

16



Descriptive Statistics

■ Purpose

- Summarise large data sets
- Describe shape of distribution
- Allow data to be presented/interpreted in a simple way

17



Measures of Central Tendency

- Mean - most commonly used measure, traditionally referred to as the Average
- Median - The mid-point of the distribution
- Mode - The most frequently occurring data value

18



Measures of central tendency: The Mean

- Calculated by summing the values and dividing by the number of values:

$$\bar{X} = \frac{\sum x}{N}$$

For example: **15, 20, 21, 19, 36, 15, 25, 15**

The sum of these 8 values is 166, so the mean is $166/8=20.75$

19



Measures of central tendency: The Median

- To calculate the median, rank the data from lowest to highest, then choose the mid-point value from the sample.

1. For the ordered values

15, 15, 15, 19, 20, 21, 25, 36, 37

↑
the mid-point (median) is **20**.

2. For the ordered values

15, 17, 18, 19, 23, 24, 25, 36

↑
the mid-point (median) is between 19 and 23 and hence the median is the average of the two = $(19+23)/2 = 21$.

20

Measures of central tendency: Mode

- Least commonly used measure

For the values

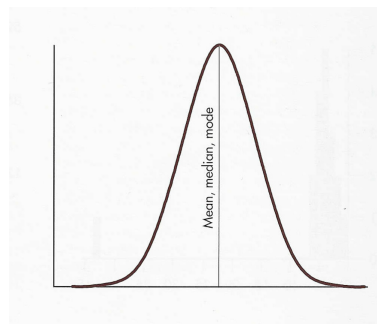
15, 15, 15, 19, 20, 21, 25, 36, 37

the most commonly occurring value (mode value) is **15**.

21

The Normal distribution

- If Mean = Median = Mode
then the data is truly Normally distributed



22



Exercise 2

- a) Using the "Births Data.xls" dataset calculate the Mean, Median and Mode values for each of the following
 - Gestational Age (Weeks)
 - Maternal Age
 - Weight at Birth (KG)
- b) Compare the effect of excluding outlier records (those flagged in exercise 1) and the Means and Medians of these variables
- c) Produce a chart showing the average Maternal age by Year of Birth

23



Measures of dispersion

- Dispersion refers specifically to the spread of the values away from the measure of central tendency.
- The most common measures of dispersion are the:
 - Standard deviation
 - Percentiles
 - Range and inter-quartile range

24



Measures of dispersion: Standard deviation (SD)

- Commonly used measure of dispersion
- Generally used in conjunction with the mean.
- A large SD will arise if values are widely spread about the mean; a small SD if they are generally close to the mean.
- To calculate SD in Excel use the function STDEVPA on the full list of data values
- For the data points **15, 17, 18, 19, 23, 24, 25, 36**
STDEVPA() = **6.2**

25



Measures of dispersion: Percentiles

- Used to more fully describe shape (distribution) of data
- The x^{th} percentile is the value under which $x\%$ of the data points lie
- 50th percentile more commonly known as the median
- 25th and 75th percentiles known as lower and upper quartiles respectively

26



Measures of dispersion: Percentiles

- From the values
15, 17, 18, 19, 23, 24, 25, 36

Lower
Quartile

Median

Upper
Quartile
- The lower quartile is thus 17.5, the median is 21 and the upper quartile is 24.5.
- In general, to work out a percentile, count the number of data points, add 1 to the count and then multiply by the percentage in question

e.g. if there are 43 data points the lower quartile (25%) is $44 \times 0.25 = 11\text{th}$ data point

if there are 44 data points the lower quartile (25%) is $45 \times 0.25 = 11.25\text{th}$ data point which is between the 11th and 12th (need an average - possibly a weighted average)

27



Measures of dispersion: Range

- **Range**
 - Easy to calculate
 - Difference between highest and lowest value.
 - For the data set

15, 17, 18, 19, 23, 24, 25, 36,

the range is simply $36 - 15 = 21$

28



Measures of dispersion: Inter-quartile range

■ Inter-quartile range

- Generally used in conjunction with median value
- Deals only with the middle 50% of the data.
- Difference between the upper and lower quartile.
- In our example the inter-quartile range is
 $24.5 - 17.5 = 7$

29



Max, Min, Median, Mode, Lower Quartile, Upper Quartile, Percentiles in Excel

- Excel has a built in functions to calculate these values from a list of data points such as
 - QUARTILE
 - MEDIAN
 - MODE
 - MIN
 - MAX
 - PERCENTILE

30



Exercise 3

- a) Using the “Births Data.xls” dataset calculate the Standard Deviation, Lower and Upper Quartile values (and Interquartile range) for each of the following
 - Gestational Age (Weeks)
 - Maternal Age
 - Weight at Birth (KG)
- b) For the “AE Dataset.xls” dataset, display and summarise the data in a similar way to the “Birth Data” dataset, in particular showing the shape of the distributions (and associated summary measures) for the following variables
 - Patient Age
 - Time Spent in AE (Minutes)
- c) What summary measures would you best use to describe the distribution of these two variables?

31



Which Measures are best?

- There is no one best measure of central tendency or dispersion
- The “best” measures will depend on the type of data being analysed
- It is always better to visualise the data graphically before adopting the summary measure of choice

32

Shape of the data

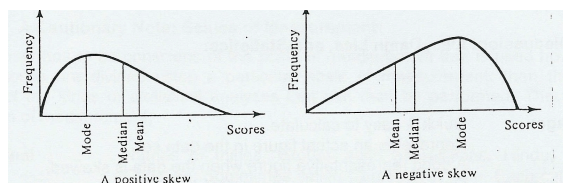
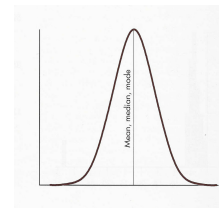
- Distributions can differ from each other in three ways:
 - Central Tendency
 - Dispersion
 - **Skewness**

33

Shape of the data: Skewness

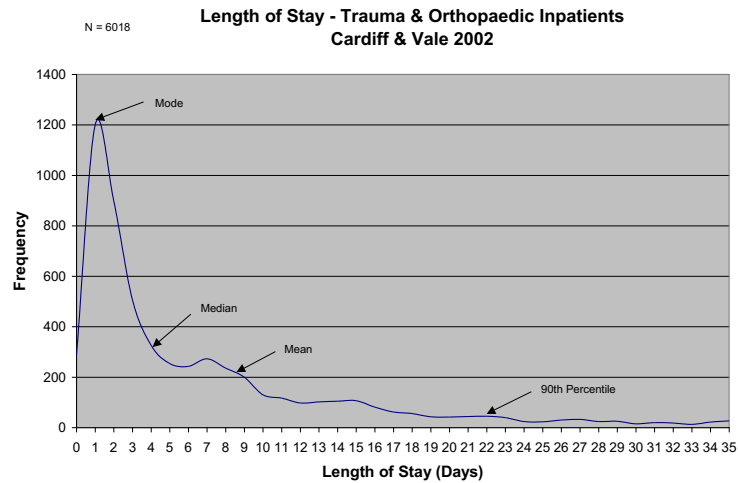
If data is evenly spread each side of the mode, the distribution has no skew.

If clustered to the right of the mode or the left of the mode then the data is positively skewed or negatively skewed respectively.

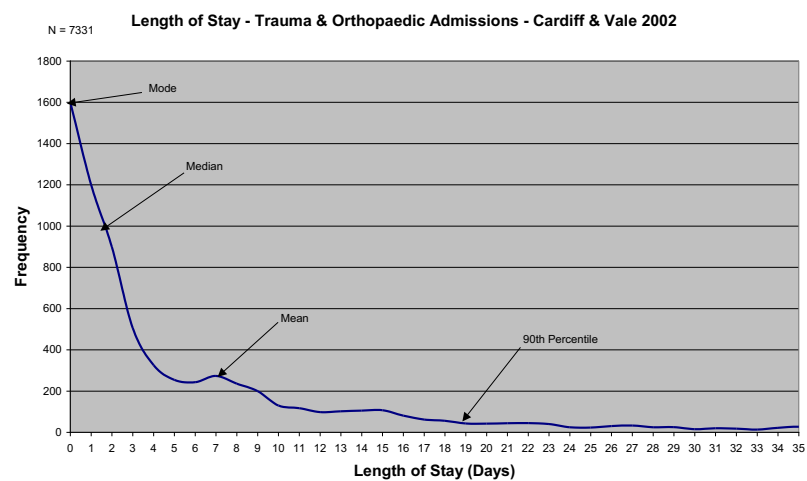


34

Numerical Data: Continuous Data (with a hint of discreteness)



35



36



Summary: When to use what measure

- Continuous or discrete data:
 - Mean, median or mode
 - Range, IQ range or SD
- Ordinal:
 - Median or mode
 - Range or IQ range
- Nominal:
 - Mode

37



Summary: When to use what measure

- If the data is not skewed, generally use mean and standard deviation.
- If the distribution is skewed (positively or negatively), consider using median, inter-quartile range or any other percentile.
- WARNING – data with anomalies (outliers) will have most impact on the mean and SD. In such cases, might be worth considering percentiles such as median and inter-quartile range.

38