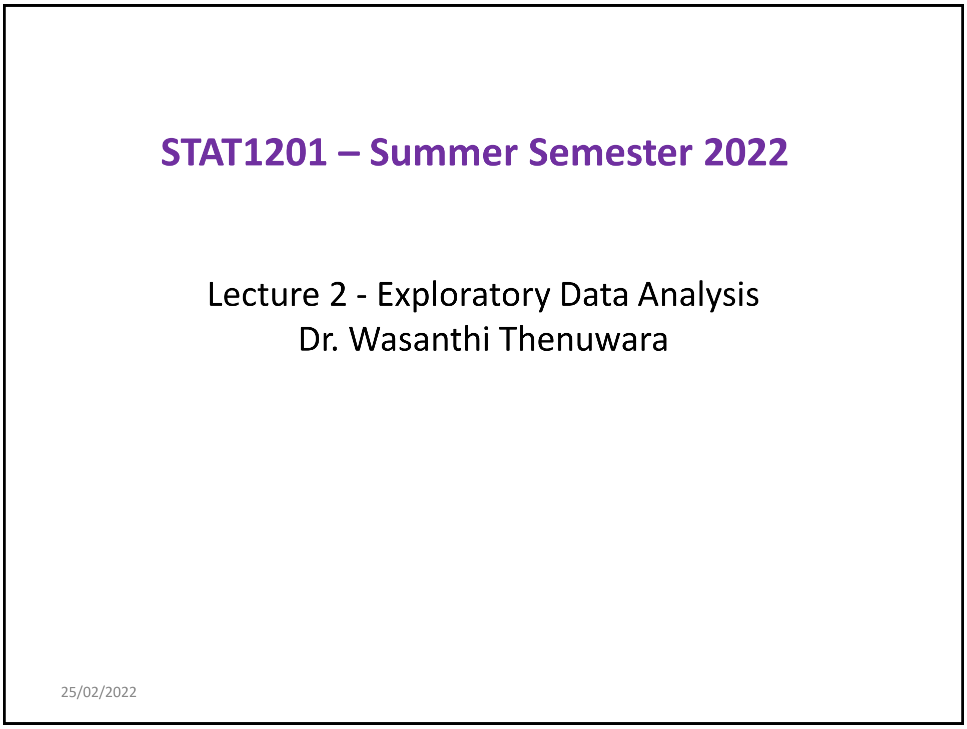
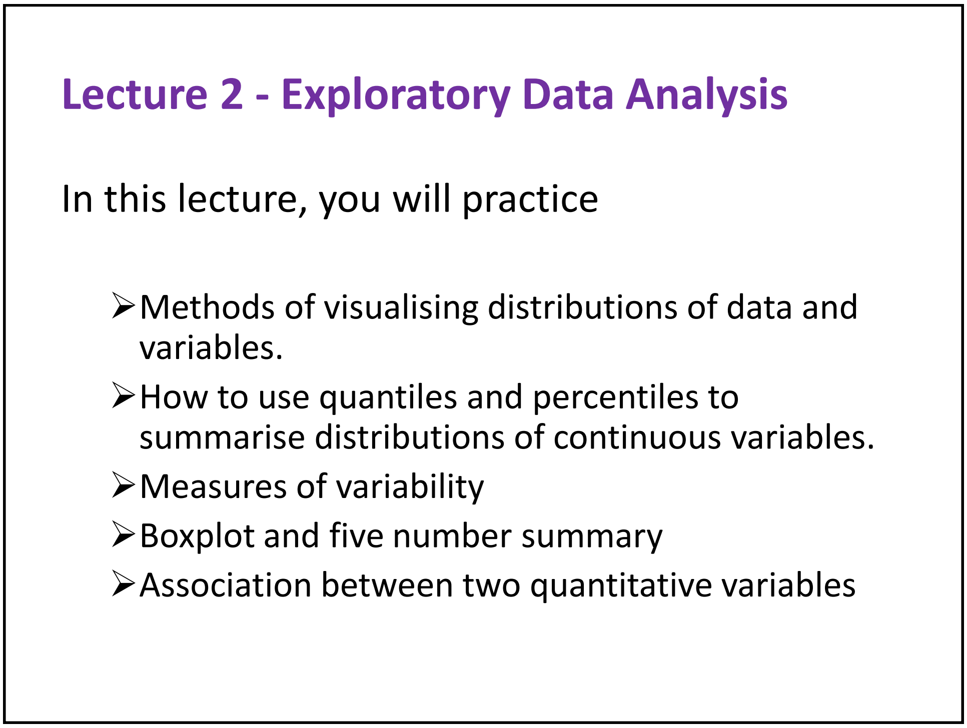


STAT1201 – Summer Semester 2022

Lecture 2 - Exploratory Data Analysis
Dr. Wasanthi Thenuwara

25/02/2022

1

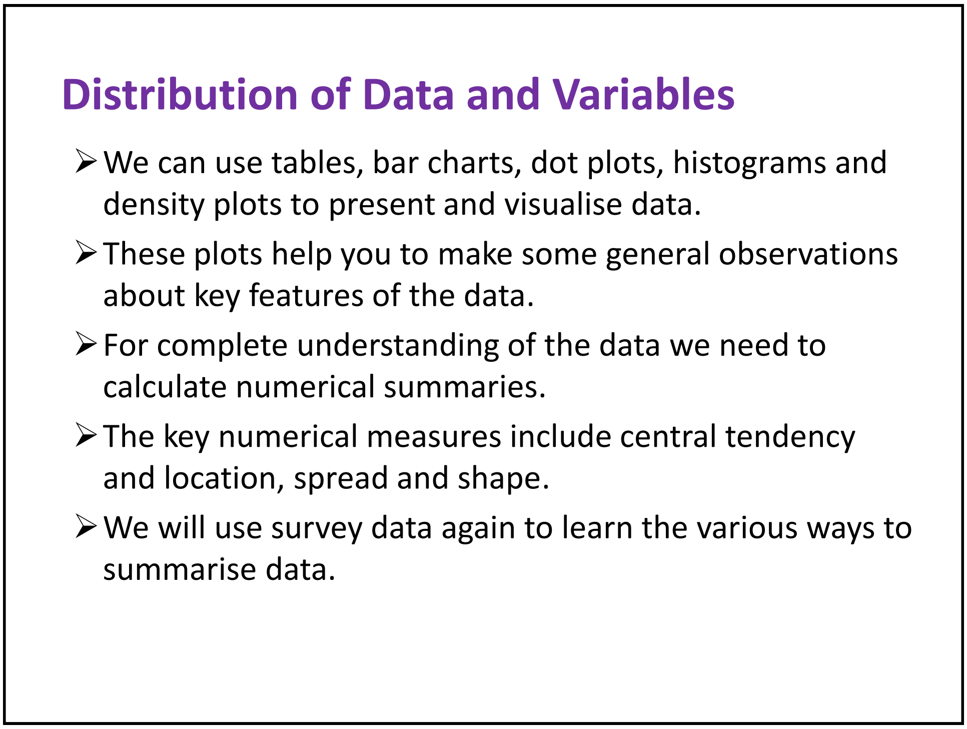


Lecture 2 - Exploratory Data Analysis

In this this lecture, you will practice

- Methods of visualising distributions of data and variables.
- How to use quantiles and percentiles to summarise distributions of continuous variables.
- Measures of variability
- Boxplot and five number summary
- Association between two quantitative variables

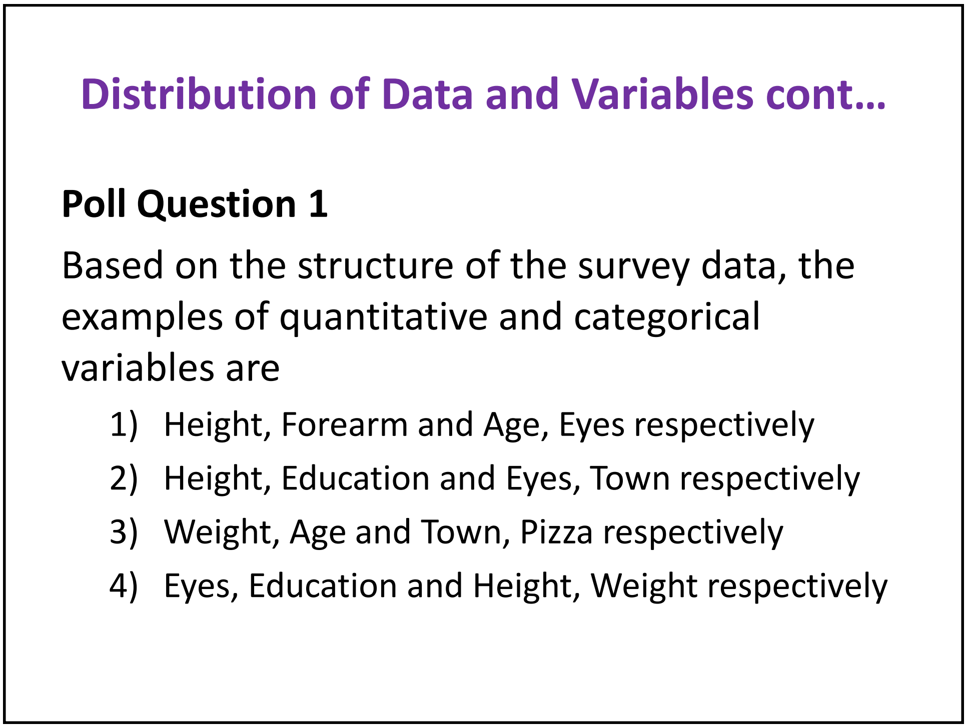
2



Distribution of Data and Variables

- We can use tables, bar charts, dot plots, histograms and density plots to present and visualise data.
- These plots help you to make some general observations about key features of the data.
- For complete understanding of the data we need to calculate numerical summaries.
- The key numerical measures include central tendency and location, spread and shape.
- We will use survey data again to learn the various ways to summarise data.

3



Distribution of Data and Variables cont…

Poll Question 1

Based on the structure of the survey data, the examples of quantitative and categorical variables are

1) Height, Forearm and Age, Eyes respectively
2) Height, Education and Eyes, Town respectively
3) Weight, Age and Town, Pizza respectively
4) Eyes, Education and Height, Weight respectively

4

Distribution of Data and Variables cont…

Present Categorical Variables

- Consider the variable “Town” in survey data. The new data file name is "M2Survey.csv“.
- The number of people live in each town can be presented in a table.

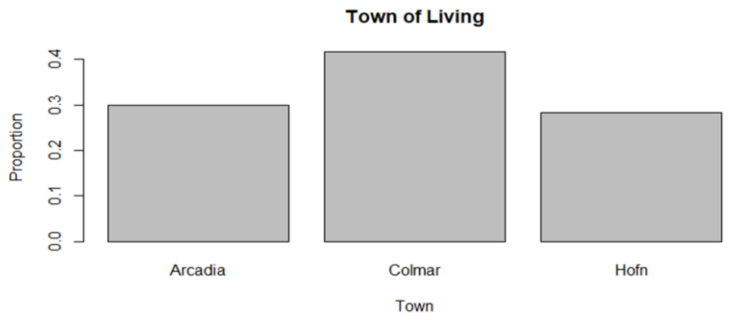
```
survey = read.csv("M2Survey.csv")
addmargins(table(survey$Town))

Arcadia  Colmar    Hofn     Sum
     18      25      17      60

options(digits=3) ## to control number of decimal
places to print

addmargins(prop.table(table(survey$Town)))

Arcadia  Colmar    Hofn     Sum
  0.300   0.417   0.283   1.000
```

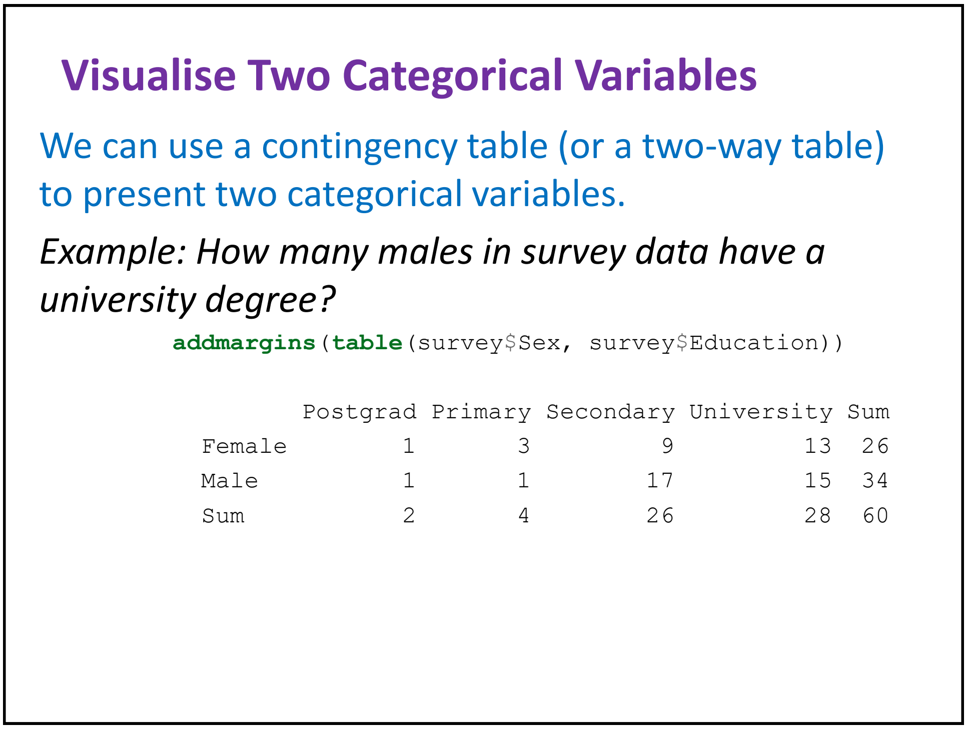
Present Categorical Variables

Bar charts

The figure below shows a bar chart for the proportion of people live in each town.

```
library(lattice)
barplot(prop.table(table(survey$Town)),
main="Town of Living", xlab="Town",
ylab="Proportion")
```





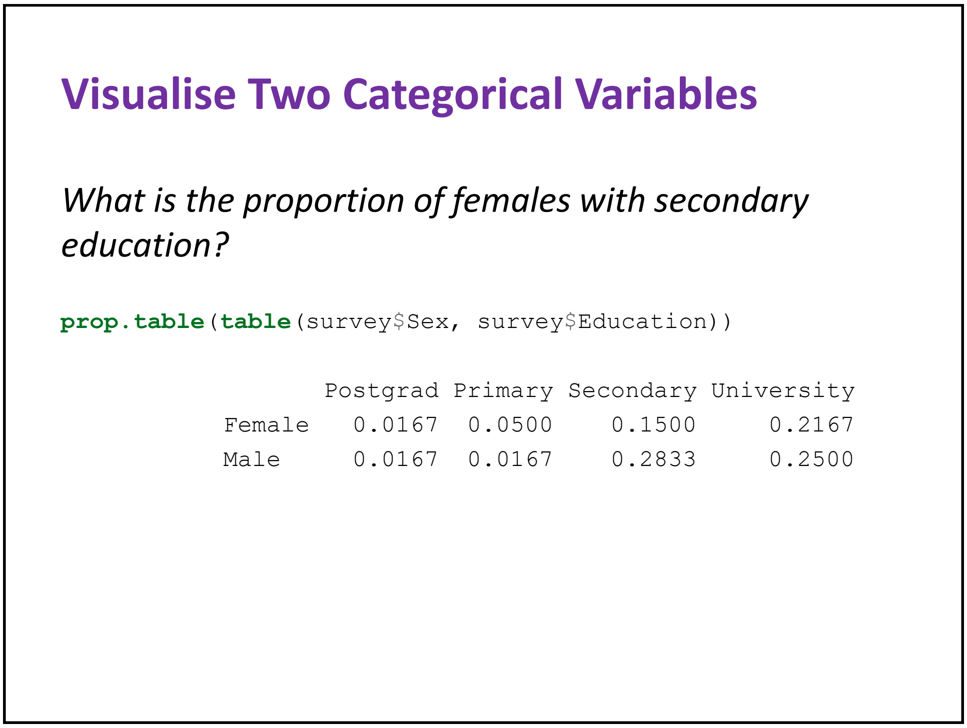
Visualise Two Categorical Variables

We can use a contingency table (or a two-way table) to present two categorical variables.

Example: How many males in survey data have a university degree?

```
addmargins(table(survey$Sex, survey$Education))

         Postgrad Primary Secondary University Sum
  Female        1       3         9         13  26
  Male          1       1        17         15  34
  Sum           2       4        26         28  60
```

7

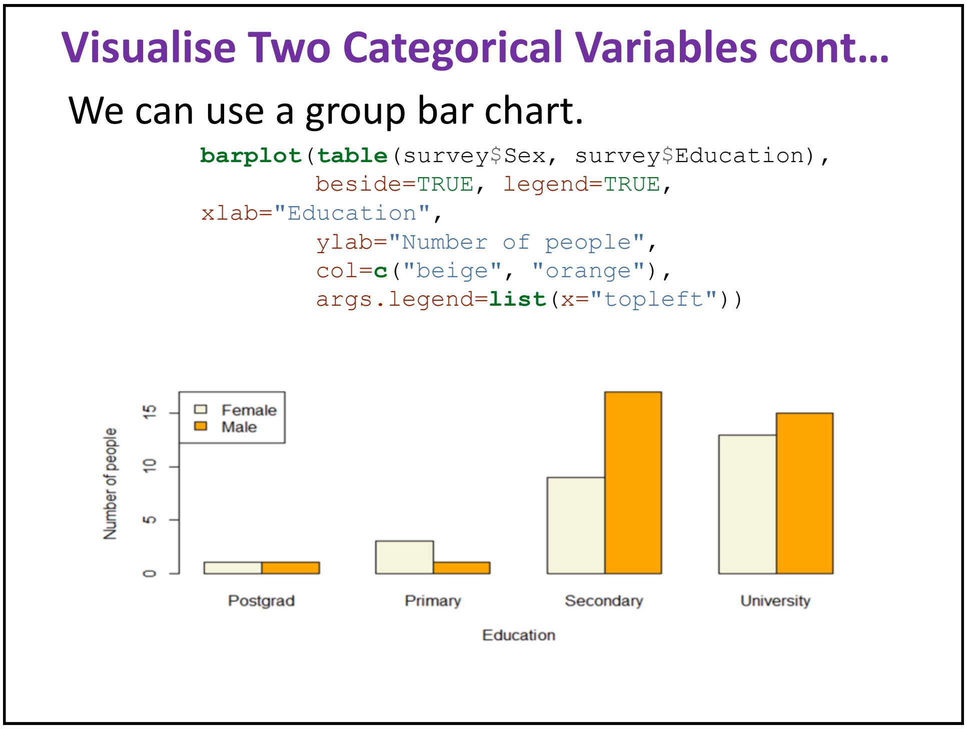


Visualise Two Categorical Variables

What is the proportion of females with secondary education?

```
prop.table(table(survey$Sex, survey$Education))

         Postgrad Primary Secondary University
  Female   0.0167  0.0500    0.1500     0.2167
  Male     0.0167  0.0167    0.2833     0.2500
```

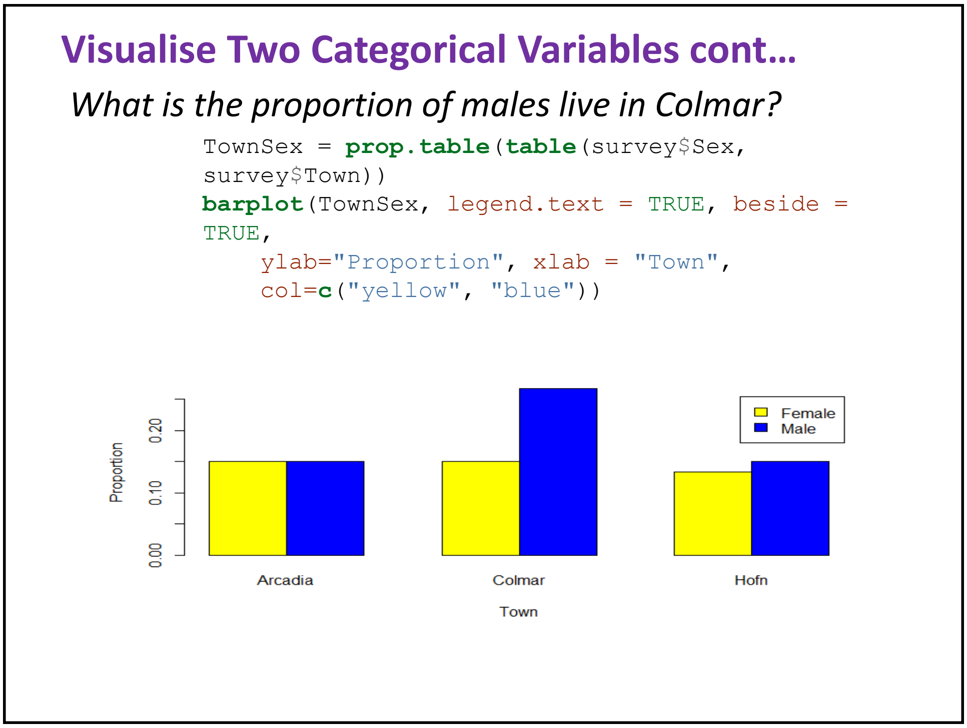
8



Visualise Two Categorical Variables cont…

We can use a group bar chart.

```
barplot(table(survey$Sex, survey$Education),
        beside=TRUE, legend=TRUE,
xlab="Education",
        ylab="Number of people",
        col=c("beige", "orange"),
        args.legend=list(x="topleft"))
```



Visualise Two Categorical Variables cont…

What is the proportion of males live in Colmar?

```
TownSex = prop.table(table(survey$Sex,
survey$Town))
barplot(TownSex, legend.text = TRUE, beside =
TRUE,
    ylab="Proportion", xlab = "Town",
    col=c("yellow", "blue"))
```

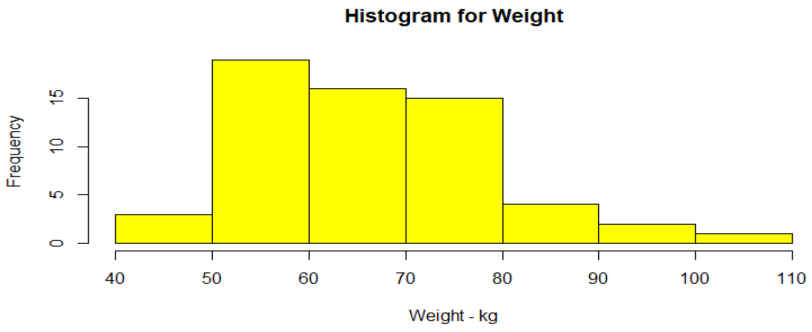
Quantitative Variables - Histograms and Density Plots

- The most useful and common graphs for displaying continuous variables.
- Histograms represent the frequency distribution of continuous variables with no gaps between bars. To construct a histogram, you first need to split the range of possible values into intervals, called bins, and then count the number of observations falling in each bin. Let’s see the histogram for Weight.

11

Quantitative Variables - Histograms and Density Plots

```
hist(survey$Weight, main="Histogram for Weight",
     col="yellow", xlab="Weight - kg")
```



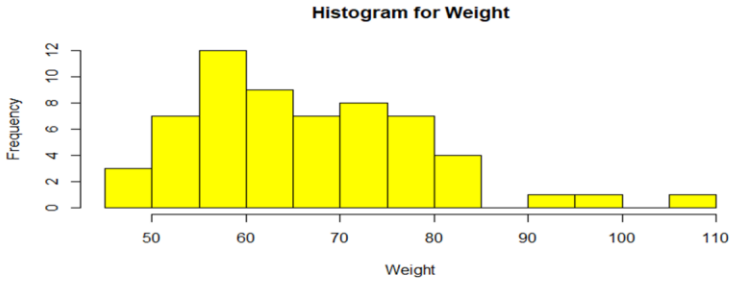
What can you see from this histogram? Shape? Outliers?

12

Histograms and Density Plots cont…

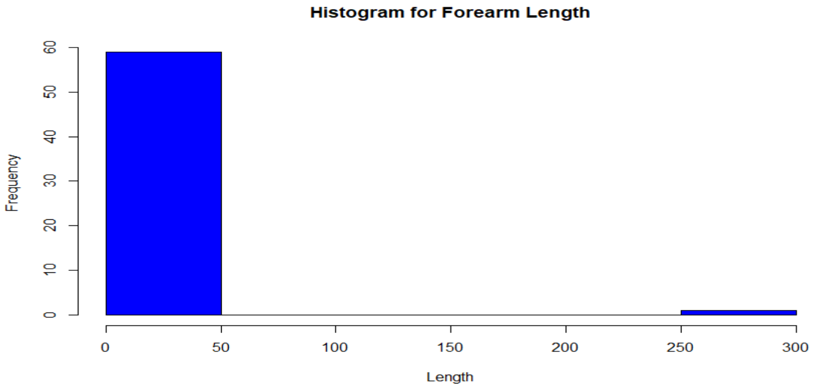
In the histogram (previous slide) R has automatically chosen to split the range of weights into 7 bins, each roughly 10 kgs wide. You can change the number of bins.

```
hist(survey$Weight,  breaks = 12,col="yellow",
     main="Histogram for Weight", xlab="Weight")
```



Histograms and Density Plots cont…

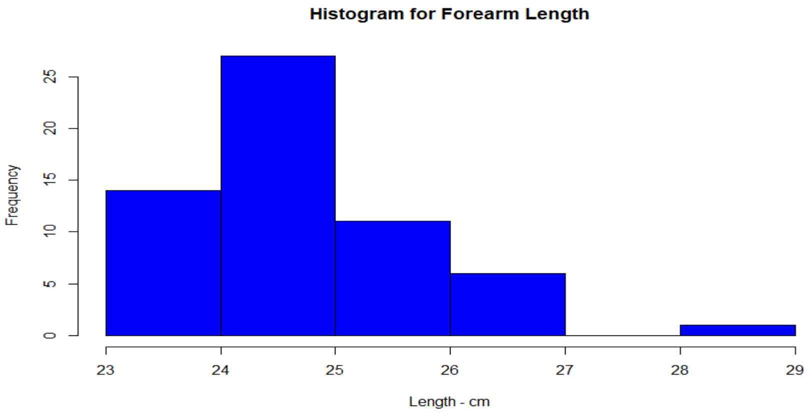
Now see the histogram for Forearm length.



What can you see from this histogram? Shape? Outliers?

Histograms and Density Plots cont…

We can remove the outlier of Forearm = 260cm and draw a histogram again.

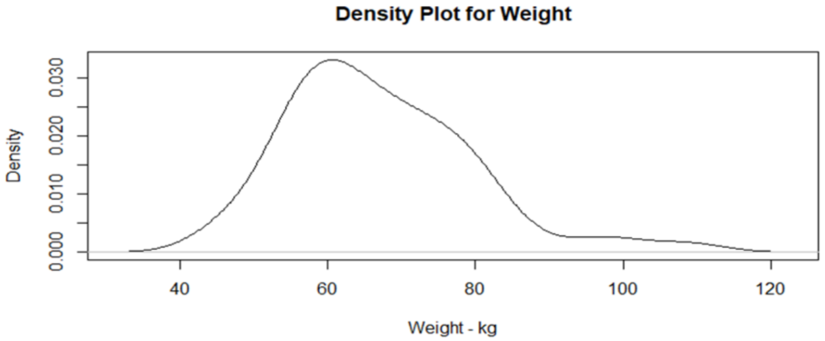


15

Histograms and Density Plots cont…

We are interested in density rather frequency. Density plots are better at determining the shape of the distribution.

```
plot(density(survey$Weight),
main="Density Plot for Weight", xlab = "Weight -
kg")
```



16

Summary Measures of Quantitative Data

Graphical representation tells you key features of data, such as the shape and spread. For complete understanding of the data, you need numerical summaries. We will discuss following descriptive measures briefly.

- Central tendency and Location
- Measures of variability
- Measures of shape

17

Measures of Central Tendency

Provides information about the centre, or middle part of a quantitative variable. Briefly focus on Mode, Median and Mean.

Mode - The most frequently occurring value in a set of data.

Median - middle value in the ordered (or ranked) data and can be used to measure the centre of the distribution. 50% of observations are to the left of the median.

- If the number of observations is odd, the median is the middle number.
- If the number of observations is even, the median is the average of two middle numbers.

18

Measures of Central Tendency

Poll Question 2

Consider the pulse rates of the first 8 people in survey data.

```
Pulse: 80 70 66 50 66 74 78 58
```

The mode and median respectively are

1) 66 and 68
2) 66 and 66
3) 66 and 58
4) 80 and 66

19

Measures of Central Tendency con…

Mean - the average of a set of numbers. Also called the arithmatic mean.

$$\text{mean} = \frac{\sum_{\mathrm{i}=1}^{n} x_i}{n}$$

What is the average Weight of people in survey data? We can use mean() in R to find this.

```
MeanWeight = mean(survey$Weight)
cat("Mean Weight:", MeanWeight, "kg")
Mean Weight: 67 kg
```

Do males are taller than females in this data?

```
aggregate(Height~Sex, data=survey, mean)
```

20

Measures of Location - Percentiles (or Quantiles) and Quartiles

Percentiles and quartiles are measures of location.

Percentiles divide a set of ranked data so that a certain fraction of data is falling on or below this location.

Example: 10th percentile is the value such that 10% of the data are equal to or below that value.

Quantiles are labeled between the values of 0 to 1. 10th percentile is same as the 0.1 quantile.

We can use R to find percentiles.

What is the 13th percentile of Weight in survey data?

```
quantile(survey$Weight, probs =0.13)

13th Percentile of Weight: 54 kg
```

That is, in survey data 13% of peoples' weight is 54kg or less.

Percentiles (or Quantiles) and Quartiles cont…

Quartiles divide a set of ranked data into four subgroups or parts. Denoted by Q_1, Q_2 and Q_3

Q_1 separates the first 25% of ranked data to its left. Same as the 25th Percentile (or 0.25 Quantile).

Q_2 separates the first 50% of ranked data to its left. Same as the 50th Percentile (or 0.5 Quantile). Also called the median.

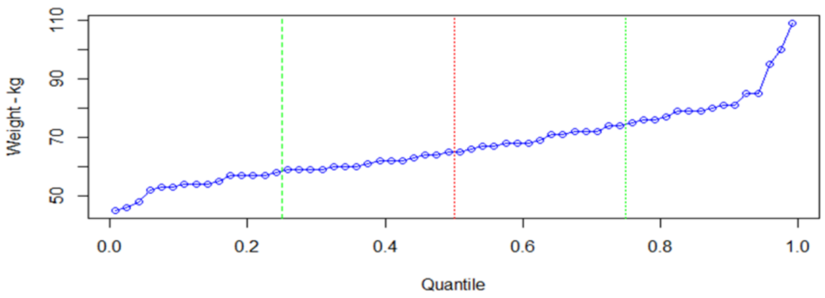
Q_3 separates the first 75% of ranked data to its left. Same as the 75th Percentile (or 0.75 Quantile).

We can use R to find Q_1; Q_2 and Q_3

```
quantile(survey$Weight, probs = c(.25, .5, .75))
 25%  50%  75%
58.8 65.0 74.2
```

Percentiles (or Quantiles) and Quartiles - Quantile Plot

```
y=ppoints(length(survey$Weight))
qqplot(y, survey$Weight, type="o", col="blue",
       xlab="Quantile", ylab="Weight - kg")
       abline(v = 0.50, lty="dotted", col =
"red")
abline(v = 0.25, lty="dashed", col = "green")
abline(v = 0.75, lty="dotted", col = "green")
```



23

Measures of Variability

The variability measures can be used to describe the spread or the dispersion of a set of data.

The most common measures of variability are range, the interquartile range, variance and standard deviation.

Range

Range = Maximum - Minimum

Range is affected by extreme values.

```
survey=read.csv("M2Survey.csv")
max(survey$Weight) - min(survey$Weight)

Range of Weights:  64 kg
```

24

Measures of Variability

IQR (Interquartile Range)

IQR measures the distance between the first and third quartiles. This is the range of the middle 50% of the data.

IQR = Q_3 - Q_1

```
IQR = IQR(survey$Weight)

IQR of Weights: 15.5 kg
```

25

Measures of Variability

Variance and Standard Deviation

Considers how far each data value is from the mean.
Sample variance is denoted by s^2

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

(n-1) is called the degrees of freedom.

What is the value of $\sum_{i=1}^{n}(x_i - \bar{x})$?

Standard deviation (SD) is the square root of variance. SD is the most useful and most important measure of variability.

What is the SD of weight in survey data? Use the function sd(survey$Weight)

We can calculate the standard deviation of weight for males and females of survey data using R.

```
aggregate(Weight~Sex, data=survey, sd)
     Sex    Weight
1 Female  8.872169
2   Male 13.017779
```

26

Measures of Variability

Variance and Standard Deviation

Considers how far each data value is from the mean.
Sample variance is denoted by s^2

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

(n-1) is called the degrees of freedom.

What is the value of $\sum_{i=1}^{n}(x_i - \bar{x})$?

Standard deviation (SD) is the square root of variance. SD is the most useful and most important measure of variability.

We can calculate the standard deviation of weight for males and females of survey data using R.

```
aggregate(Weight~Sex, data=survey, sd)
     Sex    Weight
1 Female  8.872169
2   Male 13.017779
```

27

Measures of Variability

Variance and Standard Deviation

Poll Question 3

The standard deviation of the pulse rates of the first 8 people in survey data is 10.11bpm. The researcher added 1bpm for each person’s pulse rate. The standard deviation of the new pulse rates of the same first 8 people will

1. stay the same
2. increase by 1bpm
3. decrease by 1bpm
4. no sufficient information to determine

28

Five Number Summary and Boxplot

The five-number summary and Boxplot give a compact description of a distribution, including a rough picture of its shape.

Five-number summary and boxplot for Height in survey data

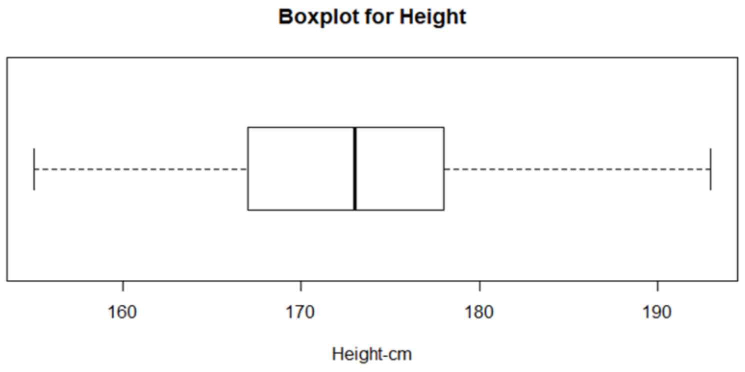
Minimum, Q_1 ,Median, Q_3 ,Maximum

```
fivenum(survey$Height)
[1] 155 167 173 178 193
```

29

Five Number Summary and Boxplot cont…

```
boxplot(survey$Height, horizontal=TRUE,
  main = "Boxplot for Height", xlab="Height-cm")
```



The distribution of Height is roughly symmetric.

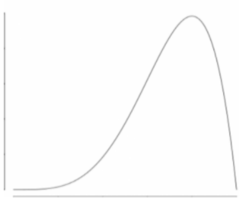
30

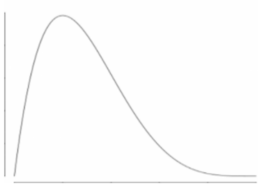
Five Number Summary and Boxplot cont…

Skewed Distributions

Skewness measures the shape of a distribution.

Left or Negatively skewed – A "tail" on the left side of the distribution (Mean < Median). E.g. Distribution of age of deaths

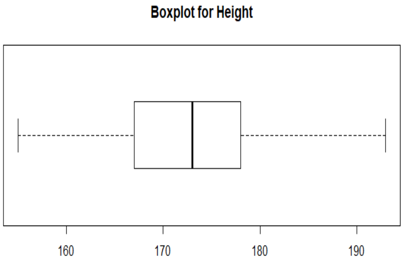


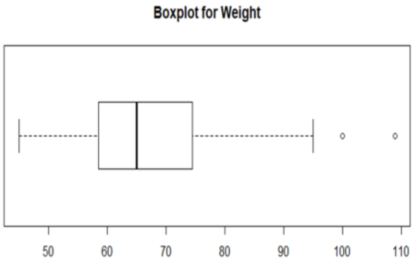


Right or Positively skewed - A "tail" on the right side of the distribution (Mean > Median). E.g. Average annual income

31

Five Number Summary and Boxplot cont…





Poll Question 4

What is the shape of the Weight distribution?

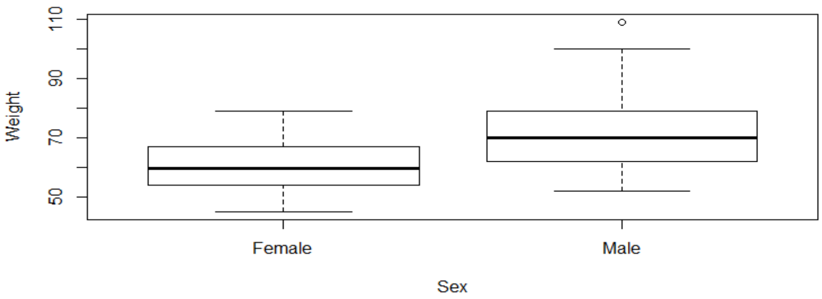
1. Roughly symmetric
2. Right or Positively skewed
3. Left or Negatively skewed
4. Exactly symmetric

32

Five Number Summary and Boxplot cont…

We can use Boxplot to compare two distributions. For example, compare weight distributions of males and females.

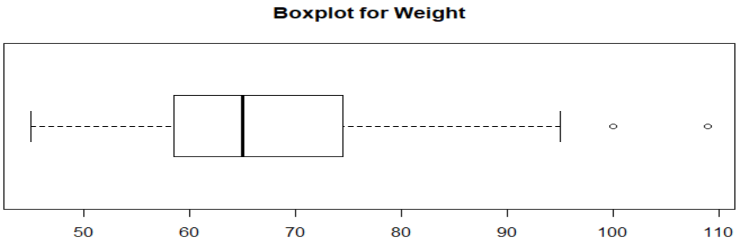
```
boxplot(Weight~Sex, data=survey)
```



What can you see from this?

Outliers in the Data

An outlier is an observation that lies an abnormal distance from the other values in a set of data. We can use a Boxplot to detect outliers in a single variable. Consider the Boxplot for the Weight variable again.



The two weights (100kg and 109kg) flagged as unusual. Are they really outliers?

Outliers in the Data

Whether you should remove outliers from your dataset depends on what causes the outliers.

The causes of outliers come from different ways.

- Data entry or measurement errors.
- Sampling problems and unusual conditions (accidentally collect an item that falls outside your target population and it might have unusual characteristics).
- Natural variation.

35

Outliers in the data

Ed Discussion

Suppose you are doing a survey to collect data to examine the bone density growth in pre-adolescent girls with no health conditions. You noticed two outliers in your data. You discovered that one subject had diabetes. The other outlier was due to a reporting error. How do you treat these outliers in your analysis?

36

Detecting outliers using 1.5xIQR rule

IQR can be used to find outliers in a dataset. The 1.5 x IQR rule is a standard method for flagging unusual observations in data. Such observations are plotted separately in box plots.

Flagged as outliers

If an observation < Q_1 - 1.5xIQR or

If an observation > Q_3 + 1.5xIQR

37

Detecting Outliers Using 1.5xIQR Rule

Poll Question 6

The quartiles and IQR for the Heights of STAT1201 students are as follows.

Q_1 = 167; Q_2 = 173; Q_3 = 178; IQR = 11

Which of the following height is flagged as an outlier?

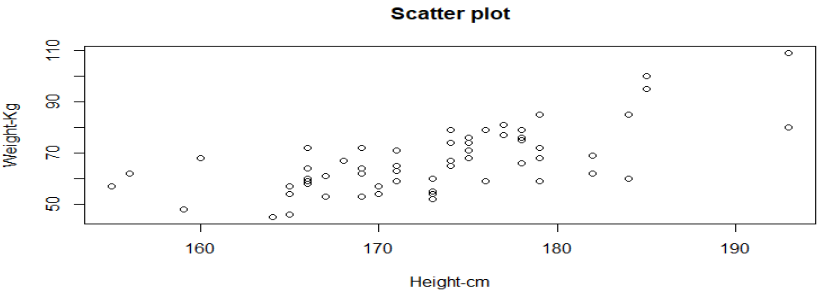
1. 172
2. 180
3. 148
4. 151

38

Association between two quantitative variables

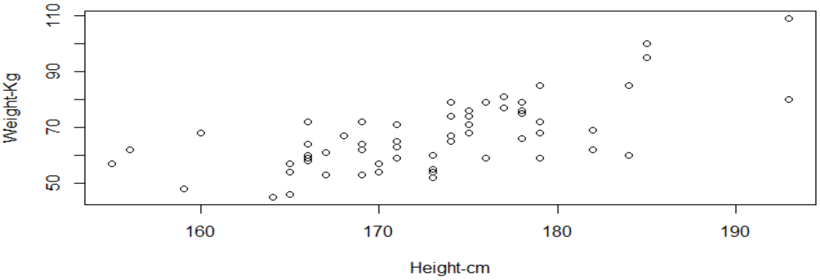
- We can use scatter plots to visualise two quantitative variables.
- Graph one variable (Y) against the other variable (X). For example, we can plot Weight against Height for survey data. That is we plot (Height, Weight) pairs for survey data.

```
plot(survey$Height, survey$Weight,
main="Scatter plot", xlab="Height-cm",
ylab="Weight-Kg")
```



39

Poll Question 7



How do you describe the Weight-Height relationship?

1. Negative linear relationship
2. Positive linear relationship
3. Quadratic relationship
4. No apparent relationship

40

Correlation coefficient (or Pearson’s correlation coefficient)

- Measures the strength of the linear relationship between two quantitative variables.
- Sample correlation coefficient is denoted by “r”.
- r is between -1 and +1 ($-1 \leq r \leq +1$).
- r close -1 implies a strong negative linear relationship
- r close +1 implies a strong positive linear relationship
- r close to 0 implies a weak linear relationship
- r=0 implies no linear relationship

41

Correlation coefficient (or Pearson’s correlation)

```
cor(survey$Height, survey$Weight)
```

```
r between Height and Weight: 0.6834002
```

There is a moderately strong linear relationship between height and weight.

42

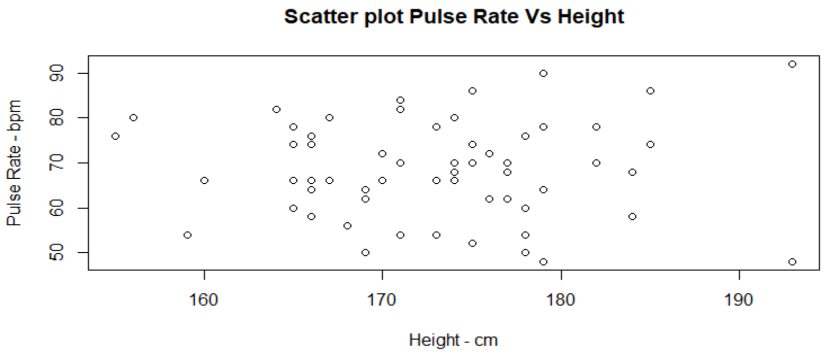
Correlation coefficient (or Pearson's correlation)

- ***r*** does not depend on which variable is considered as X and which variable is considered as Y.
- Correlation does not imply a causal effect of the two variables.
- If a scatter plot shows a correlation between two variables, we cannot infer that a change in one variable causes a change in the other variable. The correlation between two variables, X and Y, can mean several things such as
  - X causes Y,
  - Y causes X,
  - it is a pure coincidence that the two variables move together or
  - X and Y are both influenced by the same third variable.
- We will use linear regression to identify the causal effect.

43

Correlation coefficient (or Pearson's correlation)

Consider the scatter plot of Pulse Rate Vs Height for survey data.



What would be the sample correlation coefficient between Pulse Rate and Height? Guess a value ☺.

44

Correlation coefficient (or Pearson's correlation)

Poll Question 8

The correlation coefficient between pulse rate and weight of people in survey data is

1. 0.1272
2. 0.0153
3. 0.8827
4. 0.4413

45

Survey Data from Quiz 1

(For Ed Discussion – The data and the question will be made available on Ed after next week Monday (05 Dec)

"Survey_2022.csv" includes new data based on your responses for Quiz 1. Answer the following questions based on this data. Use bar charts, histograms, density plots, tables and functions used in the lecture as required to answer the questions.

1) How many observations in the new survey data?
2) Do you find outliers for the variable height?
3) Do left-handed students taller than right-handed students after removing outliers? If so, by how much?
4) Do you find a positive linear relationship between height and forearm length?
5) What is the proportion of students preferring to go to the beach for holidays?
6) What is the proportion cat person born in Summer?

46

Next …

Lecture 3 - Randomness and Probability Theory

Tuesday, 06 December 2022 at 12:00 noon via Zoom (818 1453 7986)

Reminders

Quiz 1 will be closed on Monday, 05 Dec at 3:00 pm

Thank you.