COMP320

RESEARCH PRACTICE

Ethics, Testing, Statistical Data Analysis III

Deadlines are coming

- **Peer review** of proposal and artefact on **Monday**
- **Final deadline** soon after! (check MyFalmouth)
- Make sure you go through the criteria in the **Research Dissertation Handbook** – don't forget to include anything that's required!

Ethics

Why is ethics important in research?

- Already discussed in detail earlier in the module (weeks 1 and 5)
- We must abide by Falmouth University's **Research Integrity and Ethics Policy**
- Participants in your research must give **informed consent**
- You must minimise potential **risks** arising from your research

High risk categories

- Will your project involve clinical trials?
- Will your project involve the use of human blood or other human tissue?
- Will your project involve administering any drugs, placebos, food stuffs or drink to participants?
- Will your project involve the participation of NHS and/or Social Services staff, patients, equipment and/or facilities?
- Will your project involve participants who are particularly vulnerable? (e.g. refugees, prisoners, victims of violence)
- Will your project involve participants who are unable to give informed consent? (e.g. children, people with learning disabilities)
- Will your project risk causing psychological stress or anxiety or other harm or negative consequences beyond that normally encountered by the participants in their life outside research?
- Will your project involve actively deceiving the participants? (e.g., will participants be deliberately falsely informed, will information be withheld from them or will they be misled in such a way that they are likely to object or show unease when debriefed about the study)
- Will your project involve accessing and/or storing data that comes under the Official Secrets Act and/or poses a risk to National security?
- Is there potential for your project to have unintended harmful consequences (e.g. military use of technology / 'weaponisation' of artificial intelligence)?

Medium risk categories

- Will your project involve participants?
- Will it be necessary for participants to take part in the study without their knowledge and consent at the time? (e.g. covert observation of people in non-public places)
- Will financial inducements (other than reasonable expenses and compensation for time) be offered to participants?
- Will your project involve collecting participant data (e.g. personal and/or sensitive data referring to a living individual)?
- Will your project involve accessing secondary data that is not in the public domain (e.g. personal data collected by another user)?
- Will your project involve accessing commercially sensitive information?
- Could your project have negative environmental impacts (e.g. disturbance of natural habitats; damage to, or contamination of, buildings/artefacts/wildlife)

GAMES ACADEMY | FALMOUTH UNIVERSITY

Consent

- Participants in your research must give **informed consent**
- You must typically provide two documents
  - Information sheet
  - Consent form
- (In the case of an online study these could be electronic)
- See **Appendices 5 and 6** of the **Handbook for Research Integrity and Ethics** (on LearningSpace)

Information sheet

- Should enable a potential participant to decide **whether or not** they want to take part
- Describe the project – answer the question ”why are you asking me to take part?”
- Explain what you are asking the participant to do
- Provide contact details for if participants have further questions

Consent form

- I confirm that I have read and understand the information provided above for the research study. I have had the opportunity to consider the information, ask questions and I have had these answered satisfactorily.
- I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason, without my legal rights being affected.
- I agree to take part in the above study.

Personal data

- Handling of **personal data** is covered by the **General Data Protection Regulation (GDPR)**
- Personal data = any data that relates to a **specific living person**
- If data is **anonymous** then it doesn't count as personal data – as long as it **can't be de-anonymised**

GAMES ACADEMY | FALMOUTH UNIVERSITY

Data handling best practices

- Only collect data that is **necessary** for the research
- Particularly, **avoid** collecting personal information (e.g. names, contact details) unless it’s essential
- Ensure that data is stored and handled **securely** and **GDPR-compliantly** (e.g. on university network drives)

Testing

Testing

- One of the requirements for the dissertation is a **plan** to **validate**, **verify** and **test** your computing artefact
- Validation: are you building the **right product**?
- Verification: are you building the **product right**?

Validation

- Are you building the right product?
- What are the **requirements**?
- How will you determine **whether the requirements have been met**?

Verification

- Are you building the product right?
- Does your artefact **work correctly**?
- How will you verify this?
  - Unit tests?
  - Playtesting?
- If your artefact has a **data collection** component, how will you test this in particular?

GAMES ACADEMY | FALMOUTH UNIVERSITY

Data management

Data management plan

- You must describe how you plan to **collect**, **analyse** and **present** data
- Basically demonstrate a **pipeline** – from initial data collection to statistical analysis and presentation
- (In the next few slides I will refer to ”R code” – if using other tools e.g. Python then the principles still apply)

Data collection

- Advisable to collect data in a format that R can read, e.g. Comma Separated Values (CSV)
- **Survey data** – can your survey tool export into an appropriate format?
- **Paper surveys** – how will you transcribe them?
- **Telemetry data** – how will the artefact output it and how will you collect it?

Data analysis

- Once the data is in R, how will you analyse it?
- What **statistical test(s)**?
- Provide **sample code** for how these will be run

Data presentation

- How will data be **presented**?
- Again provide R code

Information Presentation

Illustrating Your Findings

Information Presentation

- There are various techniques for reformatting and reducing data to make the analysis more interpretable or to illustrate a key point
- Graphical representations will also assist in decision making and reinforce the justification for those decisions --- e.g., has a hypothesis been falsified? To what extent is it clearly falsified?
- An overall picture of the data can be gleaned and initial conclusions drawn

Information Presentation

- It is important to select the **most** effective ways to illustrate your findings in the dissertation
- Your communication skills are under assessment --- keep all graphical depiction meaningful to justifying your analysis and/or your intellectual decisions
- Provides an overall picture of the data underlying your findings to reach and support your conclusions
- Be wary of delegating charts solely to important data:
    - Depictions can distort message of original data
    - Concise, but often lacks precision
    - Ensure adequate support in body of text
    - Leverage explicit references (e.g., “as shown in Figure 1”)

Information Presentation

- There are many ways of creating graphs in R and Rstudio!

- We will use a library called **ggplot2**, which you may need to install and load

```
> install.packages("ggplot2", dependencies=TRUE)
> library(ggplot2)
```

- Among its functions should be a **qplot()**, which covers most of the common charts.

Information Presentation

Common formats for presenting information include:

- Bar Chart
- Histogram
- Frequency Polygon & Ogive
- Pie Chart
- Scatter Plot
- Box Plot

Do You Know Your Charts and Graphs?

Which chart or graph is associated with which description?

Frequency distribution		A circular chart with slices presenting percentage breakdown
Histogram		A summary of data presented as classes and frequencies
Ogive		A two-dimensional graph of data from two variables
Pie chart		A cumulative frequency polygon
Scatter plot		A vertical bar chart presenting a frequency distribution

Bar Chart

- A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.
- The bars can be plotted vertically or horizontally.
- Bar charts are useful for displaying data that are classified into nominal or ordinal categories in order to make comparisons.

Bar Chart

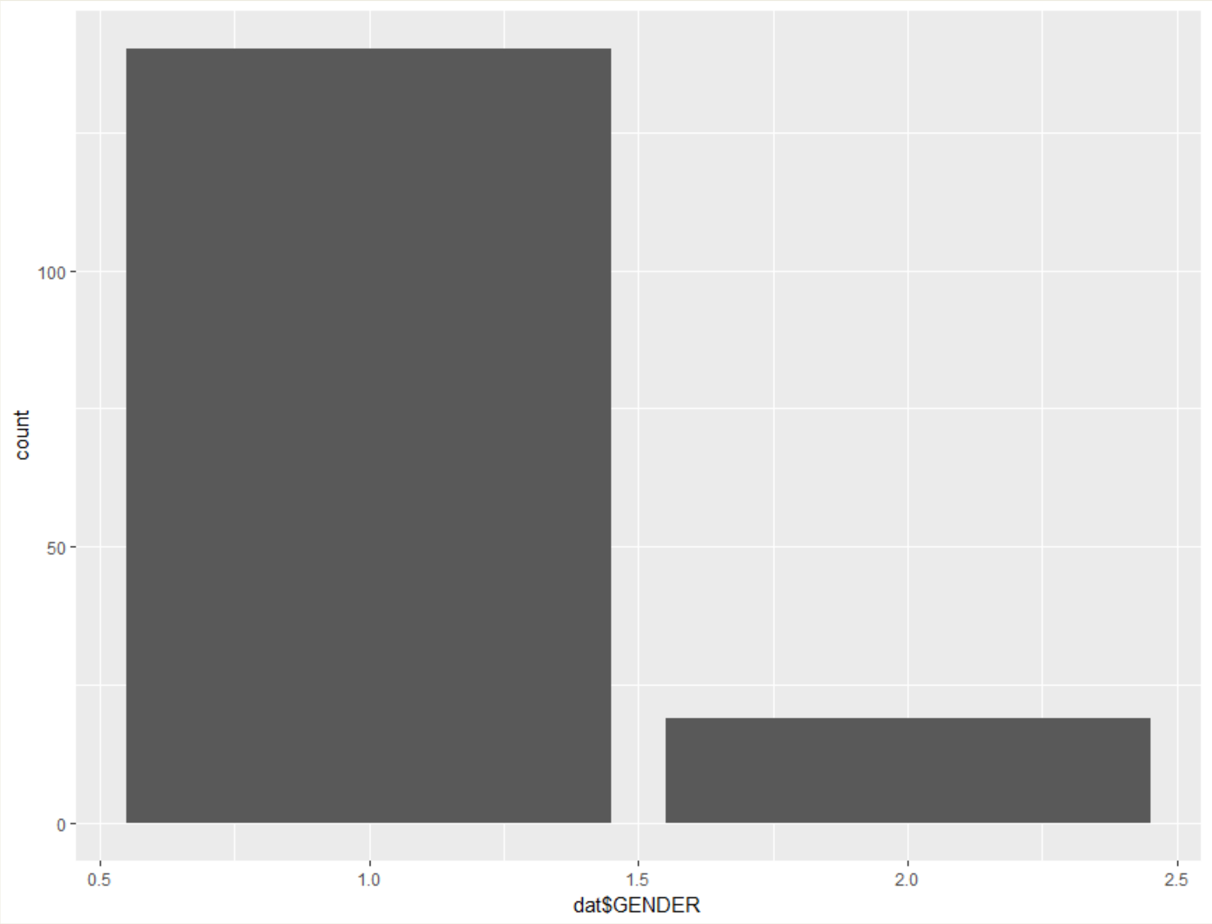
To a simple bar chart:

```
> qplot(dat$GENDER, geom="bar", stat="identity")
```

The **geom** argument refers to the type of chart that **qplot** will produce.

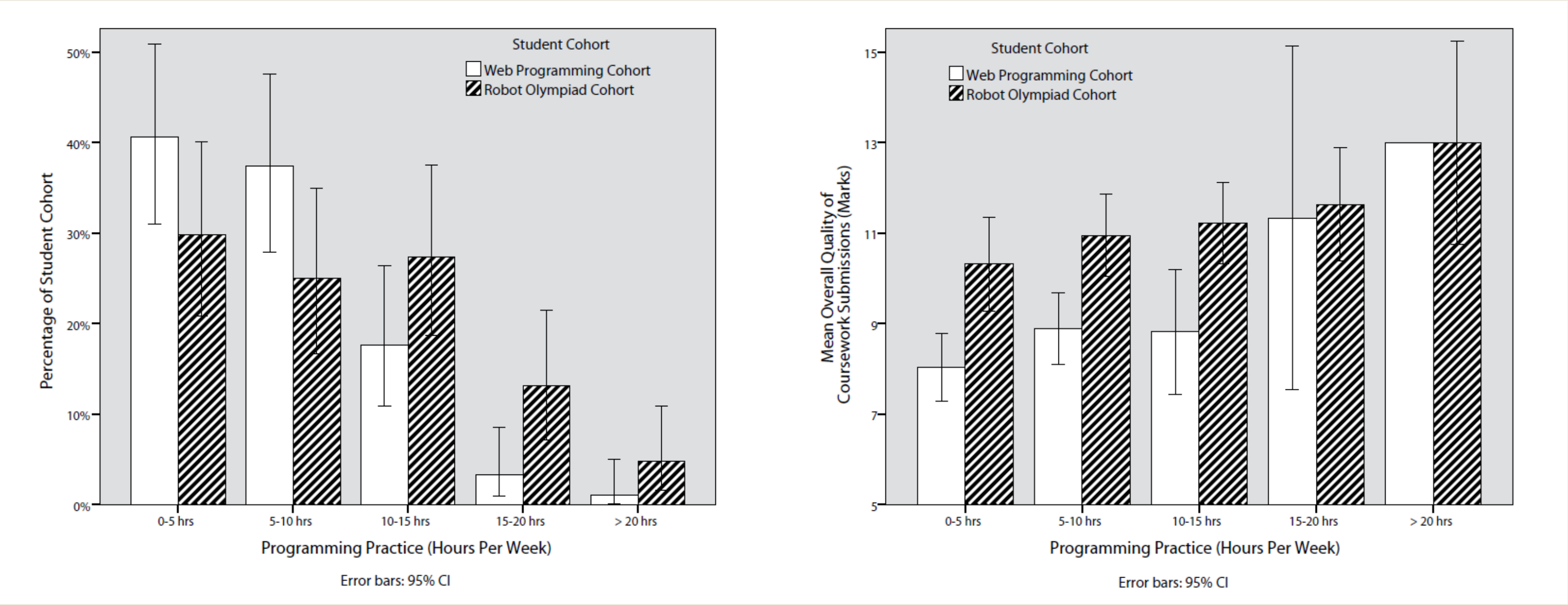
The **stat** argument is depreciated, but useful when no statistical analysis or summary statistic like a mean is needed. For a bar chart, “identify” defaults to a count.

Bar Chart



Bar Chart

- Bar charts can be made much more complex using other ggplot functions:



http://www.cookbook-r.com/Graphs/Bar_and_line_graphs_(ggplot2)/

Histogram

- A type of vertical bar chart used to depict a frequency distribution
- Construction steps:
  - Label the **x** axis with the class endpoints
  - Label the **y** axis with the frequencies
  - Label the chart with an **appropriate** title, i.e. **not** ‘bar chart’
- A quick look at the histogram reveals which class intervals produce the highest frequency totals
  - E.g. which age group most often enrols in undergraduate computing courses?

Histogram

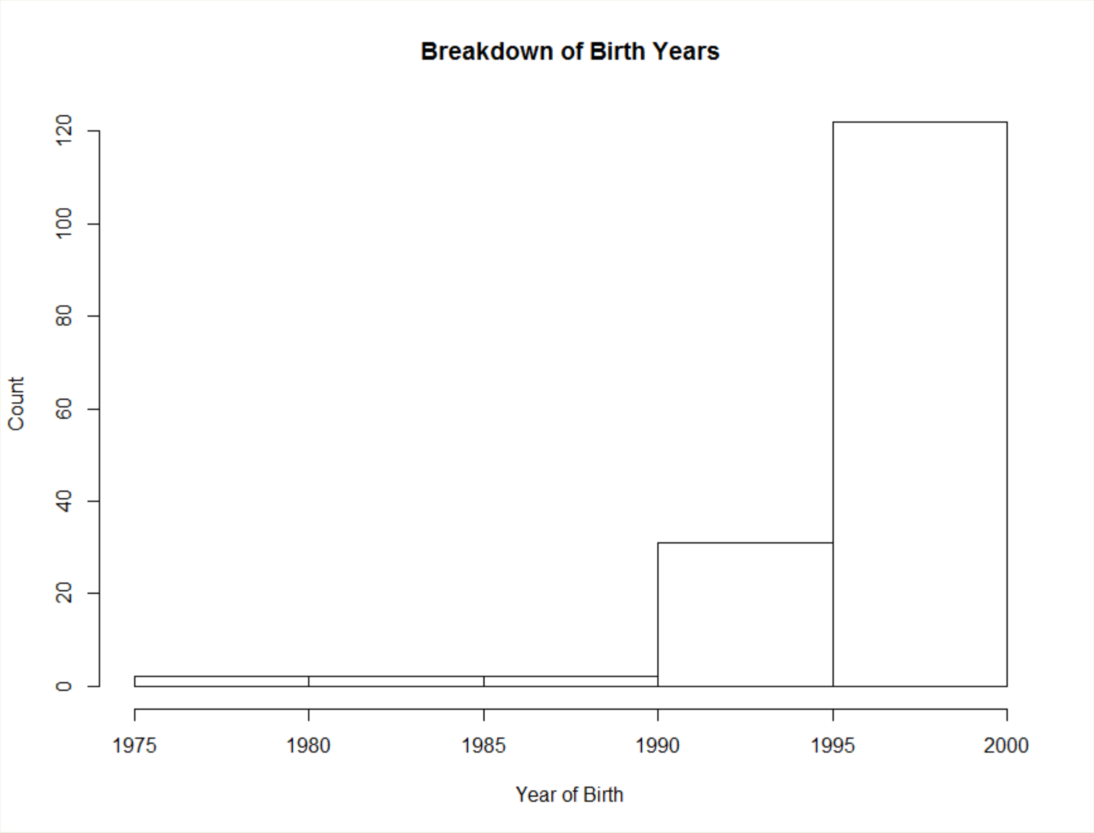
To display a frequency table, examining birth years from 1975 to 2000 in 5-year intervals, use the following command:

```
> summary(cut(dat$BIRTH_YEAR, c(1975,seq(1980,2000,5)),
include.lowest=T,right=FALSE))
```

To display a histogram based on this data:

```
> hist(dat$BIRTH_YEAR, xlab="Year of Birth",
ylab="Count", main="Breakdown of Birth Years", breaks=5)
```

Histogram



Breakdown of Birth Years
Count
0
20
40
60
80
100
120
1975
1980
1985
1990
1995
2000
Year of Birth

Frequency Polygon

- A graph in which line segments connecting the dots depict a frequency distribution

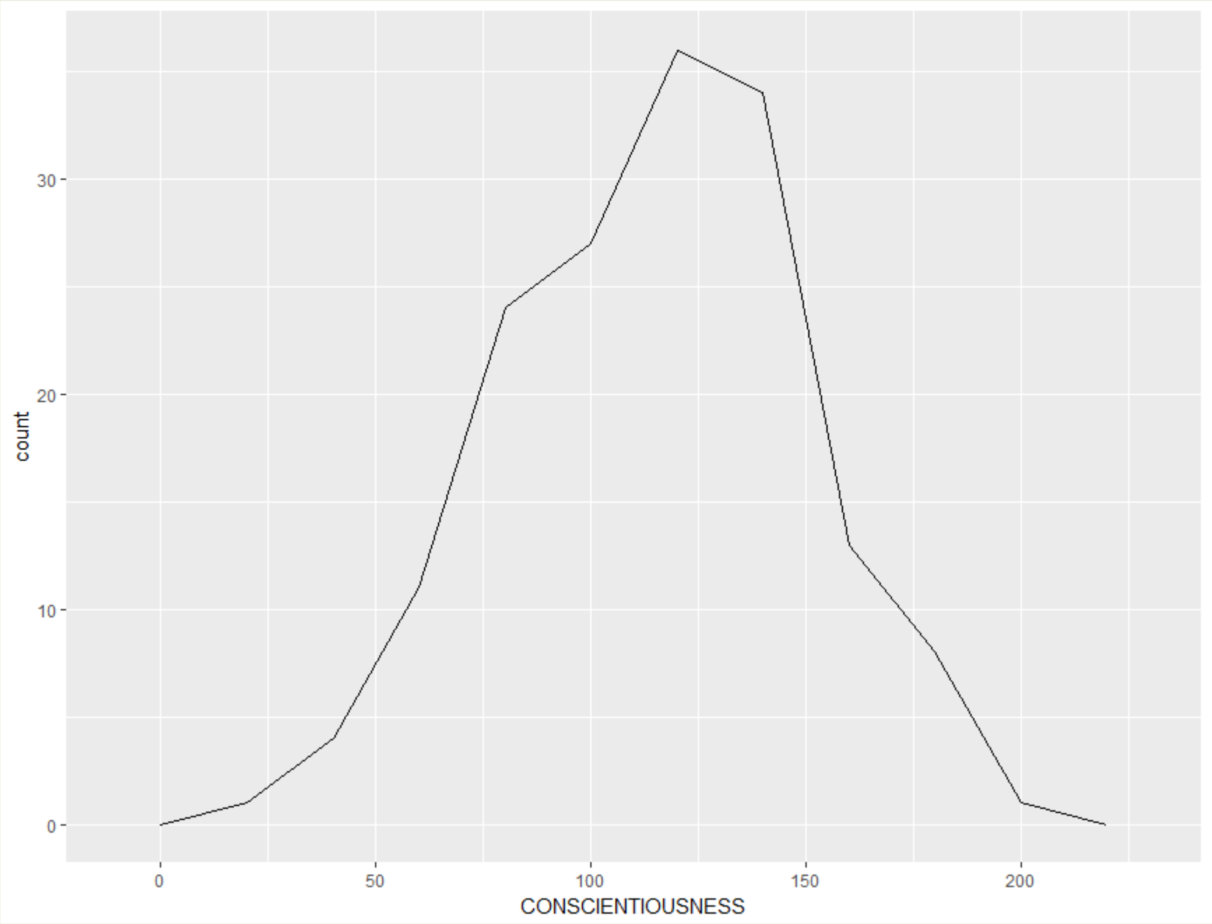
- Construction steps:
    - Label the **x** axis with the class endpoints
    - Label the **y** axis with the frequencies
    - Plot a dot for the frequency value at the midpoint of each class interval (different to Ogive)
    - Connect these dots with a line

Frequency Polygon

To display a histogram of students' conscientiousness, use the following command:

```
> ggplot(dat, aes(CONSCIENTIOUSNESS), stat="count") +
     geom_freqpoly(binwidth = 20)
```

Frequency Polygon



Ogive

- A Cumulative Frequency (CF) polygon
- Construction steps:
    - Label the **x** axis with the class endpoints
    - Label the **y** axis with the cumulative frequencies
    - A dot of ‘0’ is placed at the beginning of the first class
    - Mark a dot for the CF value at the end of each class interval
    - Connect these dots with a line

Ogive

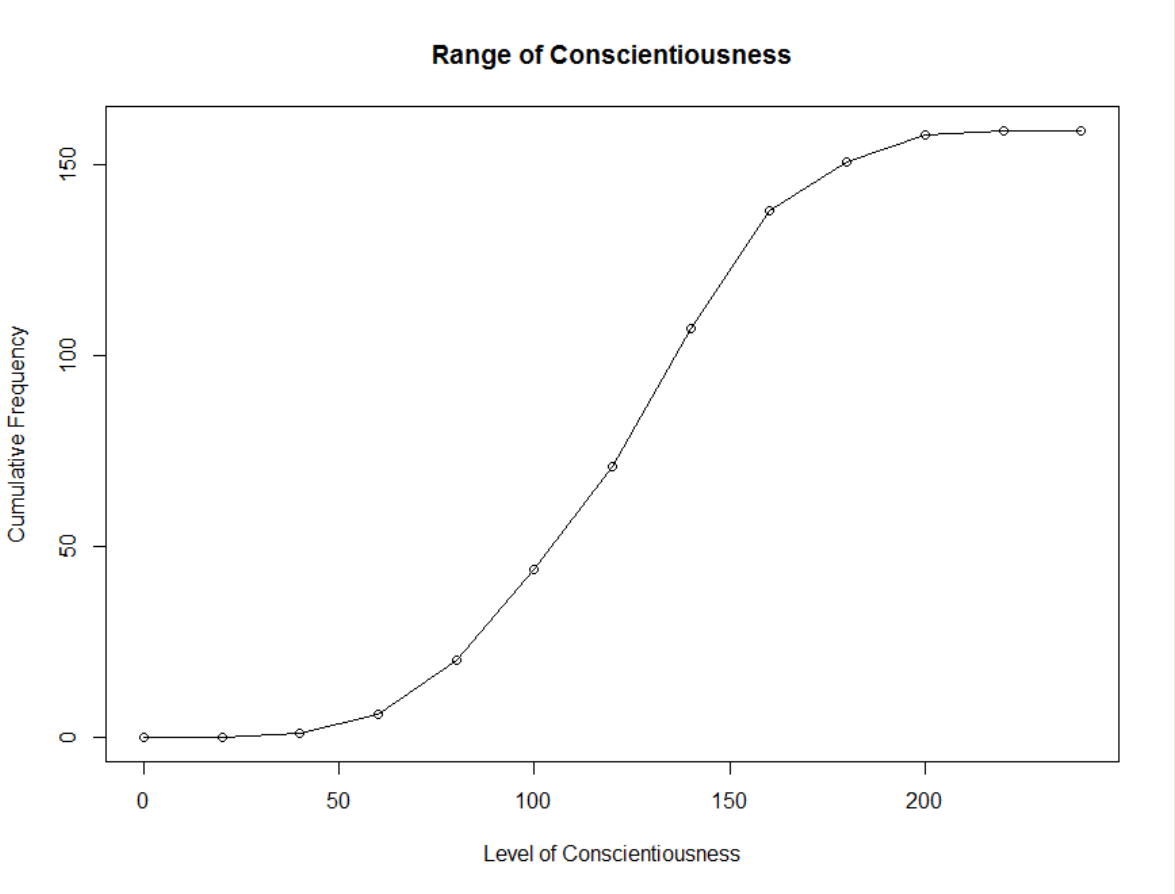
To construct an ogive, you need to format the data into cumulative frequencies:

```
> cumfreq0 <- c(0, cumsum(table(cut(dat$CONSCIENTIOUSNESS,
seq(0, 240, by=20), right=FALSE))))
```

Then plot the chart based on this data:

```
> plot(seq(0, 240, by=20), cumfreq0, main="Range of
Conscientiousness", xlab="Level of Conscientiousness",
ylab="Cumulative Frequency") + lines(seq(0, 240, by=20),
cumfreq0)
```

Ogive



Range of Conscientiousness
Cumulative Frequency
0
50
100
150
0
50
100
150
200
Level of Conscientiousness

Pie Chart

- A circular depiction of data where the area of the whole pie = 100% of the data being studied. Slices represent a % breakdown of each of the values

- Business uses: e.g. for depicting budget categories, market share, time and resource allocation

- Generally more difficult to interpret the size of the slices compared to the bars in a histogram. But- usage of ‘%’ can clarify slice size

Pie Chart

Construction steps:

1. Convert each toothpaste brand amount to a proportion by dividing each individual amount by the total

   E.g. 102 / 200 = 0.51

2. Convert each proportion to degrees by multiplying by 360°

   E.g. 0.51 * 360° = 183.6°

Pie Chart

To construct a pie chart, you first need to label the categories:

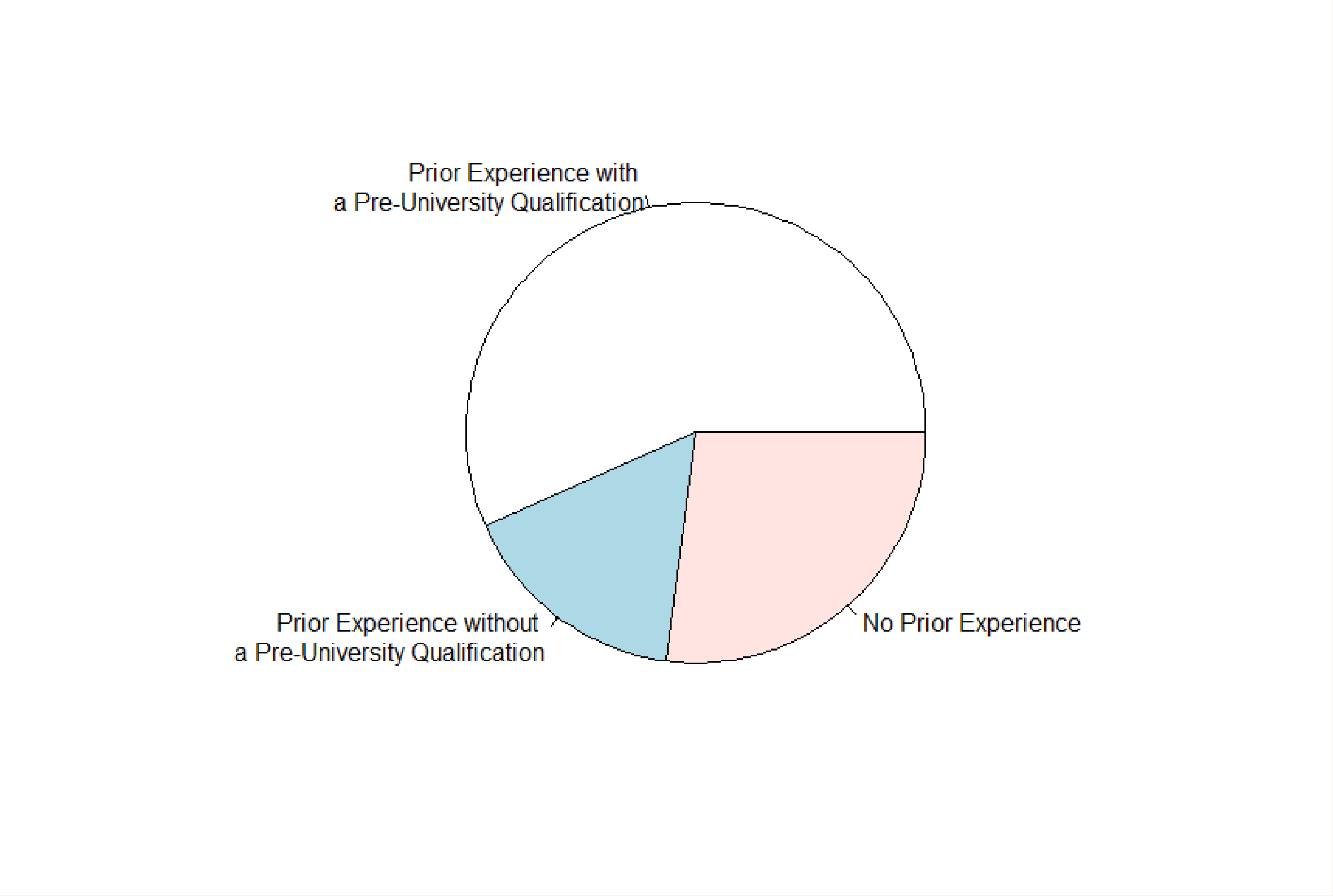
```
> Lbls <- c("Prior Experience with \na Pre-University Qualification", "Prior Experience without \na Pre-University Qualification", "No Prior Experience")
```

Then plot the pie chart based on this data:

```
> pie(table(dat$PRIOR_EXP), labels = lbls)
```

GAMES ACADEMY | FALMOUTH UNIVERSITY

Pie Chart



Pie Chart

Wrapping up commands into a single line:

```
> qplot(factor(dat$PRIOR_EXP, labels=c("Prior Experience with
\na Pre-University Qualification", "Prior Experience without \na
Pre-University Qualification", "No Prior Experience")),
dat$ANXIETY, geom = "boxplot", main="Anxiety of Students from
Different Backgrounds", xlab="Background", ylab="Level of
Programming Anxiety")
```

Take note of the **factor()** command --- useful for distinguishing between different nominal characteristics or members of experimental and non-experimental groups!

The labels struct is setup in ascending order.

Scatter Plot

- Illustrates the relationship between two variables based on its underlying data points
- E.g. the link between neurotic personality traits and programming anxiety
- Scatter graph - a two-dimensional graph plot of pairs of points from two variables
- Relationships will vary in strength, line of best fit used to indicate magnitude through slope

GAMES ACADEMY | FALMOUTH UNIVERSITY

Scatter Plot

For the line of best fit, going to use the **rlm()** function for a robust linear model to mitigate the influence of dataset outliers:

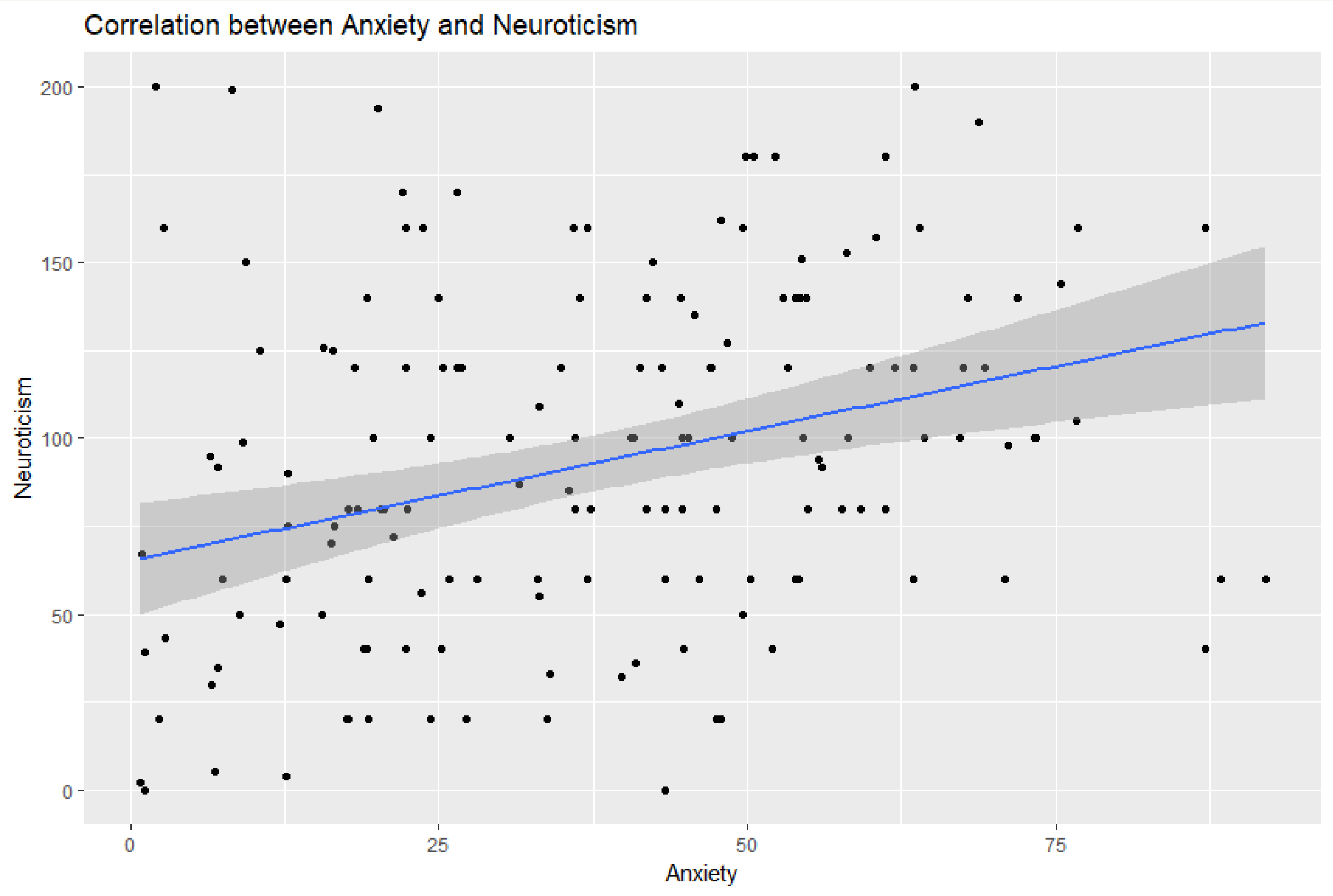
```
> library(MASS)
```

Then using the **qplot()** command with a combination of **geom** arguments including “point” (the scatter dots) and “smooth” (the line of best fit.

```
> qplot(dat$ANXIETY, dat$NEUROTICISM, geom = c("point",
"smooth"), method="rlm", main="Correlation between Anxiety and
Neuroticism", xlab="Anxiety", ylab="Neuroticism")
```

Try without the **method** argument. What happens?

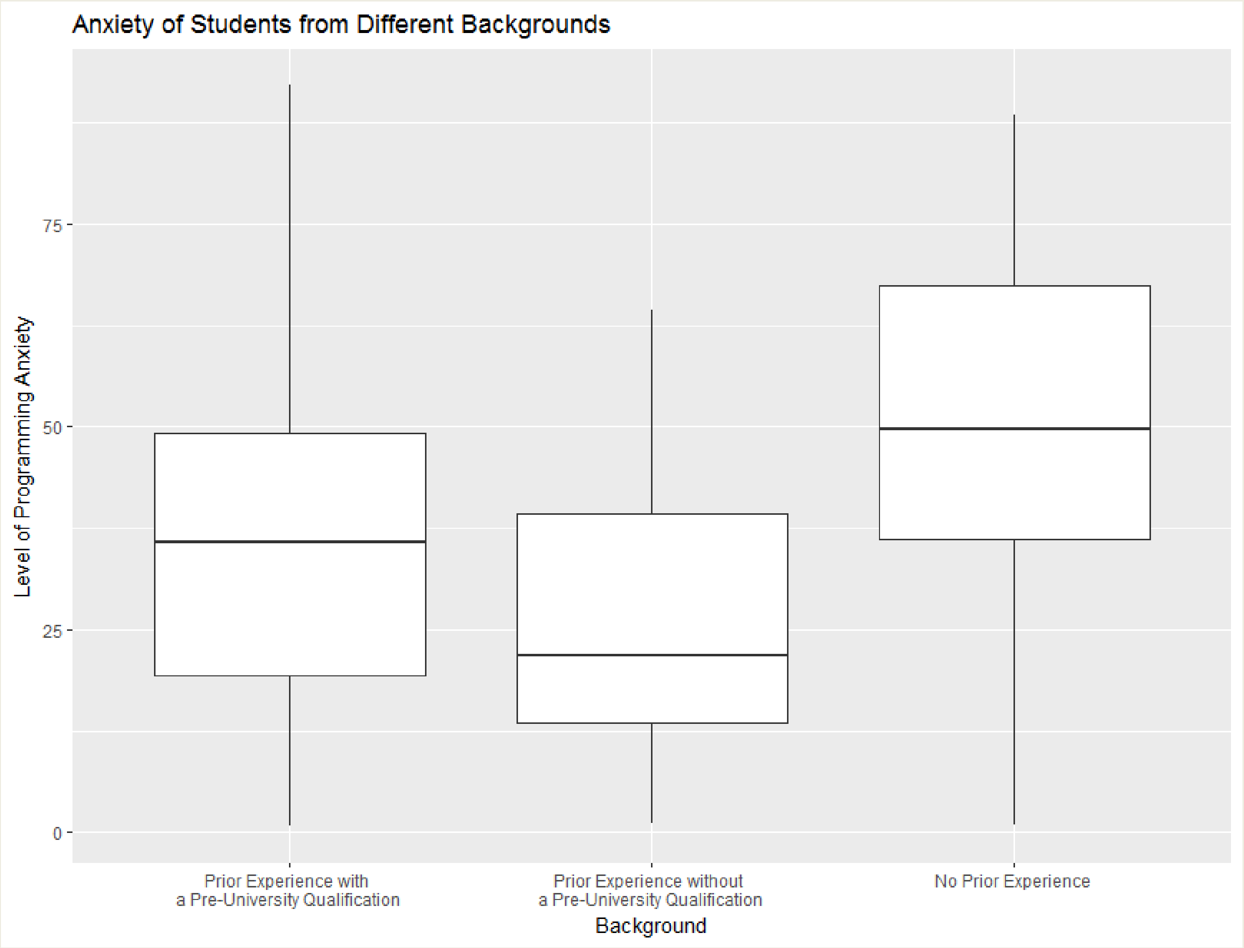
Scatter Plot



Box Plot

- Illustrates proportions of a distribution, and useful for comparing groups
- Looking “down” on the distribution from the top, like viewing hills on a plane
- Lines outside the box represent range
- Box represents the lower and upper quartiles
- Line inside the box represents the median

Box Plot


Anxiety of Students from Different Backgrounds
Level of Programming Anxiety
75
50
25
0
Prior Experience with a Pre-University Qualification
Prior Experience without a Pre-University Qualification
No Prior Experience
Background