COMP320

RESEARCH PRACTICE

Statistical Data Analysis II

Presentations

Presentations

- Scheduled for Monday and Tuesday next week – check timetable for individual slots
- Assessed on a **threshold** basis – i.e. turn up and present to pass
- Prepare a **10 minute** presentation
- Use whatever you like to prepare slides
  - But please ensure you have your slides readily accessible on the day – I strongly advise bringing them on a USB stick rather than relying on the cloud!
- Consider this a **practice run** for your proposal vivas after Christmas

What to include

- (E) Deliver a 10-minute practice presentation that will:
    i. explain the context of your project
    ii. identify and discuss the scientific literature relevant to your project
    iii. propose one or more research questions for your project
    iv. articulate the ethical considerations you have made
    v. illustrate your approach to collecting, analysing, and presenting data

What to include

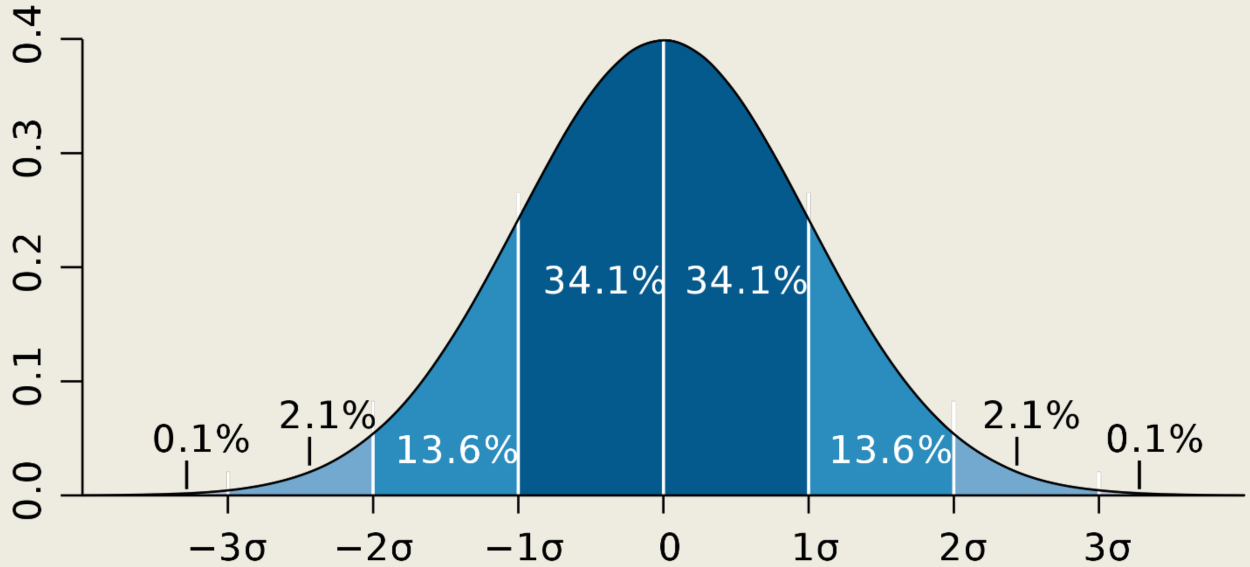
- Part E consists of a single formative submission. This work is individual and will be assessed on a threshold basis. To pass, answer the following questions:

i. What is the **context** of your project? How does it fit into the research field of computing for games?
ii. What are the **key results from the literature** upon which your project will be built?
iii. What is the **current state of knowledge** in the field? What are the open questions and challenges?
iv. What is (are) the key **research question(s)** that you will seek to answer in your project?
v. What are the key **legal, social, ethical, and/or professional issues** associated with your project?

Effect Size

Standard Deviation

- Denoted σ
- A measure of the "width", "spread", "variability" of a normal distribution

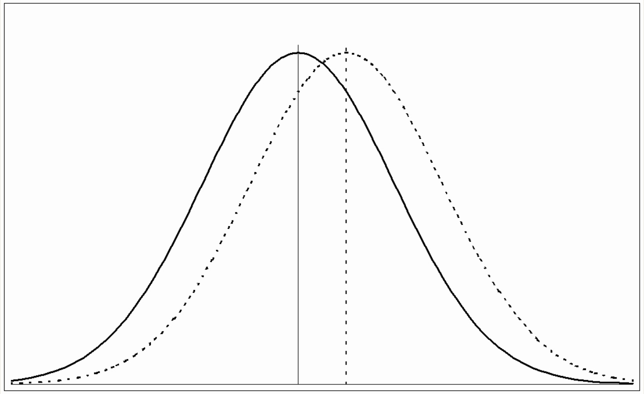


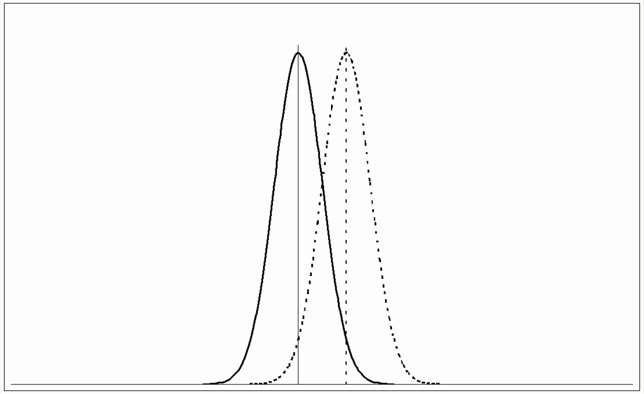
https://commons.wikimedia.org/wiki/File:Standard_deviation_diagram.svg

GAMES ACADEMY | FALMOUTH UNIVERSITY

Effect Size

- Consider two samples with means μ_1 and μ_2
- Is the difference between μ_1 and μ_2 statistically significant?





https://www.leeds.ac.uk/educol/documents/00002182.htm

Effect Size

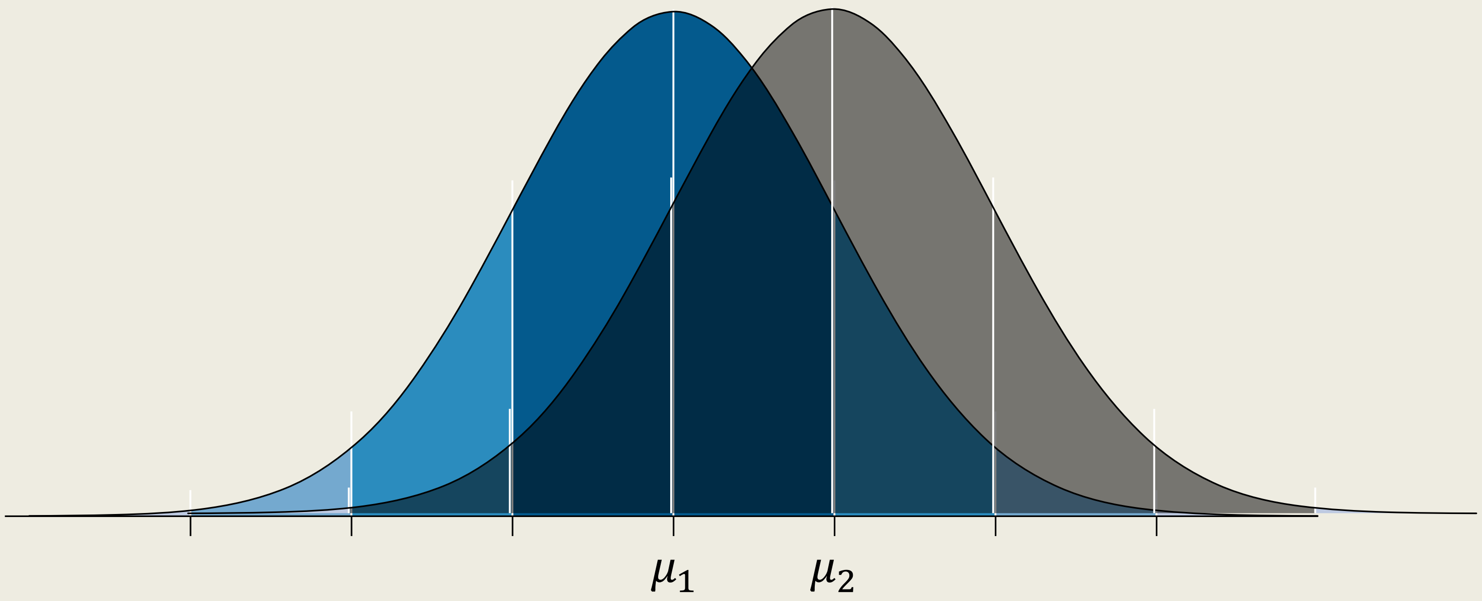
- The means don't tell us much without also considering the standard deviation
- What matters is the **effect size** – the standardised difference between means

$$\text{effect size} = \frac{|\mu_1 - \mu_2|}{\sigma}$$

- Here σ should be the population standard deviation – in practice this is probably not known so must be estimated from the samples

Effect Size

- So an effect size of 1 means that μ_1 and μ_2 are one standard deviation apart



Effect Size – Rules of Thumb for t-tests

- First proposed by Cohen (1988), expanded by Sawilowski (2009)

Effect Size d	Description
0.01	Tiny
0.2	Small
0.5	Medium
0.8	Large
1.2	Very large
2.0	Huge

Effect Size – Rules of Thumb for ANOVA

- ANOVA uses a different effect size measure, so Cohen has different guidelines:

Effect Size f	Description
0.1	Small
0.25	Medium
0.4	Large

Power Analysis with G*Power

What is power analysis?

- Statistical significance depends on several factors
  - α – probability of type 1 error
  - β – probability of type 2 error
  - d – effect size
  - N – sample size
- Power analysis allows us to determine one of these given the others
- Most useful for us: determine what N gives us a desired α, β, d
- I.e. determine how many participants you need for your experiments!

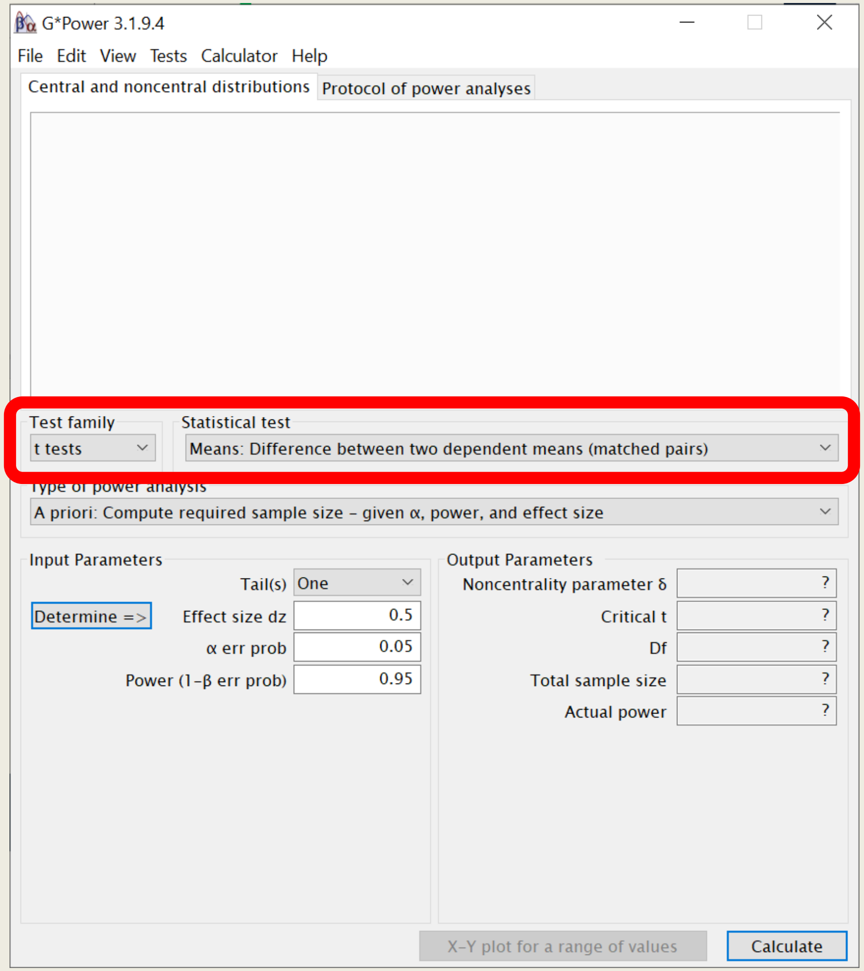
G*Power

- Download from http://www.gpower.hhu.de/
- Available for Windows and MacOSX

GAMES ACADEMY | FALMOUTH UNIVERSITY

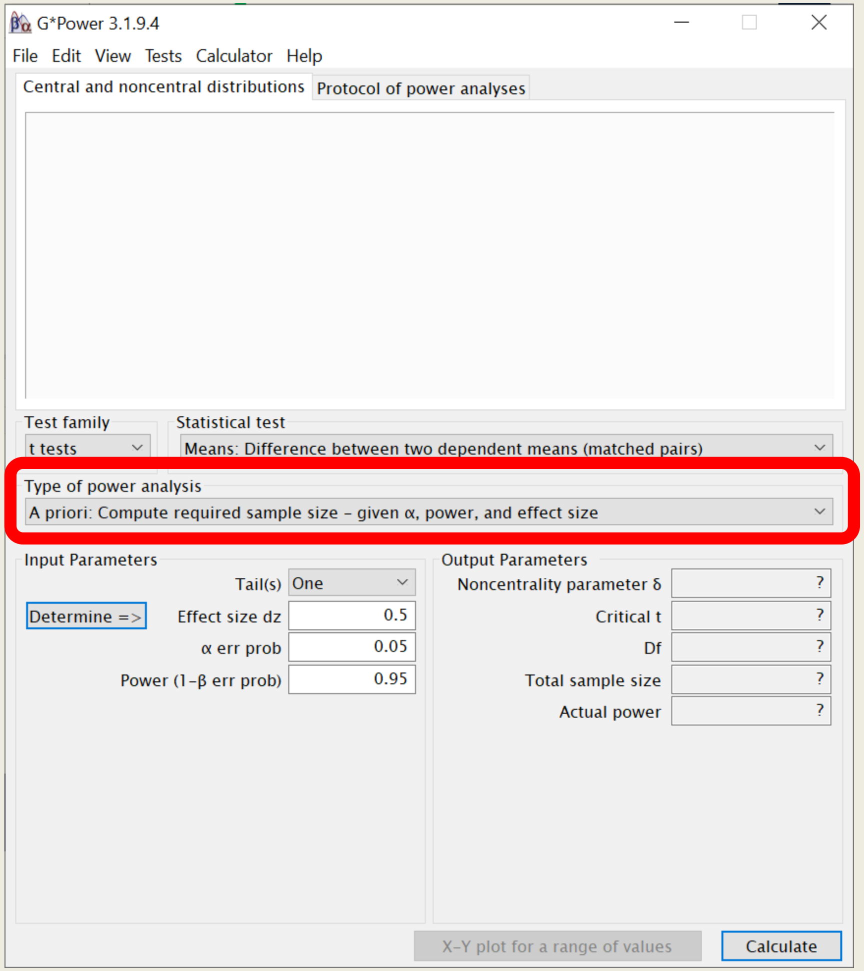
Common Settings

- G*Power supports many different statistical tests – it's vitally important to choose the correct one!
- More on this in a few slides' time



Common Settings

- *A priori* analysis (to calculate sample size) is the most useful for us
- Other settings are useful e.g. for analysing results *a posteriori*, for designing replication studies, etc.



GAMES ACADEMY | FALMOUTH UNIVERSITY

Common Settings

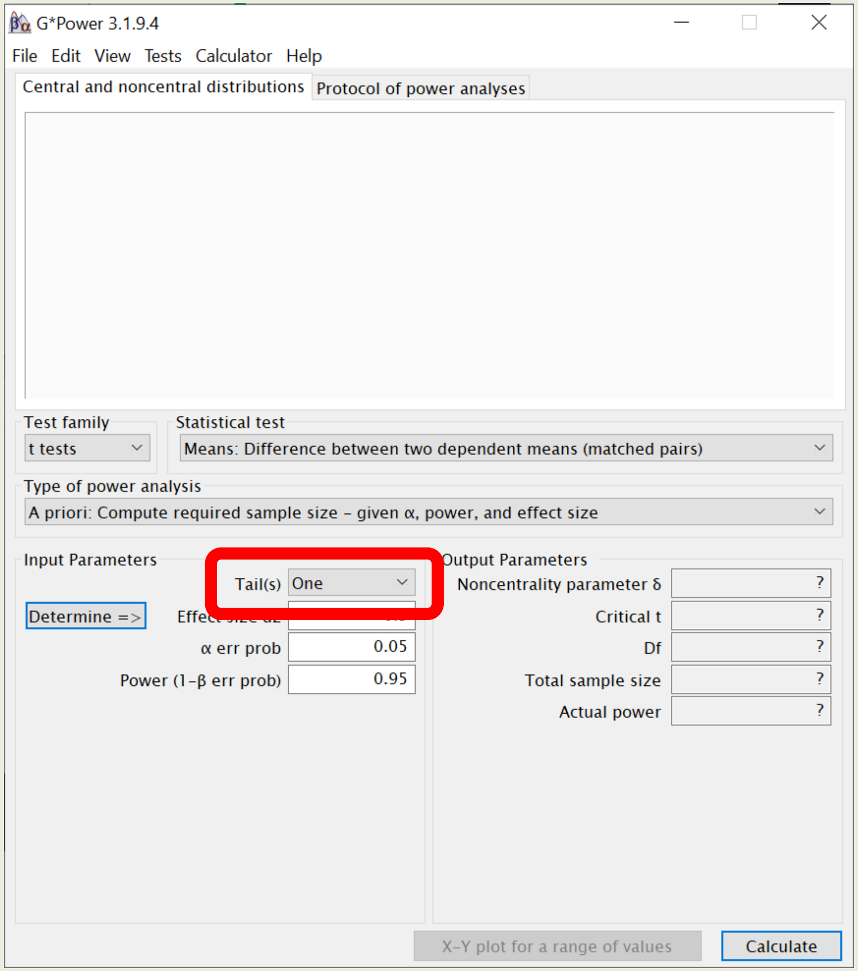
- One-tailed test:

$$\mu_1 < \mu_2$$

- Two-tailed test:

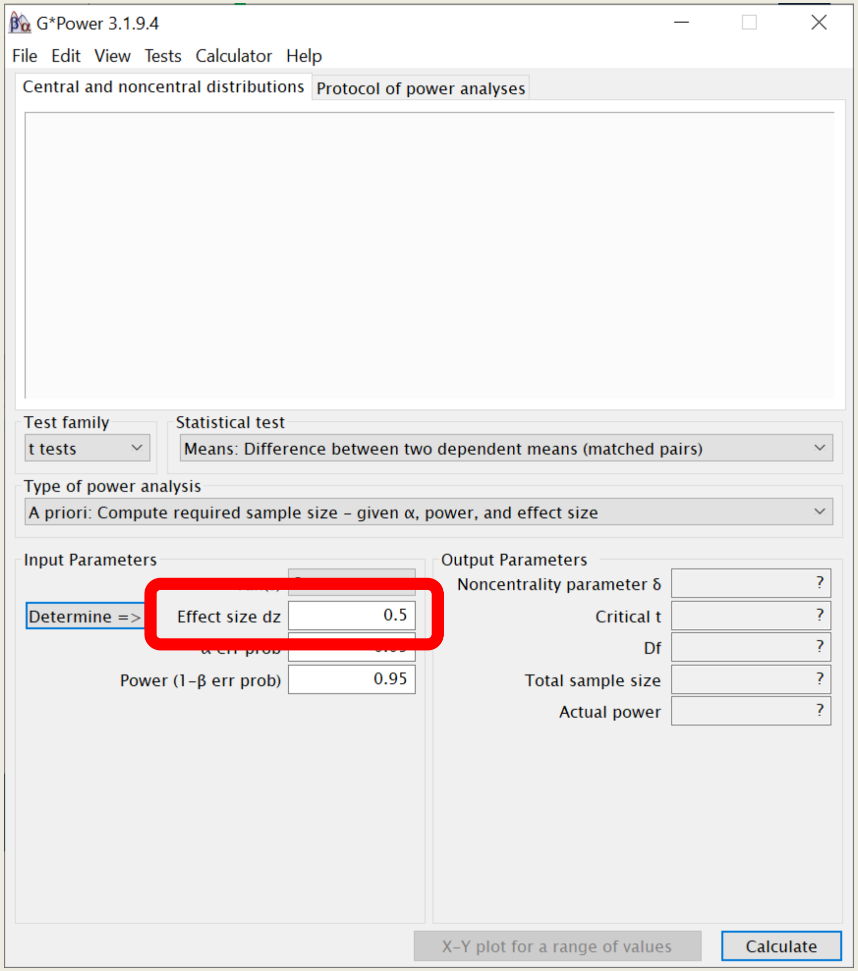
$$\mu_1 \neq \mu_2$$

- If your hypothesis states whether something will increase or decrease, that's one-tailed
- If it only states that something will change, that's two-tailed



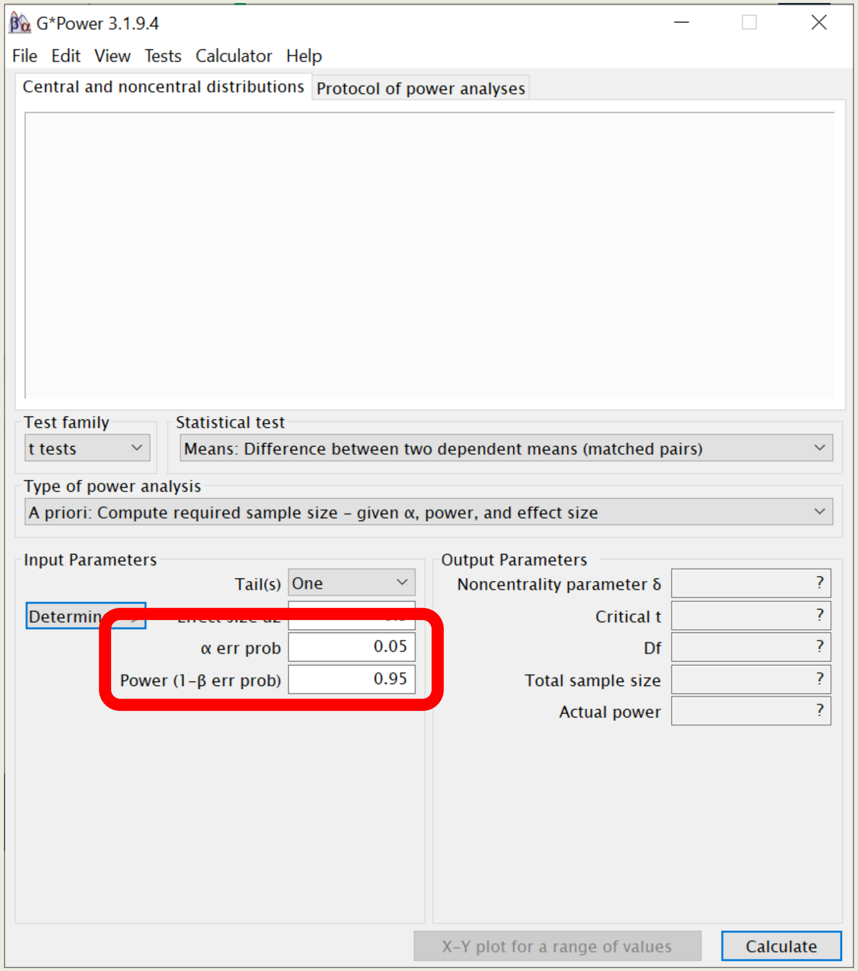
Common Settings

- I recommend using Cohen’s guidelines to set effect size
- For the scope of your dissertation project, best to stick to looking for “medium” (0.5) or “large” (0.8) effects
- The “Determine” button is for calculating effect size from existing data



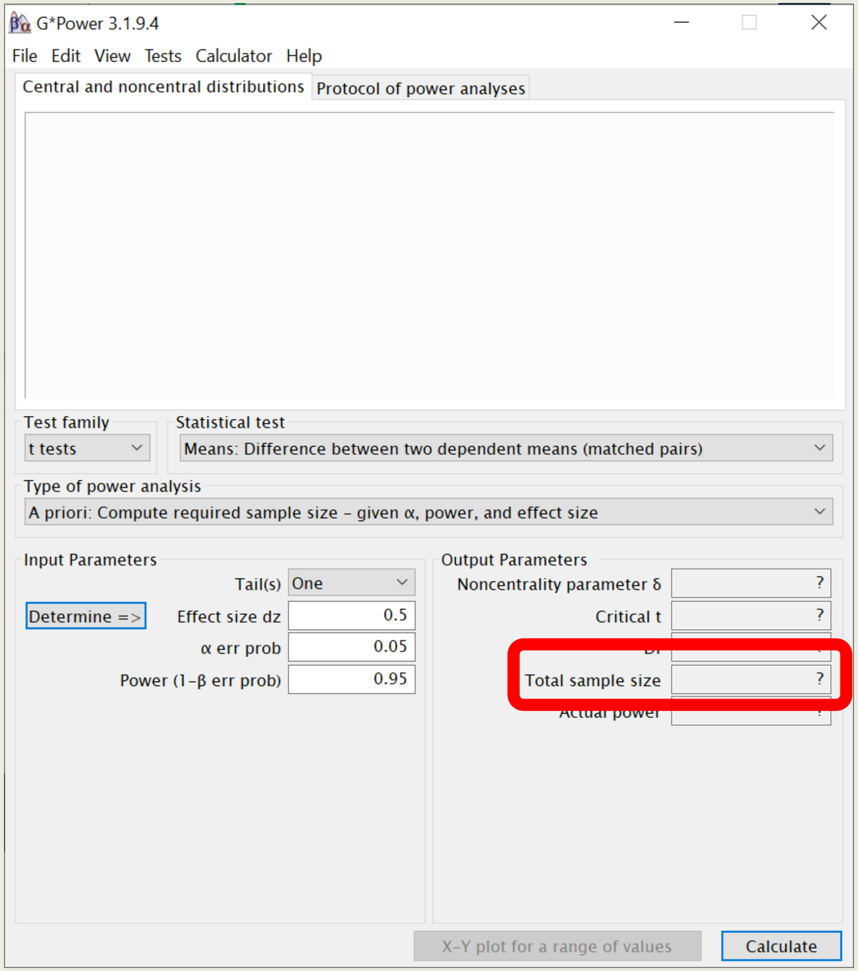
Common Settings

- $\alpha = \beta = 0.05$ (i.e. 95% confidence) is standard
- Advisable to reduce α if testing multiple hypotheses to mitigate for Type 1 error inflation (“green jelly beans cause acne”)
- Bonferroni correction: divide α by the number of hypotheses



Common Settings

- The total sample size is what we're trying to find – the number of participants!



Example 1

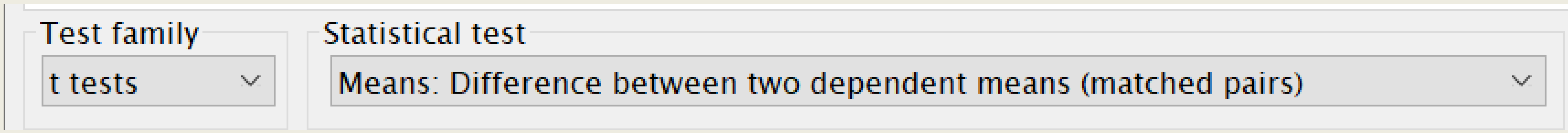
- Hypothesis: the presence of background music in a game affects the length of time the player takes to complete the level
- Experiment design: A-B test (A=music, B=no music), each participant plays both variants in a random order

Example 1

- We are comparing two means so we should use a t-test
- How many tails?
- Samples in the two groups are **paired** – same participants for both groups

Example 1

- Set the test:



- All other settings as discussed on previous slides

Example 2

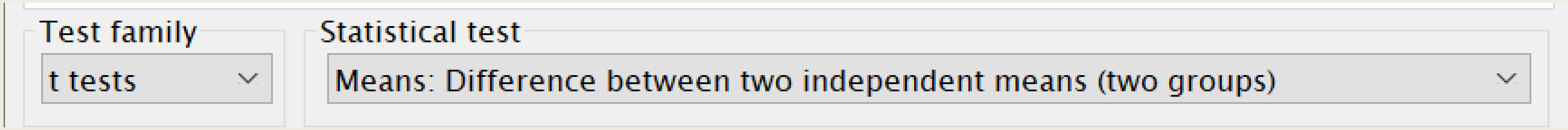
- Hypothesis: Games Academy students will achieve a higher score in the game than students from other courses
- Experiment design: two groups (GA students, non-GA students)

Example 2

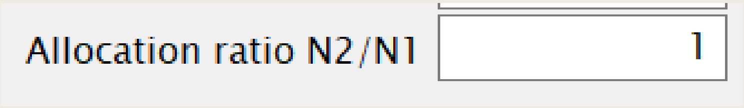
- We are comparing two means so we should use a t-test
- How many tails?
- Samples in the two groups are **independent** – different participants for each group

Example 2

- Set the test:



- Allocation ratio = 1 for equal number of participants in each group, or another value if not:



- All other settings as discussed on previous slides

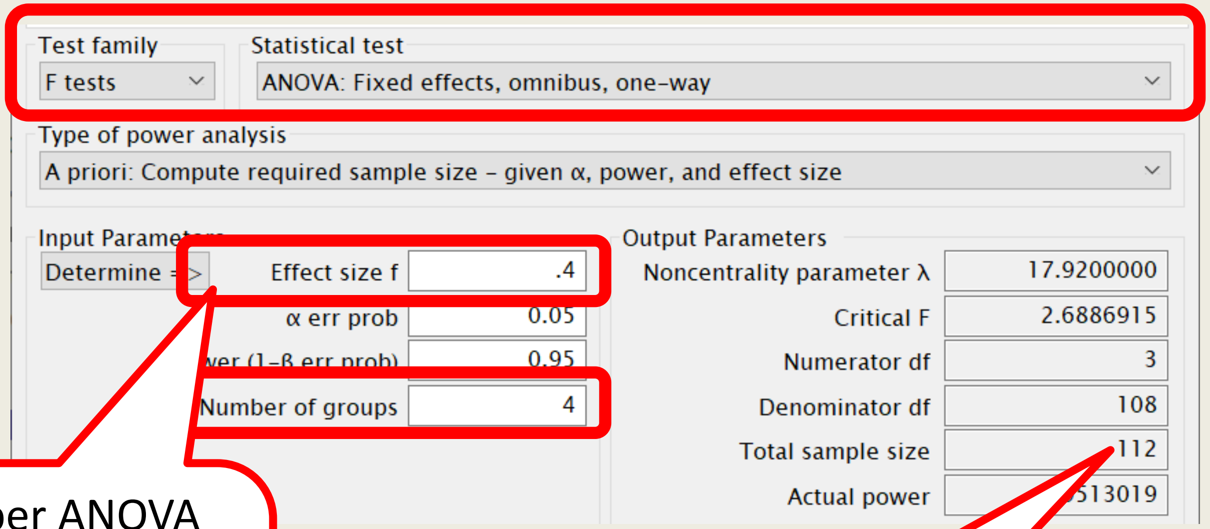
Example 3

- Hypothesis: the colour of the protagonist’s hair in the game will influence the player’s enjoyment
- Experiment design: multiple groups, each participant assigned to a group at random

Example 3

- We are comparing more than two means so we should use ANOVA
- ANOVA is always two-tailed (with more than two means we can’t make a one-tailed hypothesis)
- One independent variable (colour)
→ one-way ANOVA

Example 3



Remember ANOVA uses a different effect size measure to t-tests so Cohen's guidelines are different!

To be divided equally between groups

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

GAMES ACADEMY | FALMOUTH UNIVERSITY

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

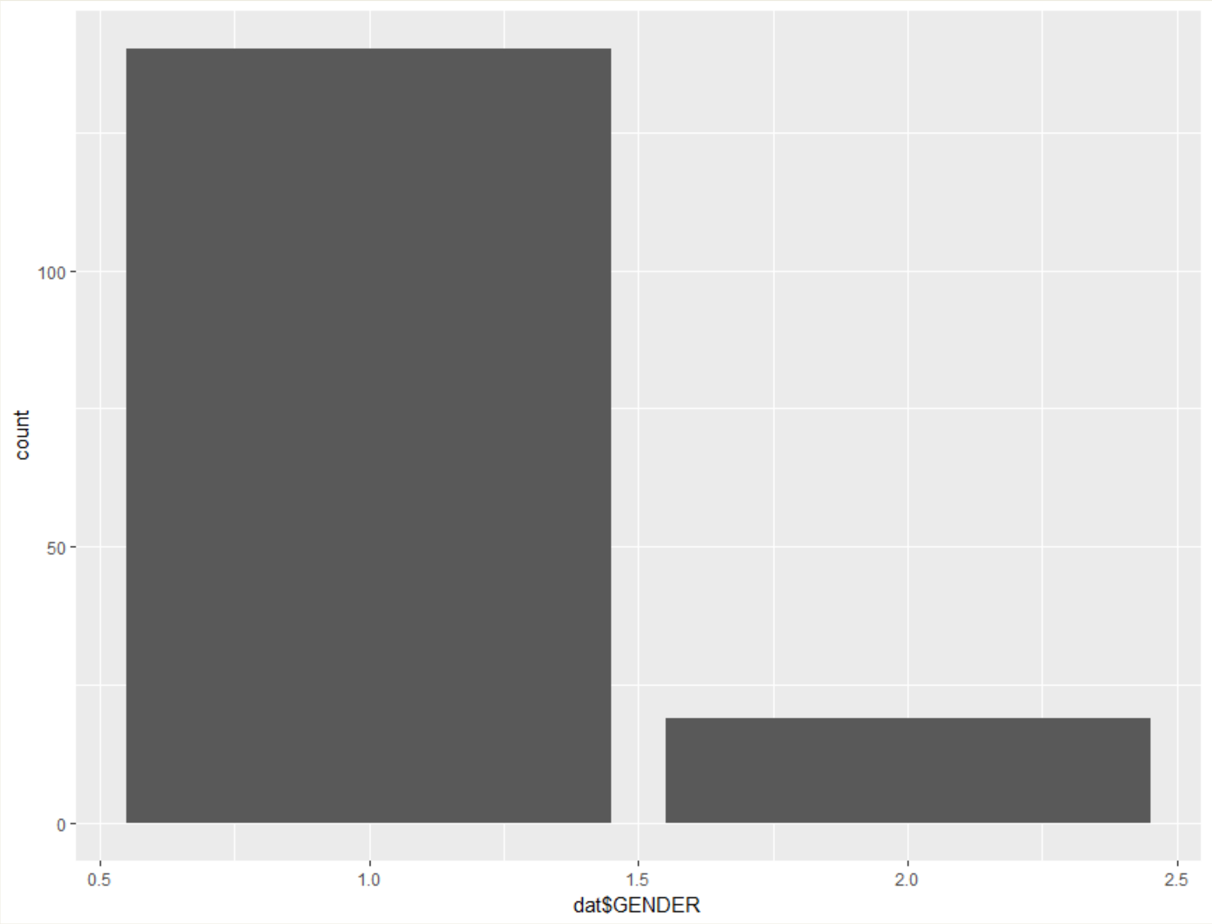
To a simple bar chart:

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

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

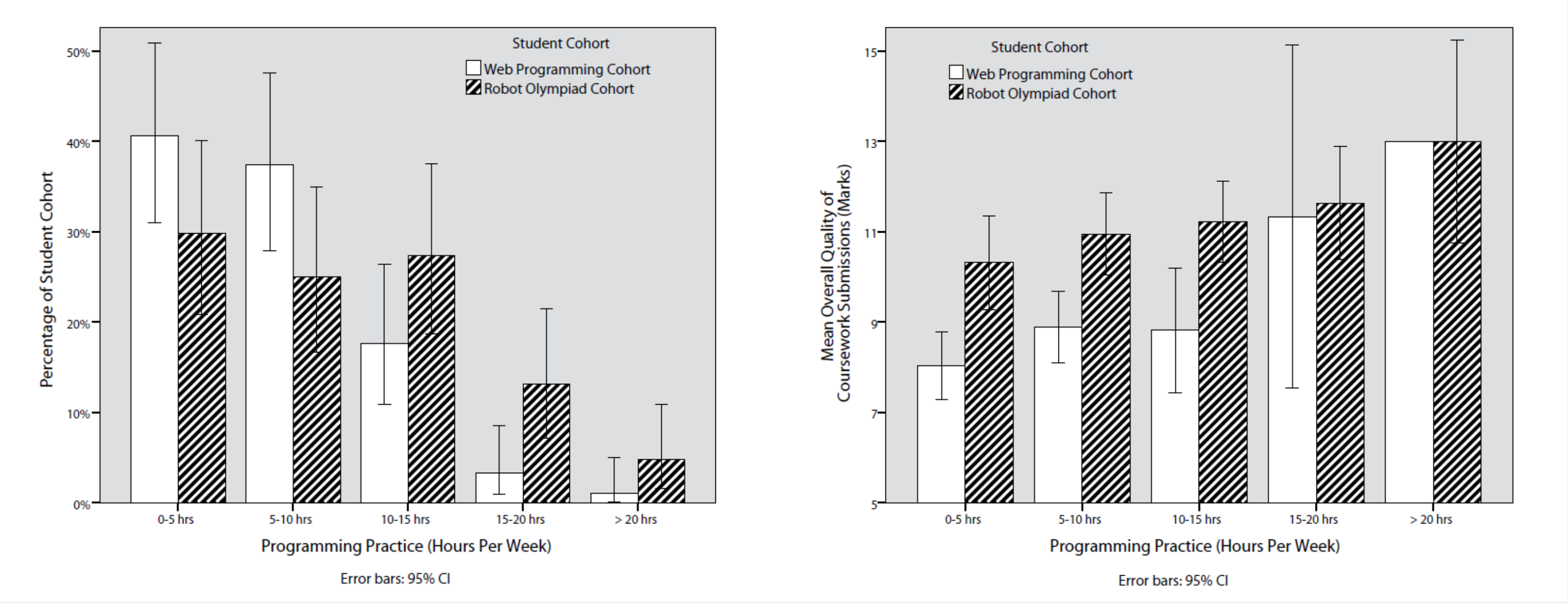
The **stat** argument is depreciated, but useful when no statistical analysis or summary statistic like a mean is needed. For a bar chart, “identify” defaults to a count.

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

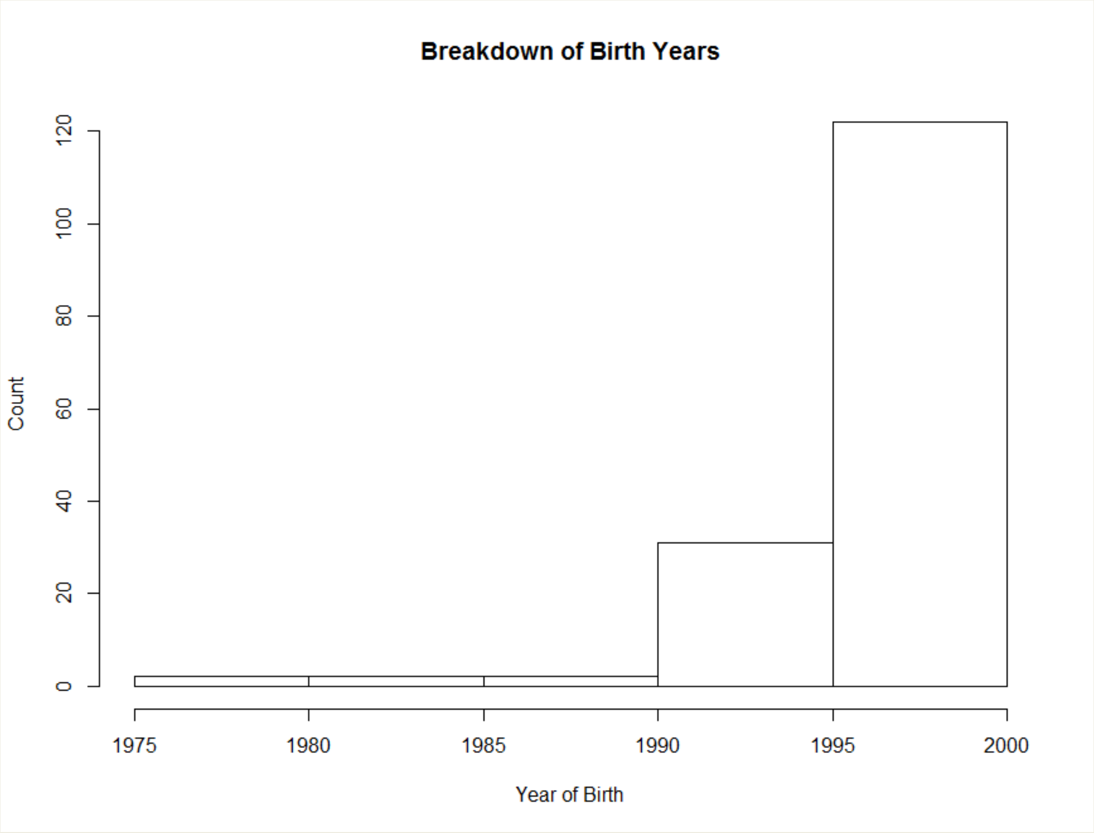
To display a frequency table, examining birth years from 1975 to 2000 in 5-year intervals, use the following command:

```
> 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

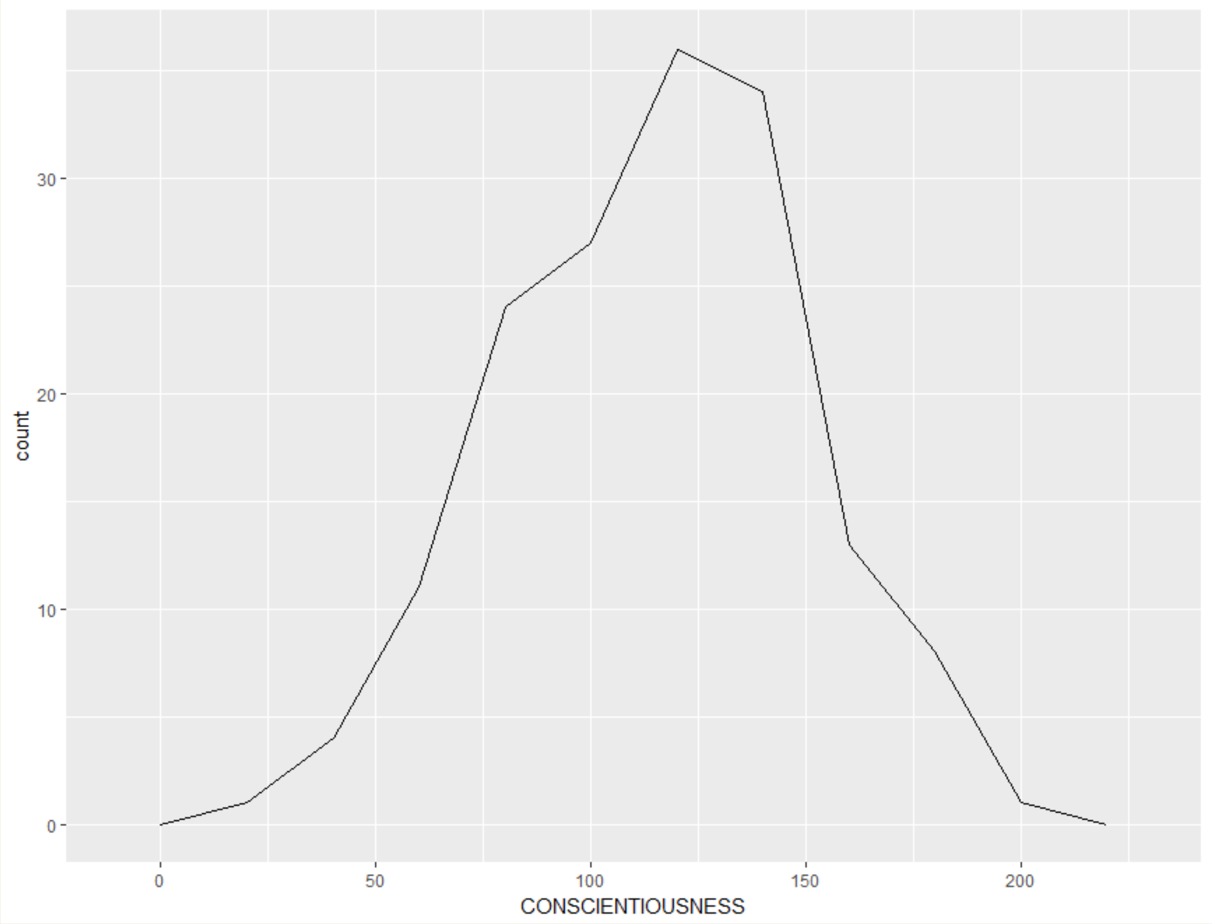
- Construction steps:
  - Label the **x** axis with the class endpoints
  - Label the **y** axis with the frequencies
  - Plot a dot for the frequency value at the midpoint of each class interval (different to Ogive)
  - Connect these dots with a line

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

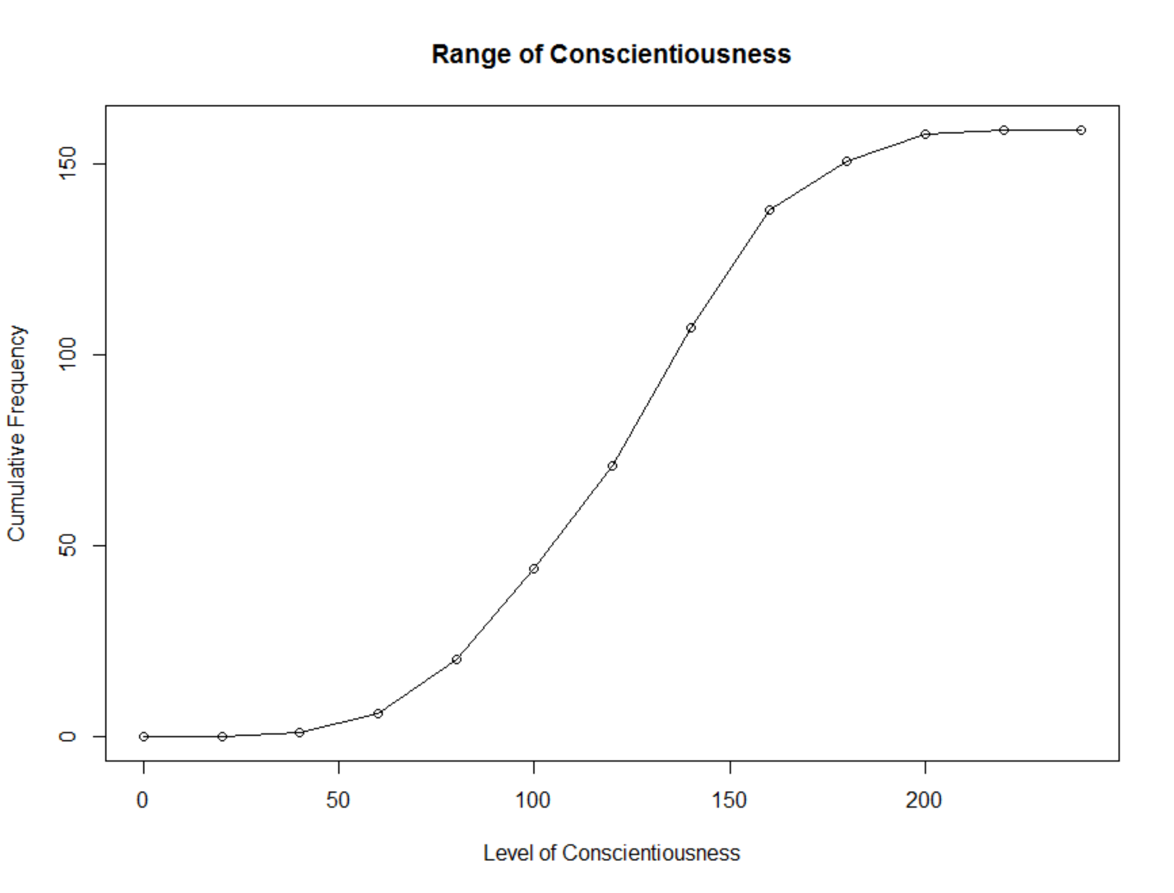
To construct an ogive, you need to format the data into cumulative frequencies:

```
> 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

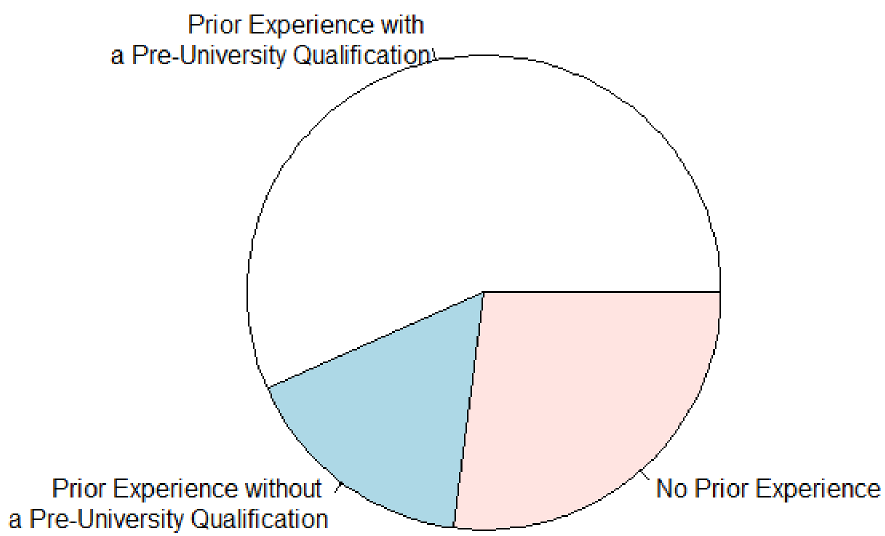
To construct a pie chart, you first need to label the categories:

```
> 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)
```

Pie Chart


Prior Experience with
a Pre-University Qualification
Prior Experience without
a Pre-University Qualification
No Prior Experience

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

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)
```

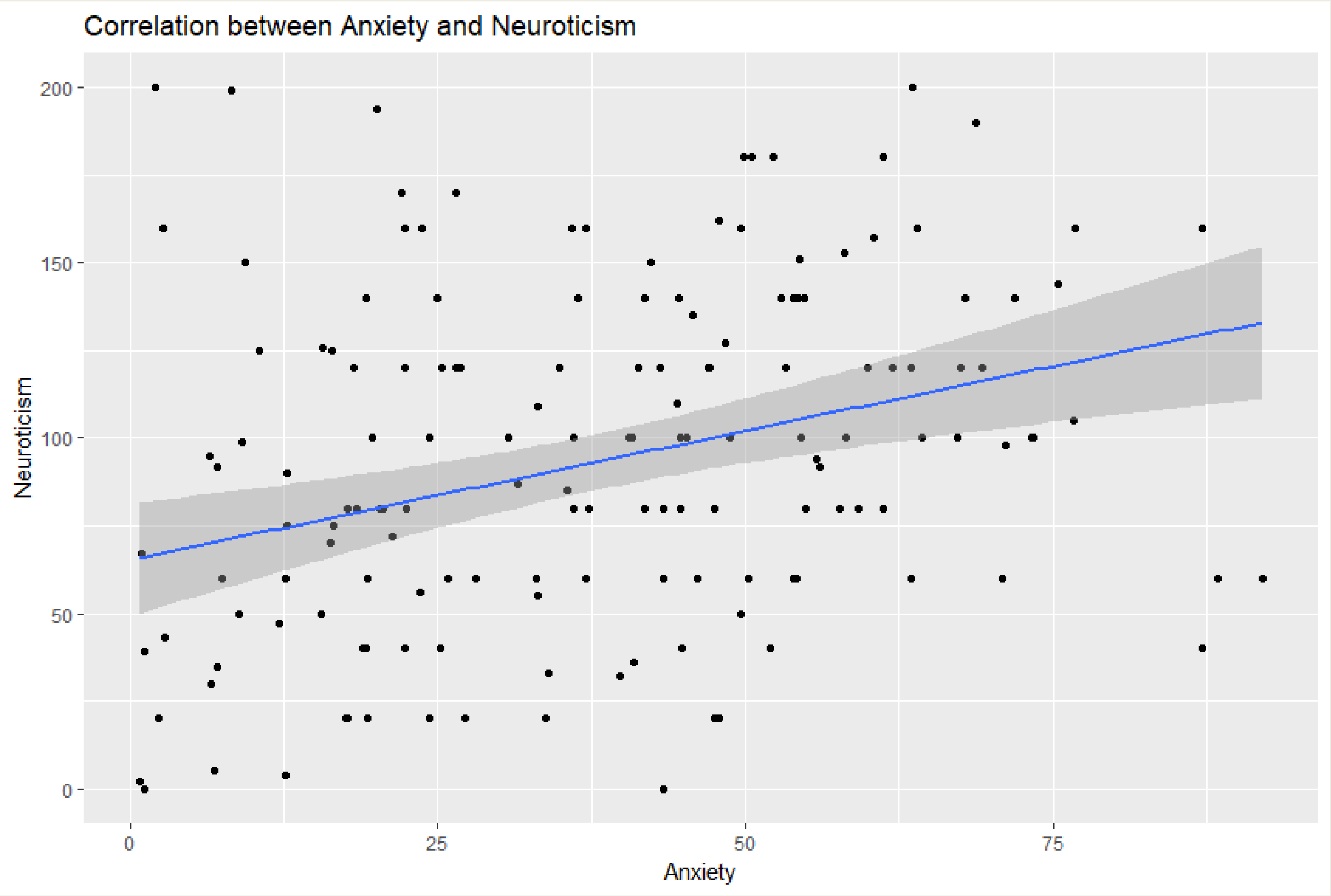
Then using the **qplot()** command with a combination of **geom** arguments including “point” (the scatter dots) and “smooth” (the line of best fit.

```
> 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?

GAMES ACADEMY | FALMOUTH UNIVERSITY

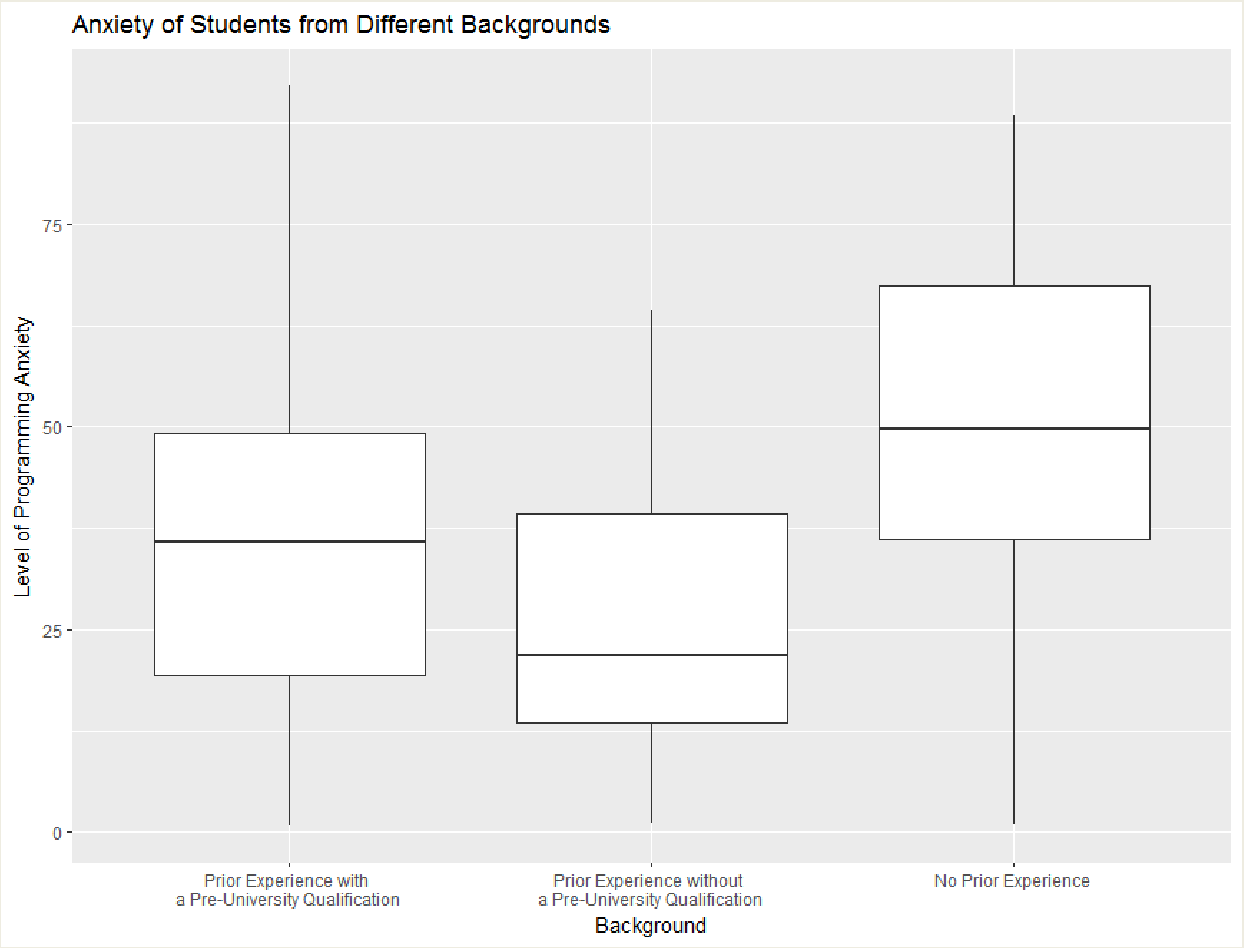
Scatter Plot



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