good structure

Assignment 1

Code

Student 2

Hide

```
knitr::opts_chunk$set(echo = TRUE)
```

Introduction

In August 2021, the DATA2X02 cohort at the University of Sydney was surveyed across 24 variables. 211 surveys were completed our of the 754 students in the cohort. For the purpose of this report, ‘student’ refers to DATA2X02 students enrolled in the course in Semester 2, 2021.

This report uses these 211 survey responses as a sample to conduct four hypothesis tests:

1. Does the number of COVID tests a student has taken in the past two months follow a Poisson distribution? Data suggests **it doesn’t.**
2. Is average self-assessed mathematical ability of male students higher than that of female students? Data suggests **it is.**
3. Are perceived stress levels of students independent of gender? Data suggests that **they are not.**

4. Do students score themselves higher on average in mathematical ability than in R coding ability? Data suggests that **they do.**

Additionally, this report investigates whether the data is a representative, random sample of the population. It also explores biases and methods for improving the survey.

Sampling Method, Biases and Potential Improvements

Given the challenges with surveying the population as a whole, sampling allows researches to infer information about the population based on the results of a subset of the population. To control for confounding variables, probability sampling is ideal because it means all eligible individuals in the population have an equal chance of being sampled. This is known as random sampling. **Random sampling is a key assumption underlying the hypothesis tests in this report.**

However, to the contrary, **this survey used non-probability (non-random) sampling.** Instead, individuals self-selected into the survey which was posted through the DATA2X02 discussion boards. This is known as convenience sampling. It introduces several biases.

- **Volunteer / selection bias:** Students who volunteer to participate might be different from those who do not. For example, volunteers might be those who are less stressed because they have more free time. They might be those who spend less time exercising and more time at their laptops. The survey might be biased towards one gender. Those who complete the survey might be demonstrating conscientiousness which correlates with mathematical ability and r coding ability.
    - Nevertheless, the possibility of **sampling error and the lack of representation should not be overstate**d for this particular study. Across the variables examined in this report, it is not expected that the confounding variables that could arise from the non-random sampling method will have large impact. For this reason, the assumption of random sampling is maintained for the purpose of hypothesis tests. Any areas where bias might occur are flagged and addressed at the relevant point.

- **Lack of mutual exclusivity in the survey mechanism:** A further source of error is that no unique identifier was used to ensure that students only completed the survey once. Regardless, due to the time and effort required in survey completion, it is unlikely that many, if any, students completed the survey more than once. For the purpose of this report, it is assumed that each survey response corresponds to a unique student. and these students are independent of each other (eg. not siblings)
- **Response bias (subjectivity):** Any question that requires subjective assessments is influenced by one's emotional current state (Kahneman, 2011). Indeed, it has been shown that we remember events by the peak emotion felt and the final emotion felt (Geng, Chen, Lam, & Zheng, 2013). This has implications for several questions. For example, one's self-assessment of mathematical ability would differ depending on whether they had last succeeded or struggled with a math problem. The same can be said of R coding ability. The survey's temporal location throughout the progression of the DATA2X02 semester could therefore bias responses. ✓
- **Response bias (anchoring):** Numerical questions are biased by anchoring effects (Jacowitz & Kahneman, 1995). Anchoring effects are when exposure to prior value (typically high or low) acts as a reference point for subsequent responses. Questions like the average entry salary in AUD of a data scientist might be influenced by the magnitude of preceding survey responses. The questions were presented in a random order to try to minimise this kind of bias.
- **Recall bias:** Memory is fallible. This might influence questions such as the COVID testing and emails questions. In particular, it could be harder for someone to recall a count of 8 COVID tests rather than 1 test. This is due to the diminishing marginal salience in recollecting multiple events of the same kind (Nadel, Hupbach, Gomez, & Newman-Smith, 2012). ✓
- **Design flaw bias:** Certain questions were not conducive to receiving a valid response. For example:
    - **Email question:** Asking participants to recall how many non-spam emails they received in their university emails last Friday is hard to recall; it's also unrealistic to expect participants to check.
    - **Salary question:** The salary question should have specified that *annual* salary is the object of interest.
    - **Stress question:** Given the wide subjectivity in ascribing a quantitative result to stress levels, it might have been prudent to classify stress as either 'low', 'moderate', or 'high'. ✓

- **Data validation**: Across multiple questions, data validation would have been very useful to avoid tedious data cleaning of variables such as height, gender and salary.
    - Sliding scales for salary and height should have been used.
    - For the gender variable, a unique validation input of 'male', 'female', or 'other' should have been used instead of free text input.
    - All 1-10 scales should have given subjective guidance around what each score meant - for example, that 5 means average.
    - The email sign-off question could have benefit from a ~5-8 option data validation drop-down box, with 'other' as an option.

Preliminary Data Setup

Package Import

First, we import R packages that **give greater functionality** to the code used in this report. Most of these are packages from the main repository of R packages (CRAN), however a lesser-known package, **`gendercoder`** , was imported to facilitate efficient re-coding of free-text gender responses.

Hide

```
library(tidyverse)
library(janitor)
library(ggplot2)
library(pwr)
library(gendercoder)  # not a CRAN package. From Github. Permits efficient
        re-coding of free-text gender responses.
library(kableExtra) # used to improve the visualisation of tables
library(knitr)
library(plotrix) # used for graphing distributions over histograms
library(RColorBrewer)
#library(MKmisc)
library(broom)
library(powerAnalysis)
library(MKpower)
```

Data import and cleaning

The CSV data file was imported and names were cleaned to make them easier to use in the code.

Abridged column names were assigned to the 24 columns.

The `visdat` package was used to visualise **missing data entries.** Missing data entries might be due to non-response errors or other errors in the collection process.

No global data cleaning was conducted. Missing values, 'absurd' entries and other data cleaning will be conducted on a case-by-case basis for each hypothesis test.

Hide

```
data <- readr::read_csv("DATA_assignment1_csv.csv") %>%
  janitor::clean_names()

column_names = c("time","covid_tests","living_arrangements","height",
                 "2_days_forward","in_aus","math_ability","r_ability",
                 "data2002_difficult","Uni_year","webcam","vaccination_stat
        us","social_media",
                 "gender","steak_preference","dominant_hand","stress",
                 "lonely","emails","sign_off","salary","unit","major","exer
        cise")
colnames(data) = column_names


total_students <- c(696,58)
```

Hide

```
visdat::vis_miss(data)
```

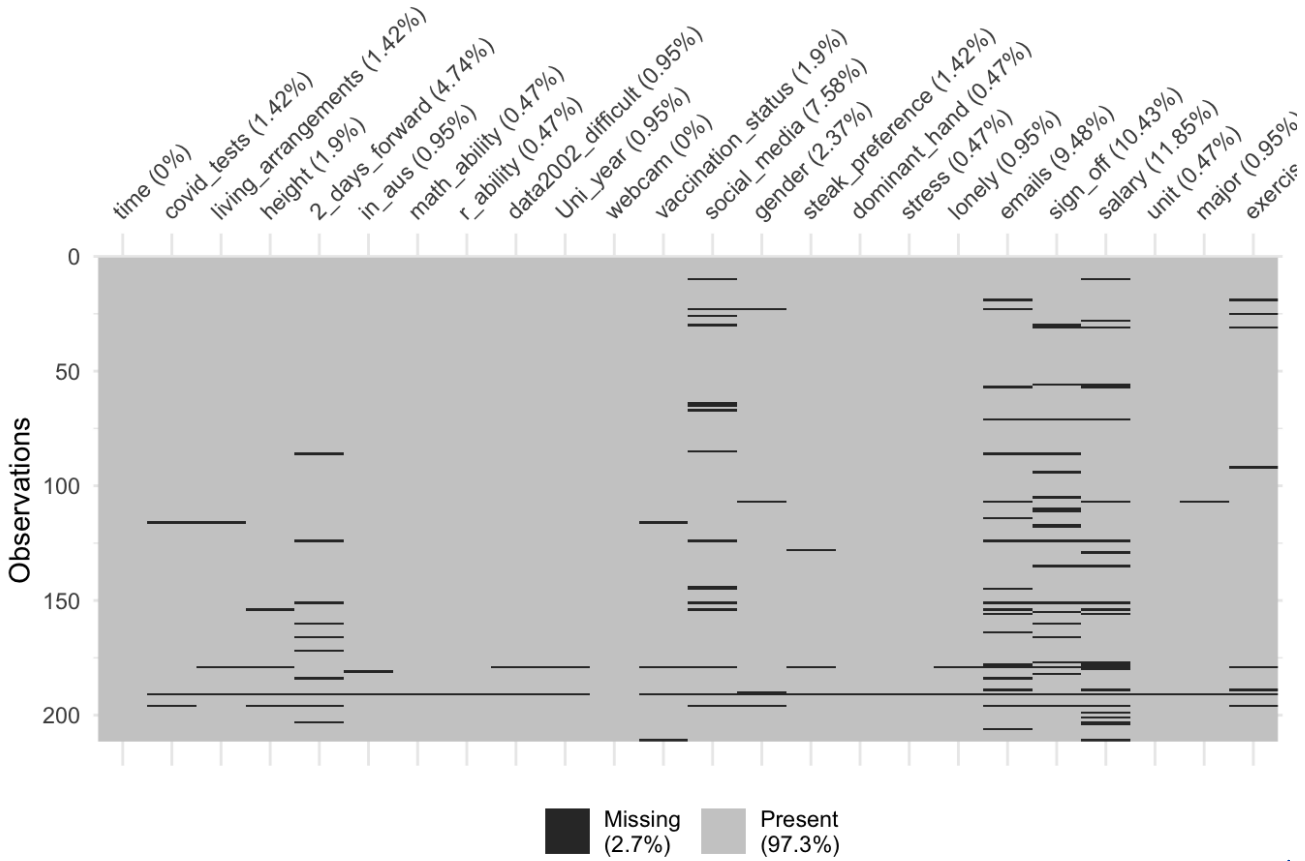


Figure caption with brief explanation of what this plot represents.

Test 1: COVID Testing Distribution

1: Introduction: *Does the number of COVID tests a student has taken in the past two months follow a Poisson distribution?*

A Poisson distribution is a discrete probability distribution that models the number or times an event is likely to occur over a fixed period. If an event happens both independently and randomly over time, and the mean rate of occurrence is constant, then a Poisson distribution might model the event.

In the case of COVID tests, it is not unreasonable to assume that COVID tests *in the past two months* have a **constant mean rate of occurrence**. The two month time condition is important because over the long-run the mean rate of occurrence of COVID tests certainly changes: two years ago we had less COVID tests than today, as we presumably will also two years from now. ✓

Whether COVID tests are **independent and random**, however, is disputable. This will be examined in the discussion, along with an investigation of **problems in data collection and validation.**

The Poison distribution is a discrete distribution, making it apt to the **discrete integer data** of the number of COVID tests.

Throughout this study, the words ‘class’, ‘cell’, and ‘group’ are used interchangeably to refer to the categories of 0, 1, 2, …, 10 COVID tests.

2: Hypotheses

- H_0 : The number of COVID tests students have taken in the past two months **follows** a Poisson distribution.
- H_1 : The number of COVID tests students have taken in the past two months **does not follow** a Poisson distribution.

3: Assumptions

A chi-squared test used with the standard approximation that a chi-squared distribution is applicable, has several assumptions:

1. **The data are counts** rather than percentages or some other transformation of the data.
2. The **expected cell count should be 5 or more** in at least 80% of the cells, and no cell should have an expected count of less than 1 (Bewick, Cheek, Ball, 2003).
    - The stricter construction of this test is that every class must have a count of 5 or more. Whilst this is prudent when there are fewer degrees of freedom, the chi-squared distribution is relatively robust under the looser assumption stated at (2) which will be used here.
    - For comparability, the assumption at (2) will be compared to the stricter assumption of *every* class having count of 5 or more, and the attractiveness of one method versus the other will be discussed.
3. The classes are both **ordinal and mutually exclusive**. Each student fits into one and only one class.
4. The observations are **independent**.

Under the aforementioned assumption that survey responses are unique, mutually exclusive and that the data are counts, assumptions 1, 3, and 4 hold. Assumption 2 will hereafter be examined.

This might be better described as checking the sensitivity of the results to the analysis design.

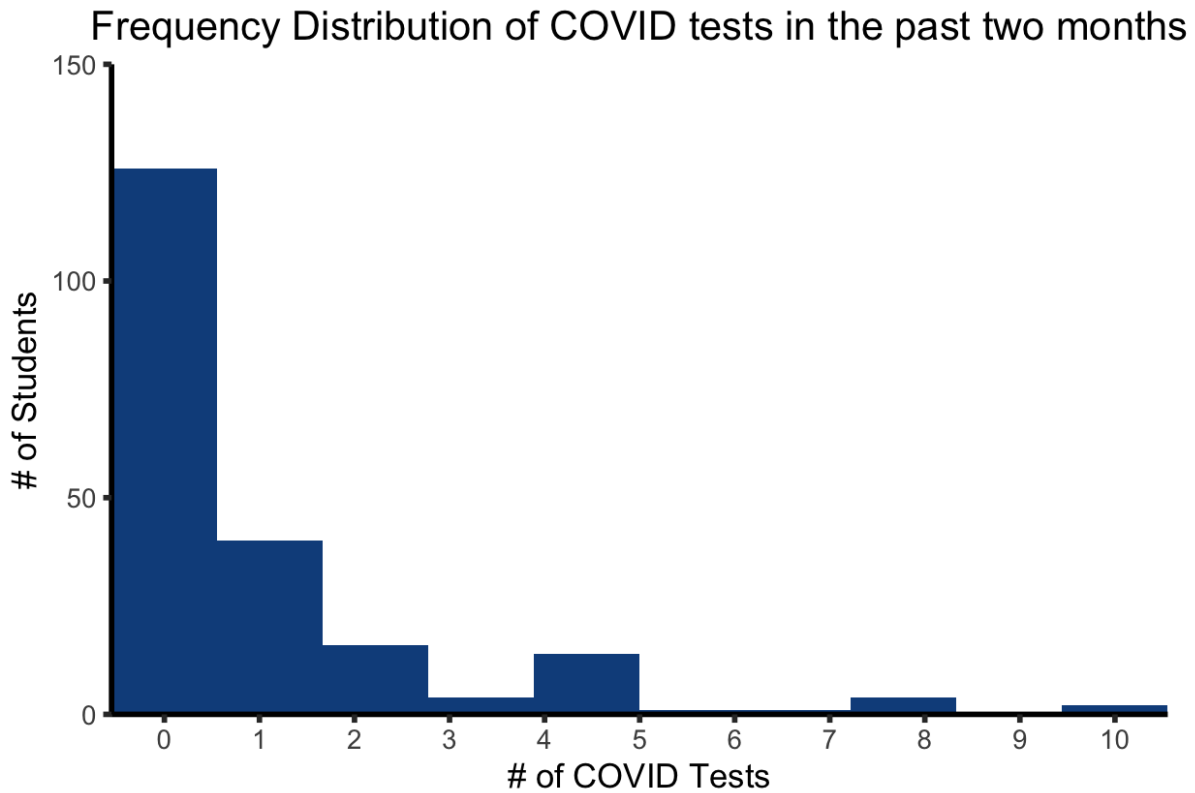
4: Analysis

Visualising the distribution of the data, it seems to resemble the basic structure of a Poisson distribution: it demonstrates strong positive skewness, lots of density around 0 to 1 and is left-modal. However, at least visually, there are reasons to believe that the distribution might be **too dense at 0**, and the right tails might demonstrate **too much kurtosis** (i.e. fat tails).

Hide

```
data1 <- data %>%
  select(covid_tests) %>%
  na.omit()

ggplot(data1, aes(x = covid_tests)) +
  geom_histogram(bins = 10, fill = "dodgerblue4") +
  labs(title = "Frequency Distribution of COVID tests in the past two mont
       hs", y = "# of Students", x = "# of COVID Tests") +
  theme_classic(base_size = 13, base_line_size = 1) +
  theme(plot.title = element_text(hjust = 0.5), plot.margin = margin(1,1,1
       ,1,"cm")) +
  scale_x_continuous(
    breaks = scales::pretty_breaks(n = 8), #add axis ticks for every data
         point
    expand = c(0,0)) + #remove the gap between the plot and the y-axis
  scale_y_continuous(expand = c(0,0), limits = c(0,150))
```



only need one of these frequency distribution plots, I'd suggest keeping the next one which compares the observed data with the expected cell counts.

Hide

```
#other changes to be made: Scale titles and axis labels and have a plot bo
        rder and perhaps change the plots background to be more readable,
         and perhaps plot a poisson distribution over the plot if this is
         somehow possible
```

Next, we extract the observed counts into a vector, `observed_counts` , and remove any NA values.

Hide

```
observed_counts <- c()
range = 0:max(data$covid_tests, na.rm = TRUE)

for (i in range){
  observed_counts[i+1] <- length(which(data$covid_tests == i))
}
df0 <- data.frame(0:10,observed_counts)
df0 <- df0 %>% t()
rownames(df0) <- c("# of COVID Tests", "Observed Counts")

kable(head(df0)) %>%
  kable_styling(htmltable_class = "lightable-classic")
```

# of COVID Tests	0	1	2	3	4	5	6	7	8	9	10
Observed Counts	126	40	16	4	5	9	1	1	4	0	2

Assigning R code variables to the parameters that are to be used in the Poisson distribution.:

- **The sample size, n = 208**, which is given by the sum of `observed_counts` .
- **The number of classes, k = 11**, which given by the length of `observed_counts` .
    - The number of classes is subject to change if classes are combined to meet the assumption that the expected counts are ≥ 5.
- **Population parameter**, **λ = 1.028846,** which is estimated from the sample.
    - Using sample data to estimate a population parameter will **reduce the degrees of freedom** in the chi-squared distribution.
    - Under the Poisson distribution, λ is equal to the average number of events in the fixed time window. In this situation, λ is equal to the average number of COVID tests in the two month period.

Hide

```
x <- 0:10 # define the groups (# of COVID tests) corresponding to the obse
        rved_counts vector
n <- sum(observed_counts) # find the sample size
k <- length(observed_counts) # find the number of groups
lam <- sum(observed_counts * x)/n # estimate the lambda parameter
```

Next, we obtain the probability of the event occurring in each class by drawing from the Poisson probability mass function. We define the 11th element in the probability vector as the probability of $P(\geq 11)$ rather than $P(11)$ so that the probability vector sums to one.

To find the expected counts (e_i), we multiply the probability vector by the sample size.

Testing the assumption that $e_i \geq 5$, we observe a violation for some expected counts.

This is seen below as our test returns `TRUE` if $e_i = np_i \geq 5$ and `FALSE` if not.

Hide

```
#Now finding the expected counts

p = dpois(x, lambda = lam) # obtain the p_i from the Poisson pmf
p[11] = 1 - sum(p[1:10]) # redefine the 11th element P(>=11) NOT P(11)

p = round(p, 5)
ey = n * p  # calculate the expected frequencies
ey >= 5
```

```
##  [1]  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

It is possible to group the data such that this assumption is met.

To achieve $e_i \geq 5$ it is necessary to amend both the observed cell counts and expected cell counts into four groups giving the following table.

Hide

```
# Grouping into 4 classes

yr = c(observed_counts[1:3], sum(observed_counts[4:11]))
eyr = c(ey[1:3], sum(ey[4:11]))

cellcounts4 <- data.frame(c(0:3), yr, eyr)
colnames(cellcounts4) = c("COVID Tests", "Observed Counts", "Expected Coun
        ts")


kable(head(cellcounts4)) %>%
  kable_styling(htmltable_class = "lightable-classic")
```

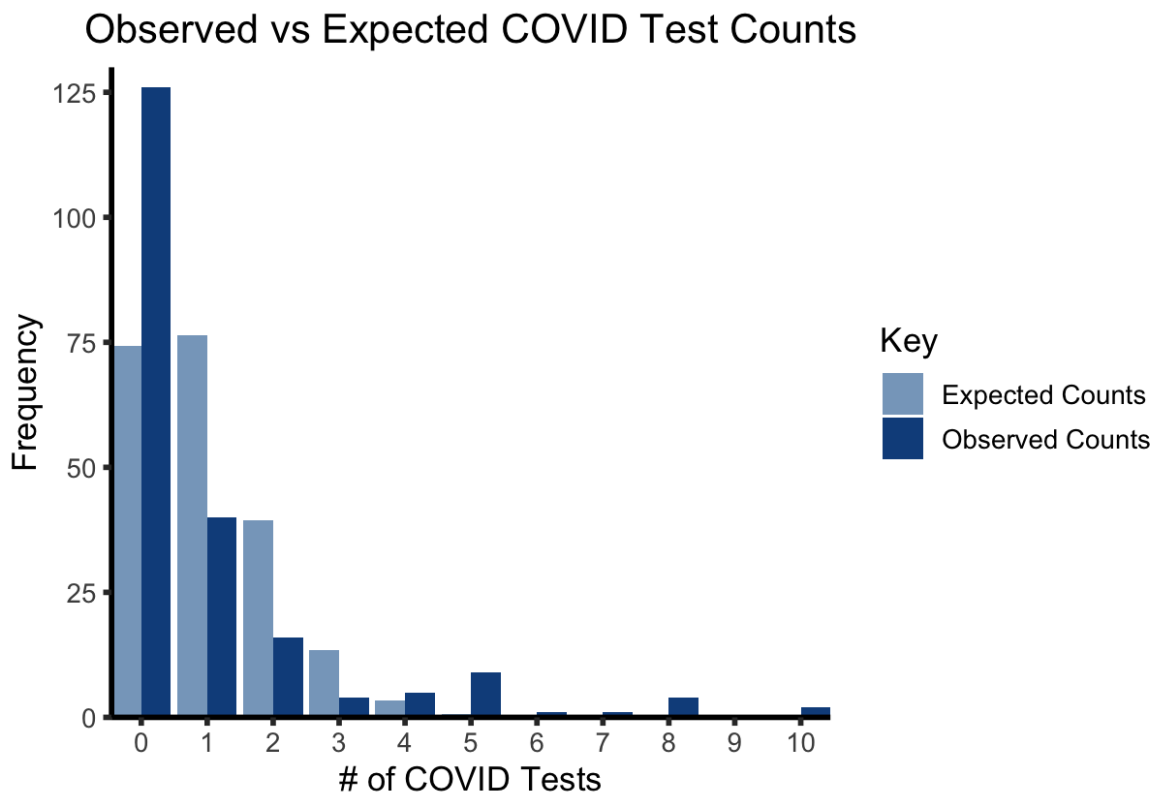
COVID Tests	Observed Counts	Expected Counts
0	126	74.34336
1	40	76.48784
2	16	39.34736
3	26	17.82352

However, **using 4 groups has a fundamental problem**. To see this, we visualise the observed vs expected counts below. Doing so, it's clear that the **observed counts have fatter tails.** This creates the problem that **when grouping classes, we are ignoring the positive kurtosis** of the observed counts. This makes the expected counts more closely resemble a Poisson distribution which is misleading.

Hide

```
df1 <- data.frame(ey, observed_counts, tests = 0:10)
df2 <- tidyr::pivot_longer(df1, cols=c('ey', 'observed_counts'), names_to=
        'Key',
values_to="Frequency")

ggplot(df2, aes(x=tests, y= Frequency, fill= Key)) +
    geom_bar(stat='identity', position='dodge') +
  labs(title = "Observed vs Expected COVID Test Counts", x = "# of COVID T
        ests") +
  theme_classic(base_size = 13, base_line_size = 1) +
  theme(plot.title = element_text(hjust = 0.5), plot.margin = margin(1,1,1
        ,1,"cm")) +
  scale_x_continuous(
    breaks = scales::pretty_breaks(n = 8), #add axis ticks for every data
         point
    expand = c(0,0)) + #remove the gap between the plot and the y-axis
  scale_y_continuous(
    expand = c(0,0),
    limits = c(0,130)) +
  scale_fill_manual(
    values = c("#87A6C5", "dodgerblue4"),
    labels = c("Expected Counts", "Observed Counts"))
```



The workaround is to group into 5 classes.

Hide

```
# Grouping into 5 classes

yr5 = c(observed_counts[1:4], sum(observed_counts[5:11]))
eyr5 = c(ey[1:4], sum(ey[5:11]))

cellcounts5 <- data.frame(c(0:4), yr5, eyr5)
colnames(cellcounts5) = c("COVID Tests", "Observed Counts", "Expected Coun
        ts")


kable(head(cellcounts5)) %>%
  kable_styling(htmltable_class = "lightable-classic", position = "center"
        )
```

COVID Tests	Observed Counts	Expected Counts
0	126	74.34336
1	40	76.48784
2	16	39.34736
3	4	13.49504
4	22	4.32848

In this situation, we better reflect the underlying structured of our observed counts.

Problematically, we violate the assumption that $e_i \geq 5$.

However, $e_5 = 4.33 \geq 1$ and also $\frac{4}{5} = 80\%$ of the expected counts are greater than 5, satisfying the looser, but still robust, assumption that was made.

5: Test Statistic

Under H_0 the test statistic is:

$$T = \sum_{i=1}^{5} \frac{(Y_i - np_i)^2}{np_i} \sim \chi^2_3$$

Where Y_i is the observed counts and e_i is the expected count in each class under the null hypotheses.

Under the assumptions, the test statistic will follow a chi-squared distribution with 3 degrees of freedom (since estimating the λ parameter from the sample reduced the degrees of freedom by 1).

Also note that Yates' (1934) **continuity correction will not be used.** Whilst it aims to correct error by assuming that the discrete frequencies can be modeled by the continuous chi-squared distribution**, it can adjust too far** (Conover, 1974).

Estimating this test statistic using R, we get that $T = 145.9794$ which is shown on the chi-squared distribution using the purple line.

Hide

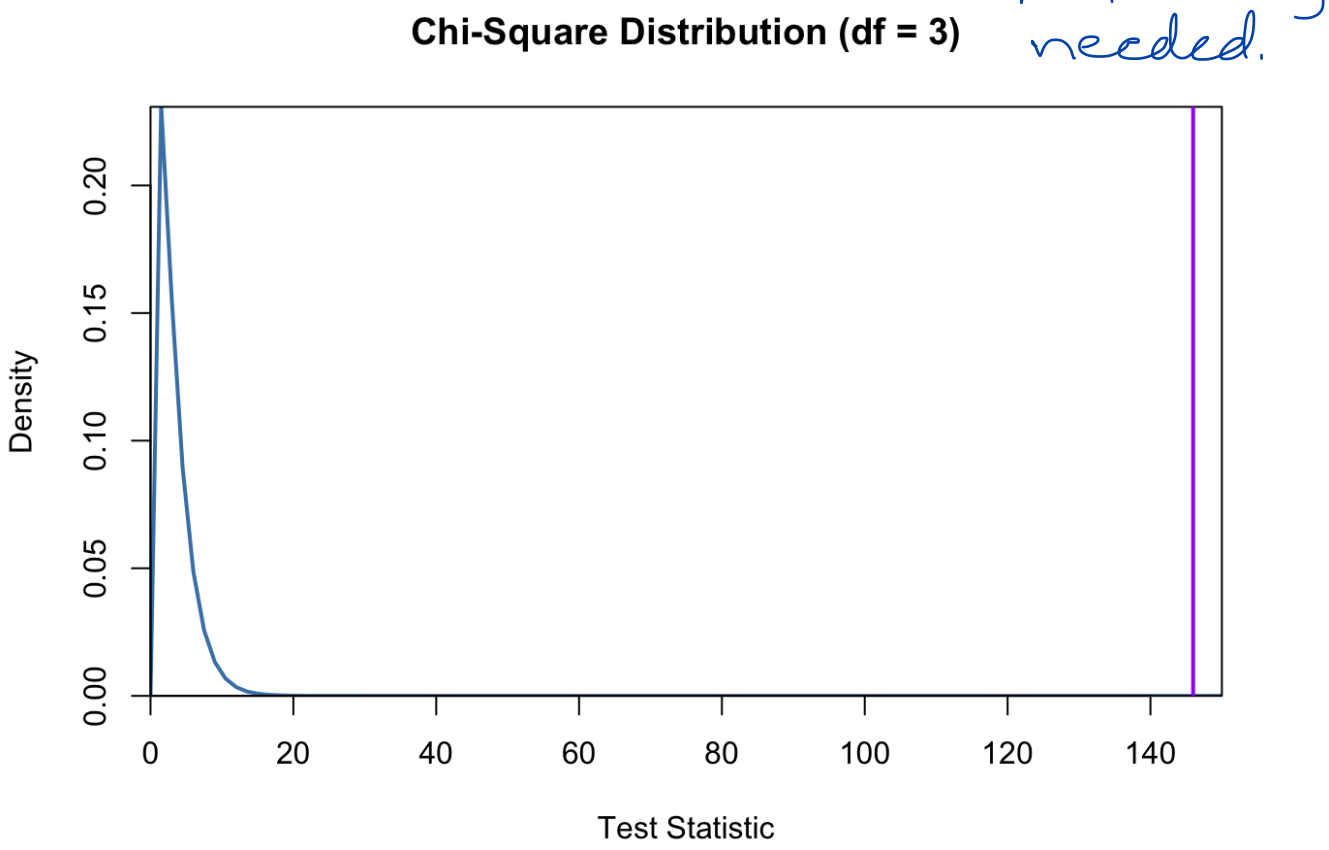
```
kr = length(yr) # with 4 combined classes
kr5 = length(yr5) # with 5 combined classes

(t0 = sum((yr - eyr)^2/eyr))  # test statistic with 4 classes
(t0_5 = sum((yr5 - eyr5)^2/eyr5))  # test statistic with 5 classes
```

Hide

```
curve(dchisq(x, df = (kr5-1-1)), from = 0, to = 150,
      main = 'Chi-Square Distribution (df = 3)', ylab = 'Density',
      lwd = 2, col = "steelblue", xlab = "Test Statistic",
      xaxs="i", yaxs="i")
abline(v = t0_5, col = "purple", lwd = 2)
```

not really needed.



Hide

```
#axis(1, round(t0_5,2), col.axis = "purple")
```

6: P-value

The p-value is the probability of having a test statistic at least as extreme as this under the assumption that the null is true. Mathematically:

$$P(T \geq t_0) = P(\chi^2_3 \geq 145.9794) \approx 0 \quad < 0.0001$$

Hide

```
(pval=1-pchisq(t0,df=kr-1-1)) #p-value with 4 classes
(pval=1-pchisq(t0_5,df=kr5-1-1)) #p-value with 5 classes
```

7: Decision

The level of significance is the predetermined probability of rejecting the null hypothesis given that it is true.

Using a level of significance of $\alpha = 0.05$ or even $\alpha = 0.01$, since $p < 0.01 < 0.05$ we have sufficient evidence to reject the null hypothesis that the number of COVID tests that students at the University of Sydney have taken in the past two months follows a Poisson distribution.

8: Discussion

The near-zero p-value means that the **null is readily rejected.** In fact, given the fat tails of the observed counts, grouping into 5 groups might underestimate just how far the observed counts differ from the Poisson distribution.

Yet, even being more lenient with grouping, and **using four groups** such that the $e_i \geq 5$ assumption holds, there is **still sufficient evidence to reject the null.**

Indeed, we get a test statistic $T = 70.90367$.

Calculating the p-value, we get:
$P(T \geq t_0) = P(\chi^2_2 \geq 70.90367) = 4.440892e^{-16} \quad < 0.001$

Problems with the assumption of an underlying Poisson distribution

Logically, it is somewhat unsurprising that the null hypothesis is rejected because several factors might undermine the assumption of an underlying Poisson distribution:

government intervention can also impact the result if someone lived in a LGA that required testing every few days to go to work

- The Poisson distribution assumes that the average time between events is known and constant. In reality, the **time between events changes** as severity of the COVID contagion fluctuates throughout time.
- The Poisson distribution assumes that the arrival of an event is independent of the event before it. This does not happen in the case of COVID tests. If someone has had say 3 COVID tests, it might be a sign that they are in a hotspot area and therefore are more likely to get another COVID test. This might explain the observation that **just as many people had 3 COVID tests as 8 COVID tests.**
- There is a geographical divergence in COVID outcomes. Problematically, only ~65% of the respondents said they were in Australia, which undermines the assumption of a constant and known average time between events.

This is a very good point.

Problems in data collection and validation

Throughout this investigation, the observation of fatter tails was made on several occasions, but little was done to **question the validity of this positive kurtosis.**

This validity assessment is crucial, because merely a few false '8-10' scores would result in a very low chance that the data will follow a Poisson distribution for two reasons, one obvious, and one non-obvious:

1. [Obvious] A Poisson distribution has few observations in the tails.
2. [Non-obvious] Fat tails will result in an **over-assessment of the mean of the observed counts**, which in turn results increases λ (when λ is estimated from the sample as done here). With a higher λ, the Poisson distribution will not predict as many observations around '0' as were observed in the data. This further diverges the observed distribution from a Poisson distribution.

Given that very few false responses at the 8-10 end of the scale can massively impact whether the data follows a Poisson distribution, this survey is not robust to false responses. A better design might be to cap the responses at 5 or more, which would be counted as a score of '5' in the estimation of λ.

Power of the test When examining the power of our Chi-Square Goodness of Fit it indicates a power level of 99% (See below). This indicates the test had a 99% probability of rejecting the null hypothesis (i.e. in this case, that COVID Test Results follow a Poisson Distribution) if the alternative hypothesis was true. Ultimately, then, we can be rather confident in the final outcome.

and all the underlying assumptions were met with no data issues.

Hide

```
es = ES.chisq.gof(p1=yr5, p0=eyr5)$w
broom::tidy(pwr.chisq.test(w=es, N = 208, df = 3, sig.level = 0.05)) %>% k
        bl() %>%  kable_styling()
```

sig.level	power
0.05	1

Test 2: To investigate self-assessed mathematical ability by gender

1: Introduction: Is average self-assessed mathematical ability of male students higher than that of female students?

There is a widely documented phenomena of women being less self-assured than men on self-assessments of ability (Rain, Neyse, David-Barett, & Schmidt, 2016). This is endemic, psychological and subconscious. It can lead to self-perpetuating cycles of lower confidence which in turn curb one's drive to pursue future opportunities (Kay & Shipman, 2014).

The purpose of this investigation is to see whether this theoretical concept can be observed from the data.

2: Hypotheses

Let X_i be the self-assessed mathematical ability of the i^{th} male student, and Y_j the self-assessed mathematical ability of the j^{th} female student.

- $H_0 : \mu_X - \mu_Y = 0$
- $H_1 : \mu_X - \mu_Y > 0$

μ_X is the population mean of male's self-assessed mathematical ability and μ_Y is the population mean of female's self-assessed mathematical ability.

Setting up the assumptions in this manner allows for a **one-sided** t-test. This **increases the power** of the test which increases the probability of rejecting the null hypothesis when the null is false (Perlman, 1969).

3: Data Inspection

First, genders were cleaned into 3 categories: “male”, “female”, and “non-binary”.

NA values, empty data entries were omitted.

“Non-binary” gender results were omitted since there was insufficient data and the hypotheses concern male vs female.

The **mean, standard deviation and count** of males and females were then summarised in a new data frame.

Hide

```
data2 <- data %>%
  mutate(gender = recode_gender(gender, dictionary = narrow, fill = FALSE
         )) %>%
  na.omit()


gender_math <- data2 %>%
  filter(gender == "male" | gender == "female") %>%
  select(gender, math_ability)


#Create a vector of male math ability
Male <- data2 %>%
  filter(gender == "male") %>%
  pull(math_ability) %>%
  na.omit()

# create a vector of female math ability
Female <- data2 %>%
  filter(gender == "female") %>%
  pull(math_ability) %>%
  na.omit()

#We can also create individual data frames of the male and female vectors
          for graphing later
datf <- data.frame(Female)
datm <- data.frame(Male)

#combine into data frame that allows two boxplots to be graphed side by si
         de
dat2 = data.frame(
  math_ability = c(Male, Female),
  gender = c(rep("Male",
                 length(Male)),
             rep("Female",
                 length(Female))))
sum2 = dat2 %>%
  group_by(gender) %>%
  summarise(Mean = mean(math_ability, na.rm = TRUE),
            SD = sd(math_ability, na.rm = TRUE),
            n = n())

kable(head(sum2)) %>%
  kable_styling(htmltable_class = "lightable-classic", position = "center"
         )
```

gender	Mean	SD	n
Female	6.2[illegible]	1.5[illegible]	53
Male	6.7[illegible]	1.8[illegible]	98

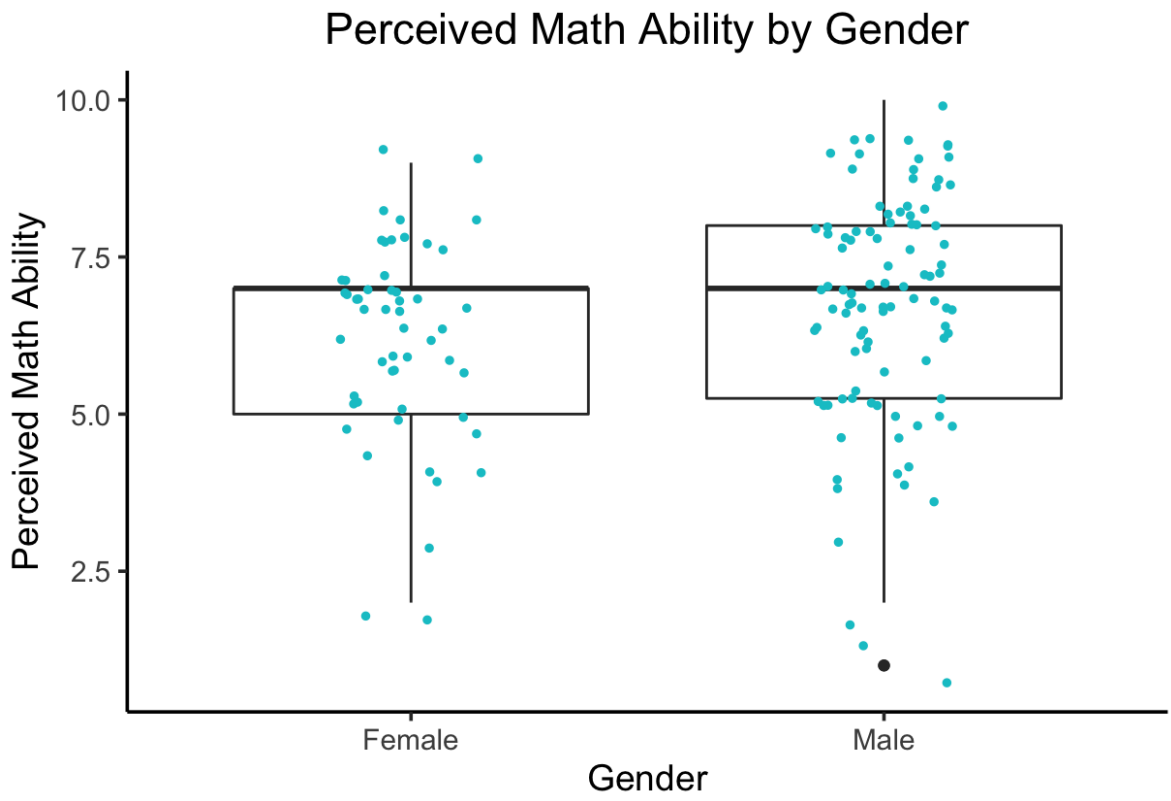
digits = 1 in the kable function

The distribution of male and female math ability can also be plotted on a boxplot, to inspect for normality. Both distributions appear somewhat **negatively skewed** and the distribution of female's self-assessed math ability appears to **lack symmetry**.

Hide

```
ggplot(dat2, aes(x = gender, y = math_ability)) +
  geom_boxplot() +
  geom_jitter(width=0.15, size = 1, colour = "turquoise3") +
  theme_classic(base_size = 14) +
    theme(
      plot.title = element_text(hjust = 0.5),
      plot.margin = margin(1,1,1,1,"cm")) +
  labs(
    title = "Perceived Math Ability by Gender",
    x = "Gender",
    y = "Perceived Math Ability")
```

set height = 0 or height = 0.01
↳ the vertical jittering in the plot below makes the scores look continuously measured but they were discrete.



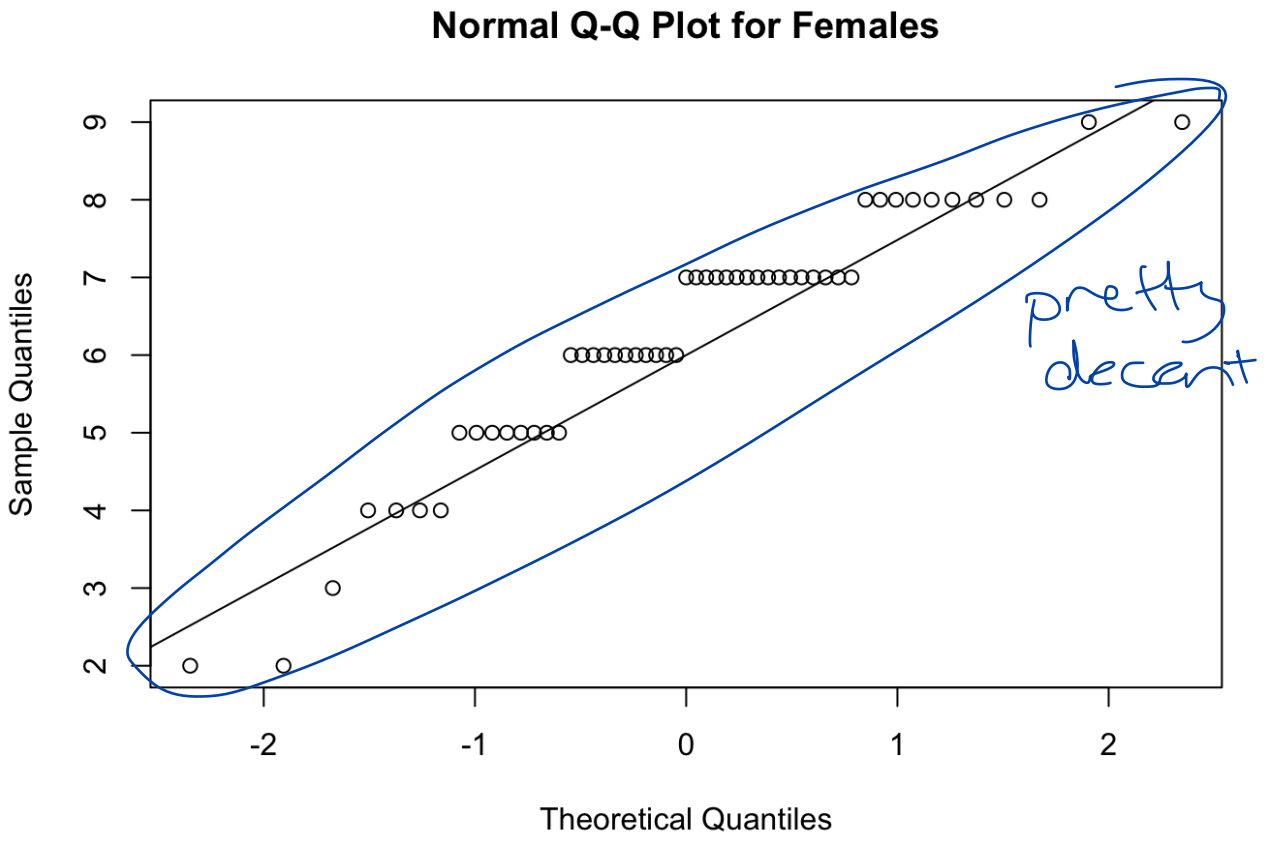
Testing for normality using quantile-quantile plots, the **observed data does not sit on the normal line**. Even accounting for the fact that the data is discrete (versus the continuous normal distribution), the data still demonstrates left skewness and heavy tails.

The QQ plot for females looks pretty ok to me (apart from the discreteness).

Hide

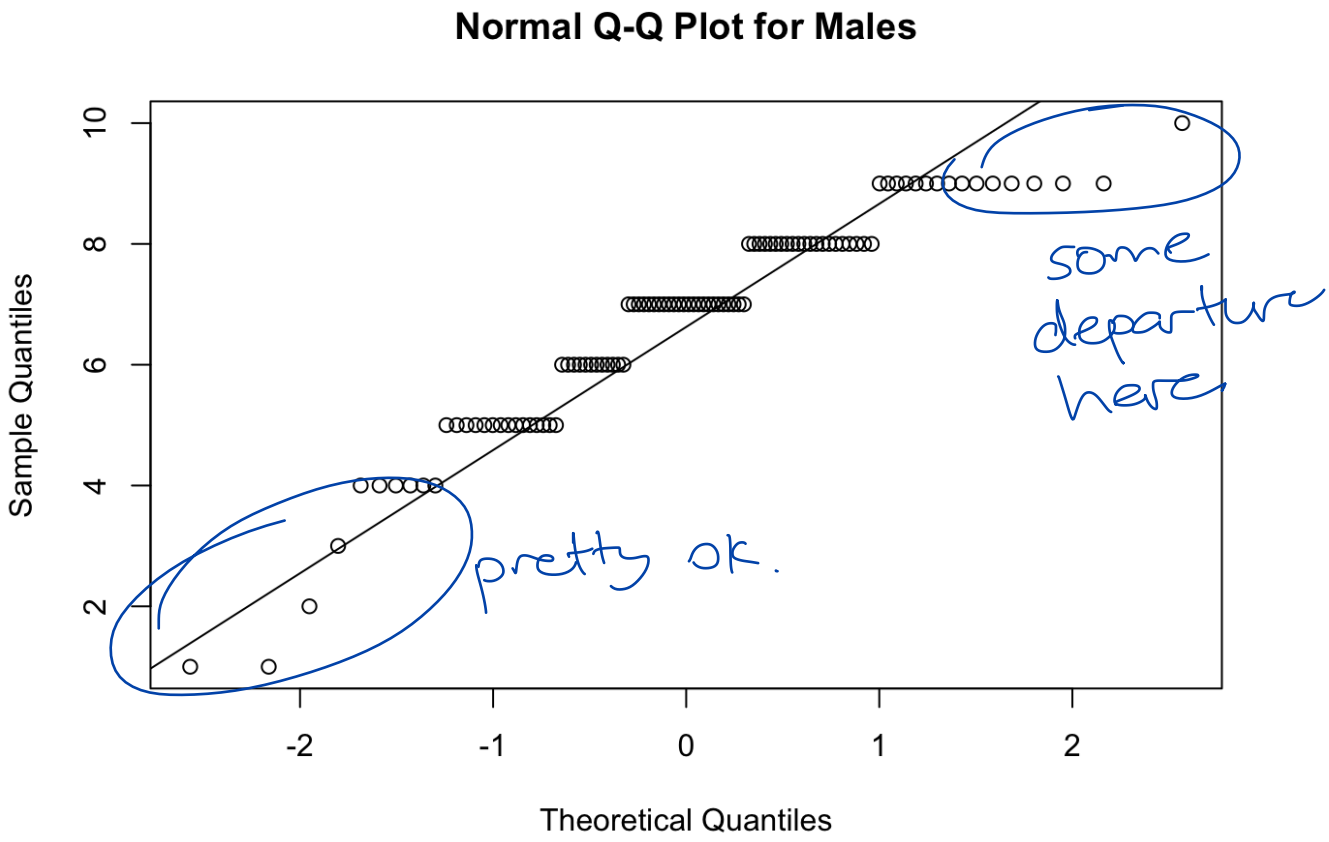
```
datf <- datf %>% t()
datm <- datm %>% t()

qqnorm(datf, main = "Normal Q-Q Plot for Females")
qqline(datf)
```



Hide

```
qqnorm(datm, main = "Normal Q-Q Plot for Males")
qqline(datm)
```



4: Assumptions

Two-sample t-tests make the following assumptions:

- $X_1, \ldots, X_{n_x}$ are iid and $Y_1, \ldots, Y_{n_y}$ are iid.
    - The aforementioned assumptions are that data is collected from a representative, independent, randomly selected portion of the total population. Male and female scores each are assumed to follow their own parametric distribution.
- The $X_i's$ are independent of the $Y_j's$
    - The response of a male doesn't depend on any of the responses of a female, and vice versa, satisfying this assumption.
- Parametrising this distribution we assume: $X \sim N(\mu_X, \sigma^2)$ and $Y \sim N(\mu_Y, \sigma^2)$
    - However, because of the central limit theorem and large (>30) number of samples, the **t-test is robust to most departures from normality** (Brosamler, 1988). It is therefore not necessary for the underlying data here to be normally distributed.
    - However, the t-test is not always as robust to **departures from homogeneity of variance** (Moser & Stevens, 1992)

- The assumption of equal variance is relatively robust when the sample size of the two groups are equal (Markowski & Markowski, 1990).
- However, given that the sample sizes of X_i and Y_j are 129 and 74 respectively, **this assumption might not hold** since 'roughly equal sample size' has been shown to mean, at most, a ratio of 1.5x of the large sample to small sample (Blanca, Alarcon, Arnau, Bono, & Bendayan, 2018).
- Thus, **Welch's t-test**, which is insensitive to equality of variances, will likely be the **more powerful, and more relevant test.**

The variances of the two populations weren't wildly different, it probably would have been fine to go with a 'classical' two-sample t-test (also fine to use Welch)

5: Statistical Testing & Decisions

5.1: Assuming equal variance

First, a two-sample t-test assuming equal population variance is used such that:

$$X_i \sim N(\mu_X, \sigma^2) \text{ and } Y_j \sim N(\mu_Y, \sigma^2)$$

The **pooled estimator** of σ is $S_p^2 = \frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}$, letting n be the sample size of males and m of females.

If you decided Welch is more appropriate there is no need to implement the less appropriate test.

The **test statistic** under H_0 is:

$$T = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{1/n + 1/m}} \sim t_{n+m-2}$$

The observed test statistic is $t_0 \approx 1.6112$

The p-value is $P(t_{n+m-2} \geq t_0) = 0.05463$

Given a level of significance of $\alpha = 0.05$ there is insufficient evidence to reject the null hypothesis.

Hide

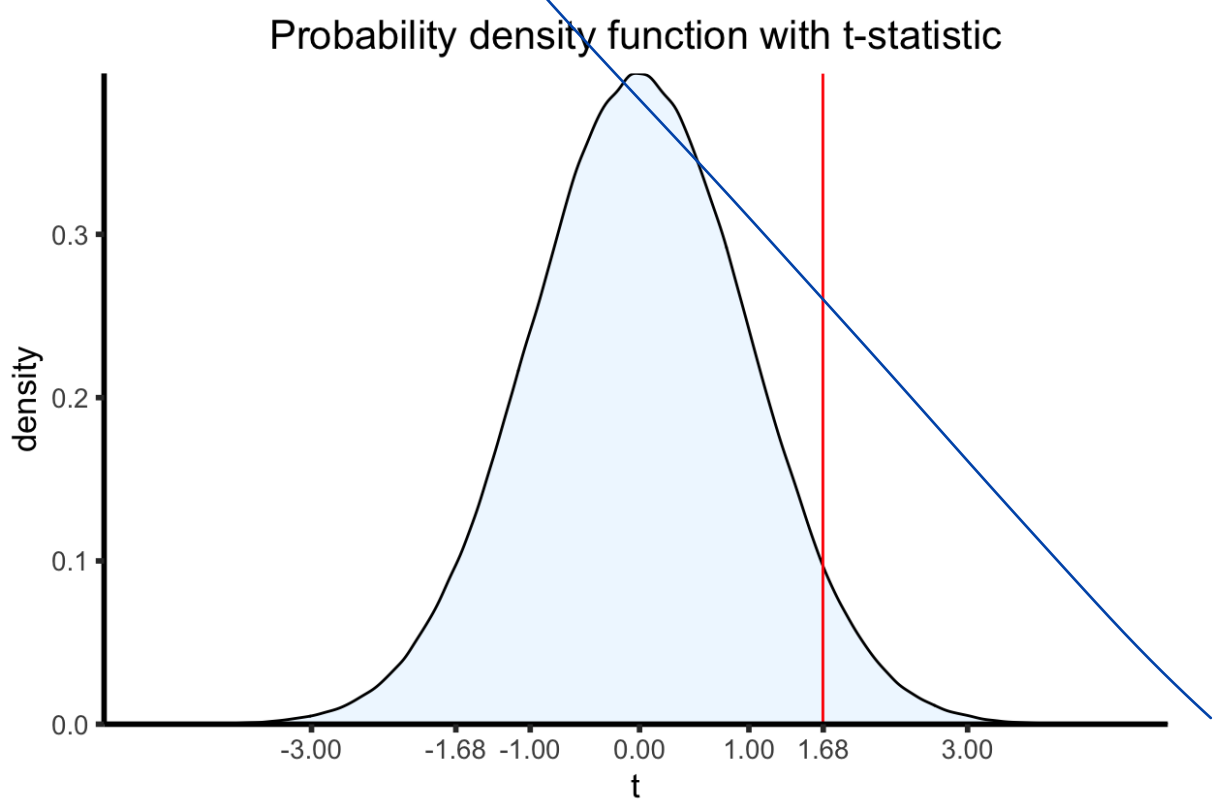
```
# Two-sample t-test

t.test(Male, Female, alternative = "greater", var.equal = TRUE)
```

The simulation below shows the test statistic on a student's t distribution.

Hide

```
ggplot(data.frame(t=rt(1000000,149))) + geom_density(aes(x=t), fill = 'ali
        ceblue') +
  geom_vline(xintercept = 1.677, colour = "red") +
  ggtitle("Probability density function with t-statistic") +
  theme_classic(base_size = 13, base_line_size = 1) +
    theme(plot.title = element_text(hjust = 0.5), plot.margin = margin(1,1
        ,1,1,"cm")) +
  scale_x_continuous(
    breaks=c(-3,-1.68,-1,0,1,1.68,3),
    expand = c(0,0)) + #remove the gap between the plot and the y-axis
  scale_y_continuous(expand = c(0,0))
```



5.2: Assuming unequal variance – Welch's t-test

The result differs under the assumption of unequal variance.

In this case, the observed t-value is 1.677 and the corresponding p-value is 0.04808.

Given a level of significance of $\alpha = 0.05$ there is now sufficient evidence to reject the null hypothesis.

Hide

```
#Welch's t-test is insensitive to equality of the variances regardless of
         whether the sample sizes are similar.

t.test(Male, Female, alternative = "greater",var.equal = FALSE) #this does
        Welch's t-test
```

The unequal sample sizes and large samples suggest that **Welch's unequal variance t-test is ideal** (Zimmerman, 2004).

Under Welch's t-test, there is sufficient evidence to reject the null hypothesis that a 5% level of significance.

Nevertheless, this report recommends further investigation to formulate a more definitive conclusion.

6: Discussion

Equal vs unequal variance

It is ***not recommended*** **to pre-test for equal variances** to help select for the Student's *t*-test or Welch's *t*-test because two-stage testing procedure is seen to **increase type I error rates** (Zimmerman, 2004). Instead, when the sample size is large and unequal, the distributions are somewhat skewed, and the sample variance are not equal, Welch's t-test should be applied directly (Fagerland & Sandvik, 2009).

Subjectivity in perception

Whilst there seems to be be evidence to suggest that females perceive their mathematical ability to be lower than males, perhaps this merely underscores gender-derived differences in the subjective perception of the rating scale. Perhaps a '7' for females means something different than a '7' for males. This test does not examine this.

use women rather than females

when using women pair it with men for males

Further research

This test does not make any attempt at causal induction. Further research would be needed to explore causal hypotheses and gather more evidence in favour of rejecting the null hypothesis presented in this study.

Power of the test Under Welch's two-sample t-test, the power is 66%. This is quite low which makes sense given that the Welch t-test only just passes the critical value threshold to conclude in favour of rejecting the null. The only moderate power of this test adds weight to conducting further studies rather than concluding that this existing one is sufficient.

(Some authors discourage this kind of post hoc power analysis)

Hide

```
es = ES.t.two(m1=6.245283, m2=6.724490, sd1 = 1.592116, sd2 = 1.820732, n1
        = 53, n2 = 98, t = 1.677)$d
power2 <- power.welch.t.test(n = 151, delta = es)$power
power2
```

✓

```
## [1] 0.6625535
```

Test 3: To investigate independence between gender and stress levels

1: Introduction: Are perceived stress levels of students independent of gender?

Following the gender-oriented theme of the previous test, this report now investigates whether perceived stress levels of students are independent of gender.

Whilst this has been examined in the workplace (Lundberg, 2005; Spielberger & Reheiser, 1994), in adolescents (Rudolph & Hammen, 1999), and in the case of post-traumatic stress (Olff, Langeland, Draijer, & Gersons, 2007), the studies on the independence between stress levels and gender in students are less widely cited (Calvarese, 2015), and at times, inconclusive (Anbumalar, Dorathy, Jaswanti, Priya, & Reniangelin, 2017).

The population is construed as all DATA2X02 students. While it is tempting to expand the population definition to *all* men and women, doing so would undermine the representativeness of the sample and introduce confounding variables between men and women such as different jobs and different hormonal responses later in life.

2: Data Inspection and Setup

Visualising the distributions of perceived stress levels, it looks as if the sampled female students perceive, on average, themselves to be experiencing higher stress levels than males.

Hide

```
data3.0 <- data2 %>%
  select(stress, gender) %>%
  na.omit() %>%
  filter(gender %in% c("male","female"))

ggplot(data3.0, aes(y = stress, x = gender)) +
  geom_boxplot() +
  geom_jitter(width=0.15, size = 1, colour = "turquoise3") +
  theme_classic(base_size = 14) +
    theme(
      plot.title = element_text(hjust = 0.5),
      plot.margin = margin(1,1,1,1,"cm")) +
  labs(
    title = "Perceived Stress Level by Gender",
    x = "Gender",
    y = "Perceived Stress Level")
```

However, this could readily be from a **confounding variable** which is the difference between how women perceive the meaning of a *score* of 7 versus men assigning a score of seven. Indeed, there is perhaps an element of **arbitrariness in a 1-10**

rating scale - and the more options there are, the greater room there is for subjectivity.

Thus, the 2x10 contingency table is **simplified into three categories**: “low stress”, “moderate stress” and “high stress”. This **always improves the readability of the data** as well as **increasing they likelihood that the expected values in each cell would be** ≥ 5**.**

Hide

isn't there also some subjectivity in the thresholds for low vs moderate and moderate vs high? How did you pick 3 and 7 as the thresholds? This should also be made clear in the text (I think it's only in the code).

```
# Creating groups for the different stress levels to improve comprehensibi
        lity
data3 <- data2 %>%    #recall from earlier that data2 has already cleaned
         genders
  select(stress, gender) %>%
  mutate(
    stress_clean = data2$stress,
    stress_clean = case_when(
      stress_clean <= 3 ~ "Low stress",
      stress_clean <= 7 ~ "Moderate stress",
      stress_clean <= 10 ~ "High stress")
)

stress_male <- data3 %>%
  filter(gender == "male") %>%
  pull(stress_clean) %>%
  na.omit()

stress_female <- data3 %>%
  filter(gender == "female") %>%
  pull(stress_clean) %>%
  na.omit()

categories <- c("Low stress","Moderate stress","High stress")
male_stress_count <- c()
female_stress_count <- c()

# loop to create vector of the counts in each group for men and women
x <- 1
for (i in categories){
  male_stress_count[x] <- length(which(stress_male == i))
  x=x+1
}
x <- 1
for (i in categories){
  female_stress_count[x] <- length(which(stress_female == i))
  x=x+1
}
stress_counts <- c(male_stress_count, female_stress_count)


# constructing the observed frequency table
mat_stress <- matrix (stress_counts, ncol = 2)
colnames(mat_stress) = c("Male", "Female")
rownames(mat_stress) = categories
mat_stress <- mat_stress %>%
  t()

kable(head(mat_stress)) %>%
  kable_styling(htmltable_class = "lightable-classic", position = "center"
        )
```

	Low stress	Moderate stress	High stress
Male	22	61	15
Female	8	26	19

3: Hypothesess

- H_0 : Perceived stress levels by students are **independent** of gender.
- H_1 : Perceived stress levels by students are **not independent** of gender.

We can restate the null hypothesis in terms of the definition for independence.

Let p_{ij} denote the probability of an observation falling into the i^{th} row and j^{th} column (i.e. the $(i,j)^{th}$ cell) in the 2x3 contingency table defined by gender and perceived stress levels.

Then, under the null:

$$p_{ij} = p_{i\cdot}p_{\cdot j} \quad \text{for all} \quad (i,j) \in (1,2) \times (1,2,3)$$

Where $p_{i\cdot}$ and $p_{\cdot j}$ refer to the row total and column totals respectively.

This comes from the definition of independence that:

$$P(X = x|Y = y) = P(X = x)$$

Stated in words, the probability of observing X given Y, is equal to the probability of X – in other words, the occurrence of Y has no impact on the occurrence of X.

4: Assumptions

A chi-squared test used with the standard approximation that a chi-squared distribution is applicable, has several assumptions:

1. The data are counts rather than percentages or some other transformation of the data.
2. The expected count in each class should be 5 or more.
3. The classes are both ordinal and mutually exclusive. Each student fits into one and only one class.

4. The observations are independent.

Assumptions 1, 3, and 4 are satisfied by the aforementioned assumptions on the data being ordinal, independent and mutually exclusive.

We then test assumption 2, that $e_{ij} \geq 5$.

Under the null hypothesis, the probability of each cell is a product of the marginal probabilities, and hence the expected frequency for each cell is given by:

$$e_{ij} = \frac{y_{i\cdot} y_{\cdot j}}{n}$$

Hide

```
# finding the expected cell counts to test the assumption
r = 2 #setting the number of rows
c = 3 #setting the number of columns
yr = apply(mat_stress, 1, sum)  # rowSums(y.mat)
yc = apply(mat_stress, 2, sum)  # colSums(y.mat)
emat = yr %*% t(yc) / sum(mat_stress) #using matrix multiplication to find
        the expected counts in each cell
rownames(emat) = c("Male", "Female")

kable(head(emat >= 5)) %>%
  kable_styling(htmltable_class = "lightable-classic", position = "center"
        )
```

	Low stress	Moderate stress	High stress
Male	TRUE	TRUE	TRUE
Female	TRUE	TRUE	TRUE

Caption or description in words of what this table of TRUEs represents.

5: Statistical Testing & Decisions

The test statistic under the null hypothesis is given by:

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(Y_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

Given a 2x3 contingency table (i.e. r = 2 and c = 3), we get an observed test statistic of $t_0 = 8.4217$

The p-value given by $P(\chi^2_2 \geq 8.4217)$ is 0.01483.

Thus, given a level of significance of $\alpha = 0.05$, there is sufficient evidence to reject the null hypothesis that gender and stress levels are independent.

Hide

```
chisq.test(mat_stress)

#comparing this to the manual calculation

t0 <- sum((mat_stress - emat)^2/emat)
```

The test statistic is shown on the chi-squre distribution using the purple line.

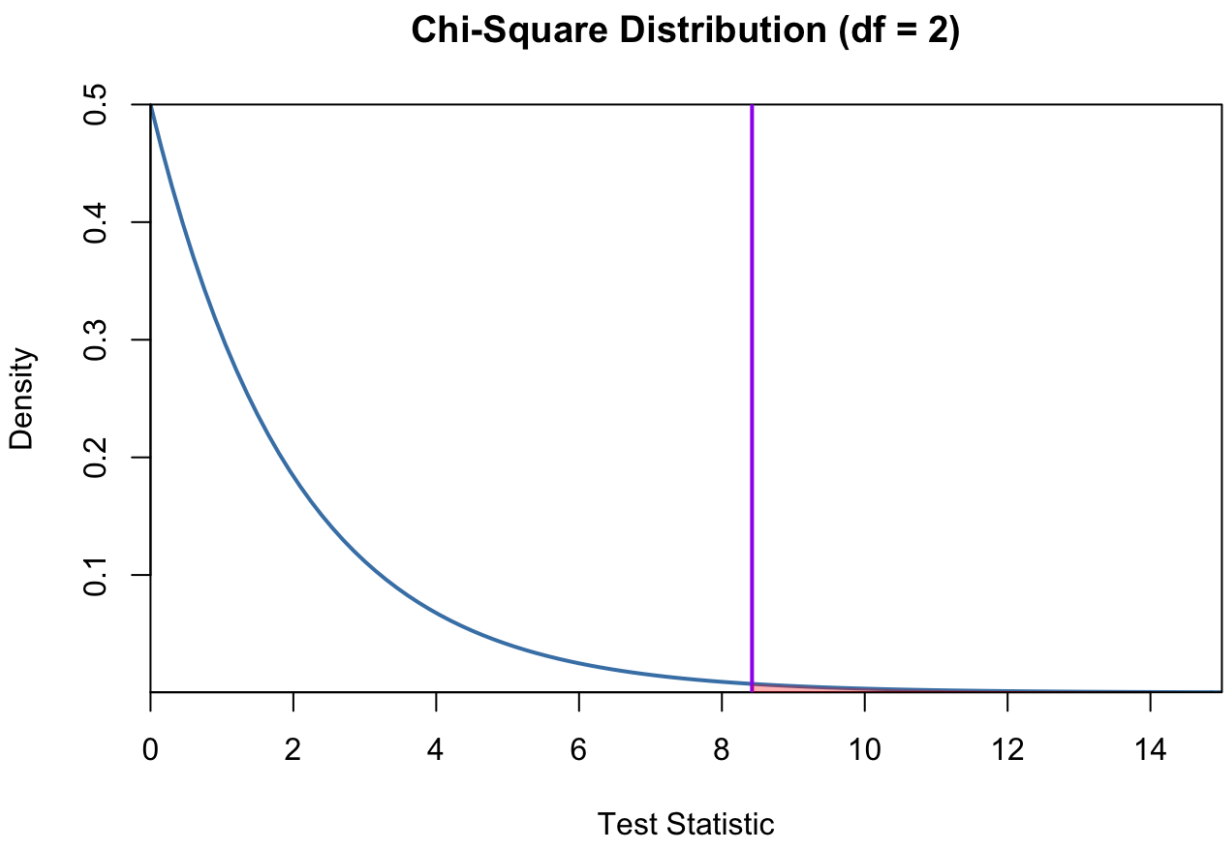
Hide

```
curve(dchisq(x, df = (2)), from = 0, to = 15,
      main = 'Chi-Square Distribution (df = 2)', ylab = 'Density',
      lwd = 2, col = "steelblue", xlab = "Test Statistic",
      xaxs="i", yaxs="i")
abline(v = t0, col = "purple", lwd = 2)

#create vector of x values
x_vector <- seq(t0, 15)

#create vector of chi-square density values
p_vector <- dchisq(x_vector, df = 2)

#fill in portion of the density plot from 0 to 15
polygon(c(x_vector, rev(x_vector)), c(p_vector, rep(0, length(p_vector))),
        col = adjustcolor('red', alpha=0.3), border = NA)
```



This is consistent with the distribution of the proportion of stress levels by gender as visualised in the chart below.

Hide

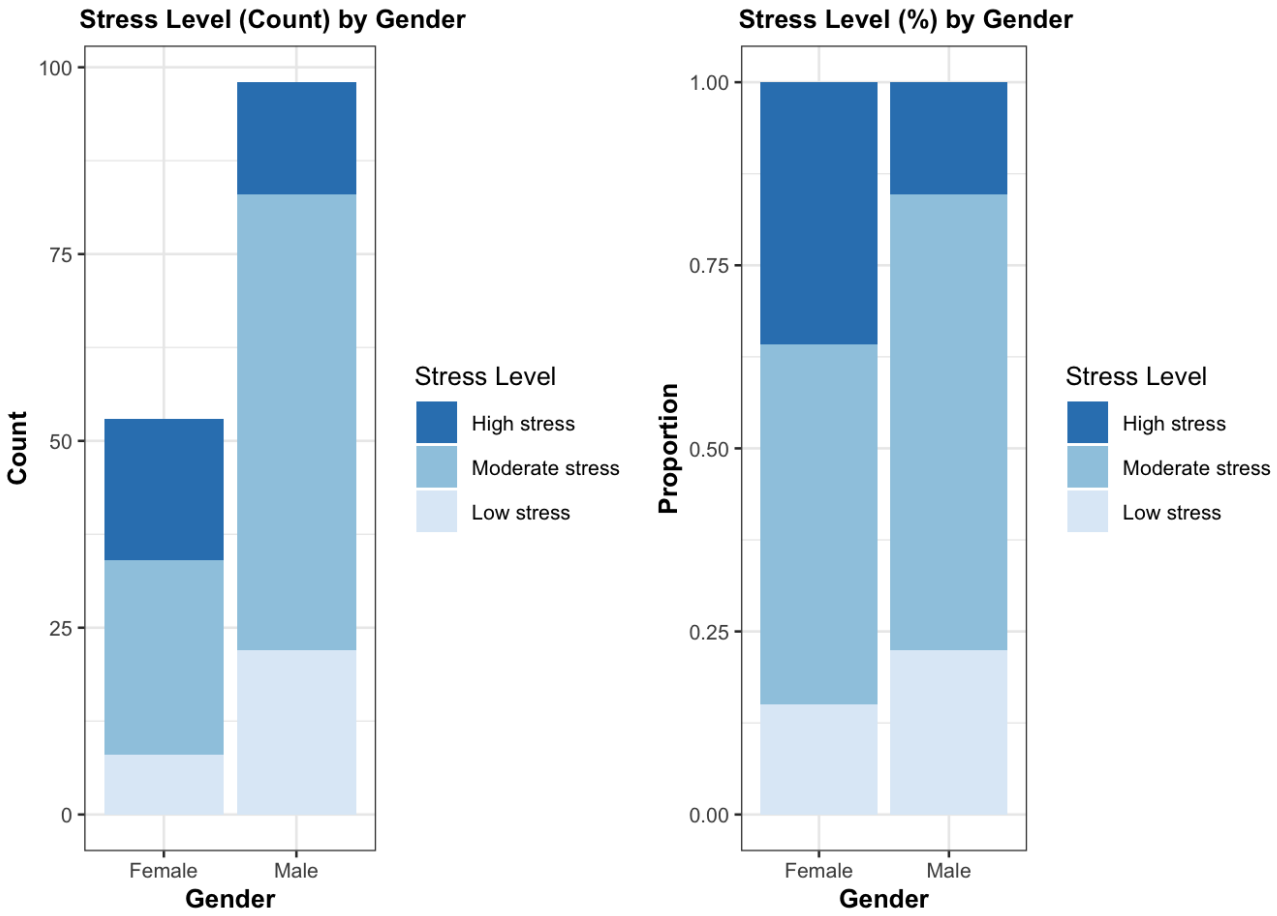
```
mat_stress_t <- mat_stress %>% t() # take the transpose of the matrix

y <- mat_stress_t %>% as.data.frame() %>% # encode the matrix as a data fr
       ame
  tibble::rownames_to_column(var = "stress") %>% # create a stress label f
       or the column
  tidyr::pivot_longer(cols = c("Male", "Female"), #turns the male and fema
       le columns into rows
                      names_to = "gender", values_to = "count") # specifie
       s the name of the columns of gender and the count

stress_name = categories[3:1]

p_base = ggplot(y, aes(x = gender, y = count, fill = factor(stress, levels
       = stress_name))) +
  theme_bw(base_size = 10) +
  scale_fill_brewer(palette = "Blues", direction = -1) +
  labs(fill = "Stress Level", x = "Gender") +
  theme(legend.position = "right") +
  theme(plot.title = element_text(hjust = 0.05, face = "bold", size = 10))
       +
  theme(axis.title = element_text(face = "bold"))
p_count = p_base +
  geom_bar(stat = "identity") +
  labs(y = "Count",
       title = "Stress Level (Count) by Gender")
p_proportion = p_base +
  geom_bar(stat = "identity", position = "fill") +
  labs(y = "Proportion",
       title = "Stress Level (%) by Gender")
gridExtra::grid.arrange(p_count, p_proportion, ncol = 2)
```

While it might be self evident from the plot, still better to spell out in words that a greater proportion of women selected high stress than men.



6: Discussion

Whilst the numerical stress values were grouped into three categorical stress levels (visualised above), it might have been prudent to use these categorical variables in the survey design.

There is perhaps selection bias at play wherein men who are stressed become tunnel-visioned and ignore surveys, vs women who are stressed who still complete surveys asked of them under a greater belief in social duty.

An alternative explanation might merely be that women are more comfortable in externalising and accepting when they are feeling stressed, an idea which is supported in the literature (Bianchin & Angrilli, 2012).

Evidence that women are merely more comfortable in externalising when they are feeling stressed. incomplete thought?

Power of the test However, when examining the power of our Chi-Square Test of independence it indicates a power level of 80%. This is reasonably powerful, however improvements could naturally be made by increasing the sample size.

Hide

```
es = ES.chisq.assoc(mat_stress)$phi
broom::tidy(pwr.chisq.test(w=es, N = 172, df = 2, sig.level = 0.05)) %>% k
        bl() %>% kable_styling()
```

sig.level	power
0.05	0.7981728

Test 4: To investigate the mean of the differences between self-assessed mathematical ability and self-assessed coding ability

1: Introduction: Do data science students score themselves higher on average in mathematical ability than in R coding ability?

There are two theses that underlie the motivation for this experiment:

1. There is, perhaps, a **positive correlation** between how one scores himself or herself on mathematical ability and R coding ability. These are **not independent** for two reasons:
    1. If someone is overconfident in their self perceptions of one ability, they might demonstrate overconfidence in other abilities.
    2. If a student is good at mathematics, then they are probably both intelligent and hardworking, which increases the likelihood of them being good at R.
2. Data science students will, on average, score their mathematical ability *higher* than their R coding ability since most students have had more experience of mathematics (studying since they were young children) than of R.

These underlying theses give rise to a certain hypothesis that will be examined through the data.

2: Hypotheses

Let D_i be the self-assessed mathematical ability score less the self-assessed R coding ability score for each student.

- $H_0 : \mu_D = 0$
- $H_1 : \mu_D > 0$

The population is all second-year data science students. The rationale is examined in the discussion.

DATA2002 is broader than "data science" students

3: Data Inspection and Setup

Before determining which test to use and which assumptions to make, it is prudent to look at the underlying structure of the data.

The boxplot shows a **positive difference** between perceived math ability and R ability, with the null hypothesised difference, zero, only just making the lower bound of the interquartile range.

Hide

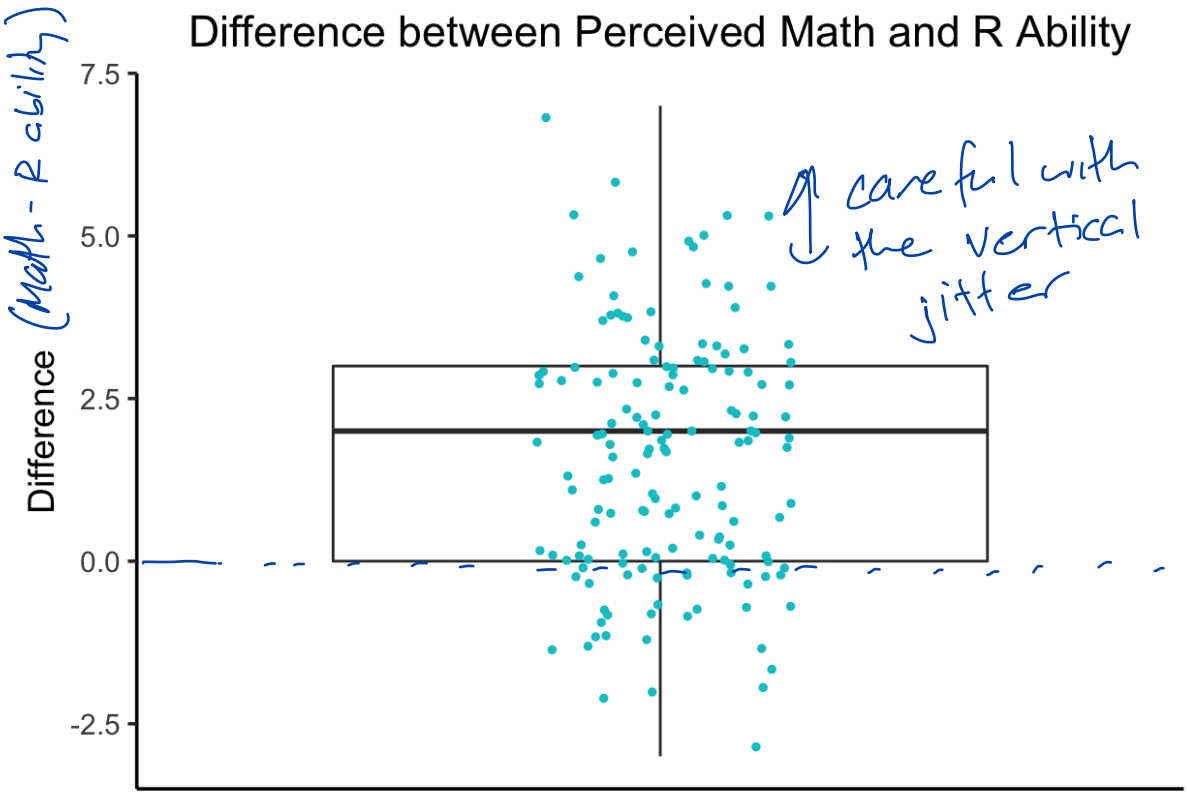
```
data4 <- data2 %>%
  select(math_ability, r_ability) %>%
  mutate(difference = math_ability - r_ability) %>%
  na.omit()

ggplot(data4, aes(x = "difference", y = difference)) +
  geom_boxplot() +
  geom_jitter(width=0.15, size = 1, colour = "turquoise3") +
  theme_classic(base_size = 14) +
    theme(
      plot.title = element_text(hjust = 0.5),
      plot.margin = margin(1,1,1,1,"cm")) +
  labs(
    title = "Difference between Perceived Math and R Ability",
    y = "Difference") +
   theme(axis.title.x = element_blank(), #remove all x-axis lebsl, ticks a
         nd text
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

+ geom_hline(yintercept = 0,
linetype = "dashed",
colour = grey(0.5)).

+ coord_flip()

with that then use fig.height=2 in the chunk option to save space

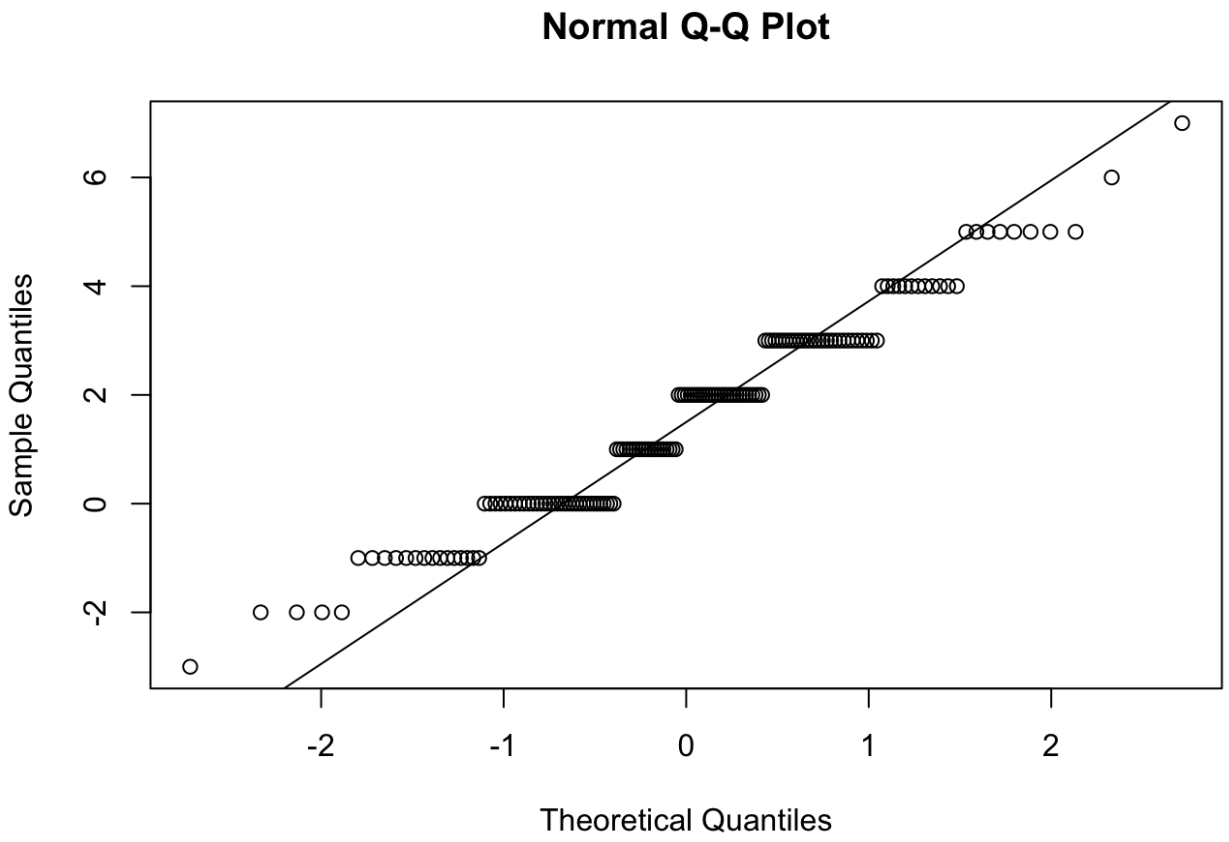


The quantile-quantile plots shows that while the distribution appears relatively symmetric, it has strongly positive kurtosis such that it differs from a normal distribution.

I think this is a bit harsh, it's not that far from being approximately normal.

Hide

```
qqnorm(data4$difference, main = "Normal Q-Q Plot")
qqline(data4$difference)
```



Hide

```
#this is purposely eval = FALSE, not valid code for what I want here.
ggplot(data4, aes(x = difference)) +
  geom_histogram(bins = 10, fill = "dodgerblue4") +
  labs(title = "Frequency Distribution of COVID tests in the past two mont
        hs", y = "# of Students", x = "# of COVID Tests") +
  theme_classic(base_size = 13, base_line_size = 1) +
  theme(plot.title = element_text(hjust = 0.5), plot.margin = margin(1,1,1
        ,1,"cm")) +
  scale_x_continuous(
    breaks = scales::pretty_breaks(n = 12), #add axis ticks for every data
        point
    expand = c(0,0)) + #remove the gap between the plot and the y-axis
  xlim(-8,8) +
  scale_y_continuous(expand = c(0,0))
```

4: Assumptions

Given that the data do not appear normally distributed, the Wilcoxon signed-rank test will be used which has greater statistical power for symmetric distributions ~~that are non-parametric (i.e. not from a probability distribution parametrised by mean and variance like a normal distribution)~~ (Shieh, Jan & Randles, 2007).

The Wilcoxon signed-rank test still makes several assumptions:

- D_i follows a symmetric distribution.

When the normality assumption isn't met.

    - The boxplot shows that this is approximately satisfied.
- D_i are random variables which are independently distributed.
    - This is satisfied by the design of the survey and assumed independence of the survey responses.
- The differences are continuous in theoretical nature and have an ordinal level of measurement.
    - Whilst the measurements are ordinal, the differences are not continuous. Regardless, continuity is assumed and the continuity correction is not applied, since it has been shown that the continuity correction frequently adjusts too far (Bergmann, Ludbook, & Spooren, 2000).

5: Statistical Testing & Decisions

The test, statistic is given by:

$$W^+ = \sum_{i:D_i>0} R_i$$

where R_i are the ranks of $|D_1|, |D_2|, \ldots, |D_n|$ under H_0, $W \sim WSR(n)$.

$p < 0.001$

Conducting the Wilcoxon signed-rank test yields a test statistic of $W^+ = 6581.5$

This corresponds to a p-value of $p = 4.21e^{-16}$ suggesting that there is a near-zero chance of having a test statistic as or more extreme to that observed given that null is true.

Using a level of significance of $\alpha = 0.05$ or even $\alpha = 0.01$, since $p < 0.01 < 0.05$ we have sufficient evidence to reject the null hypothesis that there is no difference between self-assessed math score and self-assessed R coding ability score for second-year data science students.

Hide

```
d <- data4$difference

wilcox.test(d, alternative = "greater", correct = FALSE)
```

Hide

```
n <- length(d)

q <- 0:(n*(1+n)/2)
probs = dsignrank(q, n)
names(probs) = q

plotrix::barp(dsignrank(q,n),names.arg = q, col="blue", border = "purple")
abline(v = 6581.5, col = "purple", lwd = 2)
```

Comparing with the t-test

Comparing the Wilcoxon ranked-sign test with the one-side, paired t-test, the same decision to reject the null is arrived at. The test statistic is 10.233.

The p-value is 2.2e-16.

Hide

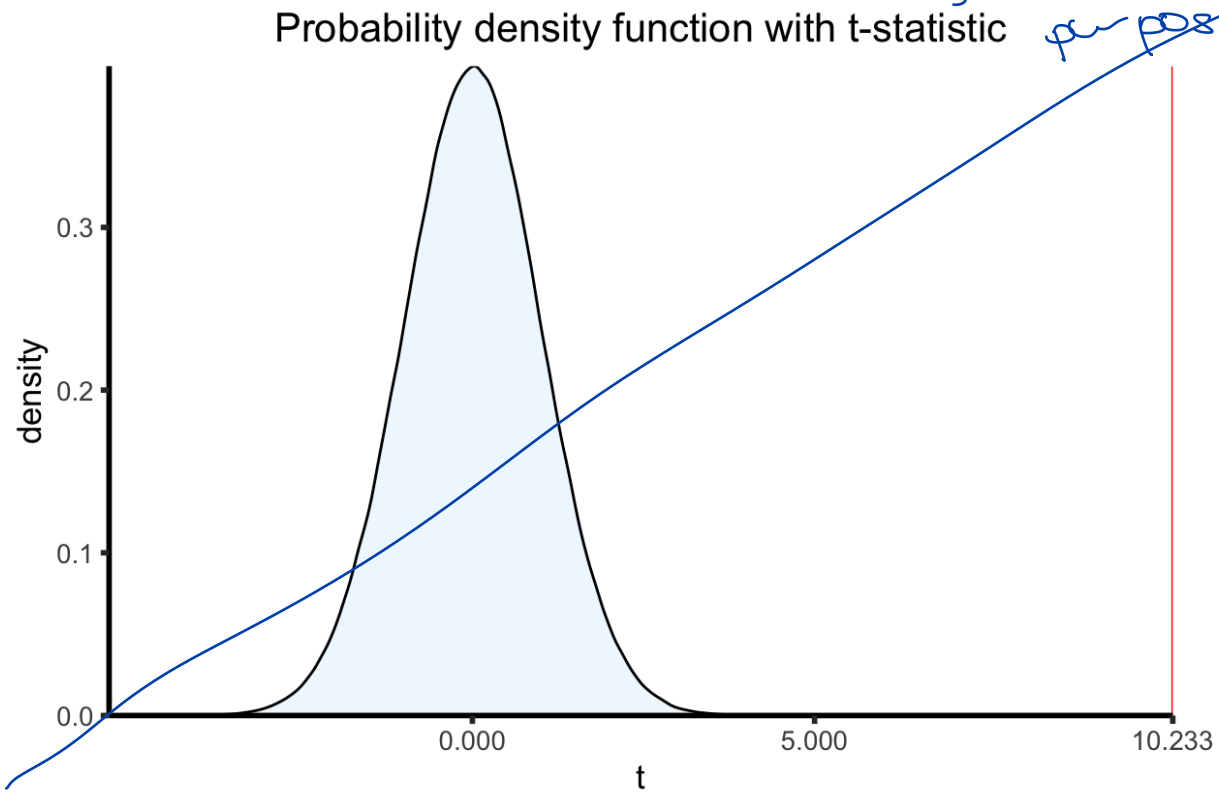
```
t.test(d, alternative = "greater", correct = FALSE)
```

The test statistic on the probability density function of the student's t-distribution is shown below.

Hide

```
ggplot(data.frame(t=rt(1000000,151))) + geom_density(aes(x=t), fill = 'ali
        ceblue') +
  geom_vline(xintercept = 10.233, colour = "red") +
  ggtitle("Probability density function with t-statistic") +
  theme_classic(base_size = 13, base_line_size = 1) +
    theme(plot.title = element_text(hjust = 0.5), plot.margin = margin(1,1
        ,1,1,"cm")) +
  scale_x_continuous(
    breaks=c(0,5,10.233,12),
    expand = c(0,0)) + #remove the gap between the plot and the y-axis
  scale_y_continuous(expand = c(0,0))
```

not necessary particularly given the t-test is only for comparison purposes.



This is unsurprising considering that the underlying distribution didn't tend far from a normal distribution, and given the large sample size, the sampling distributions of the mean of the paired differences will tend towards a normal distribution. ✓

What this shows is that whilst the Wilcoxon sign-rank test had a p-value over 2 times as small, both tests readily rejected the null hypothesis.

I would caution against comparing p-values in this way.

6: Discussion

Population

DATA2X02

The population was construed as all second-year data science students. While it is tempting to extend the population further, it would be naive to do so given that mathematical and coding ability vary dramatically between different groups. Given that the sample data was taken from second-year data science students taking DATA2X02 at USYD, it is apt to keep the population second-year data science students more broadly.

→ even within DATA2X02 we have heterogeneity:
- computational data science
- data science
- statistics
- financial maths + stats ---

Musings on perceptions

The results of this study give interesting insights into a supposed difference between perception and reality.

In reality, students are probably much closer to becoming experts at R coding than they are at mathematics, given the ever-increasing complexity of mathematics. This perhaps underscores the fallibility of perception. Exploring this, one potential reason is that the 0-10 ordinal scales are subjective and self-assessed rather than objective. Under subjective scales like this, students' self-assessment of ability likely derives from benchmarks with others of similar age. This might explain students' higher self-assessment of mathematical ability. After all, it is easy to get an illusion of competence with mathematics after mastering a few concepts and receiving good marks. — this may be true for R too — illusion of competence

Further study

Given that the null was rejected with high confidence from the sample data, this might suggest a further investigation under the assumption that the mean difference is not zero. It could be interesting to examine the percentage increase in self-assessed scores of mathematics relative to self-assessed R coding ability. other underlying factors that haven't been accounted for in this analysis.

References

Academic References

Anbumalar, C., Dorathy, A. P., Jaswanti, V. P., Priya, D., & Reniangelin, D. (2017). Gender differences in perceived stress levels and coping strategies among college students. *The International Journal of Indian Psychology*, *4*(4), 22-33.

Bianchin, M., & Angrilli, A. (2012). Gender differences in emotional responses: A psychophysiological study. *Physiology & behavior, 105*(4), 925-932.

Bergmann, R., Ludbrook, J., & Spooren, W. P. (2000). Different outcomes of the Wilcoxon—Mann—Whitney test from different statistics packages. *The American Statistician*, *54*(1), 72-77.

Bewick, V., Cheek, L., & Ball, J. (2003). Statistics review 8: Qualitative data–tests of association. *Critical care*, *8*(1), 1-8.

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2018). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit?. *Behavior Research Methods*, *50*(3), 937-962.

Bland, Martin (1995). An Introduction to Medical Statistics. Oxford University Press. p. 168. ISBN 978-0-19-262428-4.

Brosamler, G. A. (1988, November). An almost everywhere central limit theorem. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 104, No. 3, pp. 561-574). Cambridge University Press.

Calvarese, M. (2015). The effect of gender on stress factors: An exploratory study among university students. *Social Sciences*, *4*(4), 1177-1184.

Conover, W. J. (1974). Some reasons for not using the Yates continuity correction on 2× 2 contingency tables. *Journal of the American Statistical Association*, *69*(346), 374-376.

Fagerland, M. W., & Sandvik, L. (2009). Performance of five two-sample location tests for skewed distributions with unequal variances. *Contemporary clinical trials*, *30*(5), 490-496.

Geng, X., Chen, Z., Lam, W., & Zheng, Q. (2013). Hedonic evaluation over short and long retention intervals: The mechanism of the peak–end rule. *Journal of Behavioral Decision Making*, *26*(3), 225-236.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161-1166.

Kay, K., and Shipman, C. (2014) *The Confidence Gap*. Retrieved from, https://www.theatlantic.com/magazine/archive/2014/05/the-confidence-gap/359815/ (https://www.theatlantic.com/magazine/archive/2014/05/the-confidence-gap/359815/)

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Lundberg, U. (2005). Stress hormones in health and illness: the roles of work and gender. *Psychoneuroendocrinology*, *30*(10), 1017-1021.

Markowski, Carol A.; Markowski, Edward P. (1990). “Conditions for the Effectiveness of a Preliminary Test of Variance”. The American Statistician. 44 (4): 322–326.

Moser, B. K., & Stevens, G. R. (1992). Homogeneity of variance in the two-sample means test. *The American Statistician*, *46*(1), 19-21.

Nadel, L., Hupbach, A., Gomez, R., & Newman-Smith, K. (2012). Memory formation, consolidation and transformation. *Neuroscience & Biobehavioral Reviews*, *36*(7), 1640-1645.

Olff, M., Langeland, W., Draijer, N., & Gersons, B. P. (2007). Gender differences in posttraumatic stress disorder. *Psychological bulletin*, *133*(2), 183.

Perlman, M. D. (1969). One-sided testing problems in multivariate analysis. *The Annals of Mathematical Statistics, 40*(2), 549-567.

Ring, P., Neyse, L., David-Barett, T., & Schmidt, U. (2016). Gender differences in performance predictions: Evidence from the cognitive reflection test. *Frontiers in psychology*, 7, 1680.

Rudolph, K. D., & Hammen, C. (1999). Age and gender as determinants of stress exposure, generation, and reactions in youngsters: A transactional perspective. *Child development*, *70*(3), 660-677.

Shieh, G., Jan, S. L., & Randles, R. H. (2007). Power and sample size determinations for the Wilcoxon signed-rank test. *Journal of Statistical Computation and Simulation*, *77*(8), 717-724.

Spielberger, C. D., & Reheiser, E. C. (1994). The job stress survey: Measuring gender differences in occupational stress. *Journal of Social Behavior and Personality*, *9*(2), 199.

Yates, F.(1934). Contingency tables involving small numbers and the χ 2 test. *Supplement to the Journal of the Royal Statistical Society*, *1*(2), 217-235.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, *57*(1), 173-181.

Package References

Erich Neuwirth (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer (https://CRAN.R-project.org/package=RColorBrewer)

Hao Zhu (2021). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra (https://CRAN.R-project.org/package=kableExtra)

Jennifer Beaudry, Emily Kothe, Felix Singleton Thorn, Rhydwyn McGuire, Nicholas Tierney and Mathew Ling (2021). gendercoder: Recodes Sex/Gender Descriptions Into A Standard Set. R package version 0.0.0.9000. https://github.com/ropenscilabs/gendercoder (https://github.com/ropenscilabs/gendercoder)

Lemon, J. (2006) Plotrix: a package in the red light district of R. R-News, 6(4): 8-12.

Sam Firke (2021). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.1.0. https://CRAN.R-project.org/package=janitor (https://CRAN.R-project.org/package=janitor)

Stephane Champely (2020). pwr: Basic Functions for Power Analysis. R package version 1.3-0. https://CRAN.R-project.org/package=pwr (https://CRAN.R-project.org/package=pwr)

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686 (https://doi.org/10.21105/joss.01686)

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686 (https://doi.org/10.21105/joss.01686)

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag: New York.

Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.33.

R Core Team (2021)... cite R itself.

Hide

```
citation(package = "tidyverse")
citation(package = "janitor")
citation(package = "ggplot2")
citation(package = "janitor")
citation(package = "pwr")
citation(package = "gendercoder")
citation(package = "kableExtra")
citation(package = "knitr")
citation(package = "plotrix")
citation(package = "RColorBrewer")
```

This is an exceptionally well written and referenced report. It's not as concise as it could be, occasionally duplicating analyses and a couple of extra plots but that's being quite pedantic. Well done!