Survey Report

Ishaan Singh

Introduction

A survey was conducted which contained a range of questions in order to gather data from a sample of students. There are in total 172 participants and 20 questions that were asked to those who participated in the survey. The survey questions are the following.

1. How many times have you been tested for COVID?
2. Gender
3. Postcode of where you live during semester
4. How long has it been since you last went to the dentist?
5. On average, how many hours per week did you spend on university work last semester?
6. What is your favourite social media platform?
7. Did you have a dog or a cat when you were a child?
8. Do you currently live with your parents?
9. How many hours a week do you spend exercising?
10. What is your eye colour?
11. Do you have asthma?
12. On average, how many hours per week did you work in paid employment in semester 1?
13. What is your favourite season of the year?
14. What is your shoe size?
15. How tall are you?
16. How often do you floss your teeth?
17. Do you wear glasses or contacts?
18. What is your dominant hand?
19. How do you like your steak cooked?
20. On a scale from 0 to 10, please indicate how stressed you have felt in the past week.

Investigations and Results

Is this a random sample of students?

This data has been collected through a survey that was posted on a discussion forum. Responding to the survey was not mandatory and hence, it does not contain the data for all the students enrolled in this unit.

This leads us to question about whether this is a **random sample** of students enrolled in the unit. It is important to first understand what exactly the population is.

Given that there was only a short period of time (a few days), during which the survey was available, the responses recorded were from people who volunteered to answer the questions. This may mean that only students who frequently check-in on the discussion forum to read, ask or answer questions may have responded to the survey. Additionally, some students may have simply not wanted to share their data. This group of students could possibly have a different social media usage compared to others in the population. Additionally, the survey asks a question about the amount of hours spent every week in doing paid work during Semester One of 2020. Students who have significant work commitments may be less likely to go through and complete a voluntary survey, which could make the sample non-random.

Even among those who have responded to the survey, there are blank or NULL values in certain questions. Hence, in conclusion, we cannot say that the sample is truly randomly selected. However, we shall assume that it is in certain cases. A better method would have been to make the survey mandatory, along with all the questions in it, and then selecting a random sample from it (if not using all the data, since the population in this case is not too large).

What are the potential biases? Which variables are most likely to be subjected to this bias?

There are a number of ways that the results of this survey may indicate bias.

Sampling Bias. As discussed before, it may turn out to be the case that the subset of the population that has been sampled through this survey is different in major characteristics to those individuals who have not been sampled. For example, it might be true that those who have responded to the survey do not have very rigorous or time-consuming work commitments outside of University or even at University in this semester. Hence, the variable that indicates the *average hours worked in Semester One per week*, may indicate such bias. This may also be the case for the variable that indicates the *average hours spent doing University work in Semester One per week*.

Non-Response Bias. Upon inspecting the data, it is evident that not all questions have been answered by everyone who has responded to the survey. This may cause certain groups of people, or individuals with certain characteristics to be under-represented in the results. For example, there were a number of non-responses (10) for the question requesting for the participant's height. People who do not feel comfortable sharing physical characteristics about themselves, may be disinclined to respond to such questions about height and/or shoe size, which is a cause for bias. This can also be seen in the responses for the question that asks for the postcode from where the students study. Location data is quite sensitive, especially when it indicates the participant's home location and therefore, the large number of non-responses (18) for this question is most certainly a source of bias.

Measurement Bias. The method of data collection is a survey. This means that there are no measurements recorded. Rather, one must rely on the responses which are given, even if they haven't been properly measured or measured at all. For example, not all participants may have the equipment to measure their own height accurately, hence they may simply report an old measurement that they remember or might simply estimate it. This means that it is possible that certain participants overstate their height. Overstating measurements may also be a source of bias for the amount of paid work done per week, the amount of University work done per week, the amount of exercise done per week etc. Additionally, in the questions requesting the participants height and shoe size, no particular scale or units are specified, thus creating inconsistencies in the measurements gathered.

Are there any questions that needed improvement to generate useful data?

There are numerous improvements that could have been made to questions to achieve better quality responses. These include:

- The question requesting the gender of the participant, being a free response question has allowed for multiple ways of presenting the same response. For example, males may respond with "Male", or "M", or may entirely misspell the word as has occurred in the data.
- The question requesting the participant's postcode is somewhat vague, as it contains postcodes from different countries altogether.
- The question regarding the time spent University work may not be very informative. University work could be broken down into many different types of activities such as studying, leadership work or organisation in clubs and societies, undertaking paid work, attending seminars etc. Hence, we may expect certain students to include such commitments in their response.
- The question regarding favourite social media platforms has to be parsed carefully. The first step would be to have a clear definition of which services would classify as being social media. This may involve separating a video service such as YouTube from a chat service such as Whatsapp, or Messenger. Even though the question implies that only one service must be provided, many responses have listed multiple services. Hence, options to choose from or only accepting a single service could improve this question.
- The question regarding eye colour would work better if participants had bins in which to classify their eye colour, rather than provide a free response. This makes it hard to distinguish between what is for example: Hazel, Brown or Dark Brown; definitions which may be subjective.
- The question on height and shoe-size have similar flaws in that they fail to specify a scale. For example students may (and have) answer in 'cm' and 'm', if the preferred units are not specified. Additionally, they may provide a shoe-size measured on a different scale e.g. US, UK etc.

Does the number of COVID-19 tests follow a Poisson distribution?

Let us consider our hypotheses:

- **Null Hypothesis**: The number of COVID-19 tests comes from a Poisson distribution.
- **Alternative Hypothesis**: The number of COVID-19 tests does not come from a Poisson distribution.

In order to determine this, we shall first clean the column containing the number of COVID-19 tests that the participants have taken. We then can create a frequency table for this column as shown below. The top row indicates the number of tests taken, while the row beneath represents the frequency of that observation.

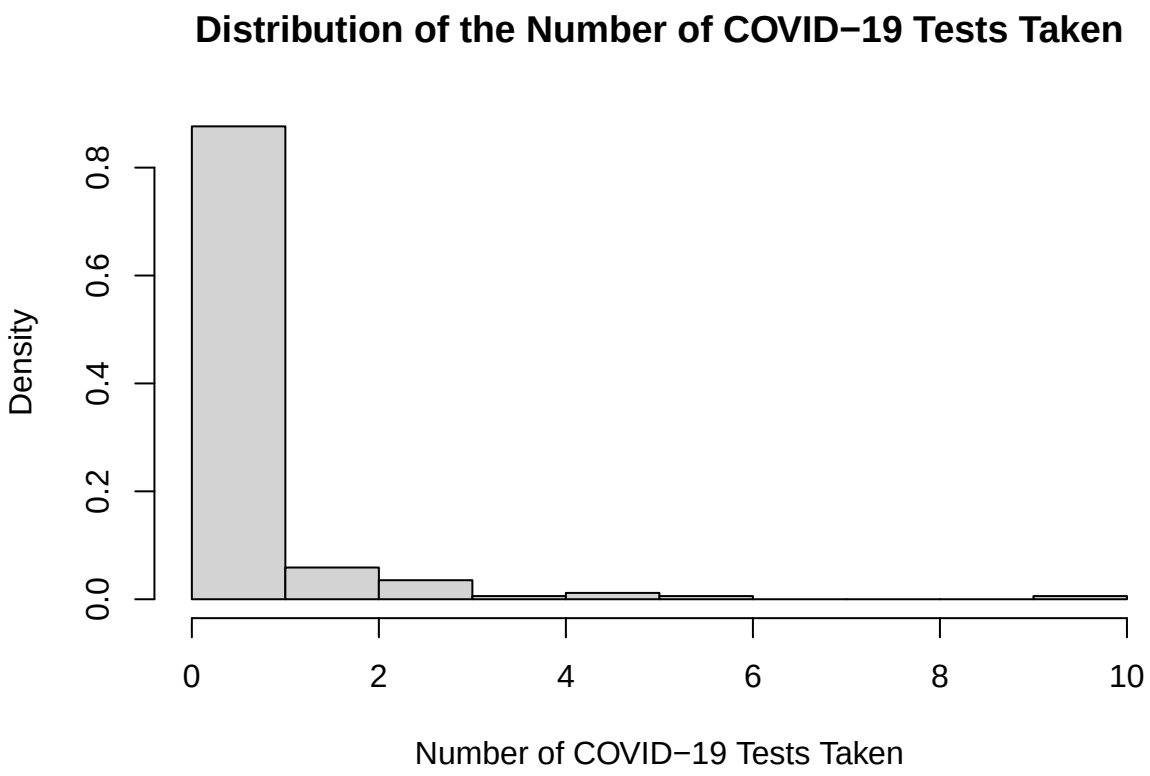
```
# Creating a frequency table for the number of COVID-19 tests
covid_tests = survey_data$Covid_Tests %>% na.omit()

covid_table = table(covid_tests)
covid_table
```

```
## covid_tests
##   0   1   2   3   4   5   6  10
## 121  28  10   6   1   2   1   1
```

```
# Converting table to a dataframe to extract the x and y vectors
covid_table = as.data.frame(covid_table)
```

Using our frequency table, we can create a histogram to view the distribution.

```
# Plotting a histogram
hist(survey_data$Covid_Tests, xlab = "Number of COVID-19 Tests Taken", ylab = "Density",
    main = "Distribution of the Number of COVID-19 Tests Taken", freq = FALSE)
```



We then use the table generated above to create a vector with the categories and the observed frequencies along with the expected probabilities.

We further need to test our **assumption** of the observations being independent, by inspecting the expected frequencies. In order to satisfy the assumption, we must have the expected frequencies for each category be greater than or equal to 5. From the histogram created above, we should expect some categories to violate this assumption, since such a large proportion of the data lies in the first few categories.

```
y = c(covid_table$Freq)  # observed frequencies
x = c(as.numeric(levels(covid_table$covid_tests))[covid_table$covid_tests])
# converting the factor to numeric variables
n = sum(y)
l = sum(y * x)/n  # estimating lambda
p = dpois(x, lambda = l)  # expected probabilities
ey = n * p  # expected counts
paste("Independence of Observations Assumption Satisfied:", all(ey >= 5))  # checking assumption
```

```
## [1] "Independence of Observations Assumption Satisfied: FALSE"
```

From our results above, we can see that our assumption about the independence of observations was violated. Hence, we deal with this issue by combining the data for the participants who have had more than two tests into a single category.

```
# Combining categories ey
yr = c(y[1:2], sum(y[3:8]))  # updated observed frequencies
```

```r
pr = c(p[1:2], 1 - sum(p[1:2]))  # updated expected probabilities
eyr = sum(yr) * pr  # updated expected counts

# testing our assumptions
paste("Independence of Observations Assumption Satisfied:", all(eyr >= 5))  # checking assumption
```

```
## [1] "Independence of Observations Assumption Satisfied: TRUE"
```

Testing our assumption against this new sample satisfies the assumption about independence.

Hence, we proceed to calculate the test-statistic and p-value.

```r
k = length(yr)

t0 = sum((yr - eyr)^2/eyr)  # test-statistic
paste("The test statistic equals:", round(t0, 3))
```

```
## [1] "The test statistic equals: 19.343"
```

```r
pval = 1 - pchisq(t0, df = k - 1 - 1)  # p-value
paste("The p-value equals:", pval)
```

```
## [1] "The p-value equals: 1.09238043553006e-05"
```

```r
paste("Reject Null Hypothesis at 5% Significance? :", pval < 0.05)
```

```
## [1] "Reject Null Hypothesis at 5% Significance? : TRUE"
```

Testing our hypothesis at the 5% significance level shall cause us to reject our null hypothesis, since our obtained p-value (1.09e-05) is smaller than 0.05.

Conclusion: Hence, we cannot say, based on the data we have, that the number of COVID-19 tests comes from a Poisson distribution.

Further Hypothesis Tests

Is the average heights of males and females similar to the general Australian population? While the sample we have is a subset of students completing the DATA2002 unit in 2020, this subset is likely to be comprised of legal adults (18+ in age). Hence, we can compare whether the mean heights for males and females are representative of the Australian adult population.

We must make the assumption that the heights have been correctly reported. This assumption can be challenged; since, for reasons discussed earlier, some participants have not provided their heights, and among those who have it is hard to determine how many participants have simply estimated their heights or provided something like an upper bound. Additionally, adult heights are usually approximated to have a normal distribution due to a range of genetic and environmental variance 1 . This too, is something we shall visually test. We will also have to exclude those who identify their gender as being "non-binary" due to a lack of availability of data for their heights.

We shall perform the test for males and females separately since we have data for the Australian male and female population. 2

We shall use a **one-sample t-test** to test our hypotheses as it allows us to compare the means of numerical data for samples of people. Hence, our hypotheses for the male sample are:

- **Null Hypothesis**: The mean heights of the male sample equals the mean height of the adult male population in Australia (175.6 cm).
- **Alternative Hypothesis**: The mean heights of the male sample does NOT equal the mean height of the adult male population in Australia (175.6 cm).

In order the test the hypotheses above, we shall conduct a two-sided t-test. An **assumption** we make for a standard t-test is that:

- Each observation in both samples is a random observation from our population.
- Hence, our observations from both samples are independently and identically distributed.

For reasons discussed previously, we cannot claim that we definitively have a random sample to perform the hypothesis tests on. However, we shall make that assumption nonetheless.

We shall inspect the data for males and females. Due the variety of spellings employed by participants in answering this survey, we filter our data to accept some common spellings for each gender. Due to some inconsistencies with units, some participants have provided their heights in metres. We will therefore have to apply a filter to ensure that all heights are in centimetres.

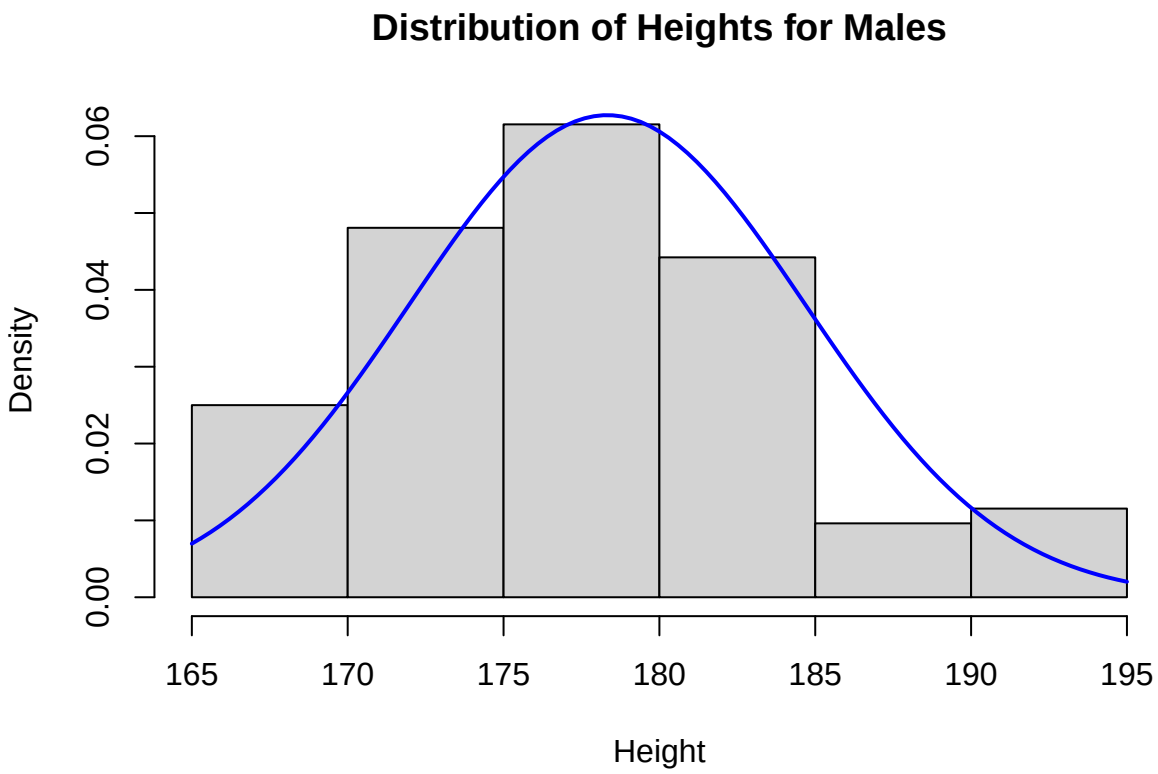
```
# unique(survey_data$Gender)

# Creating the subset of male participants
male_filtered = survey_data %>%
    filter(Gender == "Male" | Gender == "M" | Gender == "m" |
               Gender == "male" | Gender == "MALE" | Gender == "mAle",
           Height > 2.5)
male_sample = male_filtered$Height %>% # cleaning any NA values
    na.omit()

paste("Male Sample Size:", length(male_sample), "observations")
```

```
## [1] "Male Sample Size: 104 observations"
```

```
paste("Mean Height from the Male Sample:", round(mean(male_sample), 2))
```

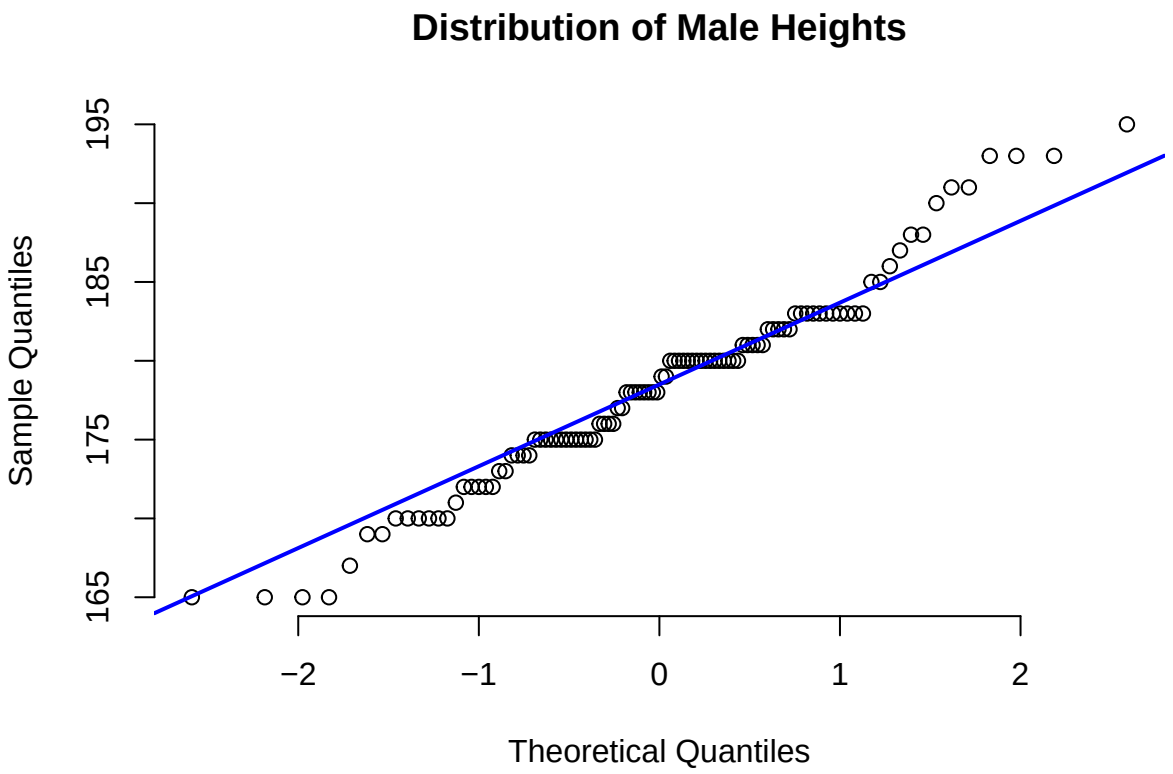
```
## [1] "Mean Height from the Male Sample: 178.33"
```

```
# Plotting a histogram
hist(male_sample, main = "Distribution of Heights for Males", xlab = "Height",
     freq = FALSE)

# Impose a normal curve with the same parameters as the male sample
curve(dnorm(x, mean = mean(male_sample), sd(male_sample)), add = TRUE, lwd = 2,
      col = "blue")
```



In order to further assess the normality of the distribution, we shall use a qq-plot for the data.

```
# qqPlots for Male Heights
qqnorm(male_sample, pch = 1, frame = FALSE, main = "Distribution of Male Heights")
qqline(male_sample, col = "blue", lwd = 2)
```



If the data were normally distributed the points on the plot would closely follow a straight line as indicated by the blue line on the plot above. However, as can be seen in the plot above, while the data largely lies close to the line, it does deviate and scatter at the end points. Hence, while the data does resemble a normal distribution, it is not exactly so.

Another interesting point to note from the plot above is how many height measurements appear very frequently. For example, a considerable number of participants have reported their height to be exactly 165 cm, 170 cm, 175 cm, 180 cm etc. This further suggests that the heights that have been reported may be mere estimations or rounded off measurements.

We can now proceed towards conducting our t-test. We shall make the assumption that our t-statistic follows a t-distribution.

```
n.males = length(male_sample)
t.males = (mean(male_sample) - 175.6)/(sd(male_sample)/sqrt(n.males))
paste("Test-Statistic:", round(t.males, 4))
```

```
## [1] "Test-Statistic: 4.3733"
```

```
p.males = pt(t.males, n.males - 1)
paste("p-Value:", round(p.males, 5))
```

```
## [1] "p-Value: 0.99999"
```

```
paste("Reject Null Hypothesis at 5% Significance? :", p.males < 0.025)
```

```
## [1] "Reject Null Hypothesis at 5% Significance? : FALSE"
```

As can be seen from our results above, since the p-value obtained (0.99999) is larger than 0.025, we cannot reject the null hypothesis at the 5% significance level.

Conclusion. From the data we have and the tests conducted we conclude that, the mean heights of the male sample equals the mean height of the adult male population in Australia (175.6 cm).

We perform a similar test for females. Our hypotheses are:

- **Null Hypothesis**: The mean heights of the female sample equals the mean height of the adult female population in Australia (161.8 cm).
- **Alternative Hypothesis**: The mean heights of the female sample does NOT equal the mean height of the adult female population in Australia (161.8 cm).

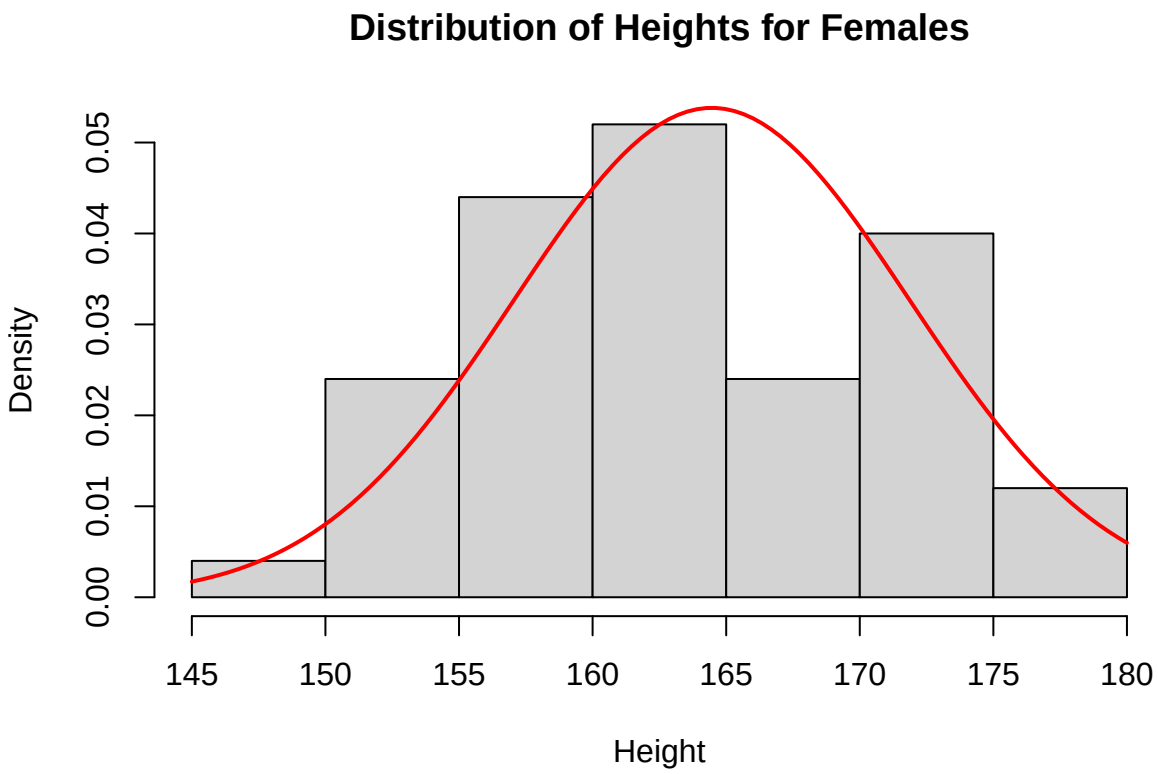
```
# Creating the subset of female participants
female_filtered = survey_data %>%
    filter(Gender == "Female" | Gender == "F" | Gender == "f" |
                Gender == "female" | Gender == "FEMALE" | Gender == "femail",
           Height > 2.5)
female_sample = female_filtered$Height %>% # cleaning any NA values
    na.omit()

paste("Female Sample Size:", length(female_sample), "observations")
```

```
## [1] "Female Sample Size: 50 observations"
```

```
paste("Mean Height from the Female Sample:", round(mean(female_sample), 2))
```

```
## [1] "Mean Height from the Female Sample: 164.46"
```

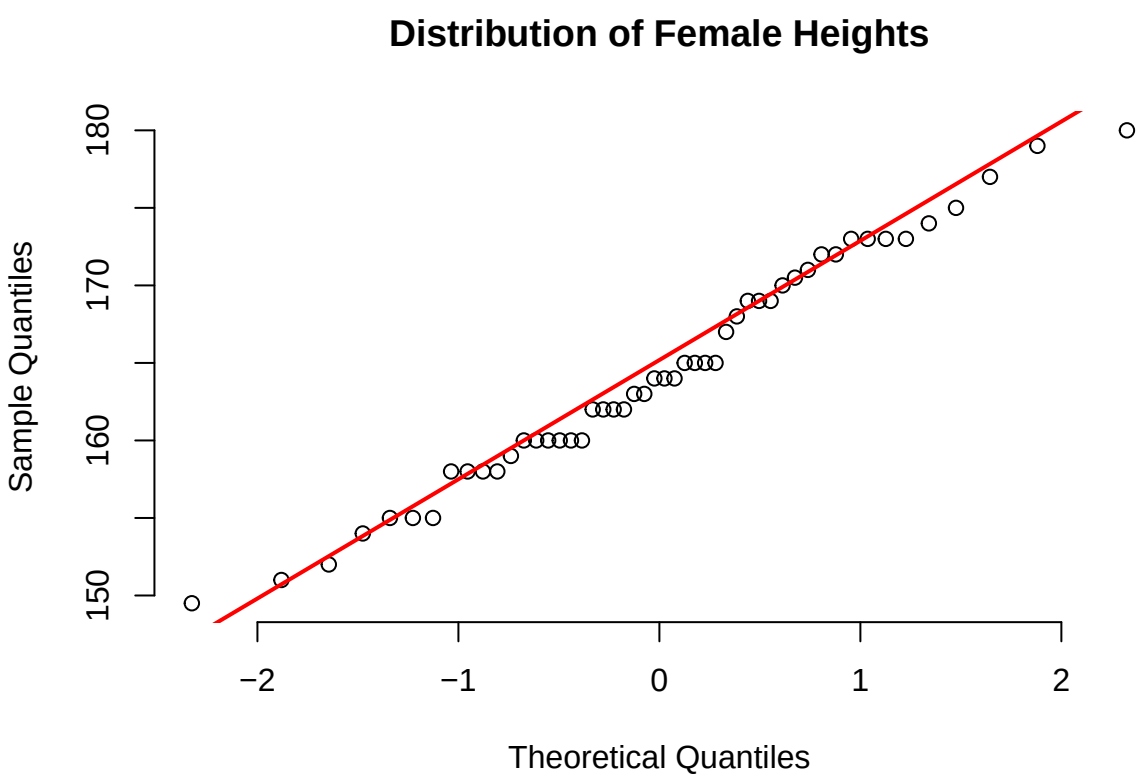
```
# Plotting a histogram
hist(female_sample, main = "Distribution of Heights for Females", xlab = "Height",
     freq = FALSE)

# Impose a normal curve with the same parameters as the male sample
curve(dnorm(x, mean = mean(female_sample), sd(female_sample)), add = TRUE, lwd = 2,
      col = "red")
```



The female sample we have is considerably smaller than the male sample. Hence, this may be a limiting factor. However, since we still have 50 observations, we can still perform the test to a reasonable certainty.

From the plot above, the distribution does appear to resemble a normal distribution. In order to clarify this we shall make use of a qq-plot.

```
# qqPlots for Male Heights
qqnorm(female_sample, pch = 1, frame = FALSE, main = "Distribution of Female Heights")
qqline(female_sample, col = "red", lwd = 2)
```



From the plot above, we see that most points are close to the line and there isn't too much deviation from the line even at the very ends of the distribution. Despite having a smaller sample of females, the data does appear to resemble a normal distribution more closely.

However, we again see some reported heights which have considerably high frequencies such as 160 cm. Hence, it may also be the case with the females that approximations or estimations of heights have been reported.

We can now proceed towards conducting our t-test. We shall make the assumption that our t-statistic follows a t-distribution.

```
n.females = length(female_sample)
t.females = (mean(female_sample) - 161.8)/(sd(female_sample)/sqrt(n.females))

p.females = pt(t.females, n.females - 1)

paste("Test-Statistic:", round(t.females, 3))
```

```
## [1] "Test-Statistic: 2.537"
```

```
paste("p-Value:", round(p.females, 4))
```

```
## [1] "p-Value: 0.9928"
```

```
paste("Reject Null Hypothesis at 5% Significance? :", p.females < 0.025)
```

```
## [1] "Reject Null Hypothesis at 5% Significance? : FALSE"
```

As can be seen from our results above, since the p-value obtained (0.9928) is larger than 0.025, we cannot reject the null hypothesis at the 5% significance level.

Conclusion. From the data we have and the tests conducted we conclude that, the mean heights of the female sample equals the mean height of the adult female population in Australia (161.8 cm).

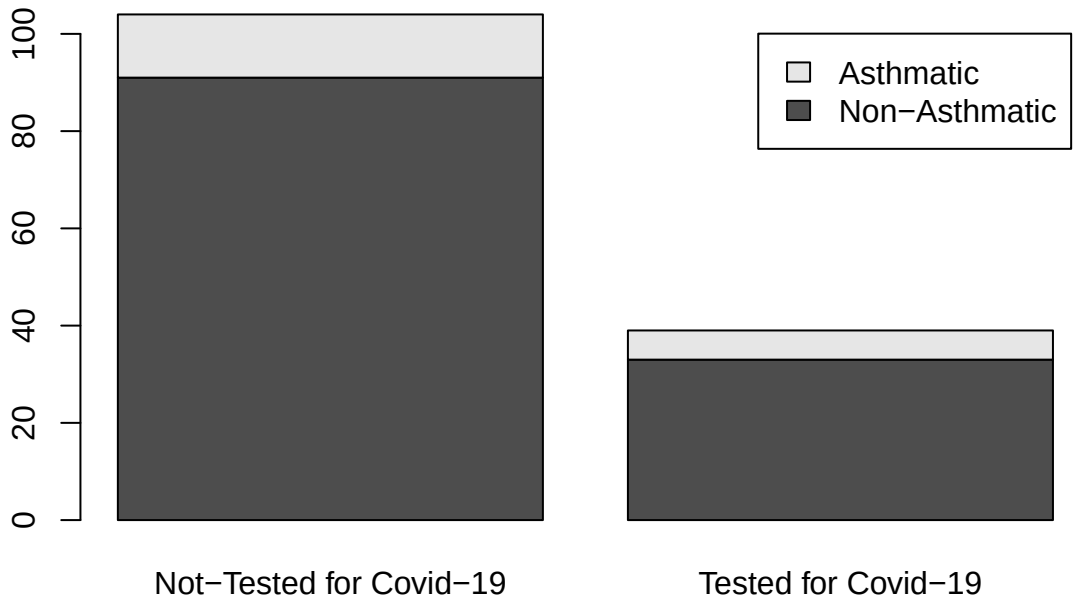
Does having Asthma increase your chances of getting tested for COVID-19? Since, COVID-19 is known to cause respiratory infections and difficulties, we may wish to consider whether participants who have asthma have been more likely to get tested for COVID-19. This will have to be a **test for independence**. Since we are comparing data for two categories and testing the existence of a relationship between them, the *test for independence* is an appropriate test to employ here. Our hypotheses are:

- **Null Hypothesis**: There is no relationship between being Asthmatic or not, and getting tested for COVID-19.
- **Alternative Hypothesis**: There is a relationship between being Asthmatic or not, and getting tested for COVID-19.

In order to properly distinguish between those who have been tested and those who haven't, we classify anyone who has been tested once or more as having been tested and the remaining as having not been tested. We must also make an **assumption** about the independence of the observations in each category which shall be tested.

Since there has been some non-response in the questions, we shall clean our data by dropping any NA values from the sample of the data used for this test.

```
# Cleaning survey_data to remove NA values
cleaned = survey_data %>% na.omit()

covid_tests = cleaned %>% mutate(tested = Covid_Tests > 0)

# Creating contingency table for Asthma and Covid tests
mat = table(covid_tests$Asthma, covid_tests$tested)

# Renaming columns and rows
colnames(mat) = c("Not-Tested for Covid-19", "Tested for Covid-19")
rownames(mat) = c("Non-Asthmatic", "Asthmatic")

# Plotting a bar chart
barplot(mat, legend.text = TRUE, )
```



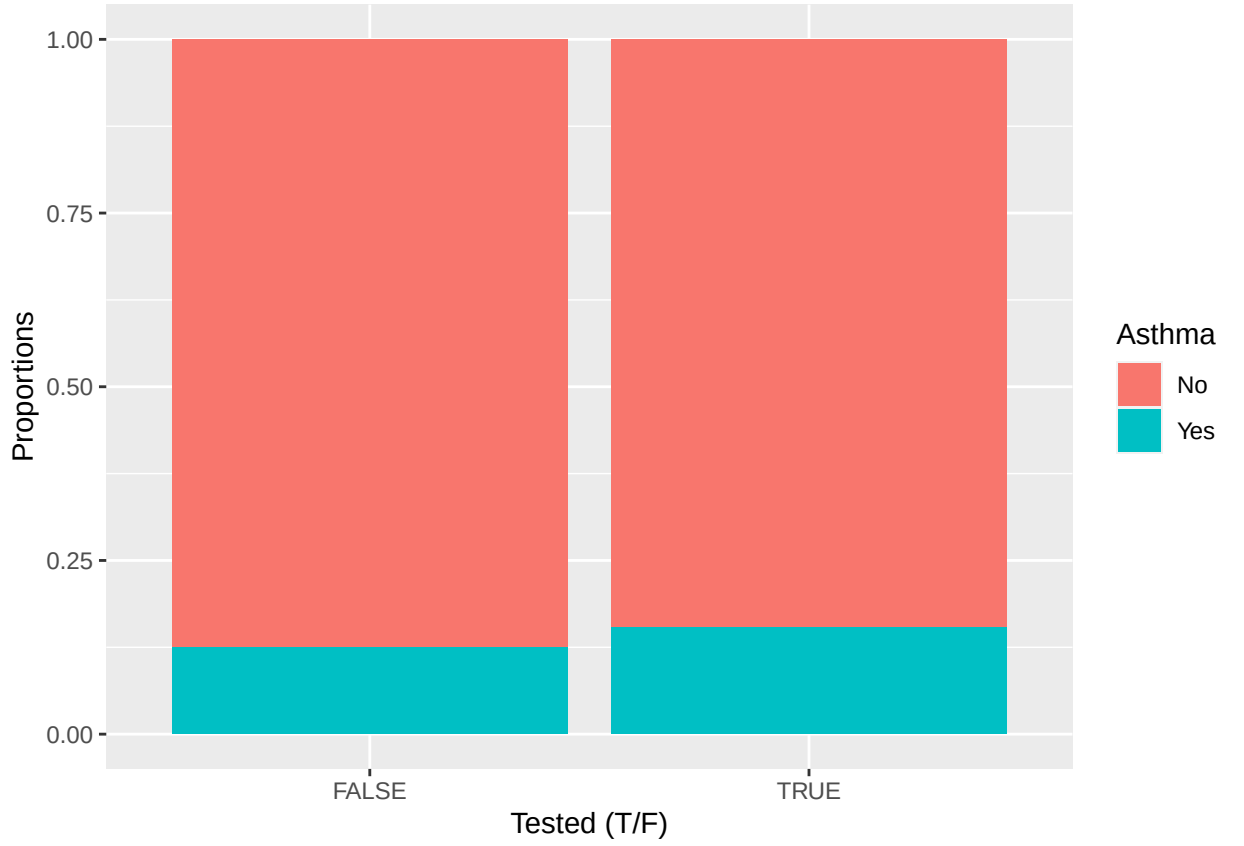

```
r = c = 2
mat_row = apply(mat, 1, sum)  # row totals
mat_col = apply(mat, 2, sum)  # column totals

mat_row.mat = matrix(mat_row, r, c, byrow = FALSE)
mat_col.mat = matrix(mat_col, r, c, byrow = TRUE)

ey.mat = (mat_row.mat * mat_col.mat)/sum(mat)  # expected counts

paste("Independence of Observations Assumption Satisfied:", all(ey.mat >= 5))  # checking assumption
```

```
## [1] "Independence of Observations Assumption Satisfied: TRUE"
```

As can be deduced from the chart above, most participants are both Non-Asthmatics and have not been tested for COVID-19. However it may be interesting to examine the proportions of Asthmatics and Non-Asthmatics in the samples of those who have been tested and those who haven't.

Additionally, we can confirm that our assumption about the independence of observations is satisfied.

```
# Plotting a bar chart
p = ggplot(covid_tests, aes(x = tested, fill = Asthma))
p = p + geom_bar(position = "fill") + xlab("Tested (T/F)") + ylab("Proportions")

p
```



From the second chart above, we see that for both; Asthmatics and Non-Asthmatics, most participants have not been tested for COVID-19. The proportions also appear to be very similar. However, we must still formally test this.

```
t0 = sum((mat - ey.mat)^2/ey.mat)  # computing test-statistic

paste("The test-statistic obtained equals:", round(t0, 3))
```

```
## [1] "The test-statistic obtained equals: 0.205"
```

```
pval = pchisq(t0, df = (r - 1) * (c - 1), lower.tail = FALSE)  # computing p-value

paste("The p-value obtained equals:", round(pval, 3))
```

```
## [1] "The p-value obtained equals: 0.651"
```

```
paste("Reject Null Hypothesis at 5% Significance? :", pval < 0.05)
```

```
## [1] "Reject Null Hypothesis at 5% Significance? : FALSE"
```

Therefore, when performing this test at the 5% significance level, we cannot reject our null hypothesis since the p-value obtained (0.651) is greater than 0.05.

Conclusion: Hence, from our test we conclude that there is no relationship between being Asthmatic or not, and getting tested for COVID-19.

Conclusion

In conclusion, despite the flaws in the data we have gathered, there are nonetheless interesting insights that have been gained about the number of COVID-19 tests taken, the average heights of males and females, and through the investigation of any relationship between if a person is an asthmatic and whether they are more likely to have been tested for COVID-19.

References

1 Komlos, J., & Kim, J. H. (1990). Estimating trends in historical heights. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 23(3), 116-120.

2 "Australian Health Survey: First Results". Australian Bureau of Statistics. 29-10-2012, viewed 21st of September 2020. https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/4338.0main+features212011-13

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

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.