# Computational Mathematics and Statistics
# Individual Assignment

Christian Jordan - 14061768

20 January 2023

## 1

**Interviews with 165 people engaged in a stressful occupation revealed that, 78 were alcoholics, 82 were depressed and 50 were both:**

**1a)** Complete the table.

| | Alcoholic | Not Alcoholic | Total |
|---|---|---|---|
| Depressed | 50 | 32 | 82 |
| Not Depressed | 28 | 55 | 83 |
| Total | 78 | 87 | 165 |

**Table 1:** Completed frequency table.

### 1b) Find the following probabilities that an individual drawn at random:

#### 1b.1. is alcoholic or depressed.

For any two events in sample space $S$, the probability of $A \cup B$ is given in equation (1).

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{1}$$

We can calculate the probabilities of $A$, $B$ and $A \cup B$, from Table 1, where $A$ is alcoholic and $B$ is depressed.

$$P(A) = P(Alcoholic) = \frac{78}{165}$$
$$P(B) = P(Depressed) = \frac{82}{165}$$
$$P(A \cap B) = P(Alcoholic \cap Depressed) = \frac{50}{165}$$

Plugging these values back into equation (1) gives the probability that an individual drawn at random is either alcoholic or depressed.

$$
\begin{aligned}
P(A \cup B) &= \frac{78}{165} + \frac{82}{165} - \frac{50}{165} \\
&= \frac{110}{165} \\
&= \frac{2}{3}
\end{aligned}
$$

**1b.2. is alcoholic given they are depressed**

$$
P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2}
$$

Again from table 1, find the probabilities below,

$$
\begin{aligned}
P(A \cap B) = P(Alcoholic \cap Depressed) &= \frac{50}{165} \\
P(B) = P(Depressed) &= \frac{82}{165}
\end{aligned}
$$

And entering these values back into equation (2) gives the probability that an individual drawn at random is an alcoholic given that they are depressed.

$$
\begin{aligned}
P(A|B) &= \left(\frac{(\frac{50}{165})}{(\frac{82}{165})}\right) \\
&= \frac{50}{82} \\
&= \frac{25}{41}
\end{aligned}
$$

**1b.3. are the events alcoholic and depressed independent?**

We can conclude that two events are independent if the following equations 3, 4 and 5 hold true.

$$
\begin{aligned}
P(A \cap B) &= P(A) \cdot P(B) \qquad (3) \\
P(A|B) &= P(A) \qquad (4) \\
P(B|A) &= P(B) \qquad (5)
\end{aligned}
$$

If we remember from question 1b.1 that, $P(A) = P(Alcoholic) = \frac{78}{165}$ and $P(B) = P(Depressed) = \frac{82}{165}$, and we know that $P(A|B) = \frac{25}{41}$ (from question 1b.2), then from equation 4.

$$
\begin{aligned}
P(A|B) &= P(A) \\
\frac{25}{41} &\neq \frac{78}{165} \qquad (6)
\end{aligned}
$$

Therefore the two events can not be independent and must be dependent.

## 2

**It is known that 2.7% of athletes take performance enhancing drugs. The test for such drugs gives a positive result in 96.2% of cases when drugs are present, but also gives a positive result in 3.5% of cases when drugs are not present. Suppose an athlete is tested at random. Find the probability that,**

**2a they are positive**

Let's start by recognising the probabilities of what we already know,

$$\begin{aligned} P(TakesDrugs) &= 0.027 \\ P(NotTakeDrugs) &= 0.973 \\ P(Positive|TakesDrugs) &= 0.962 \\ P(Negative|TakesDrugs) &= 0.038 \\ P(Positive|NotTakeDrugs) &= 0.035 \\ P(Negative|NotTakeDrugs) &= 0.965 \end{aligned} \tag{7}$$

We can now find the probability of being positive using formula,

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) + .... + P(A|B_n)P(B_n) \tag{8}$$

Where $B_1$ to $B_n$ are the exhaustive events covering all possibilities in a sample space $S$. Therefore, where $A$ = Positive, $B_1$ = Takes Drugs, and $B_2$, Not Take Drugs, find from Equation 8 filling in values from 7.

$$\begin{aligned} &P(A) = P(A|B_1) \cdot P(B_1) + P(A|B_2)P(B_2) \\ &P(A) = (0.962 \cdot 0.027) + (0.035 \cdot 0.973) \\ &P(A) = 0.060029 = 6.0029\% \end{aligned} \tag{9}$$

**2b they took drugs given that they test positive.** Using Bayes' Theorem,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{10}$$

Find that for $A$ = Takes Drugs, and $B$ = Positive,

$$\begin{aligned} P(B|A) &= 0.962 \\ P(A) &= 0.027 \\ P(B) &= 0.060029 \end{aligned}$$

And putting these values into equaton 10,

$$\begin{aligned} P(A|B) &= \frac{0.962 \cdot 0.027}{0.060029} \\ &= 0.43269 = 43.269\% \end{aligned}$$

**2c they didn't take drugs given that they tested negative.**

Similarly to part 2b, using Bayes' Theorem in equation 10 once again, where $A$ = Not Take Drugs and $B$ = Negative.

$$\begin{aligned} P(B) &= 1 - P(Positive) &&= 0.939971 \\ P(A) &= P(NotTakeDrugs) &&= 0.973 \\ P(B|B) &= P(Negative|NotTakeDrugs) &&= 0.965 \end{aligned}$$

Then putting these values again into equation 10, find that,

$$\begin{aligned} P(A|B) &= \frac{0.965 \cdot 0.973}{0.939971} \\ &= 0.99891 = 99.891\% \end{aligned}$$

# 3

**The average price of a 2018 Australian Red Wine bottle from a specific winery is assumed to follow approximately a Normal distribution with mean £35.61 and standard deviation £4.**

**3a Find the probability of a randomly selected bottle is priced:**
**3a.1. less than £36.97**

From the question, we can assume that X approximates to the normal distribution such that,

$$\begin{aligned} X &\sim \mathcal{N}(\mu, \sigma) \\ X &\sim \mathcal{N}(35.61, 4) \end{aligned} \tag{11}$$

Using probability tables for the normal distribution, the value of $Z$ can be found such that,

$$Z = \frac{X - \mu}{\sigma} \tag{12}$$

With rules for probability from distribution tables as follows,

$$P(Z > +z) = \text{(Found directly in tables)} \tag{13}$$
$$P(Z < +z) = 1 - P(Z > +z) \tag{14}$$
$$P(z_1 < Z < z_2) = P(Z > z_1) - P(Z > z_2) \tag{15}$$
$$P(Z < -z) = P(Z > +z) \tag{16}$$
$$P(Z > -1) = 1 - P(Z > +z) \tag{17}$$

Where $X$ is the $\mu$ is the mean 35.61 and $\sigma$ the standard deviation 4. Therefore from equation 12,

$$\begin{aligned} Z &= \frac{36.97 - 35.61}{4} \\ &= 0.34 \end{aligned}$$

For a normal distribution, being continuous and symmetric, from equation 14,

$$\begin{aligned} P(X < 36.97) &= P(Z < 0.34) \\ &= 1 - P(Z > 0.34) \end{aligned}$$

Finding the value of $P(Z > 0.34)$ from the MMU distribution tables,

$$\begin{aligned} P(X < 36.97) &= 1 - P(Z > 0.34) \\ &= 1 - 0.3669 \\ &= 0.6331 = 63.31\% \end{aligned}$$

**3a.2. between £31.61 and £39.61**
Similarly to question 2a.1., find the Z values as follows,

$$Z_1 = \frac{31.61 - 35.61}{4} = -1 \tag{18}$$

$$Z_2 = \frac{39.61 - 35.61}{4} = +1 \tag{19}$$

$$\begin{aligned} P(31.61 < X < 39.61) &= P(X > 31.61) - P(X > 39.61) \\ &= P(Z > -1) - P(Z > 1) \\ &= 1 - P(Z > 1) - P(Z > 1) \\ &= 1 - 2 \cdot P(Z > 1) \\ &= 1 - 2 \cdot 0.1587 \\ &= 0.6828 = 68.28\% \end{aligned}$$

**3b A wine store wants to price their wines in the top 7% or above of the market. Find the minimum price that is required.**

For a wine store to price in the top 7% or above, we must assume that the probability must be 7% or less, that is,

$$P(Z > X) = 0.07 \tag{20}$$

To achieve this, use the inverse probability tables to find value of $Z$. At probability = 0.07, $Z$ value is 1.4758, therefore to find $X$ use equation 12.

$$\begin{aligned} 1.4758 &= \frac{X - 35.61}{4} \\ X &= (1.4758 \cdot 4) + 35.61 \\ &= 41.532 \end{aligned}$$

Round up as the price needs to be above the top 7% and not below to find that price needs to be **£41.52** or higher.

## 4

**An astronomer suggests that on average 0.68 meteorites strike the earth per month. We assume that the number of meteorites that strike the earth per month follows a Poisson distribution. Find the probability:**
Assuming that this follows a Poisson distribution we can write that,

$$\begin{aligned} X &\sim Pois(\mu) \\ X &\sim Pois(0.68) \end{aligned} \tag{21}$$

With the following rules for distribution probabilities from the MMU tables,

$$P(X \leq x) = \text{(Found directly from tables)} \tag{22}$$

$$P(X \geq x) = 1 - P(X \leq x - 1) \tag{23}$$

$$P(X = x) = P(X \leq x) - P(X \leq x - 1) \tag{24}$$

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a - 1) \tag{25}$$

### 4a Of having no meteorite strikes in a month.

From equation 22, $P(X = 0) = P(X \leq 0)$ can be found directly from the tables, so look directly for 0.68 and 0 in the tables to find that,

$$P(X = 0) = 0.5066$$

### 4b It will have 2 or more strikes in a month.

From equation 23,

$$\begin{aligned} P(X \geq 2) &= 1 - P(X \leq 1) \\ &= 1 - 0.8511 \\ &= 0.1489 \end{aligned} \tag{26}$$

### 4c Will have between 2 to 4 strikes (inclusive) in a month.

From equation 25,

$$\begin{aligned} P(2 \leq X \leq 4) &= P(X \leq 4) - P(X \leq 1) \\ &= 0.9993 - 0.8511 \\ &= 0.1482 \end{aligned}$$

### 4d Will have one strike in two months.

For this question, we assume for this Poisson distribution that the probability is independent, so that the probability of 0.68 meteor strikes in 1 month, has double the probability over 2 months. The new distribution is in equation 27.

$$Y \sim Pois(1.36) \tag{27}$$

Now, from equation 24,

$$P(Y = 1) = P(Y \leq 1) - P(Y \leq 0) \tag{28}$$

Since 1.36 does not appear in the distribution tables, we can assume that the probability lies at the probability of 1.35 + $\frac{1}{5}$ of the difference to probability at 1.40.

$$
\begin{aligned}
P(Y \leq 1) &= \frac{(0.6092 - 0.5918)}{5} + 0.5918 \\
&= 0.60572 \\
P(Y \leq 0) &= \frac{(0.2592 - 0.2466)}{5} + 0.2466 \\
&= 0.25668
\end{aligned}
\tag{29}
$$

Now back to equation 28, plug in these values to find the probability of a single meteor strike,

$$
\begin{aligned}
P(Y = 1) &= 0.60572 - 0.25668 \\
&= 0.34904
\end{aligned}
$$

# 5

**The file poll84.csv contains responses from an US exit poll that took place before the 1984 US elections. It contains the following variables:**

- **vote taking the values 0 if the respondent express the intention to vote for Democrats, 1 for Republicans.**
- **income, a variable taking the values lowest, low, moderate, high, highest. ○ age, the age of the respondent.**
- **female, taking the values 0 for males, 1 for a female respondent.**

Begin by loading the data to R, and installing required packages.

```
# Read the data into dataframe poll
poll <- read.csv("~/Desktop/poll84.csv")

# Install and load useful packages
#install.packages("moments")
library(moments)
#install.packages("dplyr")
library(dplyr)
#install.packages("tidyverse")
library(tidyverse)
#install.packages("flextable")
library(flextable)
#install.packages("car")
library(car)
```

### 5a Produce a table of descriptive statistics (include the mean and the standard deviation) of the age for each sex of the responded. Comment on your findings.

Remembering that 0 and 1 correspond to ”Male” and ”Female” respectively, view the following frequency table.

```
# Summary statistics grouped by sex (Part A)
poll %>% group_by(female) %>% summarise(n=n(), min=min(age), max=max(age),
                                        mean=mean(age), sd=sd(age),
                                        median=median(age),
                                        skew=skewness(age))
```

From Table 2, notice that more females (668) responded to the poll (541 males), with a larger range of ages, however the mean respondent age is similar. With regards the standard deviation (sd), male ages are congregated more towards the mean.

| | female | n | min | max | mean | sd | median | skew |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 541 | 17 | 84 | 44.4 | 15.9 | 41 | 0.460 |
| 2 | 1 | 668 | 18 | 92 | 44.9 | 17.1 | 41 | 0.522 |

**Table 2:** Output from R generated frequency table.

**5b Investigate the voting intention per income level:**

**5b.1. by creating a contingency table and comment on your findings.**
Table 3 shows the contingency table for voting intention by income level. From this, it is difficult to assess the proportions of votes per income level for each party since

0: Voted for Democratic.
1: Voted for Republicans.

```
# Reorder income classification by factoring
poll$income = factor(poll$income,
                     levels=c("lowest", "low", "moderate", "high", "highest"),
                     ordered=TRUE)

# Cross table of frequency, vote vs income (part (b))
voteFreq <- table(poll$vote, poll$income)
voteFreq
```

| | lowest | low | moderate | high | highest |
|---|---|---|---|---|---|
| 0 | 75 | 95 | 187 | 132 | 16 |
| 1 | 46 | 82 | 247 | 255 | 74 |

**Table 3:** Contingency table of voting intention by income class.

Table 4 assigns proportions to the contingency table to make it easier to see total proportion of people from each voting category per income class. This makes it clearer that higher proportions of low income vote for Democrats (0) and higher proportion of higher income vote for Republicans.

```
# Get as proportion of total votes by sex
voteProportion <- voteFreq/rowSums(voteFreq)
voteProportion
```

| | lowest | low | moderate | high | highest |
|---|---|---|---|---|---|
| 0 | 0.14851485 | 0.18811881 | 0.37029703 | 0.26138614 | 0.03168317 |
| 1 | 0.06534091 | 0.11647727 | 0.35085227 | 0.36221591 | 0.10511364 |

**Table 4:** Proportion of total votes per voting intention by income class.

**5b.2. by creating an appropriate plot and comment on your findings.** Creating a bar-plot of the proportional data in Table 4 helps to visualise the difference in voting intention by income class as shown in Figure 1. As pointed out in part 5b.1, as income level increase, the probability from the sample of voting for the Democratic Party decrease as Republican probability increases.

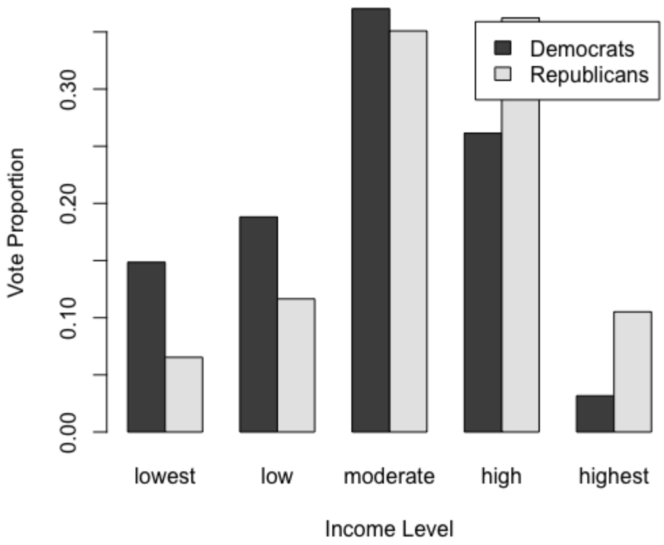

**Figure 1:** Bar plot of voting intent proportions per income class.

**5c Using the appropriate hypothesis test, decide if the voting intention of the poll is compatible with the election result (approximately 58% voted for Republicans, 40% for Democrats). State your hypothesis for your test/tests and offer your interpretation.**

The binomial test is an exact calculation from the binomial distribution, and is preferred for accurate results during statistical testing. Compared to its counterpart, the proportion test (also completed below), the binomial test can be more difficult to calculate at a larger number of observations. When using statistical software such as R, this difference in computational time can become negligible so the binomial test is preferred.

Assumptions:

- The underlying population follows the binomial distribution.
- The sample data is drawn at random from the population.
- Observations are independent.
- There are only two possible outcomes of each trial.

Use the binomial test to test if our sample of data equals the true values recorded. The binomial is a test for only two outcomes, therefore we must run two hypothesis tests, one for Republicans and One for Democrats.

**Republicans**. Does the recorded voters of 58% equal the poll data?

$H_0$ : The population proportion equals the recorded value.

$H_1$ : The population proportion does not equal the recorded value.

```
# Investigate Hypothesis Testing (Part C)
table(poll$vote)
barplot(table(poll$vote), names=c("Democrats","Republicans"),
        xlab="Vote", ylab="Frequency")

summarise(poll, mean=mean(poll$vote), n=n())
```

| 0 | 1 |
|---|---|
| 505 | 704 |

**Table 5:** Total number of voters per category in poll.

```
# Binomial test for Republicans
binom.test(704, 1209, p=0.58)
```

```
Output:
Exact binomial test

data:  704 and 1209
number of successes = 704, number of trials = 1209, p-value = 0.8842
alternative hypothesis: true probability of success is not equal to 0.58
95 percent confidence interval:
 0.5539139 0.6102837
sample estimates:
probability of success
             0.5822994
```

Since P-Value is larger than 0.05, there is not enough evidence to reject the null hypothesis that the population proportion of 0.5822994 equals the recorded value at the 95% significance level. Therefore, accept the null hypothesis and the poll is compatible with the election result for Republicans.

**Democrats**. Does the recorded voters of 40% equal the poll data?

Similarly, repeat for democrats.

$H_0$ : The population proportion equals the recorded value.
$H_1$ : The population proportion does not equal the recorded value.

```
# Binomial test for Democrats
binom.test(505, 1209, p=0.40)
```

```
Output:
Exact binomial test

data:  505 and 1209
number of successes = 505, number of trials = 1209, p-value = 0.2176
alternative hypothesis: true probability of success is not equal to 0.4
95 percent confidence interval:
 0.3897163 0.4460861
sample estimates:
probability of success
             0.4177006
```

Again, the P-Value of 0.2176 is larger than 0.05 and there is not enough evidence to reject the null hypothesis that the poll data equals the recorded data. Therefore, accept the null hypothesis that they are equal at the 95% significance level.

**One Sample Proportion Test**
Similar to the binomial test, the proportion test can be used to statistically test the difference in observed and recorded proportions.

Assumptions:

- The sample data is drawn at random from the population, therefore is unbiased.
- There are only two possible outcomes of each trial.
- Each observation is independent of the others.
- The underlying distribution of the population follows the binomial distribution.

The proportion test is an approximation to the binomial distribution and is preferred at larger

**Republicans** Does the recorded voters of 58% equal the poll data?

$H_0$ : The population proportion equals the recorded value.
$H_1$ : The population proportion does not equal the recorded value.

```
prop.test(704, 1209, p=0.58, alternative=c("two.sided"))
```

```
Output:
1-sample proportions test with continuity correction

data:  704 out of 1209, null probability 0.58
X-squared = 0.017651, df = 1, p-value = 0.8943
alternative hypothesis: true p is not equal to 0.58
95 percent confidence interval:
 0.5538657 0.6102040
sample estimates:
        p
0.5822994
```

The P-Value is larger than 0.05 so there is not sufficient evidence to reject the null hypothesis and the proportions can be considered equal.

**Democrats** Does the recorded voters of 40% equal the poll data?
Similarly, repeat for democrats.

$H_0$ : The population proportion equals the recorded value.

$H_1$ : The population proportion does not equal the recorded value.

```
prop.test(505, 1209, p=0.40, alternative=c("two.sided"))
```

```
Output:
1-sample proportions test with continuity correction

data:  505 out of 1209, null probability 0.4
X-squared = 1.5054, df = 1, p-value = 0.2198
alternative hypothesis: true p is not equal to 0.4
95 percent confidence interval:
 0.3897960 0.4461343
sample estimates:
        p
0.4177006
```

The P-Value is larger than 0.05 so there is not sufficient evidence to reject the null hypothesis and the proportions can be considered equal.

**5d n order to investigate the distribution of the age per income level, create a box-plot and comment on your findings.**

From the box-plot in Figure 2, the mean age per income class appears to decrease as income class increases (lowest-highest). There is also, perhaps surprisingly, a decrease in the range of ages as it increases. It is unclear how this sample may compare to the actual population mean and ranges and may be a cause for further investigation.

```
# Create a box-plot of age per income class
boxplot(poll$age ~ poll$income)
```

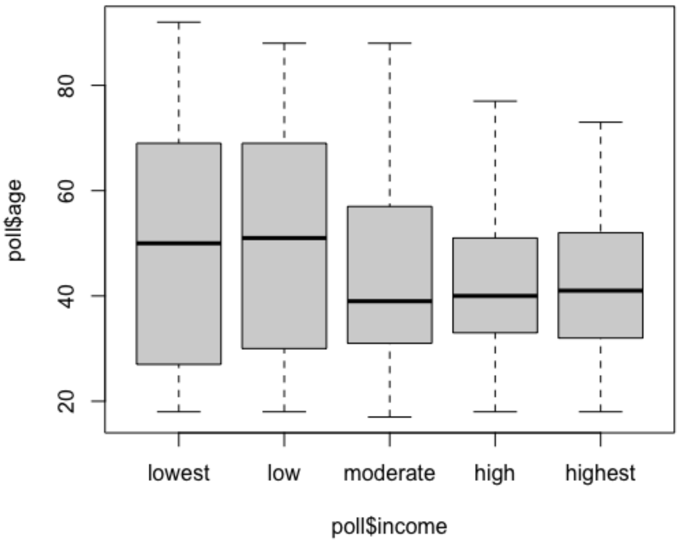

**Figure 2:** Box-plot of age per income class.

### 5e Using ANOVA, investigate if the average age is equal among the income levels.

ANOVA (Analysis of Variance) calculates statistical differences between the means of 2 or more groups.

Assumptions:

1. The responses for each group follows the normal distribution.
2. The distributions have the same variance.
3. The observations are independent.

We can check assumption 1 for normality using the Shapiro-Wilk test, however ANOVA is quite robust to deviations from this assumption.

$H_0$ : The sample population follows the normal distribution.

$H_1$ : The sample population does not follow the normal distribution.

```
# Shapiro-Wilk test for normality
shapiro.test(poll$age)
```

```
Output:
Shapiro-Wilk normality test
```

```
data:  poll$age
W = 0.95379, p-value < 2.2e-16
```

Since the P-value is less than 0.05, there is evidence to accept the null hypothesis that the data is not normally distributed. Since the ANOVA test is robust, we can continue.

Assumption 2 assumes that the groups of data have the same variance. Using the Levene test for same variance, since this test is more robust to deviations from the normal distribution than the Bartlett test.

$$H_0 : \text{The sample population variances are equal.}$$
$$H_1 : \text{The sample population variances are not equal.}$$

```
# Levene Test for variance
leveneTest(age ~ income, data=poll)
```

```
Output:
Levene’s Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)    
group    4  44.249 < 2.2e-16 ***
      1204                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The P-Value is less than 0.05, therefore would suggest accepting the null hypothesis. However, significance code of 0.05 suggests there is homogeneity amongst some groups of data and therefore ANOVA is still valid.

**ANOVA Test**
Performing the ANOVA test in R now that confirmed a valid test.

$$H_0 : \text{The sample group means are equal.}$$
$$H_1 : \text{The sample group means are different.}$$

```
fit <- aov(age ~ income, data=poll)
summary(fit)
```

```
Output:
             Df Sum Sq Mean Sq F value   Pr(>F)    
income        4  13545    3386   12.86 2.94e-10 ***
Residuals   1204 316922     263                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since the P-Value is less than 0.05, there is enough evidence to reject the null hypothesis at the 95% significance level, therefore means of the groups can be considered different.

Following on from the ANOVA test, we can compare group differences individually using Tukey's Test. The assumptions and hypothesis remain the same as the ANOVA test, however it will show statistical difference in means for group pairings per age.

From Table 6, the statistical differences where the null hypothesis can be rejected at the 95% significance level and the mean ages are different are:

- moderate-lowest
- high-lowest
- highest-lowest
- moderate-low
- high-low
- highest-low

With the means of the remaining groups considered equal.

- low-lowest
- high-moderate
- highest-moderate
- highest-high

| | diff | lwr | upr | p adj |
|---|---|---|---|---|
| low-lowest | 2.1599664 | -3.068311 | 7.3882442 | 0.7913788 |
| moderate-lowest | -5.0116731 | -9.568251 | -0.4550953 | 0.0227208 |
| high-lowest | -7.0155252 | -11.632033 | -2.3990178 | 0.0003405 |
| highest-lowest | -7.3697888 | -13.539391 | -1.2001870 | 0.0099555 |
| moderate-low | -7.1716395 | -11.124568 | -3.2187110 | 0.0000081 |
| high-low | -9.1754916 | -13.197355 | -5.1536285 | 0.0000000 |
| highest-low | -9.5297552 | -15.267980 | -3.7915300 | 0.0000615 |
| high-moderate | -2.0038522 | -5.102706 | 1.0950016 | 0.3937235 |
| highest-moderate | -2.3581157 | -7.491803 | 2.7755711 | 0.7189294 |
| highest-high | -0.3542636 | -5.541216 | 4.8326893 | 0.9997294 |

**Table 6:** Tukey's Test table of results.

This can be easily visualised in the Tukey plot in Figure 3, showing the severity of the differences and the bounds within they can be considered the same at the 95% significance level.

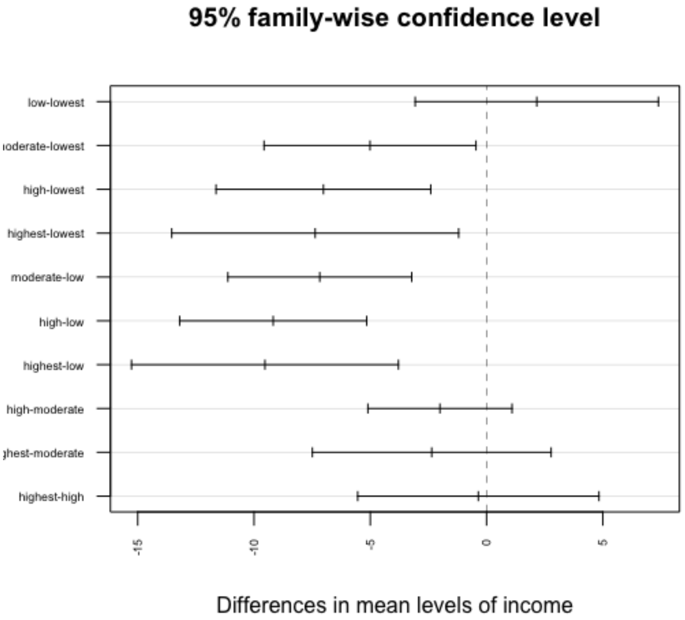

**Figure 3:** Tukey plot of pair-wise group means at 95% confidence level.

## 6 Appendix A - Full R Code

```
[frame=single]
# Read the data into dataframe poll
poll <- read.csv("~/Desktop/poll84.csv")

# Install and load useful packages
#install.packages("moments")
library(moments)
#install.packages("dplyr")
library(dplyr)
#install.packages("tidyverse")
library(tidyverse)
#install.packages("flextable")
library(flextable)
#install.packages("car")
library(car)

# Examine Data
class(poll)
sapply(poll, class)

# Factor sex category
levels(poll$female)
poll$female = factor(poll$female, levels=c(0,1), ordered=FALSE)
levels(poll$female)

# Reorder income classification by factoring
levels(poll$income)
poll$income = factor(poll$income,
                     levels=c("lowest", "low", "moderate", "high", "highest"),
                     ordered=TRUE)
levels(poll$income)

# Summary statistics grouped by sex (Part A)
poll %>% group_by(female) %>% summarise(n=n(), min=min(age), max=max(age),
                                        mean=mean(age), sd=sd(age),
                                        median=median(age),
                                        skew=skewness(age))



# Cross table of frequency, vote vs income (part (b))
voteFreq <- table(poll$vote, poll$income)
voteFreq

# Get as proportion of total votes by sex
voteProportion <- voteFreq/rowSums(voteFreq)
```

```
voteProportion

barplot(voteFreq, xlab="Income Level", ylab="Vote Frequency",
        legend = c("Democrats","Republicans"), beside=TRUE)

barplot(voteProportion, xlab="Income Level", ylab="Vote Proportion",
        legend = c("Democrats","Republicans"), beside=TRUE)

# Investigate Hypothesis Testing (Part C)
table(poll$vote)
barplot(table(poll$vote), names=c("Democrats","Republicans"),
        xlab="Vote", ylab="Frequency")

summarise(poll, mean=mean(poll$vote), n=n())

# Binomial test for Republicans
binom.test(704, 1209, p=0.58)
# alternative straight from dataframe
# binom.test(sum(poll$vote == 1), n=nrow(poll), p=0.58)

# Binomial test for Democrats
binom.test(505, 1209, p=0.40)
#binom.test(sum(poll$vote == 0), n=nrow(poll), p=0.40)

prop.test(704, 1209, p=0.58, alternative=c("two.sided"))
# prop.test(sum(poll$vote == 1), n=nrow(poll), p=0.58)

prop.test(505, 1209, p=0.40, alternative=c("two.sided"))
# prop.test(sum(poll$vote == 0), n=nrow(poll), p=0.40)

# Investigate the distribution of the age per income level (Part D)
boxplot(poll$age ~ poll$income)


# Histogram to check normality in age
hist(poll$age)

# Shapiro-Wilk test for normality
shapiro.test(poll$age)

# Bartlett Test for variance
bartlett.test(age ~ income, data = poll)

# Levene Test for variance
leveneTest(age ~ income, data=poll)

# Fit the ANOVA test
fit <- aov(age ~ income, data=poll)
```

```
summary(fit)

# Tukey’s HSD
TukeyHSD(fit)

# Tukey’s HSD
plot(TukeyHSD ( fit ), cex.axis=0.55, las=2)
```