STAT1003 Assignment

Razeen Wasif u7283652

Question 1.

```
data = heights

# Find the sample size of the dataset
sample_size <- nrow(data); sample_size
# Which are numeric variables
numeric_vars <- select_if(data, is.numeric); numeric_vars
# Which are categoric variables
categoric_vars <- select_if(data, is.factor); categoric_vars

# Draw a Random Sample of 100
set.seed(7283652)
random_sample <- data[sample(nrow(data), size=100, replace=FALSE), ]
```

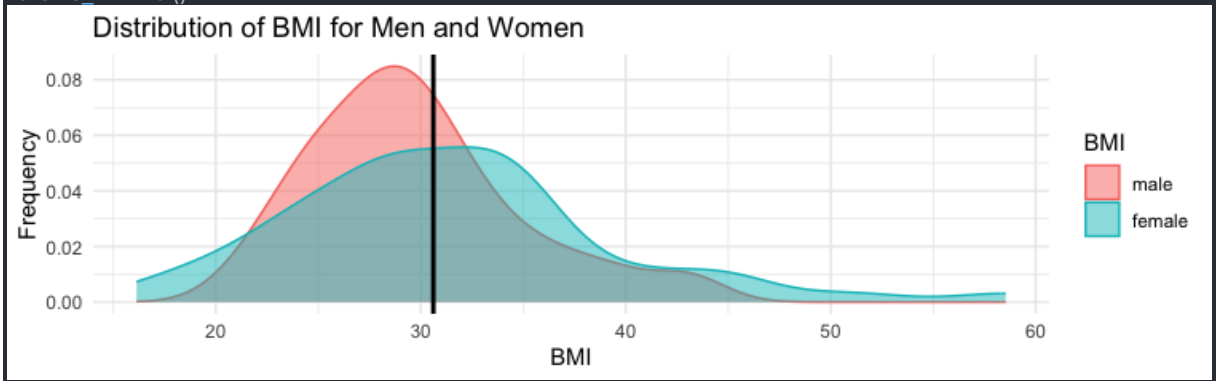
(a) The sample size 7006. The numeric variables are income, weight, height, age, education and afqt. The categoric variables are marital and sex.

```
# Convert weight to kg and height to meters
random_sample$weight_kg <- random_sample$weight * 0.453592
random_sample$height_m <- random_sample$height * 0.0254

# Define BMI, Obese and levels variable
random_sample <- random_sample %>% mutate(BMI = weight_kg / (height_m)^2)
random_sample <- random_sample %>% mutate(Obese = ifelse(BMI >= 30, TRUE, FALSE))
random_sample <- random_sample %>% mutate(Levels = case_when(
 BMI < 25 ~ "normal", BMI >= 25 & BMI < 30 ~ "overweight", BMI >= 30 ~ "Obese"
))
```

(b) The BMI calculation and the Levels assignment were followed according to the information at Centers for Disease Control and Prevention [1].

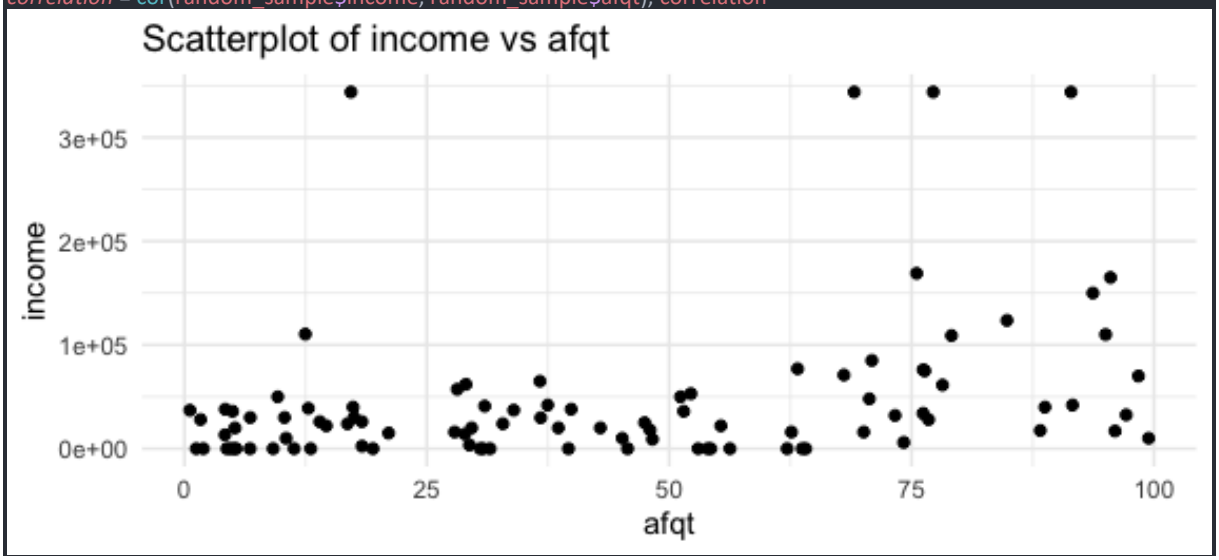
```
# Create a plot that shows the distribution of BMI for men and women
ggplot(random_sample, aes(x=BMI, color=sex, fill=sex)) +
      geom_density(alpha=0.5) +
      geom_vline(xintercept = mean(random_sample$BMI), color="black", linewidth=1) +
      labs(title="Distribution of BMI for Men and Women",
           x="BMI",
           y="Frequency",
           color="BMI",
           fill="BMI") +
 theme_minimal()
```



(c) More males have a BMI in the range of ~22 to ~32 compared to females. More females have a higher BMI (above ~35) than males. This shows that in this sample, more males have what is considered to be a normal BMI.
Both distributions are slightly right skewed with the mean BMI (represented by the black line) being pulled towards the tail. This indicates most people both male and female are in the overweight range in the sample.

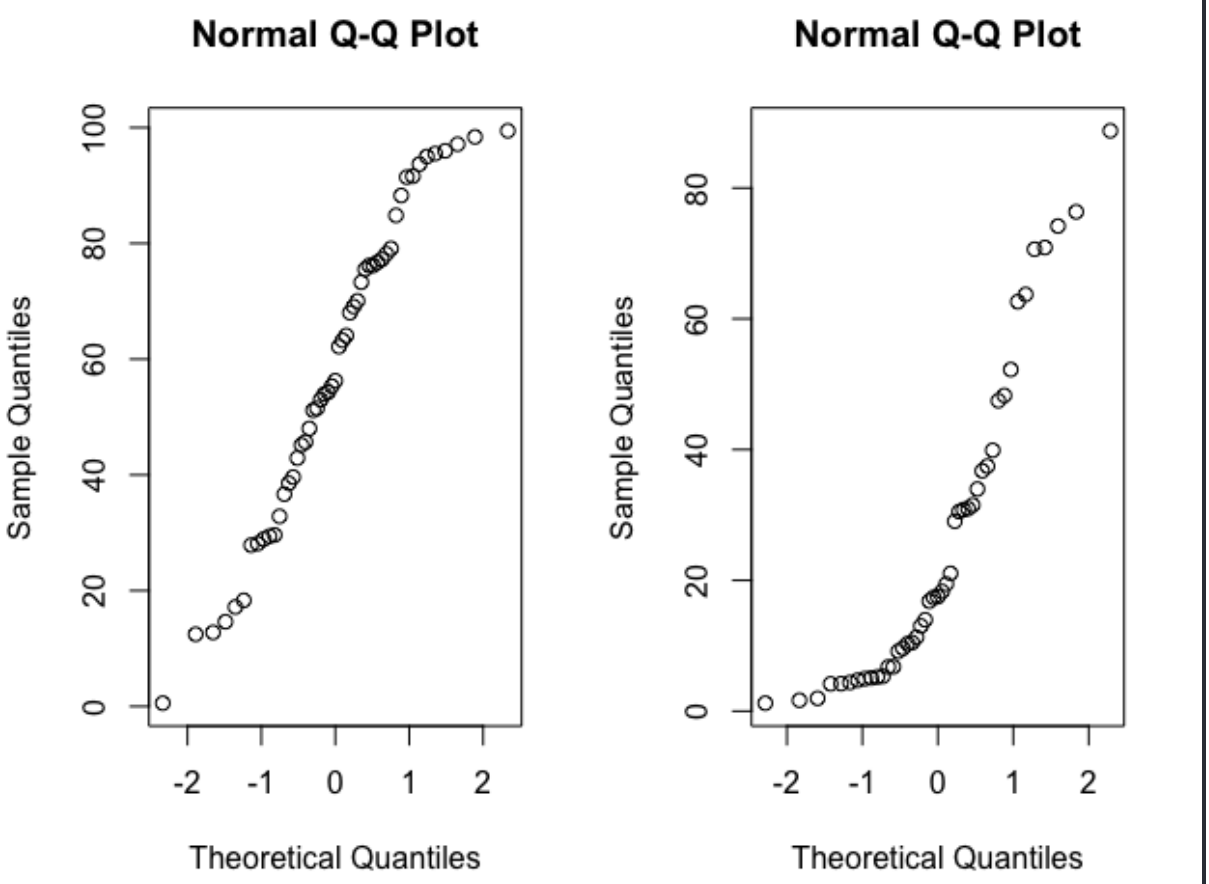
```
# Remove rows with missing values
random_sample <- na.omit(random_sample)

# Show the association between income and afqt (armed forces qualification test score)
ggplot(random_sample, aes(x=afqt, y=income)) +
  geom_point() +
  labs(title="Scatterplot of income vs afqt", x="afqt", y="income") +
  theme_minimal()
# Quantitatively measure association using correlation coefficient
correlation = cor(random_sample$income, random_sample$afqt); correlation
```



(d) From the scatterplot, there does not seem to be any apparent relationship between the two variables although it can be argued that the points show very small positive linearity between the two variables. The correlation coefficient is 0.3391 which indicates a moderately positive linear relationship. The positive coefficient shows that as one variable increases, the other tends to increase as well and the magnitude of the correlation while not very strong, still shows a discernible relationship.

```
# Create a subset of the random sample
subset_df <- random_sample[, c("afqt", "education")]
post_education <- subset_df[subset_df$education >= 13, "afqt"]
no_post_education <- subset_df[subset_df$education < 13, "afqt"]
# Null Hypothesis (H0): μ_post_edu = μ_no_post_edu
# Alternate Hypothesis (HA): μ_post_edu ≠ μ_no_post_edu
# afqt scores are independent between individuals
# Check for normality
par(mfrow = c(1, 2))
qqnorm(post_education$afqt)
qqnorm(no_post_education$afqt)
```

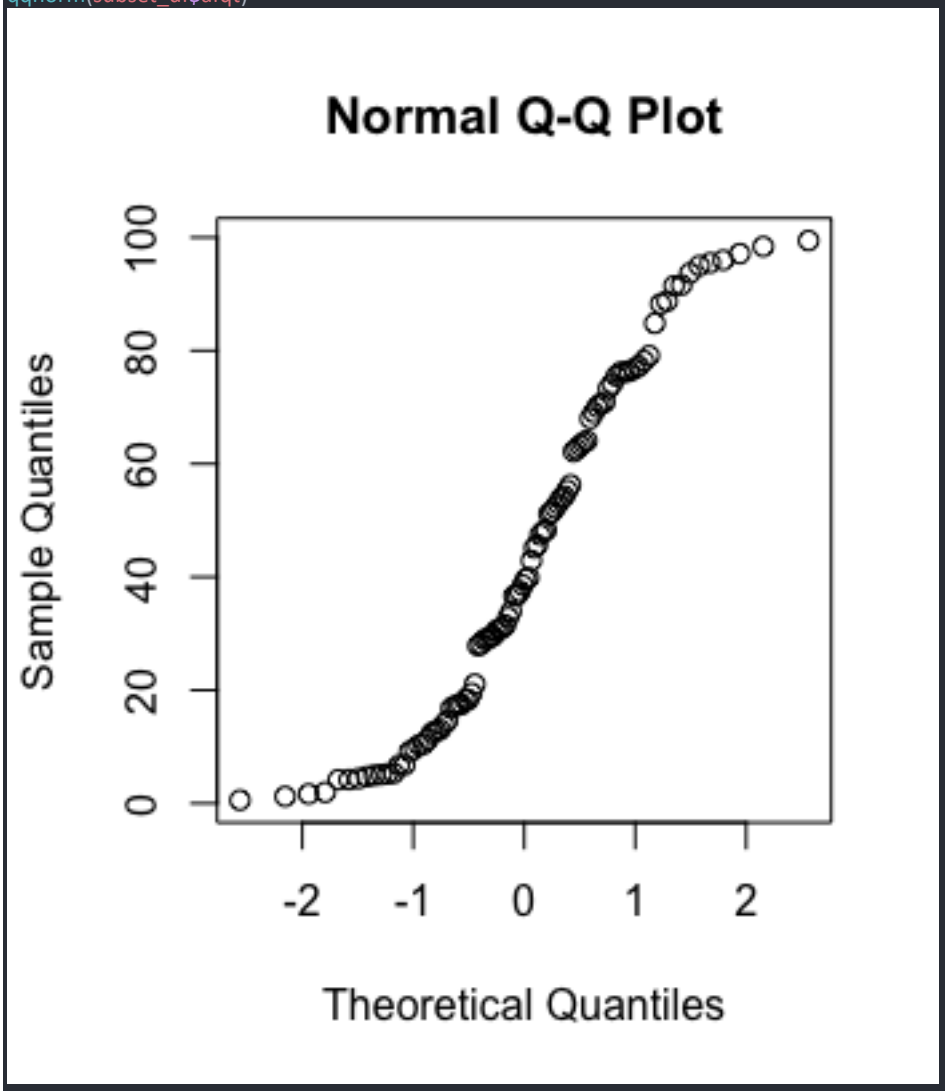



```
# Both post education and no post education approximately follow a straight line
# So we can assume they are both approximately normal
# t.test to check if there is a statistically significant difference b/w the means
test_statistic <- t.test(post_education, no_post_education)
p_value <- test_statistic$p.value
if (p_value < 0.05) {
  print("Reject the null hypothesis")
} else {
  print("Fail to reject the null hypothesis")
}
+
[1] "Reject the null hypothesis"
```

(e) The Null Hypothesis is rejected. Meaning there is a difference in afqt scores depending on a person's education level.

```
# Create subsets
# school-only = 0; some post edu = 1; more post edu = 2
subset_df$education_group <- cut(subset_df$education,
                        breaks = c(-Inf, 12, 15, Inf),
                        labels = c("0", "1", "2"),
                        include.lowest = TRUE)
# Null Hypothesis (H0): There is no difference in the mean afqt scores among the three groups
# Alternate Hypothesis (HA): There is a difference
# Assume independence and distributions seem normal approximately
par(mfrow=c(1,1))
```

```r
qqnorm(subset_df$afqt)
```




```r
# Check if variance is equal among the distributions
levene_test = leveneTest(afqt ~ education_group, data=subset_df)
print(levene_test)
anova_result <- aov(afqt ~ education_group, data=subset_df)
print(summary(anova_result))
```

(f) The distribution is approximately normal and performing the levene test we get a F value of 0.07 and p-value of 0.9322 suggesting there is no significant difference in variances across the education groups. The anova test returns a p-value of 5.54e-09 which is less than 0.05 so we reject the null hypothesis. This means that there is a difference in mean afqt scores across different education groups.

Question 2.

```r
# H_0 : μ = 0.5  # H_A : μ ≠ 0.5
set.seed(7283652)
# Define the parameters
n=30; alpha=0.05; mu=0.5;
# number of simulations to perform
num_simulations = 80000
reject_count = 0
```

```r
# Perform simulations
for (i in 1:num_simulations) {
 # Generate the uniform dist. sample with given range
 sample_data <- runif(n, min=0, max=1)
 # Perform hypothesis test
 test_statistic <- t.test(sample_data, mu=mu, alternative="two.sided")
 # Calculate the p-value for the two-sided test
 p_value = test_statistic$p.value
 # CHeck if null hypothesis should be rejected
 if (p_value < alpha) {
  reject_count <- reject_count + 1
 }
}

# Estimate type 1 error rate
type_1_error_rate = reject_count / num_simulations; print(type_1_error_rate)
```

80,000 simulations were performed, where in each simulation a sample of sample size 30 was generated from a uniform distribution with the range 0 to 1. Following that, the two-sided test statistic was calculated using the t.test function taking in the generated sample data for that simulation and μ of 0.5. From the test statistic, the p-value can be extracted and if it is less than alpha (0.05 in this case) then the null hypothesis can be rejected. The number of times the null hypothesis is rejected is stored in a variable. Using the total number of rejections, the type I error rate can be calculated which equalled 0.049525. This is extremely close to the alpha value of 0.05. In any hypothesis test, we don't want to incorrectly reject H_0 more than 5% of the time which corresponds to alpha of 0.05. So, it can be seen that out of 80000 simulations, the probability of rejecting the null hypothesis incorrectly in any given simulation is approximately 0.05, i.e. the significance level.

References:

[1] *Defining Adult Overweight & Obesity* (2022) *Centers for Disease Control and Prevention*. Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/obesity/basics/adult-defining.html (Accessed: 14 May 2024).

[2] Homogeneity of Variance Test in R: The Definitive Guide, Datanovia. *Available at: https://www.datanovia.com/en/lessons/homogeneity-of-variance-test-in-r/#:~:text=The%20F%2Dtest%20statistic%20can,test()%20. (Accessed: 15 May 2024).*