ADM 2304 - Assignment 2

Renad Gharz

16/05/2022

1. Climate Change

> The dataset **Climate** contains information on the temperatures observed in 1948 and 2018 for a set of 197 randomly sampled locations from the National Oceanic and Atmospheric Administration's records. In the dataset, the variable "dx90_1948" indicates the numbers of days in which the temperature exceeded 90 °F in 1948 for each location. Similarly, "dx90_2018" indicates the number of days in which the temperature exceeded 90 °F in 2018. We would like to determine whether the dataset provides convincing evidence that there were more days in 2018 with temperature exceeding 90 °F than in 1948. To this end, answer the following questions:

1.a

> Are the observations of the number of days in which the temperature exceeded 90 °F in 1948 and 2018 independent? Is there a relationship between these two datasets? Briefly explain your answer.

No, the observations of the number of days in which the temperature exceeded 90°F in 1948 and in 2018 are not independent because the samples are paired. The two samples are paired because the observations in one sample have a special correspondence with the other sample; in this case, it's the stations where the events were observed.

1.b

> Compute the difference in the number of days exceeding 90 °F for each station (number of days in 2018 minus number of days in 1948). Is the distribution of the differences reasonably normal? Justify your answer.

New Column

```r
##Difference between 2018 and 1948
climate['diff'] <-
  climate$dx90_2018 - climate$dx90_1948
head(as.data.frame(climate))
```

```
##       station latitude  longitude dx70_1948 dx70_2018 dx90_1948 dx90_2018 diff
## 1 USC00203823 41.93520  -84.64110       131       147        11        16    5
## 2 USC00276818 44.25800  -71.25250        80        99         1         1    0
## 3 USC00186620 39.41317  -79.40025       143       150         4         1   -3
## 4 USC00331890 40.24030  -81.87100       156       158        18        15   -3
## 5 USC00235987 37.83950  -94.37400       216       175        59        51   -8
## 6 USC00395691 45.56550 -100.44880       138       132        39        18  -21
```

Checking Normality With Histogram

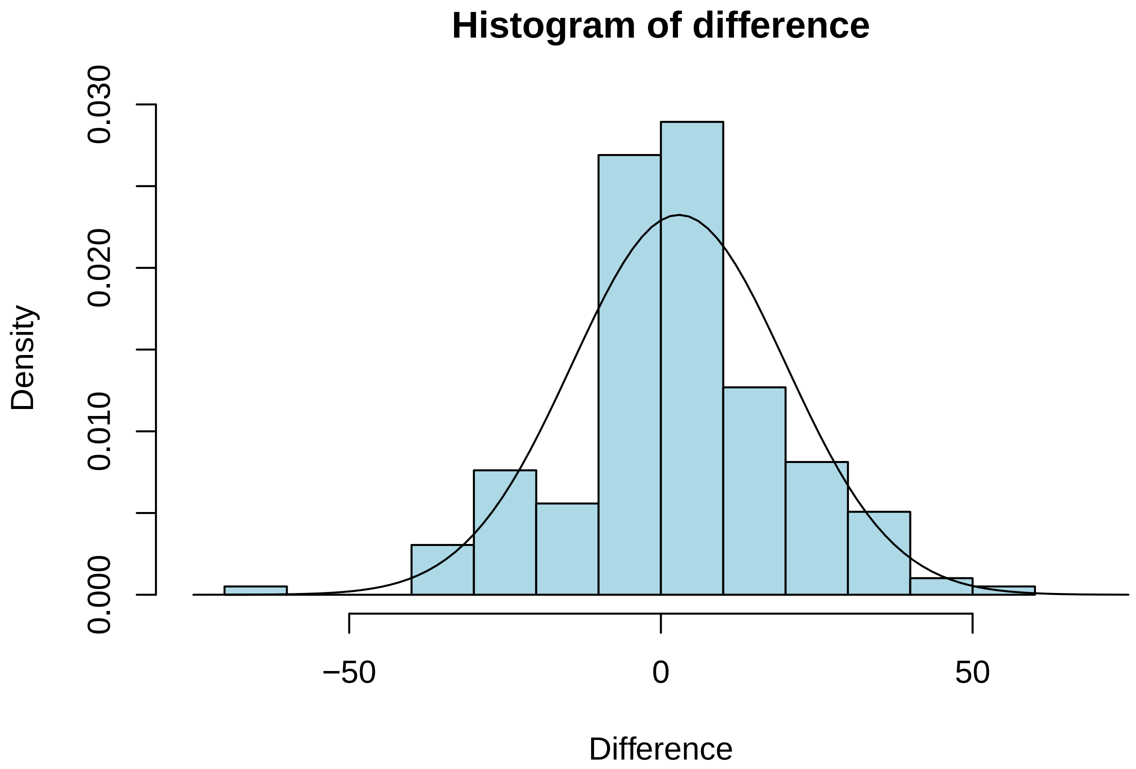
```r
##Mean of difference
climate_diff_mean <- round(
  mean(climate$diff),
  digits = 3)
climate_diff_mean
```

```
## [1] 2.898
```

```r
##Standard deviation of difference
climate_diff_sd <- round(
  sd(climate$diff),
  digits = 3)
climate_diff_sd
```

```
## [1] 17.164
```

```r
##Plotting histogram
climate_histogram <-
  hist(climate$diff,
       main = "Histogram of difference",
       xlab = "Difference",
       xlim = c(-75, 75),
       prob = TRUE,
       col = "light blue")
##Adding normal curve to histogram
curve(dnorm(
  x,
  mean = climate_diff_mean,
  sd = climate_diff_sd),
  add = TRUE)
```



By looking at the histogram above, we can tell that the distribution is approximately normal because the histogram bars roughly align with the normal curve which indicates that the distribution is approximately normal. The bars also look roughly symmetrical with minimal skew.

1.c

> Use software to find the mean and the standard deviation of the differences, and then manually test the hypothesis that there were more days in 2018 with temperature exceeding 90 °F than in 1948. Use a 5% significance level and the p-value approach. Confirm your results using statistical software.

```
##Mean of difference
climate_diff_mean <- round(
  mean(climate$diff),
  digits = 3)
climate_diff_mean
```

```
## [1] 2.898
```

```
##Standard deviation of difference
climate_diff_sd <- round(
  sd(climate$diff),
  digits = 3)
climate_diff_sd
```

```
## [1] 17.164
```

The mean of difference is 2.90 while the standard deviation of difference is 17.16.

Matched Pairs Test Assumptions

- The two samples are dependent by having a special correspondence; they measure the before (1948) and after (2018) temperature data at a certain station. However, the observations between each station are independent.
- We are assuming that the samples are random and represent less than 10% of the population.
- The distribution of differences is approximately normal as proven in part (b).

Defining Hypotheses

$$\begin{cases} H_0 : \mu_d = 0 & \text{observations in 2018 are the same as those in 1948} \\ H_A : \mu_d > 0 & \text{observations in 2018 are greater than those in 1948} \end{cases}$$

Hypothesis test

```r
##Hypothesis test
a2_q1_c_ht <- t.test(
  climate$diff,
  alternative = "greater",
  conf.level = 0.95)
a2_q1_c_ht
```

```
##
##  One Sample t-test
##
## data:  climate$diff
## t = 2.3702, df = 196, p-value = 0.009374
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  0.8774974       Inf
## sample estimates:
## mean of x
##  2.898477
```

```r
##Validating test
a2_q1_c_ht$p.value < 0.05
```

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

Because the p-value is less than the alpha value ($0.009 < 0.05$), we reject the null hypothesis in favour of the alternative. There is sufficient evidence within our sample to prove that the observations in 2018 are greater than those in 1948.

1.d

> Construct manually the corresponding one-sided 95% confidence interval for the mean difference in the number of days exceeding 90 °F and confirm your calculations using statistical software. Does the confidence interval confirm the conclusion obtained from the hypothesis test in part c)? Briefly explain your answer.

```r
a2_q1_c_ci <- round(MeanCI(
  climate$diff,
  conf.level = 0.95,
  sides = "left"),
  digits = 3)
a2_q1_c_ci
```

```
##   mean lwr.ci upr.ci
##  2.898  0.877    Inf
```

The software results confirm our results from the manual calculations above that the lower bound is 0.877 and that we reject the null hypothesis because it is not within $(0.877, \infty)$.

1.e

> Are the required conditions for the hypothesis test and confidence interval above satisfied? Briefly explain your answer.

The required conditions for the hypothesis test and confidence interval calculation for a matched pairs test are met. First, the data are matched, meaning that the come from the same population and are thus dependent or "matched". This is the case here since the data measures the before and after temperatures at each station in the dataset.

Second, the data in the sample are independent of one another. The temperatures at one station (row) will not affect the temperatures at another station, thus the stations are independent of another but the temperatures of 2018 and 1948 for each station are not independent one another.

Third, as proven with the histogram in part (b), the distribution of the differences in temperatures is approximately normal, which is another condition for the matched pairs test.

2. Fuel Economy

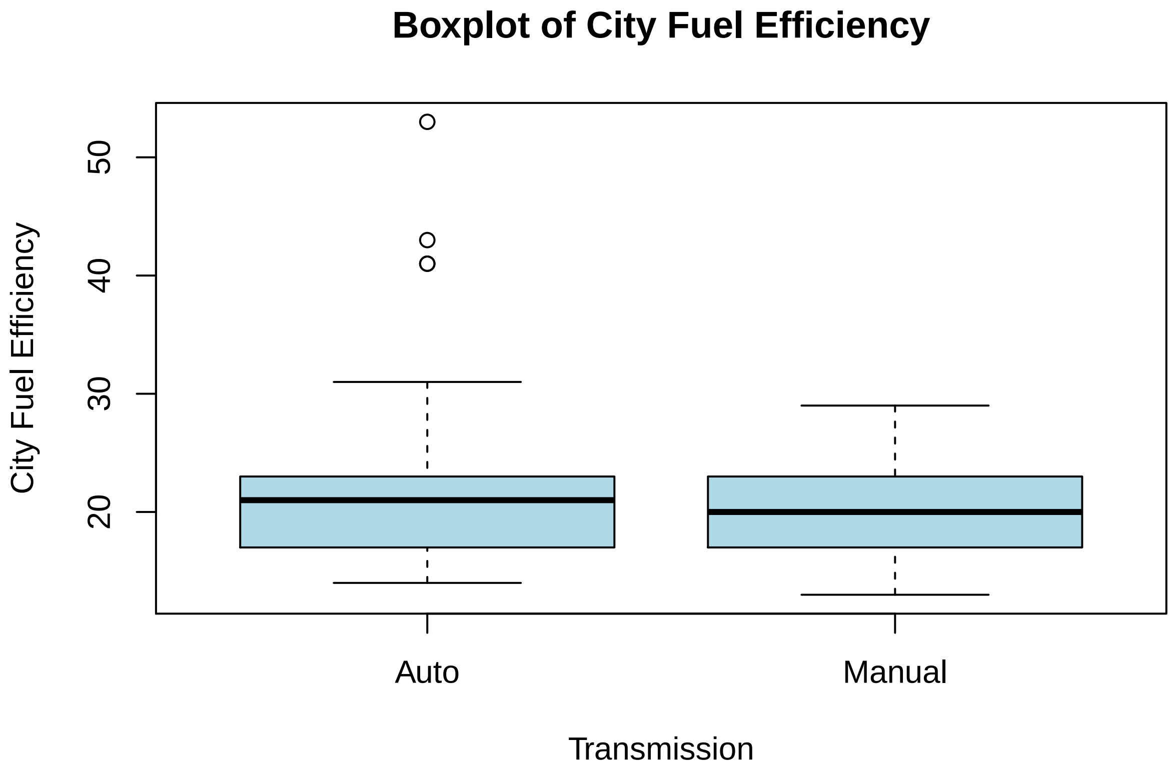
> The dataset Fuel contains information on fuel economy, in miles per gallon, published by the U.S. Department of Energy for a random sample of 100 cars manufactured in 2022.

2.a

> Create a side-by-side boxplot to examine the distribution of city fuel efficiency by transmission type (auto or manual). Comment on the normality of the distributions and on any differences you observe between them. Use software to find the corresponding mean and standard deviation values.

Checking Normality

```
##SBS boxplots
boxplot(
  fuel$`City Fuel Efficiency` ~
    fuel$`Transmission (Auto, Manual)`,
  main = "Boxplot of City Fuel Efficiency",
  xlab = "Transmission",
  ylab = "City Fuel Efficiency",
  col = "light blue")
```



The distribution for city fuel efficiency by automatic transmission is significantly skewed to the right (positively skewed) with several outliers. We can tell it is skewed by looking at how the upper whisker is much longer than the lower whisker of the box.

On the other hand, the distribution for city fuel efficiency by manual transmission is only moderately skewed slightly to the right (positively skewed). We can again tell this by looking at the upper whisker that is slightly longer than the lower whisker.

Means and Standard Deviations

```
##Mean auto transmission
fuel_auto_mean <- mean(as.numeric(
  unlist(
    fuel[fuel$`Transmission (Auto, Manual)`
         =="Auto", 'City Fuel Efficiency'])))
fuel_auto_mean
```

```
## [1] 22
```

```
##Mean manual transmission
fuel_manual_mean <- mean(as.numeric(
  unlist(
    fuel[fuel$`Transmission (Auto, Manual)`
         =="Manual", 'City Fuel Efficiency'])))
fuel_manual_mean
```

```
## [1] 20.34
```

```r
##SD auto transmission
fuel_auto_sd <- round(sd(as.numeric(
  unlist(
    fuel[fuel$`Transmission (Auto, Manual)`
         =="Auto", 'City Fuel Efficiency']))),
  digits = 3)
fuel_auto_sd
```

```
## [1] 7.856
```

```r
##SD manual transmission
fuel_manual_sd <- round(sd(as.numeric(
  unlist(
    fuel[fuel$`Transmission (Auto, Manual)`
         =="Manual", 'City Fuel Efficiency']))),
  digits = 3)
fuel_manual_sd
```

```
## [1] 4.552
```

2.b

Carry out manually an appropriate parametric hypothesis test to determine whether there exists a difference in mean city fuel efficiency between cars with manual and automatic transmissions. Assume that the city fuel efficiencies for manual and automatic cars are approximately normally distributed, the two population variances are unequal, and use a 5% significance level. Use statistical software to obtain the corresponding degrees of freedom and to confirm your results.

Defining Hypotheses

$$\begin{cases} H_0 : \mu_d = 0 & \text{no difference between city fuel efficiency of manual and automatic cars} \\ H_A : \mu_d \neq 0 & \text{difference between city fuel efficiency of manual and automatic cars} \end{cases}$$

Hypothesis Test

```r
##Hypothesis test
a2_q2_b_ht <- t.test(
  fuel$`City Fuel Efficiency`~
    fuel$`Transmission (Auto, Manual)`,
  conf.level = 0.95)
a2_q2_b_ht
```

```
##
##  Welch Two Sample t-test
##
## data:  fuel$`City Fuel Efficiency` by fuel$`Transmission (Auto, Manual)`
## t = 1.2928, df = 78.568, p-value = 0.1999
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8959606  4.2159606
## sample estimates:
##   mean in group Auto mean in group Manual
##                22.00                20.34
```

```r
##Degrees of freedom
a2_q2_b_df <- round(
  a2_q2_b_ht$parameter,
  digits = 0)
a2_q2_b_df
```

```
## df
## 79
```

```r
##Critical value
a2_q2_b_ct <- round(qt(
  0.05/2,
  a2_q2_b_df,
  lower.tail = FALSE),
  digits = 3)
a2_q2_b_ct
```

```
## [1] 1.99
```

Validating Test

```r
##Validating test
a2_q2_b_ht$statistic > a2_q2_b_ct
```

```
##     t
## FALSE
```

The test statistic is not greater than the critical value ($|t_{stat}| > t_{\alpha} \rightarrow |1.2928| \not> 1.992$), so we fail to reject the null hypothesis in favour of the alternative. There is insufficient evidence to prove that there is a difference between the means of city fuel efficiency between manual transmission cars and automatic transmission cars.

2.c

> Confirm the conclusion obtained from the hypothesis test in part b) by manually constructing the corresponding 95% confidence interval for the mean difference in city fuel efficiency between cars with manual and automatic transmissions. Use statistical software to check your results and explain your answer.

```r
##Confidence interval
a2_q2_c_ci <- round(MeanDiffCI(
  fuel$`City Fuel Efficiency`~
    fuel$`Transmission (Auto, Manual)`,
           conf.level = 0.95,
           sides = "two.sided",
           paired = FALSE),
      digits = 2)
a2_q2_c_ci
```

```
## meandiff   lwr.ci   upr.ci
##     1.66    -0.90     4.22
```

The confidence interval is (-0.9, 4.22)

This result confirms the results of the hypothesis test in part (b) that we fail to reject the hypothesis because the null value μ_0 is within the range of the corresponding confidence interval for the hypothesis test:

$$0 \in (-0.9, 4.22)$$

3. Daily Activity & Obesity

> A research team aims to compare the level of physical activity of lean and mildly obese people who don't exercise. The dataset **Activity** includes the number of minutes per day that the subjects of this study spent standing or walking over a 12-day period.

3.a

> Are the samples independent? Assuming that all the required conditions are met, what would it be the appropriate parametric test for this study? Explain your answers.

Yes, the samples are independent because there is no special correspondence between the lean subject and the obese subject. Each sample studies the walking or standing activity for different subjects that have no relation with one another. The activity time of the lean subject does not influence nor is it influenced by the activity time of the obese subject, and vice-versa.

Assuming the samples are normally distributed, the appropriate parametric test in this case would be a 2-sample independent t-test. First, we're not working with proportions, so this eliminates all the proportion tests. Second, there are 2 samples in this case so we can't use 1-sample tests thus eliminating the 1-sample z- and t-tests. Third, the sample size is $n \ngtr 30$ so we cannot use a z-test because the sample is too small. This leaves us with a matched pairs test and a 2-sample independent t-test; and in this case, because the samples are independent as previously mentioned we can't use a matched pairs test, thus, leaving us with a 2-sample independent t-test.

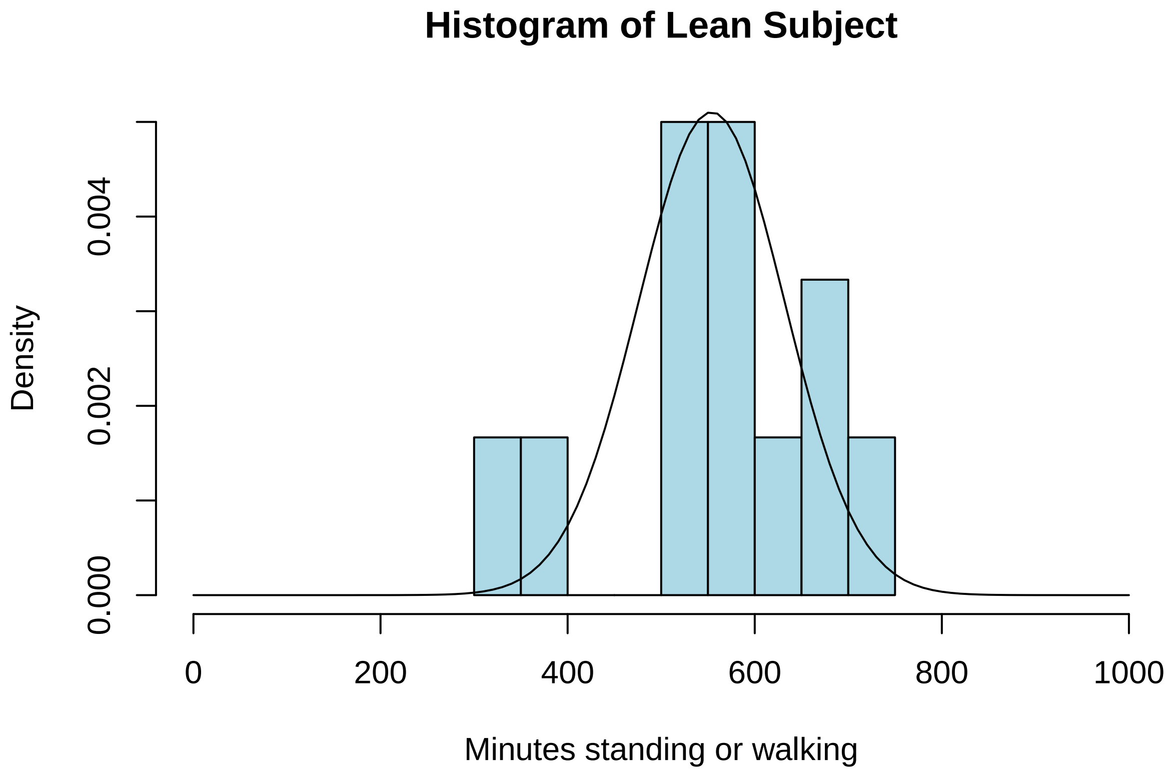
3.b

> Create histograms and boxplots to examine the distribution of the amount time per day spent standing or walking by type of subject. Comment on the suitability of using a parametric test to address the question of interest.

Plotting Histogram

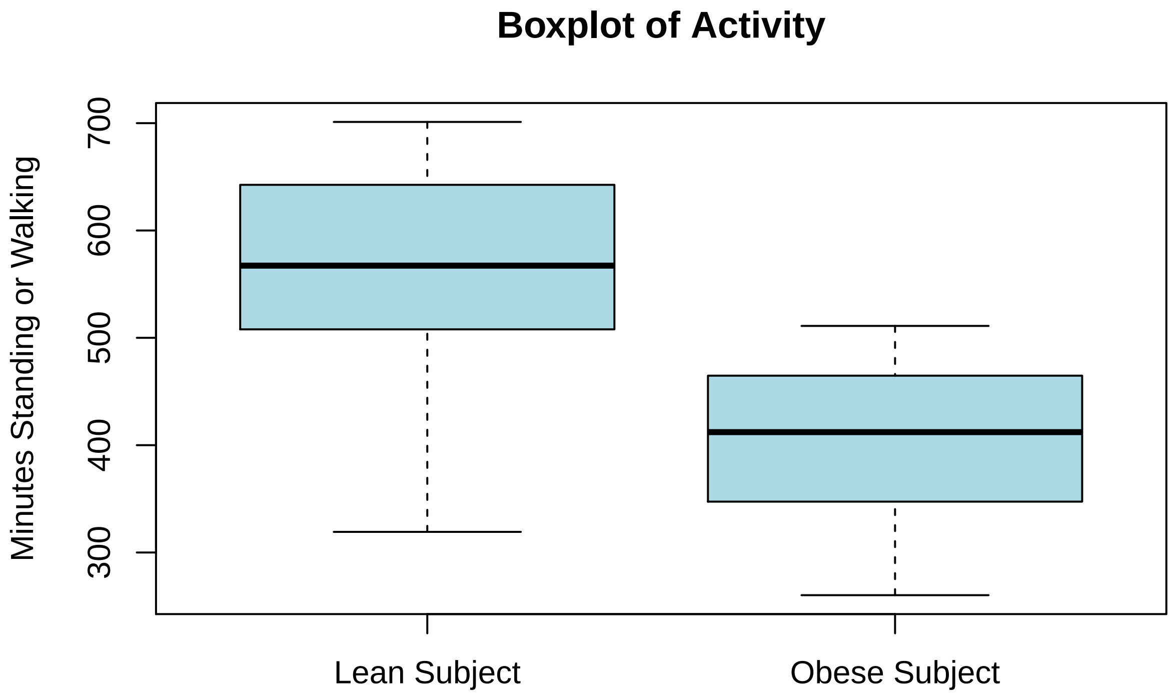
```r
##Plotting histogram (lean)
hist(activity$`Lean Subject`,
       main = "Histogram of Lean Subject",
       xlab = "Minutes standing or walking",
       xlim = c(0, 1000),
       breaks = 10,
       prob = TRUE,
       col = "light blue")

##Mean of lean subject
lean_mean <- mean(na.omit(activity$`Lean Subject`))
##SD of lean subject
lean_sd <- sd(activity$`Obese Subject`)
##Adding normal curve to histogram
curve(dnorm(
  x,
  mean = lean_mean,
  sd = lean_sd),
  add = TRUE)
```



Plotting Boxplot

```
##SBS Boxplots
boxplot(activity,
         main = "Boxplot of Activity",
         ylab = "Minutes Standing or Walking",
         col = "light blue")
```



By looking at the histograms, we can see that the bars are not symmetrical indicating the data is not normal. Furthermore, the bars are not reasonably aligned with the red normal curve further proving the lack of normality. There are also gaps in between the histogram bars showing a lack of normality. This is the case for both the lean and obese subjects' samples.

As for the boxplots, we can see that both samples are left skewed by the significantly longer lower whiskers compared to their right upper counterparts. The median lines are also not centered inside the box further indicating a lack of normality.

Because the boxplot and histograms indicate the two samples are skewed and not normal, it is not appropriate to use the 2-sample independent t-test (or any other parametric test) because the normality assumption is not there. If we use a parametric test in this case, the results will very likely not be accurate to help us answer the question of interest and possibly even lead us to the wrong conclusion; thus, we need to use a non-parametric test to make up for the lack of normality.

3.c

> Regardless of you answer to part b), specify the name of an appropriate non-parametric test for this study. State the corresponding hypotheses in terms of the median activity level per day for lean and mildly obese subjects and test for a difference at the 5% significance level.

In this case, because the data is independent as mentioned in part (a) and we have a large sample case ($n > 10$) we should use the Wilcoxon Rank-Sum Test (or the Mann-Whitney U Test) for $n > 10$.

Defining Hypotheses

$$\begin{cases} H_0 : \theta_1 = \theta_2 & \text{no difference in median activity level per day between lean and obese subjects} \\ H_A : \theta_1 \neq \theta_2 & \text{difference in median activity level per day between lean and obese subjects} \end{cases}$$

Hypothesis Test

```r
##Wilcoxon Rank-Sum Test
a2_q3_ht <- wilcox.test(
  as.numeric(unlist(activity$`Lean Subject`)),
  as.numeric(unlist(activity$`Obese Subject`)),
  alternative = "two.sided",
  conf.level = 0.95,
  paired = FALSE,
  exact = FALSE,
  na.rm = TRUE)
a2_q3_ht
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  as.numeric(unlist(activity$`Lean Subject`)) and as.numeric(unlist(activity$`Obese Subject`))
## W = 146.5, p-value = 0.001425
## alternative hypothesis: true location shift is not equal to 0
```

Validating Test

```r
##Validating test
a2_q3_ht$p.value < 0.05
```

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

Since the p-value is less than the significance level, we reject the null hypothesis in favour of the alternative. Our test results show that there is enough evidence to prove that there is a difference between the median activity time of the lean subject and the obese subject.

4. Management and Engineering Students

> Assume that, at the University of Ottawa, Management and Engineering graduate students must take a similar statistics course during their studies. The university would like to assess the association between the placement of students in the Management and Engineering faculties and their final grade in the course. For this purpose, the university has categorized students' grades in the past five years into six groups of 90-100, 80-90, 70-80, 60-70, 50-60, and less than 50. The table below summarizes the distribution of the grades by faculty. Please note that the values in the table correspond to the number of students in each category.

```
##       Management Engineering
## 90-10         30          45
## 80-90         35          50
## 70-80         20          50
## 60-70         10          40
## 50-60          5          20
## < 50           7           8
```

4.a

> Are the required conditions for a chi-square test of homogeneity (independence) satisfied? Explain your answer.

Yes, the required conditions for a chi-square test of homogeneity are satisfied. First, the sample observations are independent, because the results of one student does not influence nor does it get influenced by another student's, thus the observations are independent of each other. Second, we assume that 320 students (sample size) are less than 10% of the total student population at uOttawa. Third, each scenario (cell) has at least expected 5 cases. Thus; the conditions for a chi-square test of homogeneity are met.

4.b

> Create a table with the corresponding expected counts and having row totals, column totals, and grand total. Round each cell value to two decimal places. Show your computation of the expected count for the number of students with a grade between 90 and 100 in Management.

```r
##Expected counts
chisqe_exp_counts <-
  chisq.test(uo_students)$expected
chisqe_exp_counts
```

```
##       Management Engineering
## 90-10  25.078125   49.921875
## 80-90  28.421875   56.578125
## 70-80  23.406250   46.593750
## 60-70  16.718750   33.281250
## 50-60   8.359375   16.640625
## < 50    5.015625    9.984375
```

4.c

> Perform a chi-square test to assess the association (or independence) between faculty placement and final grade at the 5% level of significance. Show your computation of the contribution to the chi-square statistic of the cell associated with the number of students with a grade between 90 and 100 in Management.

Hypothesis Test

```r
##Hypothesis test
a2_q4_c_ht <- chisq.test(uo_students)
a2_q4_c_ht
```

```
##
##  Pearson's Chi-squared test
##
## data:  uo_students
## X-squared = 11.747, df = 5, p-value = 0.03842
```

```r
##Critical value
a2_q4_c_ct <- round(qchisq(
  0.05,
  a2_q4_c_ht$parameter,
  lower.tail = FALSE),
  digits = 3)
a2_q4_c_ct
```

```
## [1] 11.07
```

Validating Test

```r
##Validating test
round(a2_q4_c_ht$statistic, 3) > a2_q4_c_ct
```

```
## X-squared
##      TRUE
```

The critical value χ^2_α with $\alpha = 0.05, d.f. = 5$ is 11.070. Since the test statistic χ^2_{df} is greater than the critical value χ^2_α, $\chi^2_{df} > \chi^2_\alpha \rightarrow 11.746 > 11.070$, we reject the null hypothesis in favour of the alternative. The variables (faculty and grade) are not independent. There is sufficient evidence to prove that the faculty of the students that take the statistics course influences their final grade in that course.