STAT 210
Applied Statistics and Data Analysis:
Homework 4 - Solution

Due on Oct. 12/2025

You cannot use artificial intelligence tools to solve this homework.

Show complete solutions to get full credit. Writing code is not enough to answer a question. Your comments are more important than the code. Do not write comments in chunks. Label your graphs appropriately

Question 1

A car manufacturing company has developed a new exhaust system designed to reduce CO2 emissions. To evaluate its effectiveness, the company conducted an experiment using 25 identical cars, all driven under similar conditions in city traffic.

With the standard exhaust system, the average CO2 emissions were measured at 191.5 g/km. The emissions data collected using the new exhaust system are stored in the variable `emss1` within the `25Fhw4Q1` file.

Use a significance level of $\alpha = 0.01$ for all statistical tests in this question.

(a) Perform an exploratory analysis to assess whether the assumption of normality is reasonable for this dataset. You are required to produce two plots: First, combine a histogram, a graph of the estimated density and a curve for the normal density with parameters estimated from the sample. Use adequate names for the axes and add a legend. Second, do a quantile plot using the function `qqPlot` in the `car` library with the argument `line` set to `r`. What does this option do? Comment on what you observe in the plots. Do you think that the assumption of normality for the data is valid?

(b) Assuming that the variable `emss1` follows a Gaussian distribution, write down a formula for the lower one-sided confidence interval for the mean at level $(100 - \alpha)\%$ (This interval is bounded above but unbounded below). Calculate this confidence interval for the mean for the case $\alpha = 0.01$, check whether the reference value of 191.5 falls inside or outside and give an interpretation.

(c) What parametric test would be adequate for testing whether the new exhaust system decreases CO2 emissions for the car? State clearly what hypotheses you are testing and which assumptions are needed for the test. Explain why you think they are satisfied. Give a formula for the test statistic and calculate its value. Describe the sampling distribution and explicitly identify the type I and type II errors. Carry out this test and discuss the results.

(d) What non-parametric tests will be adequate for the problem in 1(c)? What assumptions are needed, and why do you think they are satisfied? Perform this test, discuss the results, and compare them with your previous results.

(e) The new exhaust system was also tested in highway driving conditions and the results are stored in the `emss2` in the same file. What parametric test is adequate for testing whether the average emission level in both driving conditions is equal? State clearly what hypotheses you are testing and which assumptions are needed for the test. Explain why you think they are satisfied. Identify explicitly the type I and type II errors. Carry out this test and discuss the results.

Solution

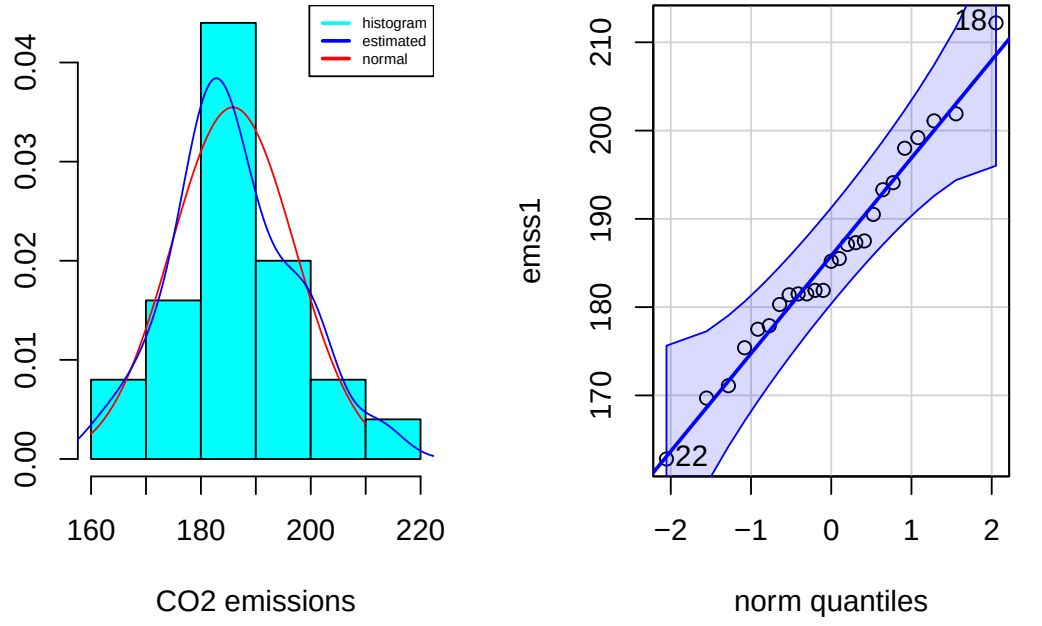
Load the data

```
data_q1 <- read.table('25Fhw4Q1', header = T)
str(data_q1)
```

```
## 'data.frame':    25 obs. of  2 variables:
##  $ emss1: num  186 187 178 170 185 ...
##  $ emss2: num  190 197 184 182 176 ...
```

(a) Exploratory plots:

```
library(MASS)
library(car)
attach(data_q1)
par(mfrow = c(1,2))
truehist(emss1, xlab = 'CO2 emissions')
curve(dnorm(x, mean(emss1),sd(emss1)), 160, 210, add = T, col = 'red')
lines(density(emss1), col = 'blue')
legend('topright', c('histogram','estimated','normal'), col = c('cyan','blue','red'),
       lwd = 2, cex = 0.5)
xyz <- qqPlot(emss1, line = 'r')
```




```
par(mfrow = c(1,1))
```

The first plot shows that the sample is approximately symmetric, and the fit of the estimated density to the reference normal density is very close. The quantile plot is very good and the fit to the reference line is excellent. Both plots support the assumption of normality for the data set. To strengthen this conclusion, we can carry out a Shapiro-Wilk test of goodness-of-fit to the normal distribution:

```
shapiro.test(emss1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  emss1
```

```
## W = 0.98202, p-value = 0.922
```

The p-value is very high and we do not have evidence to reject the null hypothesis of normality.

(b) Define $\bar{X}_n = \frac{1}{n}\sum_1^n X_i$. We saw in the videos that when the population is normal but we have to estimate the standard deviation from the sample, the sampling distribution is

$$\frac{\bar{X}_n - 191.5}{s_n/\sqrt{n}} \sim t_{n-1}$$

where s_n is the sample standard deviation, and n is the sample size, $n = 25$ in this case. Define $t_{n,\alpha}$ to be the quantile for the t distribution with n degrees of freedom, i.e., if T_n is a random variable with t_n distribution,

$$P(T_n \leq t_{n,\alpha}) = \alpha.$$

Using this, we have that

$$P\Big(\frac{\hat{\mu}_n - \mu}{s_n/\sqrt{n}} \leq t_{24,0.01}\Big) = 0.01$$

and therefore

$$P\Big(\frac{\hat{\mu}_n - \mu}{s_n/\sqrt{n}} > t_{24,0.01}\Big) = 0.99$$

Multiplying by $s_n/\sqrt{n}$ inside the probability, we get

$$P\Big(\hat{\mu}_n - \mu > \frac{s_n t_{24,0.01}}{\sqrt{n}}\Big) = 0.99$$

And from this, we get that

$$P\Big(\mu < \hat{\mu}_n - \frac{s_n t_{24,0.01}}{\sqrt{n}}\Big) = 0.99$$

We now proceed to evaluate the terms in the formula

```
n <- 25
error <- qt(0.01,df=n-1)*sd(emss1,)/sqrt(n)
mean(emss1,) - error
```

```
## [1] 191.436
```

We observe that the confidence interval does not include the reference value of 191.5. This implies that, based on this sample, a hypothesis test of the null hypothesis that the mean CO2 emission is 191.5 against the alternative that it is lower would lead to the rejection of the null hypothesis.

(c) We want to test the following hypothesis

$$H_0 : \mu = 191.5 \qquad vs. \qquad H_1 : \mu < 191.5$$

If we can assume that the distribution of CO2 emissions is normal, we can use the one-sample t-test. The quantile plot above seems to support this assumption.

The test statistic is given by

$$\hat{t} = \frac{\hat{\mu}_n - \mu}{s_n/\sqrt{n}}$$

which has a t_{24} (sampling) distribution. The value for this statistic is

```
(tstat <- (mean(emss1) -191.5)/(sd(emss1)/sqrt(25)))
```

```
## [1] -2.520599
```

A type I error would be to conclude that the new exhaust system reduces CO2 emissions when, in fact, it does not. A type II error would be to conclude that the new system does not improves CO2 emisisons when, in fact, it does.

To do the test in R, we use the `t.test` function:

```
t.test(emss1, mu = 191.5, conf.level = .99, alternative = 'l')
```

```
##
##  One Sample t-test
##
## data:  emss1
## t = -2.5206, df = 24, p-value = 0.009386
## alternative hypothesis: true mean is less than 191.5
## 99 percent confidence interval:
##     -Inf 191.436
## sample estimates:
## mean of x
##   185.832
```

Observe that the value for the test statistic and for the confidence interval coincide with our previous results. The p-value is below the 1% significance level, and we reject the null hypothesis: at the 99% confidence level, we conclude that the new exhaust reduces the CO2 emissions of the cars.

(d) We can use Wilcoxon's test, which assumes that the distribution is continuous and symmetric with respect to the mean value. Since we previously accepted that the distribution was normal, this assumptions seems to be valid.

For the test in (c) we have

```
wilcox.test(emss1, mu = 191.5,  alternative = 'l')
```

```
## Warning in wilcox.test.default(emss1, mu = 191.5, alternative = "l"): cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  emss1
## V = 76, p-value = 0.01032
## alternative hypothesis: true location is less than 191.5
```

The p-value now is slightly above 0.01 and we do not reject the null hypothesis that the new systems does not improve CO2 emissions.

However, since the assumptions for the t-test seems to be validated by the plots, we would prefer the parametric test in these conditions.

(e) We now have two samples, one for city driving and one for highway. We want to test whether the average CO2 emissions are the same. We are testing

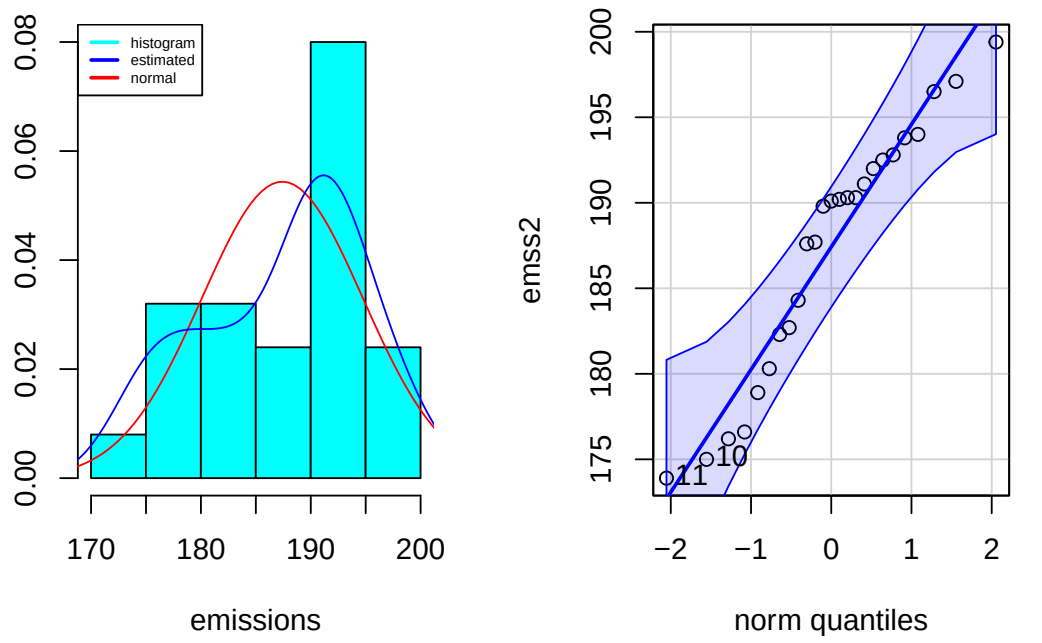
$$H_0 : \mu_1 = \mu_2 \qquad vs. \qquad H_1 : \mu_1 \neq \mu_2$$

where μ_1 is the average for the first (city) sample while μ_2 is the average for the second.

We need to check the assumption of normality for the second sample

```
par(mfrow = c(1,2))
truehist(emss2, h=5, xlab = 'emissions')
curve(dnorm(x, mean(emss2),sd(emss2)), 160, 210, add = T, col = 'red')
lines(density(emss2), col = 'blue')
```

```
legend('topleft', c('histogram','estimated','normal'), col = c('cyan','blue','red'),
       lwd = 2, cex = 0.5)
xyz <- qqPlot(emss2, line = 'r')
```




```
par(mfrow = c(1,1))
```

In this case the first plot is not very good. The sample does not look symmetric, and the estimated density is not close to the reference normal density. However, the quantile plot on the right is very good, with a good fit of the points to the refernce line. This plot supports the assumption of normality for this data set. To strengthen this conclusion, we can carry out a Shapiro-Wilk test of goodness-of-fit to the normal distribution:

```
shapiro.test(emss2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  emss2
## W = 0.93837, p-value = 0.1358
```

The p-value is above the usual levels and we do not have evidence to reject the null hypothesis of normality.

We now do the two-sample t test to compare the means:

```
t.test(emss1, emss2)
```

```
##
##  Welch Two Sample t-test
##
## data:  emss1 and emss2
## t = -0.58994, df = 41.301, p-value = 0.5584
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.00534  3.83734
## sample estimates:
## mean of x mean of y
##   185.832   187.416
```

The p-value is large and we do not have evidence to reject the null hypothesis that the means are equal.

Question 2

A parmaceutical company wants to test a new weight reduction drug on rats. The experiment uses 15 rats which were weighted before and after receiving the drug. The results are stored in the file `25Fhw4Q2`. The column `wt1` holds the measurements for the weight at the beginning of the experiment and `wt2` has the values at the end.

(a) Use graphical tools to compare the weights before and after taking the drug and comment on what you observe.

(b) You want to test whether the drug effectively reduces weight using this data. What hypotheses do you want to test? What parametric test or tests could be appropriate here? What are the assumptions? Why do you think they are satisfied in this case? Identify the type I and type II errors. Carry out the test(s) and discuss your results.

(c) The pharmaceutical company is only interested in producing this drug if the reduction in weight for the rats is more than 2 g. To simplify the test, suppose you want to test that the reduction in weight is 2 g versus the alternative that it is more. How would you carry out this test with the data that you have? What assumptions are needed? Are they justified in this case? Carry out the test or tests and comment on your results.

(d) What non-parametric tests will be adequate for the problems in 2(b) and 2(c)? What assumptions are needed, and why do you think they are satisfied? Perform these tests, discuss the results, and compare them with your previous results.

Solution

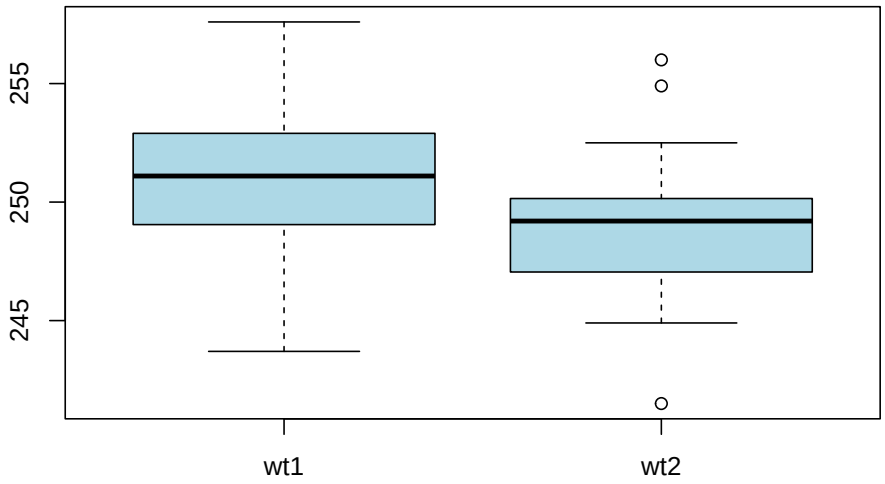
Read the data

```
data_q2 <- read.table('25Fhw4Q2', header = T)
str(data_q2)
```

```
## 'data.frame':    15 obs. of  2 variables:
##  $ wt1: num  249 251 249 258 248 ...
##  $ wt2: num  248 249 247 252 247 ...
```

(a) We do boxplots for the weights before and after

```
boxplot(data_q2, col = 'lightblue')
```



We see that the values after taking the drug are lower than the values before. The two boxes have similar sizes. There are three outliers in the `wt2` plot.

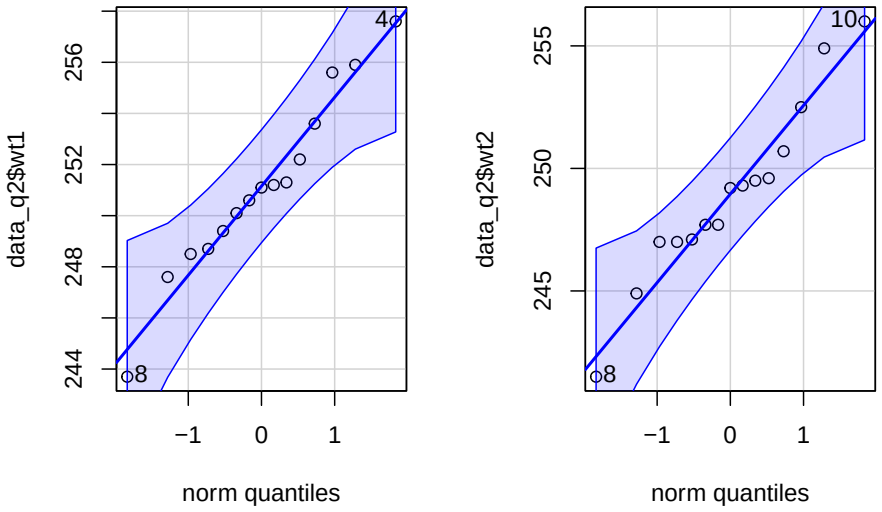
(b) The test we are interested in is

$$H_0 : \mu_1 = \mu_2 \qquad vs. \qquad \mu_1 > \mu_2$$

where μ_{C1} and μ_{C2} represent the average weight before and after taking the drug. The null hypothesis is that there is no effect, while the alternative is that the treatment reduces weight.

Since the measurements are made on the same subjects, before and after treatment, we have paired data. Therefore, if we want to use a parametric test we should use a t-test for paired data. This test assumes that the population distribution is normal, and that the measurements are correlated. The latter assumption is verified because the measurements are paired. To check normality we use a quantile plot

```
par(mfrow = c(1,2))
xyz <- qqPlot(data_q2$wt1, line = 'r')
xyz <- qqPlot(data_q2$wt2, line = 'r')
```




```
par(mfrow = c(1,1))
```

The quantile plots look good and support the assumption of normality.

We do the test with the following command. Observe that we want a one-sided test since we want to determine if the treatment reduces the cholesterol level.

```
t.test(data_q2$wt1, data_q2$wt2, paired = T, alternative = 'g')
```

```
##
##  Paired t-test
##
## data:  data_q2$wt1 and data_q2$wt2
## t = 5.684, df = 14, p-value = 2.821e-05
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  1.495283      Inf
## sample estimates:
## mean difference
##        2.166667
```

The p-value is small and we reject the null hypothesis of no effect. Observe that if we don't do a paired test, the conclusion is different:

```r
t.test(data_q2$wt1, data_q2$wt2, alternative = 'g')
```

```
##
##  Welch Two Sample t-test
##
## data:  data_q2$wt1 and data_q2$wt2
## t = 1.6447, df = 27.974, p-value = 0.05561
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.07435466         Inf
## sample estimates:
## mean of x mean of y
##  251.1400  248.9733
```

Let's look at the difference between the weight before and after treatment, and let's call this variable dif:

$$\text{dif} = \text{weight(before) - weight(after)}.$$

We want to test whether dif is equal to 2 versus the alternative that it is bigger:

$$H_0 : \text{dif} = 2 \qquad vs. \qquad H_A : \text{dif} > 2$$

We start by calculating the difference between the cholesterol levels before and after the treatment:

```r
dif <- data_q2$wt1 - data_q2$wt2
```

The test is a one-sample t-test:

```r
t.test(dif, mu = 2, alternative = 'greater')
```

```
##
##  One Sample t-test
##
## data:  dif
## t = 0.43723, df = 14, p-value = 0.3343
## alternative hypothesis: true mean is greater than 2
## 95 percent confidence interval:
##  1.495283      Inf
## sample estimates:
## mean of x
##  2.166667
```

The p-value is large and we cannot reject the null hypothesis. The reduction is not bigger than 2 grams.