STAT 210
Applied Statistics and Data Analysis:
Homework 4

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.

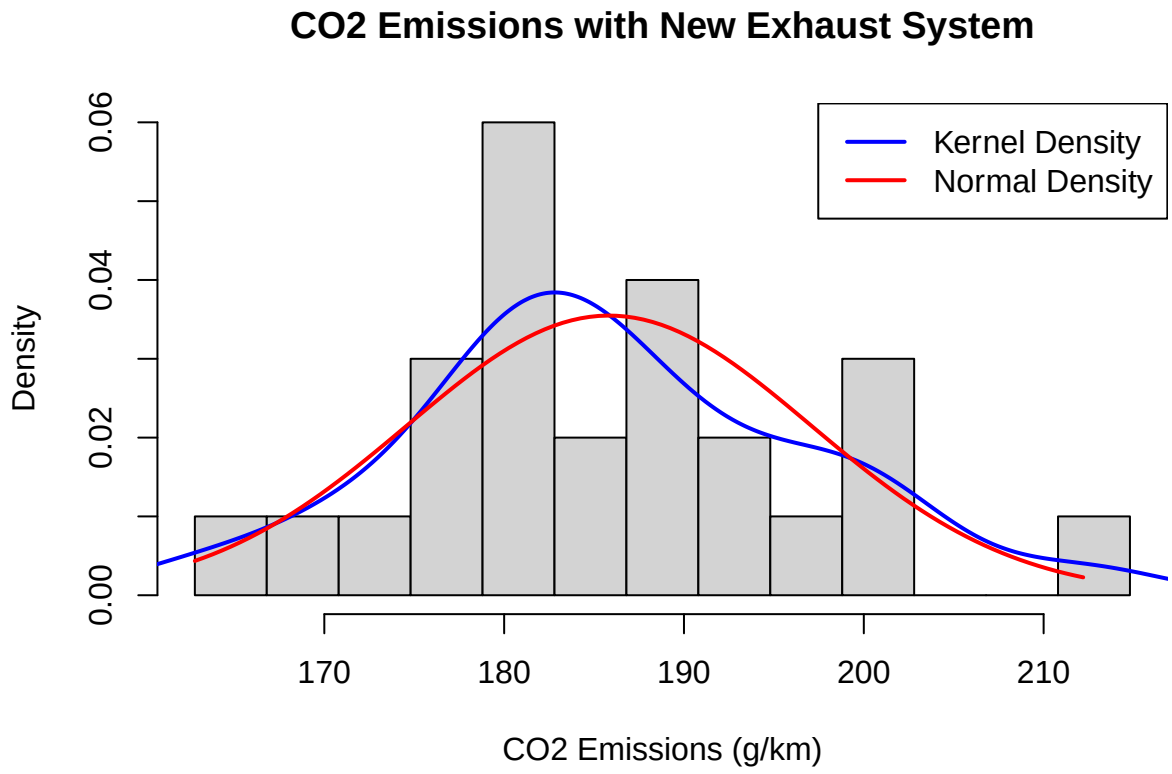
Use a significance level of $\alpha = 0.01$ for all statistical tests in this question.

(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?

We first show the data in the form of a histogram and add in an estimated density curve and a normal density curve to compare.

```
data <- read.table("25Fhw4Q1")

mean_emss1 <- mean(data$emss1)
sd_emss1 <- sd(data$emss1)

#creates a histogram of the data in emss1. probability = true normalizes the y axis so the total value
hist(data$emss1,
  probability = TRUE,
  main = "CO2 Emissions with New Exhaust System",
  xlab = "CO2 Emissions (g/km)",
  ylab = "Density",
  breaks = seq(min(data$emss1), max(data$emss1) + 4, by = 4)
)
```

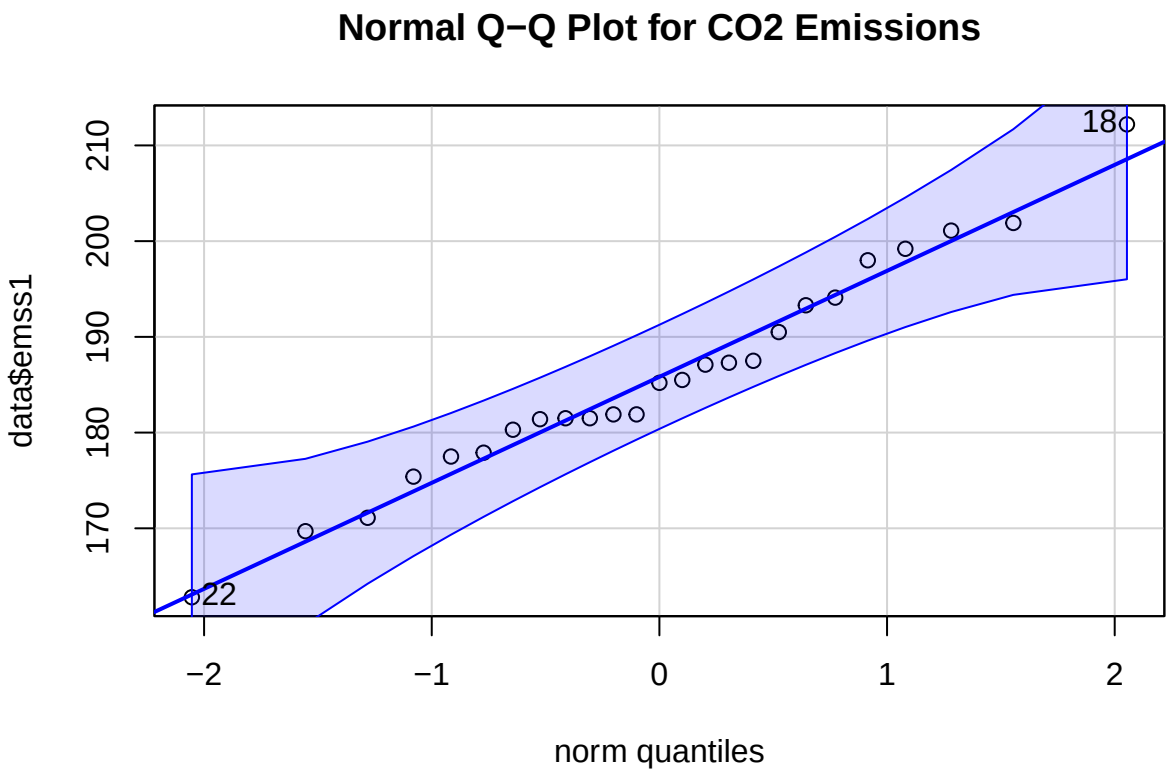
```r
#draw the line of estimated density
lines(density(data$emss1), col = "blue", lwd = 2)

#create a normal density curve
x_vals <- seq(min(data$emss1), max(data$emss1), length = 100)
lines(x_vals, dnorm(x_vals, mean = mean_emss1, sd = sd_emss1),
      col = "red", lwd = 2)


#add in a legend for the two lines
legend("topright",
  legend = c("Kernel Density", "Normal Density"),
  col = c("blue", "red"),
  lwd = 2)
```




```r
#create a Q-Q plot of the data/
library(car)
```

```
## Loading required package: carData
```

```r
qqPlot(data$emss1, line = "r", main = "Normal Q-Q Plot for CO2 Emissions")
```




```
## [1] 18 22
```

The argument line = r makes it a robust regression line. which means it becomes more resistant to outlier data. this gives a better visualisation to check normality Comparing the histogram and estimated density with the normal curve. there is some difference but normality is possible We then did the Q-Q test with the points lying close to the normality line. thus the assumption that the data is normally distributed 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.

```
alpha <- 0.01
n <- length(data$emss1)
t_critical <- qt(1 - alpha, df = n - 1)

upper_bound <- mean_emss1 + t_critical * sd_emss1 / sqrt(n)
c(-Inf, upper_bound)
```

```
## [1]    -Inf 191.436
```

to get the lower one-sided 100 - alpha confidence interval for the mean the region would be mean + t-critical value * sd / sqrt(n) applied above we found the region to be from - infinity to 191.436, and the reference value of 191.5 doesn't lie in our 99% confidence interval. that means we have strong evidence pointing that the mean of the new system is less than 191.5 g/km

(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.

e use a one-sample left tailed t-test with the hypothesis mu = 191.5 g/km and the alternative hypothesis mu < 191.5 g/km the assumptions we make are the following: Independence: each sample is measured independently since the runs shouldn't affect each other Identical conditions (low variance): assumed since we were told identical cars were used under similar conditions in city traffic. approximately normally distributed: this was further explored in part A where we used the histogram and Q-Q plot to check normality and we had strong evidence of normality and given that the sample n is small (25), the t-test is most appropriate.

the test statistic formula is t = (mean - mu) / (sd / sqrt(n)) with mu being our hypothesis (191.5).

```
(mean_emss1 - 191.5) / (sd_emss1 / sqrt(25));
```

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

and we reject the hypothesis if t <= -2.5206

Type 1 error (False Positive): concluding that the new system reduces CO2 (rejecting hypothesis) even though mu = 191.5.
Type 2 error (False Negative): failing to detect reduction (accepting hypothesis) even though mu < 191.5

```
t.test(data$emss1, mu = 191.5, alternative = "less", conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  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
```

(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.

The Wilcoxon signed rank test could be thought of as the non-parametric version of the one-sample t-test. it is used to check if the median of the population differs from a specified value (the given 191.5). Assumptions for this test are similar, but the most important difference is it does not require approximate normality. Independence: samples are measured independently (satisfied as given). Continuous distribution: the measured data is continuous. Symmetry: the data should be symmetrical around the median (shown in the histogram and qq plot in part A).

therefore we choose a H0 to be = 191.5, and H1 to be < 191.5

```r
wilcox.test(data$emss1,
            mu = 191.5,
            alternative = "less",
            conf.level = 0.99)
```

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

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

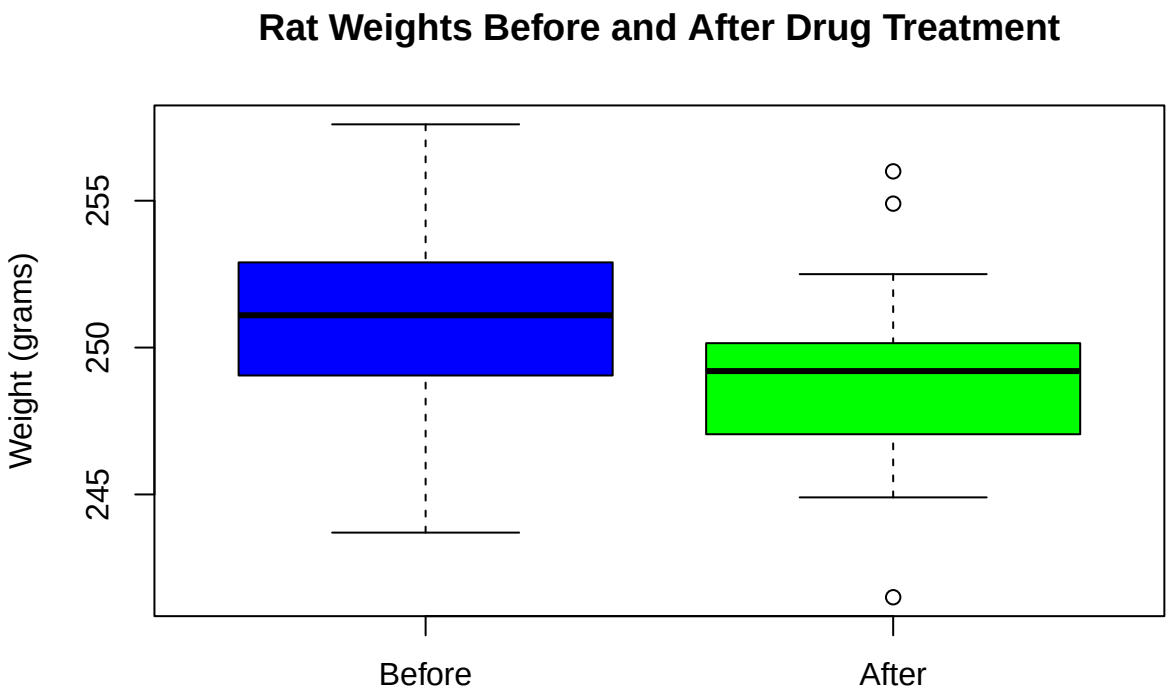
we found the p-value to be $0.01032 > 0.01$. which means we fail to reject H0. this contradicts our previous result from the t-test but given that the t-test rejected the hypothesis, and the p-value of the Wilcoxon test was very close to the p-value of rejection (0.01 or less). we can say that we have more evidence pointing at the fact that the new system does in fact lower CO2 emissions.

Question 2

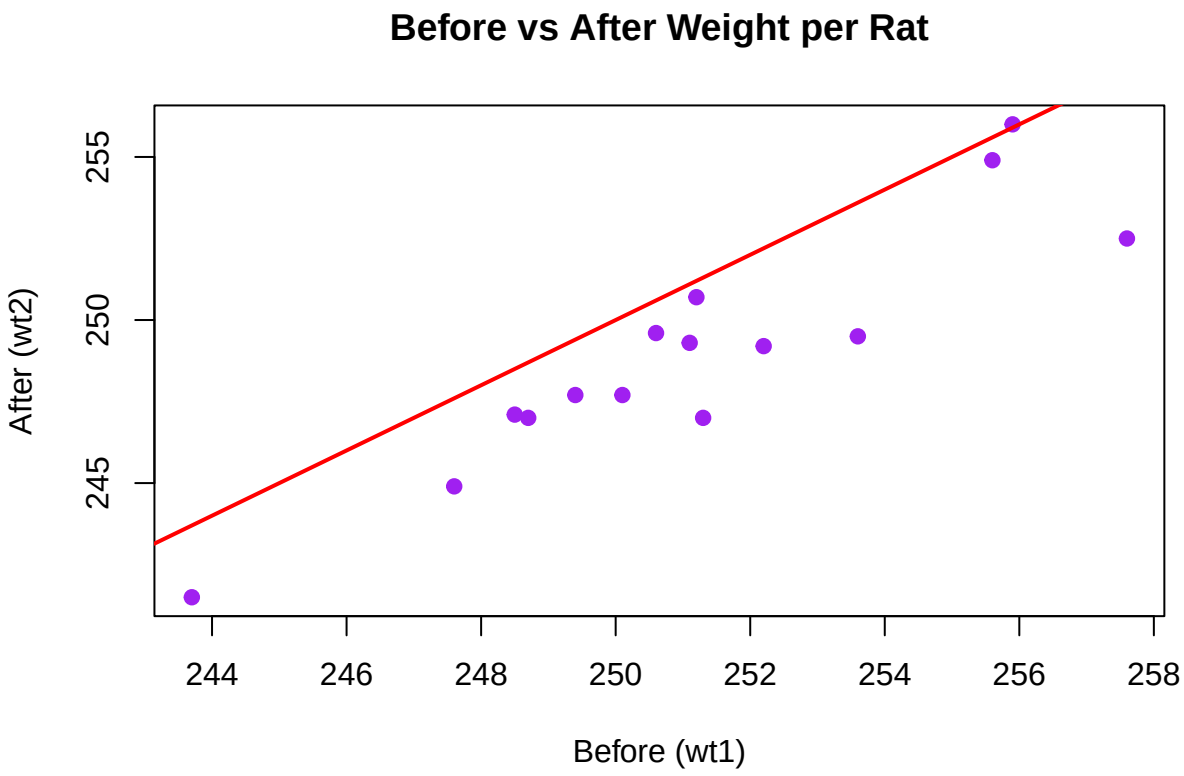
A parmaceutical company wants to test a new weight reduction drug on rats. The experiment uses 15 rats which were weighted before and one week after receiving the drug. The results are stored in the file `25Fhw4Q2`. The column `wt1` holds the measurements for the weight at the beginning of the experiment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

```r
data2 <- read.table("25Fhw4Q2")

#create a boxplot of the two sets
boxplot(data2$wt1, data2$wt2,
        names = c("Before", "After"),
        main = "Rat Weights Before and After Drug Treatment",
        ylab = "Weight (grams)",
        col = c("blue", "green"))
```




```r
#create a plot of the data using wt1 on x and wt2 on y. each rat of the 15 gets one point
plot(data2$wt1, data2$wt2,
     main = "Before vs After Weight per Rat",
     xlab = "Before (wt1)",
     ylab = "After (wt2)",
     pch = 19, col = "purple")

#create a line with slope 1 to show the no difference in weight (before = after)
abline(a = 0, b = 1, col = "red", lwd = 2)
```



The boxplot shows that almost all quartiles of the weight were reduced in the "after" sample.

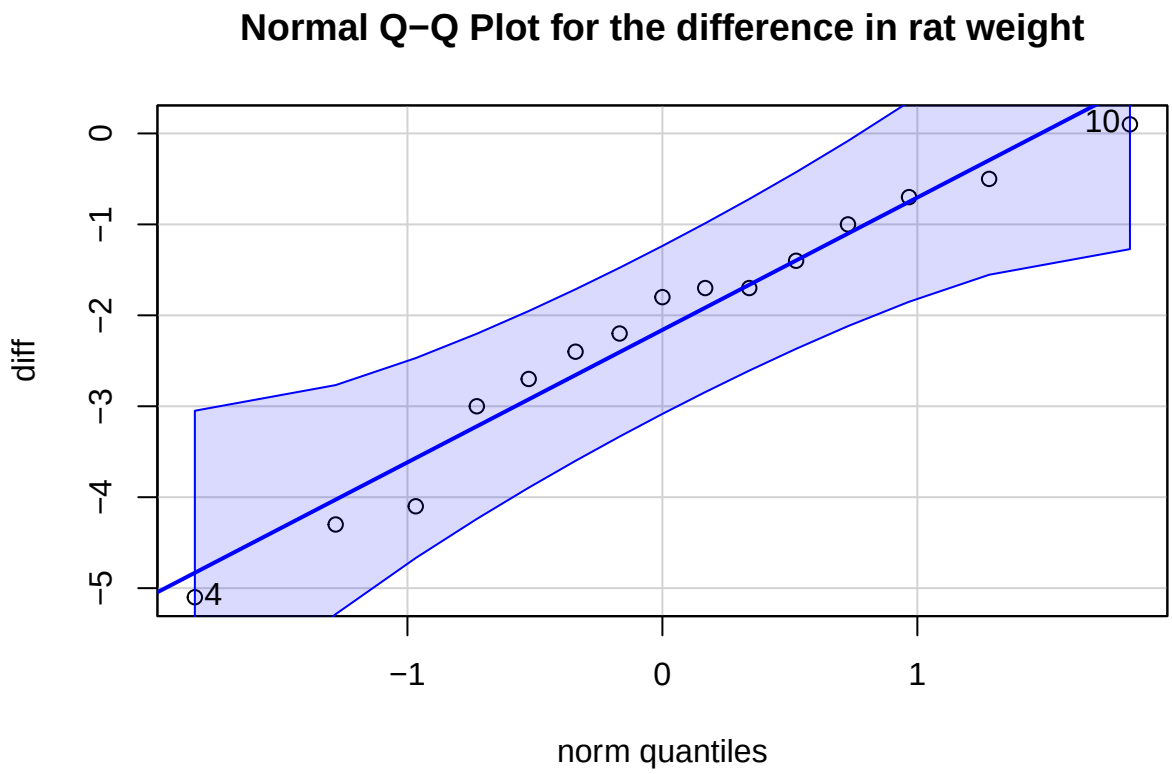
The scatter plot shows all but one of the data points being under the no-change line. which means that the majority of rats had a weight reduction.

(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.

The hypothesis H0 is that mu = 0 (no weight change), and alternative H1 is mu < 0 (weight reduction).

Since we have before and after sets, this presents what is called paired measurements. we do a quick qqplot to determine normality and choose a proper test

```
diff <- data2$wt2 - data2$wt1
qqPlot(diff, line = "r", main = "Normal Q-Q Plot for the difference in rat weight")
```




```
## [1]  4 10
```

Points are close to the qq line and do not curve away with heavy tails. therefore normality assumption is reasonable.

I decided to use a paired-sample t-test with the same required assumptions as the one-sampled version.: Independence: Rats are independent subjects and do not affect eachother. Identical conditions: rats in the experiment face similar conditions (sleep time, cage size, etc.), which is more than likely satisfied. Approximate normality: was concluded using the QQ plot above.

Type 1 error (false positive): conclude the drug reduces weight when it does not. Type 2 error (false negative): fail to reject hypothesis and conclude drug has no effect.

```
t.test(data2$wt2, data2$wt1,
       paired = TRUE,
       alternative = "less",
       conf.level = 0.99)
```

```
##
##  Paired t-test
##
## data:  data2$wt2 and data2$wt1
## t = -5.684, df = 14, p-value = 2.821e-05
## alternative hypothesis: true mean difference is less than 0
## 99 percent confidence interval:
##      -Inf -1.16625
```

```
## sample estimates:
## mean difference
##       -2.166667
```

The test gives a 99% confidence interval (-inf to -1.16625) so we can reject the hypothesis and conclude that there is strong evidence that the drug causes weight reduction in rats.

(c) The pharmaceutical company is only interested in 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.

Since nothing fundamental changed except for the tested hypothesis, we redo the t.test with the new hypotheses H0: mu = -2, H1: mu $<$ -2.

Independence: Rats are independent subjects and do not affect eachother. Identical conditions: rats in the experiment face similar conditions (sleep time, cage size, etc.), which is more than likely satisfied. Approximate normality: was concluded using the QQ plot above.

Type I error: Concluding the mean reduction $>$ 2 g when it is not. Type II error: Failing to detect a true reduction $>$ 2 g.

```
t.test(data2$wt2, data2$wt1,
       paired = TRUE,
       mu = -2,
       alternative = "less",
       conf.level = 0.99)
```

```
##
##  Paired t-test
##
## data:  data2$wt2 and data2$wt1
## t = -0.43723, df = 14, p-value = 0.3343
## alternative hypothesis: true mean difference is less than -2
## 99 percent confidence interval:
##      -Inf -1.16625
## sample estimates:
## mean difference
##       -2.166667
```

given that the p-value = 0.3343 » 0.01, we fail to reject H0 therefore there is insufficient evidence that the weight reduction exceeds 2g.

(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.

If normality is in quesiton, we can perform the paired wilcoxon sign-ranked test which has the following assumptions: Paired observations: each rat is measured before and after the use of the drug Independence: each rats response to the drug is independent Symmetry: the data points are symmetrical around the median (shown by the qq plot)

we then perform the test using the two different pairs of hypotheses.

```r
wilcox.test(data2$wt2, data2$wt1,
            mu = 0,
            paired = TRUE,
            alternative = "less",
            conf.level = 0.99)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  data2$wt2 and data2$wt1
## V = 1, p-value = 6.104e-05
## alternative hypothesis: true location shift is less than 0
```

This shows a p-value « 0.01 which means we reject the hypothesis H0: mu = 0. and the evidence strongly suggests that the drug is causing weight reduction.

```r
wilcox.test(data2$wt2, data2$wt1,
            mu = -2,
            paired = TRUE,
            alternative = "less",
            conf.level = 0.99)
```

```
## Warning in wilcox.test.default(data2$wt2, data2$wt1, mu = -2, paired = TRUE, :
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  data2$wt2 and data2$wt1
## V = 57, p-value = 0.4435
## alternative hypothesis: true location shift is less than -2
```

This shwos a p-value » 0.01 which means we fail to reject the hypothesis H0 : mu = -2, and there is no evidence suggesting that the weight reduction is more than 2g.