Homework #5

Due: Tuesday, October 26 @ 6pm [*40points*]

Please remember to give R code, as well as answers, for any problems where you used R

Problem 1: [*5points*]

For this homework assignment, we are all going to try using R markdown, which is not only great for these homework assignments, but is a great skill for doing analyses in R. For more on R markdown, check out any of the numerous websites online, including this tutorial.

a. An R markdown is a special type of R script that includes both plain text and code, which makes it a perfect choice for writing up a report or a page for a lab notebook! To create an R markdown, select `File > New File > R markdown`. A popup window will ask you to name the document and choose the output type. For now, let's stick with HTML output (default) but note you could also choose to output the document as a PDF (if you first install latex to your machine). Name the document and press "OK" to create the report.
b. Take a look at the file that is generated. It is written in "markdown syntax" with normal text, *italics*, **bold**, and headers (i.e. "problem 1"). Most importantly, you also have "code chunks" designated by the "```{r}". Any code you type in these chunks will be evaluated and the results printed.
c. If you are using RStudio 1.4 or newer, you can also view/write R Markdowns in the "visual markdown editor" (i.e. you don't really have to know markdown syntax anymore!). To switch to this mode, click the "visual markdown editor" button on the upper right of the script panel (looks like an A). This will more or less show you what the report will look like once generated.
d. Without changing anything, press the "knit" button at the top. An HTML document should be generated showing text, code, and figures.
e. Play around with adding text or code chunks in the test markdown and when you feel comfortable, create a new markdown to use for this homework assignment. **You will get full credit on this problem if your HW5 is turned in as an HTML or PDF markdown output.**

Full credit for homework assignments completed using R markdown. Half credit if there is proof that they tried R markdown but the assignment was completed in another way.

Problem 2: [*15points*]

Consider the following two strains of mice:

Strain1	42.6	57.89	45.22	45.4	40.01	48.64	48.42	46.86	49.47	52.14	46.11	43.53
Strain2	46.2	54.12	52.48	46.97	65.89	57.41	59.07	51.07	56.24			

a. Write down the t statistic for testing the hypothesis that the population means are equal. (Show your work, do not just simply give a number!) [*3points*]

To test the hypothesis $H_0 : \mu_1 = \mu_2$

$t_s = \frac{\bar{y}_1 - \bar{y}_2 - 0}{\sqrt{SE_1^2 + SE_2^2}}$ (1 pt. equation)

$\bar{y}_1 = 47.19$; $\bar{y}_2 = 54.38$ (0.5 pt. correct means)

$SE_1 = s_1/\sqrt{n} = 4.69/\sqrt{12} = 1.35$ (0.5 pt. correct SE)

$SE_2 = s_2/\sqrt{n} = 6.16/\sqrt{9} = 2.05$

$t_s = \frac{47.19 - 54.38}{\sqrt{1.35^2 + 2.05^2}} = -2.92$ (1 pt. correct t)

Grading note: means/SE don't have to be calculated separately, can all be in one equation

b. Using the value you obtained in (a) and the `rt/qt/pt/dt` functions (not `t.test`!), test the hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_A : \mu_1 \neq \mu_2$. Compute the p-value and give an interpretation (in words) of the finding. [*3points*]

```
# we want to use pt(t, df). We already know t, now we need df
# conservatively, we can use df = smaller of the two n-1: 9-1 = 8 (not 12-1 = 11)
# because it is a non directional hypothesis, we need to find the area in both tails (or find the area in
pt(-2.92, 8)*2
```

```
## [1] 0.01929128
```

```
# note we could also test the positive t value with
pt(2.92, 8, lower.tail = F)*2
```

```
## [1] 0.01929128
```

The p-value of 0.0192 suggests that there is a low probability of observing a difference as large as -7.19 under H_0, and so we reject the null hypothesis that the means are equal.

Grading note: 1 point for using pt with correct df (half point if wrong df). 1 point for getting both tails, 1 point for correct explanation.

c. How would you change the above to test $H_A : \mu_1 < \mu_2$? $H_A : \mu_1 > \mu_2$? [*2points*]

```
# for mu1 < mu2 - only lower tail
pt(-2.92, 8)
```

```
## [1] 0.009645638
```

```
# for mu1 > m2 - only upper tail
# note that this prob is very large! One directional hypothesis must be in the right direction!
pt(-2.92, 8, lower.tail = F)
```

```
## [1] 0.9903544
```

```
# alternatively, we could re-calculate t to show mu2-mu1 instead of mu1-mu2
old_t <- (47.19-54.38)/(sqrt(1.35^2 + 2.05^2))
old_t
```

```
## [1] -2.929209
```

```
new_t <- (54.38-47.19)/(sqrt(1.35^2 + 2.05^2))
new_t
```

```
## [1] 2.929209
```

```
pt(new_t, 8)
```

```
## [1] 0.9904894
```

Grading note: one point mu1 < mu2, one point for mu1 > mu2, each all or nothing

d. Now, using R's `t.test()` function, repeat parts (b) and (c) [*3points*]

2

```
# create vectors for sample 1 and sample 2
s1 <- c(42.6, 57.89, 45.22, 45.4, 40.01, 48.64, 48.42, 46.86, 49.47, 52.14, 46.11, 43.53)
s2 <- c(46.2, 54.12, 52.48, 46.97, 65.89, 57.41, 59.07, 51.07, 56.24)

# mu1 != mu2 (one point)
t.test(s1, s2)
```

```
##
##  Welch Two Sample t-test
##
## data:  s1 and s2
## t = -2.9239, df = 14.483, p-value = 0.01079
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12.451951  -1.933049
## sample estimates:
## mean of x mean of y
##   47.19083  54.38333
```

```
# mu1 > mu2 (one point)
t.test(s1, s2, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  s1 and s2
## t = -2.9239, df = 14.483, p-value = 0.005393
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -2.870082
## sample estimates:
## mean of x mean of y
##   47.19083  54.38333
```

```
# mu1 < mu2 (one point)
t.test(s1, s2, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  s1 and s2
## t = -2.9239, df = 14.483, p-value = 0.9946
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -11.51492       Inf
## sample estimates:
## mean of x mean of y
##   47.19083  54.38333
```

Grading note: one point each, no partial credit

    e. Suppose that the data above could NOT be assumed to be normal. How might you test these hypotheses? Give R code and explain what you are doing and why you chose this test. [4points]

Wilcoxon mann-whitney U test; eg for $H_A : \mu_1 \neq \mu_2$

```
wilcox.test(s1, s2)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  s1 and s2
## W = 17, p-value = 0.007253
## alternative hypothesis: true location shift is not equal to 0
```

Permutation test:

```
# one way to do this:
# combine data
both <- c(s1, s2)

# create boolean vector of T for sample 1 and F for sample2
strain1 <- c(rep(TRUE, length(s1)), rep(FALSE, length(s2)))

# create empty vector to hold perm_data
perm_data <- c()
# scramble the data
for(i in 1:1000) {
    # sample true/false vector to scramble data
    perm <- sample(strain1)
    # select fake s1 and s2 based on sample
    s1_perm <- both[perm]
    s2_perm <- both[!perm]
    # calculate fake difference of means
    perm_data <- c(perm_data, mean(s1_perm) - mean(s2_perm))
}

# p-value = num of differences >= true difference out of 1000
# note - add one to numerator and denominator to avoid p = 0
(sum(abs(perm_data) > abs(mean(s1) - mean(s2)))+1)/(1000+1)
```

```
## [1] 0.008991009
```

Grading note: 2 points for choosing a correct test (1/2 points if the test was wrong but they knew it was supposed to be non parametric). 2 points for doing the test properly (1/2 points if it was attempted with a minor error)

Problem 3: [4points]

Diastolic blood pressure measurements on American men ages 18-44 years follow approximately a normal curve with $\mu = 81$ mm Hg and $\sigma = 11$ mm Hg. The distribution for women ages 18-44 is also approximately normal with the same SD but with a lower mean: $\mu = 75$ mm Hg. Suppose we are going to measure the diastolic blood pressure of n randomly selected men and n randomly selected women in the age group 18-44 years. Let E be the event that the difference between men and women will be found statistically significant by a t test. How large must n be in order to have $Pr[E] = 0.9$: [1point/each]

    a. if we use a two-tailed test at $\alpha = 0.05$? $n = 72$

```
power.t.test(delta = 81-75, sd = 11, sig.level = 0.05, power = 0.9)
```

```
##
##      Two-sample t test power calculation
##
##              n = 71.60735
##          delta = 6
##             sd = 11
##      sig.level = 0.05
##          power = 0.9
##    alternative = two.sided
##
## NOTE: n is number in *each* group
```

    b. if we use a two-tailed test at $\alpha = 0.01$? $n = 102$

```
power.t.test(delta = 81-75, sd = 11, sig.level = 0.01, power = 0.9)
```

```
##
##      Two-sample t test power calculation
##
##              n = 101.6951
##          delta = 6
##             sd = 11
##      sig.level = 0.01
##          power = 0.9
##    alternative = two.sided
##
## NOTE: n is number in *each* group
```

    c. if we use a one-tailed test (in the correct direction) at $\alpha = 0.05$? $n = 59$

```
power.t.test(delta = 81-75, sd = 11, sig.level = 0.05, power = 0.9, alternative = "one.sided")
```

```
##
##      Two-sample t test power calculation
##
##              n = 58.25716
##          delta = 6
##             sd = 11
##      sig.level = 0.05
##          power = 0.9
##    alternative = one.sided
##
## NOTE: n is number in *each* group
```

    d. Suppose we only have a sample size of n = 45. How lare must the difference between means be in order to have 80% power if we used a two-tailed test at $\alpha = 0.05$? $difference >= 6.56$

```
power.t.test(sd = 11, sig.level = 0.05, power = 0.8, n = 45)
```

```
##
##       Two-sample t test power calculation
##
##              n = 45
##          delta = 6.568868
##             sd = 11
##      sig.level = 0.05
##          power = 0.8
##    alternative = two.sided
##
## NOTE: n is number in *each* group
```

Take away 1 point (out of 4) if they didn't round up for n - can't have non-integer n! Otherwise each part is all or nothing

Problem 4: [16*points*]

A scientist conducted a study of how often her pet parakeet chirps. She recorded the number of distinct chirps the parakeet made in a 30-minute period, sometimes when the room was silent and sometimes when music was playing. Import the data she collected into R with the following code:

```
url <- "https://raw.githubusercontent.com/katiesevans/IGP_biostatistics/main/data/chirps.tsv"
chirps <- read.delim(url)
```

    a. Construct a 95% confidence interval for the mean increase in chirps (per 30 minutes) when music is playing over when music is not playing. Be sure to show your work! [4*points*]

Since we are interested in the INCREASE of chirps with music, it is most appropriate to use a one-sided CI with a lower bound. (However, since this was unclear, a two-sided interval will also be considered)

One-sided interval

```
# we are interested in the INCREASE of chirps with music, we can use a one-sided CI with a lower bound
# CI = y1-y2 - t_0.05*SE_y1-y2

# because this is PAIRED, we can use the differences (1 point for taking difference - treating paired)
diff <- chirps$difference

# SE = s / sqrt(n) (1 point for calculating SE_d - 0.5 point with small error)
SE <- sd(diff)/sqrt(length(diff))

# t_0.025 - df = n - 1 (subtract one point if df = 28*2)
df = nrow(chirps)-1

# calculate t with qt() (1 point for calculating ts with qt() - 0 points if not one sided (i.e wrong p))
ts <- qt(0.95, df)

# confidence interval (1 point correct interval - 0 points if not one sided)
c(mean(diff) - ts*SE, Inf)
```

```
## [1] 9.973171      Inf
```

Two-sided interval

6

```
# if you did a twosided (not correct, but allowed since the Q was unclear)
ts <- qt(0.975, df)

# confidence interval (1 point correct interval - 0 points if not one sided)
c(mean(diff) - ts*SE, mean(diff) + ts*SE)
```

```
## [1]  9.777669 12.079474
```

b. What conditions are necessary for the confidence interval to be valid? Are those conditions satisfied? [*5 points*]

1. The differences must be regarded as a random sample from a large population. (1 pt)

> We can assume the scientist randomly chose 30 min intervals throughout the day, so this assumption can be valid. (1 pt)

2. The population distribution of the differences must be approximately normal (or sample size must be large to be approx normal given CLT) (1 pt)

> We can test normality of the differences by plotting the distribution, using a Q-Q plot and/or the Shapiro-Wilks test. (1 pt for explaining (need Q-Q plot and/or shapiro-wilks test for full credit, only 0.5 point for plotting distribution), 1 pt for showing at least one of these tests and results)
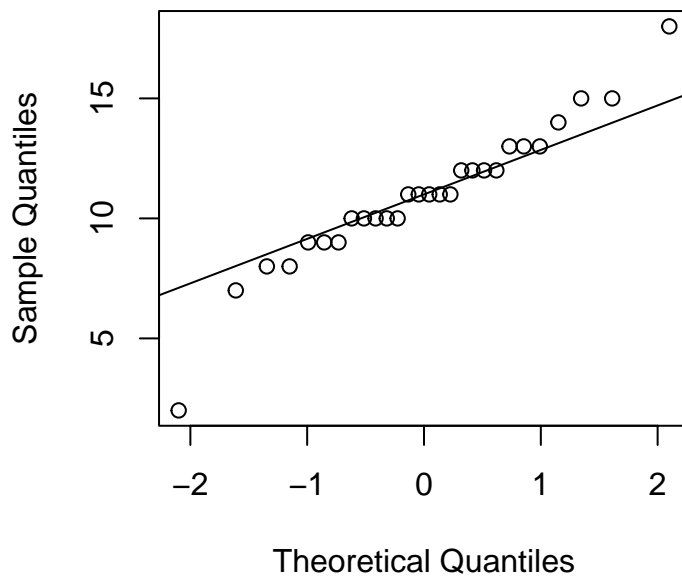
```
# distribution looks mostly normal
hist(chirps$difference)
```

Histogram of chirps$difference
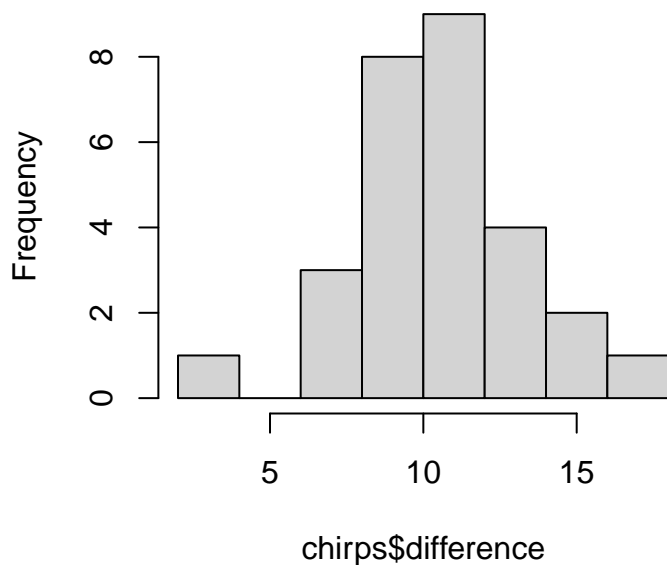

```
# Q-Q plot also seems mostly normal with tails on either end
qqnorm(chirps$difference)
qqline(chirps$difference)
```

Normal Q–Q Plot


```
# shapiro test shows pvalue of 0.167 - which is no evidence for NON-normality, meaning the data is normall
shapiro.test(chirps$difference)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  chirps$difference
## W = 0.94706, p-value = 0.167
```

   c. Use a t test to test the null hypothesis of no difference in the number of chirps with or without music against the alternative hypothesis that there is an increase in chirps when music is playing. You can calculate by hand or use R's `t.test()`. [3points]

```
# HO: diff = 0
# HA: diff > 0 (diff = music - no music)
t.test(chirps$difference, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  chirps$difference
## t = 19.483, df = 27, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  9.973171      Inf
## sample estimates:
## mean of x
##  10.92857
```

```
# OR
t.test(chirps$with_music, chirps$without_music, paired = T, alternative = "greater")
```

```
## 
##  Paired t-test
## 
## data:  chirps$with_music and chirps$without_music
## t = 19.483, df = 27, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  9.973171      Inf
## sample estimates:
## mean of the differences
##                10.92857
```

Grading note: 1 point for treating as paired, 1 point for knowing it is not a two-sided test, and 1 point for choosing the correct size or having the correct answer. OK to calculate t and p-value by hand OR with t.test()

    d. Suppose that the data above could NOT be assumed to be normal. How might you test these hypotheses? Give R code and explain what you are doing and why you chose this test. [4points]

We can test paired data with a wilcoxon signed rank test

```
wilcox.test(chirps$with_music, chirps$without_music, paired = T, alternative = "greater")
```

```
## 
##  Wilcoxon signed rank test with continuity correction
## 
## data:  chirps$with_music and chirps$without_music
## V = 406, p-value = 1.917e-06
## alternative hypothesis: true location shift is greater than 0
```

Alternatively, we could again use permutation/randomization.

```
# one way to do this:
# combine data
both <- c(chirps$with_music, chirps$without_music)

# create boolean vector of T for sample 1 and F for sample2
withmusic <- c(rep(TRUE, 28), rep(FALSE, 28))

# create empty vector to hold perm_data
perm_data <- c()
# scramble the data
for(i in 1:1000) {
    # sample true/false vector to scramble data
    perm <- sample(withmusic)
    # select fake s1 and s2 based on sample
    music_perm <- both[perm]
    silent_perm <- both[!perm]
    # calculate fake difference of means
    perm_data <- c(perm_data, mean(music_perm) - mean(silent_perm))
}

# p-value = num of differences >= true difference out of 1000
# note - add one to numerator and denominator to avoid p = 0
(sum(abs(perm_data) > abs(mean(chirps$with_music) - mean(chirps$without_music)))+1)/(1000+1)
```

```
## [1] 0.000999001
```