Assignment 7

Your name and student ID

Today's date

- Due date: November 5, 11:59pm (make sure to provide enough time for Gradescope submission to be uploaded if you include large image files).
- Late penalty: 50% late penalty if submitted within 24 hours of due date, no marks for assignments submitted thereafter.
- Submission process: Please submit a PDF of your assignment to Gradescope. You must tell Gradescope which questions are on which pages. If you can't see it properly on Gradescope, open the PDF in a PDF viewer at the same time so you can make the selections accurately, others points will be deducted since this makes grading much less efficient.

Helpful hint:

- Knit your file early and often to minimize knitting errors! If you copy and paste code, you are bound to get an error that is hard to diagnose. Hand-writing code is the way to smooth knitting! We recommend knitting your file each time after you write a few sentences/add a new code chunk, so you can detect the source of knitting errors more easily.

Question 1 [12 points total].

In two wards for elderly patients in a local hospital the following levels of hemoglobin (grams per liter) were found for a simple random sample of patients from each ward.:

Ward A:

```
ward_a <- c(12.2, 11.1, 14.0, 11.3, 10.8, 12.5, 12.2, 11.9, 13.6, 12.7, 13.4, 13.7)
```

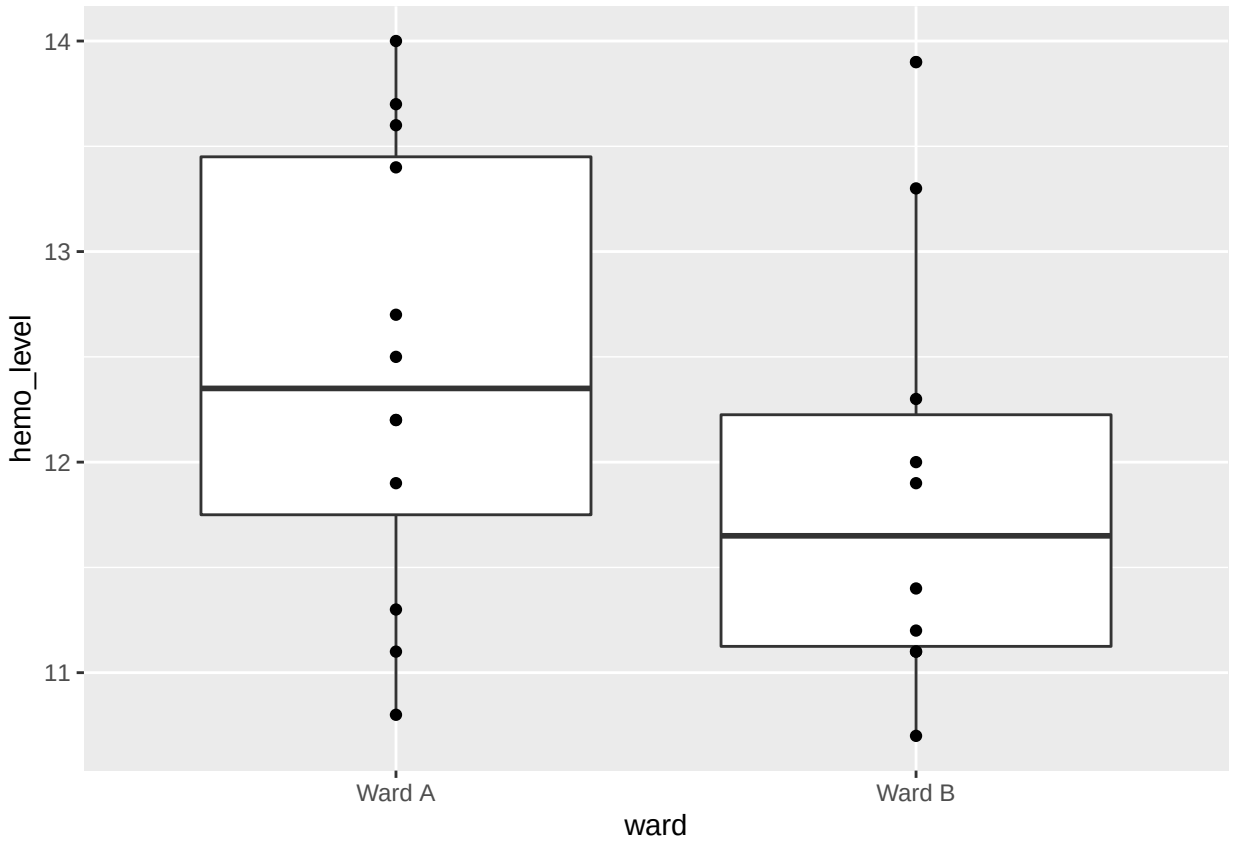
Ward B:

```
ward_b <- c(11.9, 10.7, 12.3, 13.9, 11.1, 11.2, 13.3, 11.4, 12.0, 11.1)
```

1.a) [2 points] Make two box plots to compare the hemoglobin values for Ward A and Ward B. Overlay the boxplots with their raw data. Comment on the similarities/differences portrayed by the plots, keeping in mind that the sample size is relatively small for these two wards.

```
hemoglobin <- data.frame(hemo_level = c(ward_a, ward_b),
                         ward = c(rep("Ward A", 12), rep("Ward B", 10)))
```

```
# Solution
ggplot(hemoglobin, aes(y = hemo_level, x = ward)) + geom_boxplot() + geom_point()
```



Solution: - There is some overlap in the middle 50% of the data form these two wards. There do not appeard to be outliers in either distribution. Both samples appear to be roughly symmetric. The sample median is higher in Ward A than Ward B.

1.b) [2 points] What two assumptions do you need to make to use any of the t-procedures? Because each ward has a rather small sample size (n $<$ 12 for both), what two characteristics of the data would you need to check for to ensure that the t procedures can be applied?

- Two assumptions: SRS, normality of underlying dataset
- No outliers, data has similar shapes

1. c) [4 points] Using only `dplyr` and `*t` functions, create a 95% confidence interval for the mean difference between Ward A and Ward B. You can do this by using `dplyr` to calculate the inputs required to calculate the 95% CI, and then plugging these values in on a separate line of code (or using your calculator). Use a degrees of freedom of 19.515 (You don't need to calculate the degrees of freedom, you can use this value directly). Show your work and interpret the mean difference and its 95% CI.

```
#write your code here

#solution
hemoglobin %>% group_by(ward) %>% summarise(sample_mean = mean(hemo_level),
                                            sample_var = var(hemo_level),
                                            n = length(hemo_level))
```

```
## # A tibble: 2 x 4
##   ward   sample_mean sample_var     n
##   <fct>        <dbl>      <dbl> <int>
## 1 Ward A        12.4       1.14    12
```

```
## 2 Ward B        11.9       1.07    10
```

```
#here is how you calculate degrees of freedom
deg_free <- ( (1.140909/12) + (1.065444/10) )^2/( (1/11)*(1.140909/12)^2 + (1/9)*(1.065444/10)^2)
mean_diff <- 12.45 - 11.89
se_diff <- sqrt(1.140909/12 + 1.065444/10)
t_star <- qt(p = 0.025, df = deg_free)

mean_diff + t_star*(se_diff)
```

```
## [1] -0.3781371
```

```
mean_diff - t_star*(se_diff)
```

```
## [1] 1.498137
```

Solution: The sample mean difference is 0.56 and its 95% CI goes from -0.44 to 1.56. This means that if we were to repeat this procedure 100 times, we would expect that 95 of the CIs would contain the true difference. The range of the difference goes from negative to positive indicating that at the 5% level there is no evidence against the null hypothesis of no difference.

1.d) [4 points] Perform a two-sided t-test for the difference between the two samples, where the null hypothesis is that the underlying means are the same. Start by writing down the null and alternate hypotheses, then calculate the test statistic (showing your work) and p-value. Continue to assume that the degrees of freedom is 19.515. Verify the p-value by running the t-test using R's built in function. Show the output from that test. Hint: to perform the t-test using R's built in function, you need to pass the function an x and y argument, where x includes that values for Ward A and Y includes the values for Ward B. `dplyr`'s `filter()` and `pull()` functions will be your friends.

Solution:

$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ $t = \frac{(12.45 - 11.89) - 0}{\sqrt{\frac{1.140909}{12} + \frac{1.065444}{10}}}$ $t = 0.56/0.4490213 = 1.247157$

We need to compare this t-statistic to a t distribution with 19.515 degrees of freedom:

```
pt(1.247157, df = 19.515, lower.tail = F)*2
```

```
## [1] 0.2271006
```

Thus there is a 22.7% chance of seeing a difference of the size we saw or larger under the hypothesis of no difference. This is quite probable, so we conclude that there is no evidence against the null hypothesis.

Check this against the `t.test` output:

```
t.test(x = hemoglobin %>% filter(ward == "Ward A") %>% pull(hemo_level),
       y = hemoglobin %>% filter(ward == "Ward B") %>% pull(hemo_level),
       alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  hemoglobin %>% filter(ward == "Ward A") %>% pull(hemo_level) and hemoglobin %>% filter(ward ==
## t = 1.2472, df = 19.515, p-value = 0.2271
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3781372  1.4981372
## sample estimates:
## mean of x mean of y
##     12.45     11.89
```

Question 2 [7 points total]

The time to perform open heart surgery is normally distributed. Sixteen patients (chosen as a simple random sample from a hospital) underwent open heart surgery that took the following lengths of time (in minutes).

```
op_time <- c(247.8648, 258.4343, 315.6787, 268.0563, 269.9372, 320.6821,
             280.5493, 225.3180, 243.8207, 251.5388, 304.9706, 277.3140,
             278.6247, 269.3418, 248.0131, 322.9812)
surg_data <- data.frame(op_time)
```

a) [1 point] You wish to know if the mean operating time of open heart surgeries at this hospital exceeds four hours. Set up appropriate hypotheses for investigating this issue.

Solution: $H_0 : \mu = 4$ hours (240 mins) $H_a : \mu > 4$ hours (240 mins)

b) [3 points] Test the hypotheses you formulated in part (a). Report the p-value. What are your conclusions? (Do not use the `t.test` function for this question)

Solution:

```
surg_data %>% summarise(mean = mean(op_time), se = sd(op_time)/sqrt(16))
```

```
##       mean       se
## 1 273.9454 7.305622
```

$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ $t = \frac{273.9454-240}{7.305621} = 4.646477$

```
pt(4.646477, df = 15, lower.tail = F)
```

```
## [1] 0.0001582348
```

The p-value of 0.000158, which is very small. There is only a miniscule chance of seeing the sample mean we saw (or larger) if the null hypothesis is true. Thus we reject the null hypothesis in favor of the alternative, that the operating time exceeds 4 hours.

c) [3 points] Construct a 95% CI on the mean operating time.

```
qt(p = 0.975, df = 15)
```

```
## [1] 2.13145
```

$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$ $273.9454 \pm 2.13145 \times 7.305621 = 258.3738$ to $289.517 = 4.31$ hours to 4.73 hours

Thus, using a method that includes the null value 95 times out of 100, our 95% CI is 4.31 hours to 4.73 hours.

---

3. [1 point] Suppose you were testing the hypotheses $H_0 : \mu_d = 0$ and $H_a : \mu_d \neq 0$ in a paired design and obtain a p-value of 0.21. Which one of the following could be a possible 95% confidence interval for μ_d?

a. -2.30 to -0.70
b. -1.20 to 0.90
c. 1.50 to 3.80
d. 4.50 to 6.90

Solution: b

---

4. [1 point] Suppose you were testing the hypotheses $H_0 : \mu_d = 0$ and $H_a : \mu_d \neq 0$ in a paired design and obtain a p-value of 0.02. Also suppose you computed confidence intervals for μ_d. Based on the p-value which of the following are true?

a. Both a 95% CI and a 99% CI will contain 0.
b. A 95% CI will contain 0, but a 99% CI will not.
c. A 95% CI will not contain 0, but a 99% CI will.
d. Neither a 95% CI nor a 99% CI interval will contain 0.

Solution: c