Note: This homework assignment covers Chapter 8.

Disclaimer: If you use R, include all R code and output as attachments. Do not just "write in" the R code you used. Also, don't just write the answer and say this is what R gave you. If my grader can't see how you got an answer, it is wrong. I want to see your code and your answers accompanying your code (like in the notes).

1. In a study conducted in the Department of Zoology at Virginia Tech University, data were collected on density measurements (i.e., the number of organisms per m^2) at two different locations. The goal was to compare the population mean number of organisms (per m^2) between the two locations. Independent samples were collected from each location; here are the data.

Location 1		Location 2	
5030	4980	2800	2810
13700	11910	4670	1330
10730	8130	6890	3320
11400	26850	7720	1230
860	17660	7030	2130
2200	22800	7330	2190
4250	1130		
15040	1690		

I used R to construct side-by-side boxplots. These are shown on the next page.

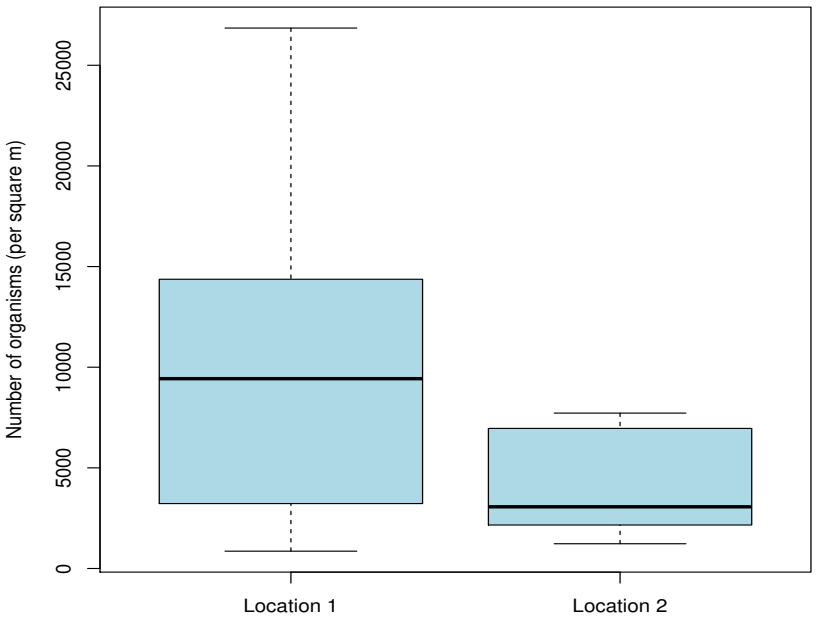
(a) If you wanted to write a confidence interval for $\mu_1 - \mu_2$, the difference of the two population mean number of organisms between locations, select the confidence interval you would use:

- the one that assumed equal population variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
- the one that did not assume the population variances were equal.

Explain why you chose the answer you did. In addition, what statistical procedure could be used to determine which assumption is more reasonable? (explanation only; no calculations are needed here).

(b) I asked R to get both intervals (each at the 95 percent confidence level). Here are the intervals:

```
> t.test(loc.1,loc.2,conf.level=0.95,var.equal=TRUE)$conf.int
[1]   914.09 10639.23
```




```
> t.test(loc.1,loc.2,conf.level=0.95,var.equal=FALSE)$conf.int
[1]  1389.00 10164.33
```

For the interval that you picked in part (a), interpret the corresponding interval here.
(c) I looked at the (normal) qq plots for both samples, and I detected some moderate departures from linearity in both plots (concentrated in the upper tail).

- Are you surprised that I detected linearity departures? Look at the boxplots above and comment.
- Does this finding affect your conclusions in part (b)? Explain how it could or why it may not.

2. Two different brands of latex (water-based) paint are being considered for use in a large construction project. To choose between the brands, one of the key factors is the time it takes the paint to dry. Engineers sampled 22 specimens of each brand and measured the drying times (in hours) of each specimen. The data collected are given below.

(a) Treat these samples as independent samples, one taken from the Brand A population and one taken from the Brand B population. The goal is to learn how the population mean drying times for the two brands compare. Perform a thorough analysis that addresses this question. **My idea of a thorough analysis includes**

Specimen	Brand A	Brand B	Specimen	Brand A	Brand B
1	3.5	4.7	12	5.2	6.2
2	2.7	3.9	13	4.0	5.1
3	3.9	4.5	14	4.1	5.4
4	4.2	5.5	15	3.4	4.8
5	4.6	4.0	16	3.3	4.9
6	2.7	5.3	17	4.2	6.6
7	3.3	4.3	18	5.3	4.3
8	5.2	6.0	19	3.7	4.9
9	4.2	5.3	20	3.0	5.1
10	2.9	3.7	21	4.0	3.9
11	4.4	5.5	22	2.8	5.2

- a complete description of the statistical assumptions as well as checking these assumptions
- showing all calculations (carried out "by hand" or preferably using R)
- (if helpful/needed) well-constructed, informative graphs which are relevant to the problem at hand
- a well-written paragraph that summarizes the entire analysis (which should include the final main conclusions).

(b) If you performed a matched-pairs analysis in part (a), then you did the analysis incorrectly because a matched-pairs analysis assumes that the samples are dependent (and I told you the samples were independent). However, would it be possible to learn about the population mean drying times using a matched-pairs design? If so, explain how you could design a matched-pairs experiment to accomplish this; make sure you tell me explicitly how you are doing the "matching." If you don't think it would be possible, then explain why.

3. One kind of dioxin, called tetrachlorodibenzodioxin (TCDD), was once thought to cause cancer and birth defects, but subsequent research showed it to have only mild toxic effects except at very high exposure levels. About 10 years ago, a study published in a leading journal in environmental science reported the levels of TCDD of 20 Massachusetts Vietnam veterans who may have been exposed to Agent Orange during the war. The TCDD levels in plasma and in fat tissue are listed below. The goal of the study was to learn how these two population mean levels compared.

(a) For this cohort of individuals, why might it be of interest to learn about how the mean TCDD levels compare in plasma and fat tissue?
(b) Acknowledging that this is a matched pairs set up (first explain why), analyze these

Veteran	Plasma	Fat Tissue	Veteran	Plasma	Fat Tissue
1	2.5	4.9	11	6.9	7.0
2	3.1	5.9	12	3.9	2.9
3	2.1	4.4	13	4.2	4.6
4	3.5	6.9	14	1.6	1.4
5	3.1	9.0	15	7.2	7.7
6	1.8	4.2	16	1.8	1.1
7	6.0	10.0	17	20.0	11.0
8	3.0	5.5	18	2.0	2.5
9	36.0	41.0	19	2.5	2.3
10	4.7	4.4	20	4.1	1.5

data to answer the research question in part (a). State all conclusions in plain English.
(c) Explain what advantages this matched pairs design has over an experiment that would require two independent samples of veterans.

4. Airplanes approaching the runway for landing are required to stay within a "localizer region" (a certain distance left and right of the runway). When an airplane deviates from the localizer region, the FAA calls this an "exceedence." At the Schiphol airport in Amsterdam, two airlines (SAS and Lufthanza) were under investigation. In a one-week period, SAS had 8 out of 86 observed flights classified as exceedences. Lufthanza had 10 out of 142 observed flights classified as exceedences.
(a) Calculate a 95 percent confidence interval for $p_1 - p_2$, the difference in the population proportions of exceedences for SAS (p_1) and Lufthanza (p_2). Interpret the interval. What does the interval suggest about the two airline's ability to stay within the localizer region?
(b) FAA officials want to plan a larger study that assumes a 99 percent confidence level and equal numbers of airplanes sampled for both airlines (so that $n_1 = n_2 = n$). Find the smallest sample size n (per airline) that will produce a 99 percent confidence interval for $p_1 - p_2$ to have margin of error equal to 0.03. You can use the information in the problem to estimate any parameters needed for this calculation. *Hint:* First, find the value of $z_{\alpha/2}$ that corresponds to 99 percent confidence. Then set the margin of error in the confidence interval

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

equal to 0.03 and solve for $n_1 = n_2 = n$. Use the data in the problem to calculate $\widehat{p}_1$ and $\widehat{p}_2$ and then use these as estimates for the larger study.