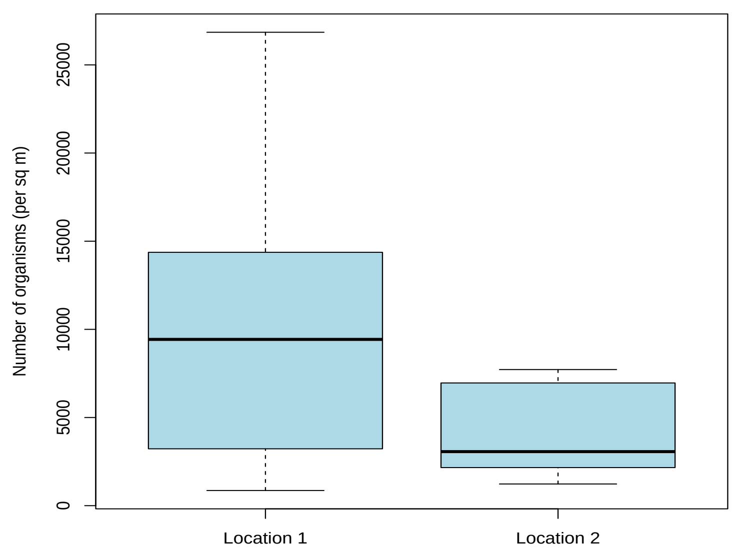
**Instructions:** This homework assignment covers **Chapter 8** of the course notes. On each part, there is opportunity for partial credit, so show all of your work and explain all of your reasoning. No work/no explanation means no credit even if your answer is correct. If you use R to answer any part or to check your work, please include all code and output as attachments. Do not just write out the code you used.

1. In a study conducted in the Department of Zoology at Virginia Tech University, biologists collected 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 producing the data below:

| 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 below:




(a) If you wanted to write a confidence interval for $\Delta = \mu_1 - \mu_2$, the difference of the two population mean number of organisms between locations, which 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% confidence level). Here are the intervals:

```
> t.test(loc.1,loc.2,conf.level=0.95,var.equal=TRUE)$conf.int
[1]   914 10639

> t.test(loc.1,loc.2,conf.level=0.95,var.equal=FALSE)$conf.int
[1]  1389 10164
```

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 and comment.
- Does this finding affect your conclusions in part (b)? Explain how it could or why it may not.
- Remember the sample sizes here are small ($n_1 = 16$ and $n_2 = 12$). It is hard to make definite conclusions about the population distributions with such small samples.

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 44 pieces of pressure-treated wood; the pieces were randomized to receive either Brand A or Brand B paint (22 per brand). The drying time (in hours) of each specimen was then measured. The data collected are given below:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Brand A | 3.5 | 2.7 | 3.9 | 4.2 | 4.6 | 2.7 | 3.3 | 5.2 | 4.2 | 2.9 | 4.4 |
| | 5.2 | 4.0 | 4.1 | 3.4 | 3.3 | 4.2 | 5.3 | 3.7 | 3.0 | 4.0 | 2.8 |
| Brand B | 4.7 | 3.9 | 4.5 | 5.5 | 4.0 | 5.3 | 4.3 | 6.0 | 5.3 | 3.7 | 5.5 |
| | 6.2 | 5.1 | 5.4 | 4.8 | 4.9 | 6.6 | 4.3 | 4.9 | 5.1 | 3.9 | 5.2 |

(a) Treat these two samples as independent, 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**

- 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 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? Explain how you could design a matched-pairs experiment to accomplish this. Start with the same 44 pieces of pressure-treated wood and make sure you tell me explicitly how you are doing the "matching." No calculations here; you are just describing how a matched-pairs experiment could be carried out.

3. 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-month period, SAS had 8 out of 86 randomly sampled flights classified as exceedences. Lufthanza had 10 out of 142 randomly sampled flights classified as exceedences.
(a) Calculate a 95% confidence interval for $\Delta = p_1 - p_2$, 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% 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% confidence interval for $\Delta = p_1 - p_2$ 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% 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.