Homework 7 Solutions

Problem 1

 # Problem 1 (a)

I would choose the one that did not assume the population variances were equal, because the box plot shows that the variances of two samples differs a lot.

Confident interval of the ratio of two population variances can be used to determine the assumption. If the confident interval does not include 1, this suggest that the population variances are different.

 # Problem 1 (b)

Confident interval: [1389.00, 10164.33]

We are 95% confident that the population mean difference of the number of organisms between location 1 and location 2 is between 1389 and 10164.33. The mean number of organism at location 1 is higher than location 2.

 # Problem 1 (c)

No, I am not surprised, because the boxplots show moderate skew to the right pattern, it might have moderate departure from normality. Thus, it's not surprising to get moderate linearity departure in qq plot.

No, this doesn't affect my finding in part (b), because t-test is robust to normal departure. A moderate normal departure won't affect our result.

Problem 2

 # Problem 2 (a)

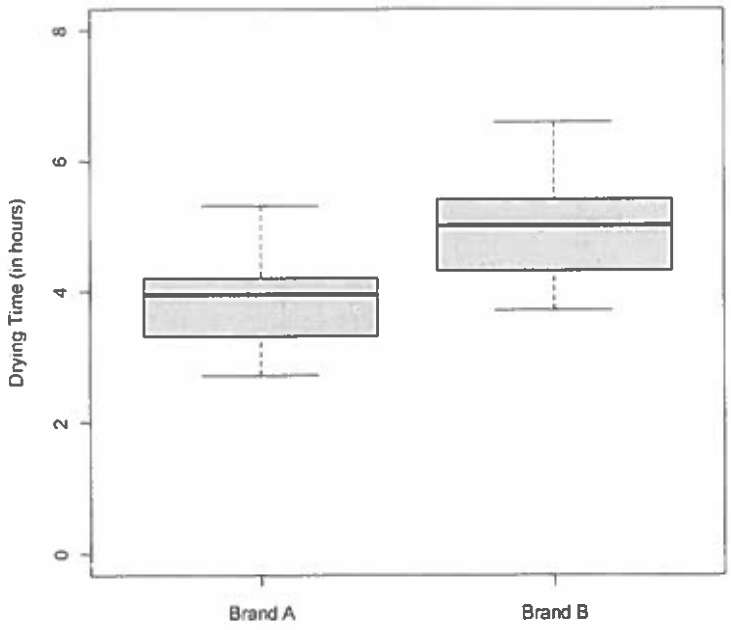
This problem involves two populations. The first population includes Brand A paint specimens; the second population includes Brand B paint specimens. The goal is to compare the **population mean drying times** for the two brands. We have two **independent** samples, one from each population.

Here is the R code I used to do this problem.

```
brand.a = c(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 = c(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)
```

The first thing is to examine the data visually; I will use side-by-side boxplots as I do in the notes.

```
# Side-by-side boxplots
boxplot(brand.a,brand.b,xlab="",names=c("Brand A","Brand B"),
      ylab="Drying Time (in hours)",ylim=c(0,8),col="grey")
```



There is nothing in the plot that says we should worry about an equal variance assumption. However, to compare the population mean drying times, I will construct both intervals: (a) one that assumes equal population variances and (b) one that does not. I will assume a 95 percent confidence level.

```
> # Confidence interval: Equal variance assumption
> t.test(brand.a,brand.b,conf.level=0.95,var.equal=TRUE)$conf.int
[1] -1.5902383 -0.6370344
> # Confidence interval: Unequal variance assumption
> t.test(brand.a,brand.b,conf.level=0.95,var.equal=FALSE)$conf.int
[1] -1.5902623 -0.6370105
```

The overall conclusion will be the same regardless of which assumption is made on the population variances (equal variance assumption/unequal variance assumption). Furthermore, the intervals themselves are nearly identical.

Interpretation for both intervals: We are 95 percent confident that the difference in the population mean drying times $\mu_A - \mu_B$ is between -1.59 and -0.64 hours. Because this interval contains only negative values, this suggests that the population mean drying time for Brand A is smaller than the population mean drying time for Brand B.

Just out of curiosity, I decided to write a 95 percent confidence interval for the ratio of the population variances. Remember that I wrote an R function to calculate this interval:

```
ratio.var.interval = function(data.1,data.2,conf.level=0.95){
df.1 = length(data.1)-1
df.2 = length(data.2)-1
F.lower = qf((1-conf.level)/2,df.1,df.2)
```

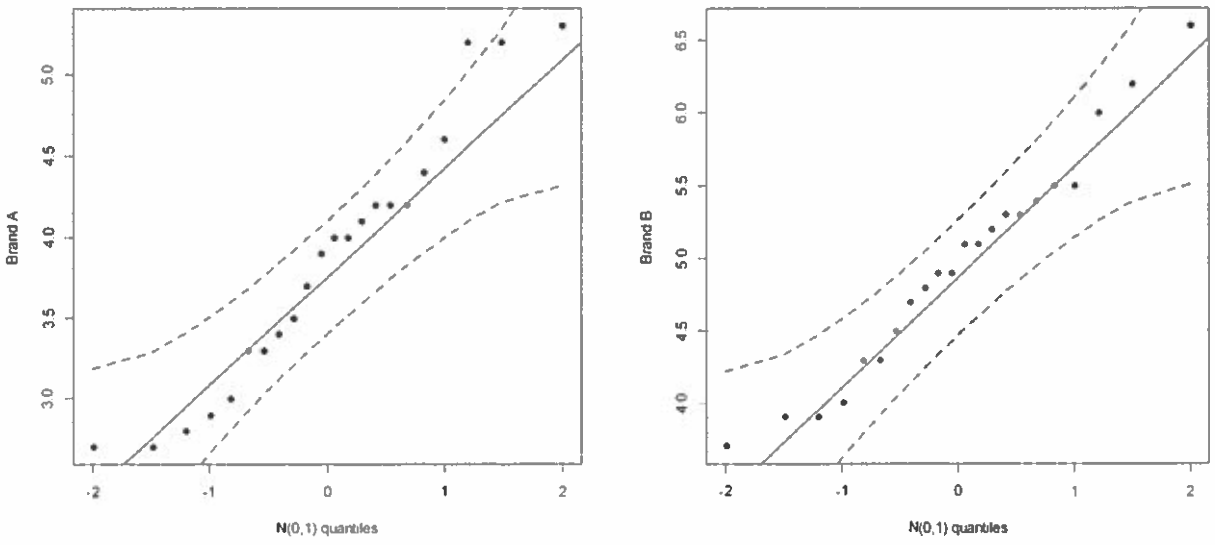
```
F.upper = qf((1+conf.level)/2,df.1,df.2)
s2.1 = var(data.1)
s2.2 = var(data.2)
c((s2.2/s2.1)*F.lower,(s2.2/s2.1)*F.upper)
}

> # CI for ratio of population variances
> ratio.var.interval(brand.a,brand.b)
[1] 0.3823315 2.2180210
```

No surprise here. The confidence interval (0.38, 2.22) does in fact contain "1." Note that because the two intervals above are nearly identical, a good argument can be made that this constructing this interval is not really needed (especially because its strict interpretation depends critically on normality).

Finally, we should check the normality assumption on both samples. Of course, we know that t confidence intervals for means are robust to normality departures (our variance-ratio interval is not). Furthermore, it is hard to get definitive conclusions on normality/not with small samples (22 in each sample). However, let's see if there are any red flags.

```
# QQ plots to check normality assumption
library(car)
qqPlot(brand.a,distribution="norm",xlab="N(0,1) quantiles",
     ylab="Brand A",pch=16)
qqPlot(brand.b,distribution="norm",xlab="N(0,1) quantiles",
     ylab="Brand B",pch=16)
```

Brand A is on the left; Brand B is on the right; see next page. I put these "side-by-side" manually in Word.



There are no major red flags here. Again, it is difficult to make such a clear-cut assessment with small samples. Remember that t procedures for population means are robust to normality departures anyway.

Overall conclusions: Our analysis suggests that there is a difference in the population mean drying times between Brand A and Brand B paints. The population mean drying time for Brand A is 0.64 to 1.59 hours shorter than that for the Brand B paint. There are no large concerns with the underlying statistical assumptions in making this claim.

 # Problem 2 (b)

The independent sample analysis assumes that the two samples are independent, which would be appropriate if the paints were applied to different pieces of material (e.g., 22 pieces of wood for Paint A; 22 pieces of wood for Paint B). In this instance, any extra physical differences in the wood materials could increase the amount of variability in the analysis.

A matched pairs design could proceed as follows:

- Select 22 pieces of wood (each piece needs to be sufficiently large)
- On one side of the wood, apply Brand A paint. On the other side, apply Brand B paint. The side that receives each brand should be determined at random so that there is no bias introduced (e.g., if one side of the wood consistently has different properties than the other side).
- Measure the drying time for each paint on each piece of wood as in the two-independent sample design.

When might this be preferred? If the wood materials are highly variable (e.g., wood taken from different piles, manufacturers, different types of wood), then a matched-pairs design would eliminate this variability. This variability cannot be "blocked out" in a two-independent sample design. If the wood materials themselves are perfectly homogenous to begin with (or nearly homogenous), then it would not matter if we used a two-independent sample design or a matched-pairs design. Of course, a matched pairs study may be cheaper because we need fewer pieces of wood to paint.

Problem 3

 # Problem 3 (a)

For treatment purposes, it might be important to know how TCDD is distributed in different parts of the body. Also, plasma materials move more freely throughout the body, whereas fat tissue is more stationary. Therefore, understanding population level differences can help us discern how to treat TCDD exposure more thoroughly; e.g., should treatment be more targeted or more globally applied?

 # Problem 3 (b)

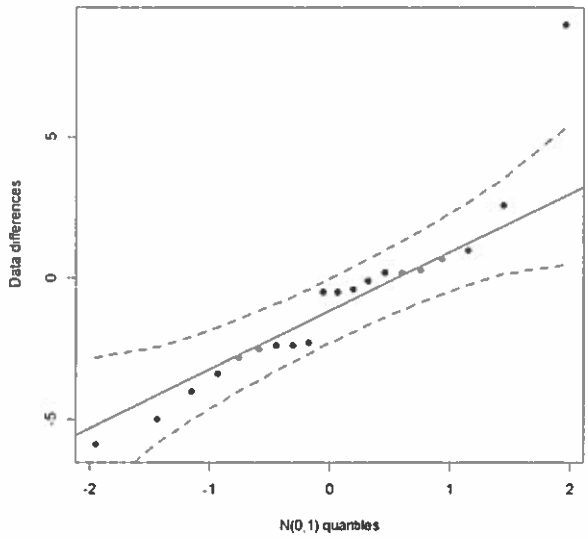
```
# Enter data
plasma =
c(2.5,3.1,2.1,3.5,3.1,1.8,6.0,3.0,36.0,4.7,6.9,3.9,4.2,1.6,7.2,1.8,
     20.0,2.0,2.5,4.1)
fat = c(4.9,5.9,4.4,6.9,9.0,4.2,10.0,5.5,41.0,4.4,7.0,2.9,4.6,1.4,
     7.7,1.1,11.0,2.5,2.3,1.5)
# Create data differences (for matched pairs analysis)
diff = plasma-fat
```

Here are the results from writing a **matched pairs** confidence interval for $\mu_1 - \mu_2$, where μ_1 is the population mean TCDD level in plasma and μ_2 is the population mean TCDD level in fat tissue:

```
> t.test(diff,conf.level=0.95)$conf.int
[1] -2.3969777  0.5769777
```

Interpretation: We are 95 percent confident that the difference in the population mean TCDD levels is between -2.40 and 0.58. Because this interval does include "0," we cannot conclude that the population mean TCDD level in plasma is different than the population mean TCDD level in fat tissue.

Assumptions: This analysis assumes that the 20 MA veterans form a random sample from a larger population (e.g., all US veterans? all veterans in MA?, etc.). Secondly, this analysis does require the data differences (TCDD level in plasma minus TCDD level in fat tissue) to be normally distributed. Remember, though, that matched pairs confidence intervals (based on the t distribution are robust to mild normality departures). Here is a qq-plot for the data differences, constructed under a normal assumption:



There is one obvious outlying observation in the sample (Veteran 17, whose difference is 9). The first thing I would do is make sure that this individual's observed concentration levels (20 for plasma; 11 for fat tissue were correctly recorded). Even just one or two outliers can distort the entire analysis, especially in studies with a small number of individuals.

Overall conclusions: At the 95 percent confidence level, we cannot conclude that the mean TCDD exposure levels are different in veterans' plasma and fat tissue at the population level. (Provided that Veteran 17's measurements are legitimate), there are no large concerns with the underlying statistical assumptions in making this claim.

```
# Problem 3(c)
```
The main advantage of the matched pairs design is that you remove variation associated with the veterans being biologically different. Because two measurements are obtained on the same individual, you are comparing plasma TCDD levels to fat tissue TCDD levels without adding in extra variation that makes one veteran different from another veteran. Another advantage is that, in general, fewer individuals (veterans) are needed to run the experiment and carry out the analysis. The only possible disadvantage that I can think of here is that, because two measurements are needed from each veteran, we could be subjecting them to more discomfort by assaying them twice (once for plasma, once for fat tissue). This could require more time from them as patients in the study. Usually, however, study protocols clearly will describe the procedures involved for the subjects (before they consent), so this "disadvantage" might be minimal, at best.

Problem 4

 # Problem 4 (a)

Here, we want to construct a confidence interval for the difference of the two population proportions:

- p_1 = population proportion of exceedences for SAS
- p_2 = population proportion of exceedences for Lufthanza.

Our **sample proportions** are 8/86 for SAS and 10/142 for Lufthanza. We assume that the two samples (one from SAS; the other from Lufthanza) are independent. Recall that I wrote an R function to calculate the confidence interval for p_1-p_2.

```
proportion.diff.interval = function(y.1,n.1,y.2,n.2,conf.level=0.95){
      z.upper = qnorm((1+conf.level)/2)
      var.1 = (y.1/n.1)*(1-y.1/n.1)/n.1
      var.2 = (y.2/n.2)*(1-y.2/n.2)/n.2
      se = sqrt(var.1+var.2)
      moe = z.upper*se
      c((y.1/n.1-y.2/n.2)-moe,(y.1/n.1-y.2/n.2)+moe)
      }
```

Here is the 95 percent confidence interval:

```
> proportion.diff.interval(8,86,10,142)
[1] -0.05182771  0.09702915
```

Interpretation: We are 95 percent confident that the difference of the population proportion of exceedences between the two airlines is between -0.05 and 0.10. Because this interval includes "0," we cannot conclude that the two airlines commit runway exceedences with different proportions at the population level.

Remark: It is easy to see that the Rules of Thumb here are verified. The number of "successes" for SAS and Lufthanza are 8 and 10, respectively, both of which exceed 5. The number of "failures" are 78 and 132, respectively, both of which exceed 5.

 # Problem 4 (b)

For a 99 percent confidence level, we want to find the 99.5th percentile of a standard normal distribution; i.e., $z_{\alpha/2} = z_{0.01/2} = z_{0.005}$:

```
> qnorm(0.995,0,1)
[1] 2.575829
```

4. (b) $$Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}=0.03$$

$$2.575829\sqrt{\frac{\frac{8}{86}(1-\frac{8}{86})}{n}+\frac{\frac{10}{142}(1-\frac{10}{142})}{n}}=0.03$$

$$n=\frac{\frac{8}{86}(1-\frac{8}{86})+\frac{10}{142}(1-\frac{10}{142})}{\left(\frac{0.03}{2.575829}\right)^2}$$

$$=1104.586$$