Homework 10

S320/520

Upload your answers as a PDF file or Word document through the Assignments tab on Canvas by 4pm, Thursday 19th November.

Note: Answers should be in your own words. Answers that are in the lecturer's words will not receive credit, and answers that are in the lecturer's words and punctuation may receive negative credit.

1. (10 points) Trosset chapter 12.6 problem set A

   (a)
   ```
   all.data = scan("http://mypage.iu.edu/~mtrosset/StatInfeR/Data/salinity.dat")
   x1 = all.data[1:12]
   x2 = all.data[13:20]
   x3 = all.data[21:30]
   boxplot(x1, x2, x3)
   qqnorm(x1)
   qqnorm(x2)
   qqnorm(x3)
   ```
   The plots reveal two outliers, one in each of sites A and C. Otherwise, the ANOVA assumptions seem reasonable.

   (b)
   ```
   n1 = 12
   n2 = 8
   n3 = 10
   grand.mean = mean(all.data)
   mean1 = mean(x1)
   mean2 = mean(x2)
   mean3 = mean(x3)
   SSB = n1*(mean1-grand.mean)^2 + n2*(mean2-grand.mean)^2 +
    n3*(mean3-grand.mean)^2
   SSW = (n1-1)*var(x1) + (n2-1)*var(x2) + (n3-1)*var(x3)
   SST = SSB + SSW
   between.meansquare = SSB/2
   within.meansquare = SSW/27
   F = between.meansquare / within.meansquare
   1 - pf(F, df1=2, df2=27)
   ```
   Since the P-value is minuscule, we reject the null hypothesis. The three sites do not all have the same salinity. (Further analysis, such as pairwise Welch t-tests, would confirm that all three locations are different from each other.)

Variation	Sum of squares	DF	Mean square	F-statistic	P-value
Between	38.80	2	19.40	66.0	4×10^{-11}
Within	7.93	27	0.294		
Total	46.73	29			

Table 1: ANOVA table to test hypothesis that three sites have the same salinity.

2. (10 points) Trosset chapter 12.6 problem set C. Note: Both times the question says $\overrightarrow{y}_1, \ldots, \overrightarrow{x}_5$, it should be $\overrightarrow{y}_1, \ldots, \overrightarrow{y}_5$.

   (a)
```
x1 = c(124,42,25,45,412,51,1112,46,103,876,146,340,396)
x2 = c(81,461,20,450,246,166,63,64,155,859,151,166,37,223,138,72,245)
x3 = c(248,377,189,1843,180,537,519,455,406,365,942,776,372,163,101,20,283)
x4 = c(1234,89,201,356,2970,456)
x5 = c(1235,24,1581,1166,40,727,3808,791,1804,3460,719)
all.data = c(x1,x2,x3,x4,x5)
y1 = log(x1)
y2 = log(x2)
y3 = log(x3)
y4 = log(x4)
y5 = log(x5)
boxplot(x1,x2,x3,x4,x5, main="Boxplots of survival data",
 ylab="Survival time in days")
boxplot(y1,y2,y3,y4,y5, main="Boxplots of logged survival data",
 ylab="Log of survival time in days")
```

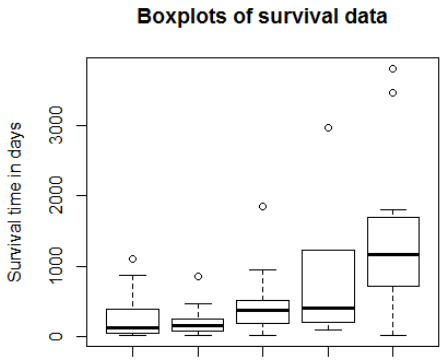

Figure 1: Problem Set C, Q1: Side-by-side boxplots of survival data.

The boxplots of the raw survival times are skewed and have vastly differing spreads, so the ANOVA assumptions are not satisfied. The boxplots of the logged survival times are reasonably symmetric and have spreads that are reasonably close given the small samples,

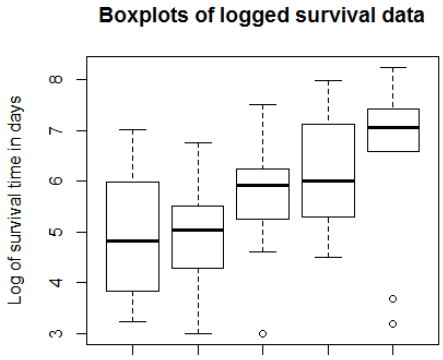

Figure 2: Problem Set C, Q1: Side-by-side boxplots of logged survival data.

though there are a few outliers (low survival times for one colon and two breast cancer cases). Still, it's at least plausible that the ANOVA assumptions are approximately satisfied.

(b)
```
n1 = length(y1)
n2 = length(y2)
n3 = length(y3)
n4 = length(y4)
n5 = length(y5)
grand.mean = mean(log(all.data))
mean1 = mean(y1)
mean2 = mean(y2)
mean3 = mean(y3)
mean4 = mean(y4)
mean5 = mean(y5)
SSB = n1*(mean1-grand.mean)^2 + n2*(mean2-grand.mean)^2 +
 n3*(mean3-grand.mean)^2 + n4*(mean4-grand.mean)^2 +
 n5*(mean5-grand.mean)^2
SSW = (n1-1)*var(y1) + (n2-1)*var(y2) + (n3-1)*var(y3) +
 (n4-1)*var(y4) + (n5-1)*var(y5)
SST = SSB + SSW
between.meansquare = SSB/4
within.meansquare = SSW/59
F = between.meansquare / within.meansquare
1 - pf(F, df1=4, df2=59)
```

Let X_{ij} be the survival time in days of patient j in sample i. Let $Y_{ij} = \log X_{ij}$. Let $\mu_i = EY_{ij}$. We test the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$. The P-value is 0.004, so we reject this null. The mean log survival times differs depends on the type of cancer.

Variation	Sum of squares	DF	Mean square	F-statistic	P-value
Between	24.49	4	6.12	4.29	0.004
Within	84.27	59	1.43		
Total	108.76	63			

Table 2: ANOVA table for log survival times of cancer patients treated with ascorbate.

3. (5 points) *In theory, the ANOVA F-test can be used when there are only two samples if the normality and homoscedasticity assumptions are met. However, it's better to use Welch's t-test because it has weaker assumptions (e.g. variances may be different) while the F-test is sensitive to violations of its assumptions when there are only two samples.*

   Recall for the untransformed stereogram data on Canvas (`stereograms.txt` in the Data folder of Files):

```
> stereograms = read.table(file.choose(), header=TRUE)
> treatment = stereograms$time[stereograms$group==2]
> control = stereograms$time[stereograms$group==1]
> t.test(treatment, control)

Welch Two Sample t-test

data:  treatment and control
t = -2.0384, df = 70.039, p-value = 0.04529
alternative hypothesis: true difference in means is not equal to 0

> t.test(treatment, control, var.equal=T)

Two Sample t-test

data:  treatment and control
t = -1.9395, df = 76, p-value = 0.05615
alternative hypothesis: true difference in means is not equal to 0
```

   So Welch's test gives a P-value of 0.04529, while Student's test gives a P-value of 0.05615.

   Find a P-value for the F-test of the null hypothesis that the treatment and control populations have the same mean. How does this compare to the Welch and Student P-values?

   Using the variables `treatment` and `control` defined above,

```
n1 = length(treatment)
n2 = length(control)
all.data = c(treatment, control)
grand.mean = mean(all.data)
mean1 = mean(treatment)
mean2 = mean(control)
```

```
SSB = n1*(mean1-grand.mean)^2 + n2*(mean2-grand.mean)^2
SSW = (n1-1) * var(treatment) + (n2-1) * var(control)
between.meansquare = SSB / 1
within.meansquare = SSW / (n1 + n2 - 2)
F = between.meansquare / within.meansquare
1 - pf(F, df1 = 1, df2 = n1 + n2 - 2)
```

This gives a P-value of 0.05615, exactly the same P-value as Student's test gives. The two tests are equivalent in this sense. This is bad! We've noted that Student's is not very good in this situation, when the normal assumption is violated and the variances are unequal. Therefore we should *not* use the F-test when there are only two samples.