Homework 10

S520

Upload your answers as a PDF or HTML file through the Assignments tab on Canvas by 4 pm, Thursday 14th April.

Trosset question numbers refer to the hardcover textbook. Draw all graphs in R and include all R code. You may work with others, but you must write up your homework independently — you should not have whole sentences in common with other students or other sources.

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

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

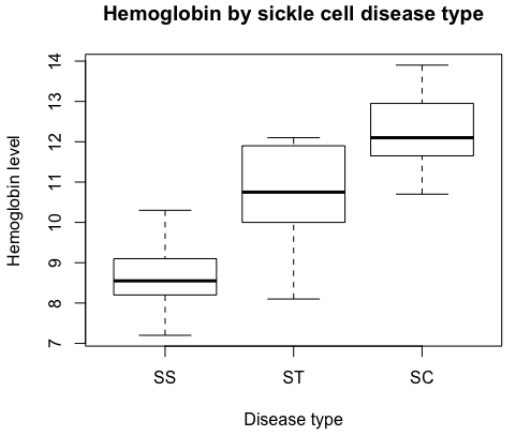
   (b)
   ```
   n1 = length(x1)
   n2 = length(x2)
   n3 = length(x3)
   grand.mean = mean(salinity)
   mean1 = mean(x1)
   mean2 = mean(x2)
   mean3 = mean(x3)
   SSB = n1*(mean1-grand.mean)^2 + n2*(mean2-grand.mean)^2 +
    n3*(mean3-grand.mean)^2
   SSW = (n1-1)*var(x1) + (n2-1)*var(x2) + (n3-1)*var(x3)
   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.

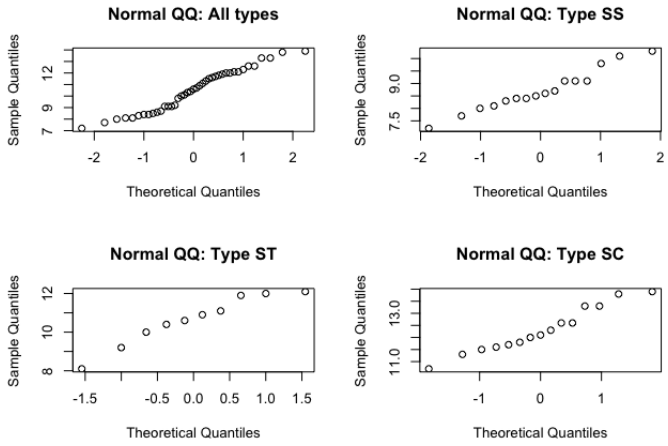
2. (10 points.) Trosset chapter 12.6 problem set B

(a)
```
hemo = scan("http://mypage.iu.edu/~mtrosset/StatInfeR/Data/sickle.dat")
SS = hemo[1:16]
ST = hemo[17:26]
SC = hemo[27:41]
boxplot(SS, ST, SC, names=c("SS","ST","SC"),
  xlab="Disease type", ylab="Hemoglobin level",
  main="Hemoglobin by sickle cell disease type")
par(mfrow=c(2,2))
qqnorm(hemo, main="Normal QQ: All types")
qqnorm(SS, main="Normal QQ: Type SS")
qqnorm(ST, main="Normal QQ: Type ST")
qqnorm(SC, main="Normal QQ: Type SC")
```



The boxplots show a little variation in spread — the range of ST is a little wider than the others — but this isn't a big enough difference to matter. The QQ plots are all reasonably close to straight lines. The ANOVA assumptions seem reasonable.

(b)
```
n1 = length(SS)
n2 = length(ST)
n3 = length(SC)
grand.mean = mean(hemo)
mean.SS = mean(SS)
```




```
mean.ST = mean(ST)
mean.SC = mean(SC)
SSB = n1*(mean.SS-grand.mean)^2 + n2*(mean.ST-grand.mean)^2 +
 n3*(mean.SC-grand.mean)^2
SSW = (n1-1)*var(SS) + (n2-1)*var(ST) + (n3-1)*var(SC)
SST = SSB + SSW
between.meansquare = SSB/2
within.meansquare = SSW/38
F = between.meansquare / within.meansquare
1 - pf(F, df1=2, df2=38)
```

Variation	Sum of squares	DF	Mean square	F-statistic	P-value
Between	99.89	2	49.94	50.0	2×10^{-11}
Within	37.96	38	0.999		
Total	137.8	40			

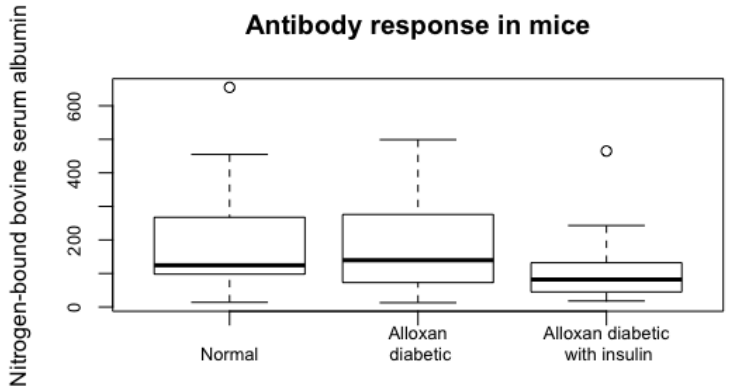
Table 2: ANOVA table to test null hypothesis that mean hemoglobin is the same for three types of sickle cell disease.

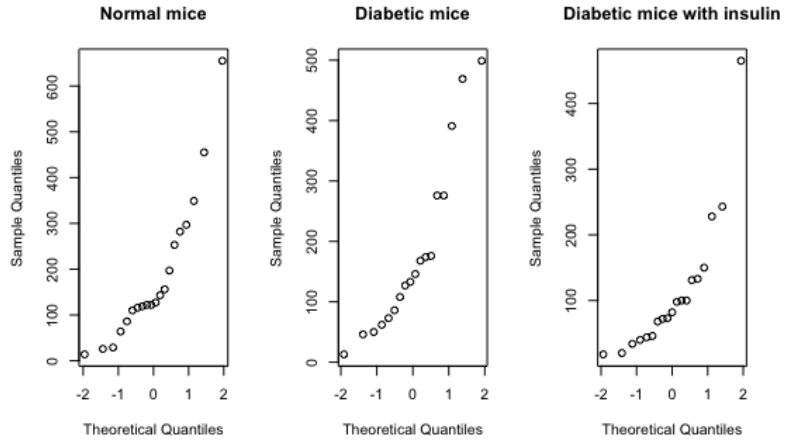
The P-value is minuscule, so we reject the null hypothesis. The three types of sickle cell disease do not have the same average hemoglobin level. From further tests or just looking at the boxplots, SS is the lowest, ST is in the middle, and SC is the highest. (The lower the hemoglobin, the worse the sickle cell disease.)

3. (10 points.) Trosset chapter 12.6 problem set G

   (a)
```
mice = scan("http://mypage.iu.edu/~mtrosset/StatInfeR/Data/mice.dat")
normal = mice[1:20]
alloxan = mice[21:38]
insulin = mice[39:57]
boxplot(normal, alloxan, insulin,
```

```
  main="Antibody response in mice",
  ylab="Nitrogen-bound bovine serum albumin",
  names=c("Normal","Alloxan\n diabetic","Alloxan diabetic\n with insulin"),
  cex.axis=0.8)
par(mfrow=c(1,3))
qqnorm(normal, main="Normal mice")
qqnorm(alloxan, main="Diabetic mice")
qqnorm(insulin, main="Diabetic mice with insulin")
```

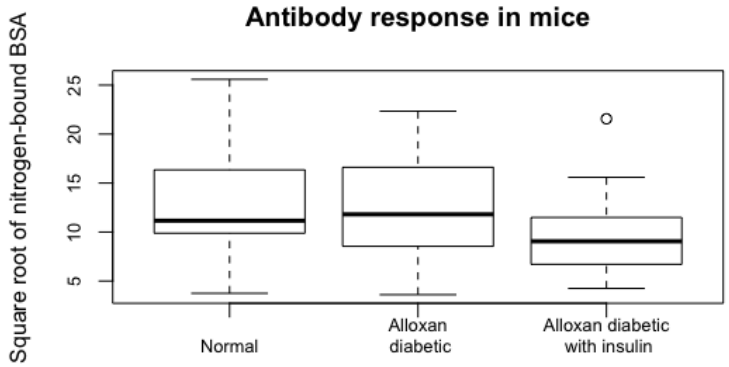


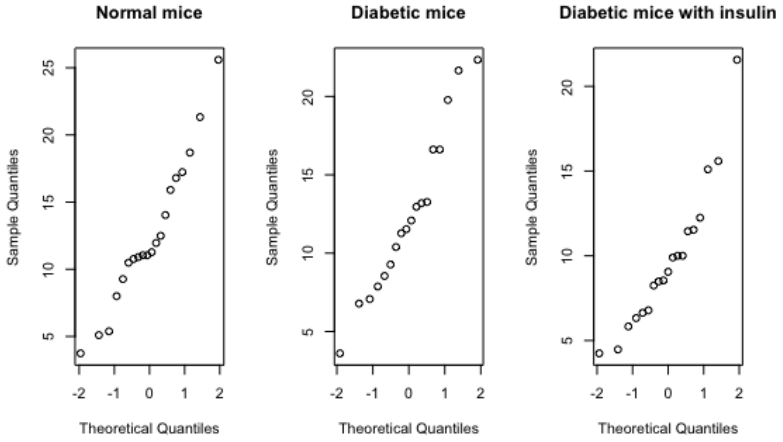


The boxplot spreads vary greatly and the QQ plots are curved. As the samples sizes are small, we can't blithely ignore the ANOVA assumptions.

(b)
```
normal.sqrt = sqrt(normal)
alloxan.sqrt = sqrt(alloxan)
insulin.sqrt = sqrt(insulin)
par(mfrow=c(1,1))
```

```
boxplot(normal.sqrt, alloxan.sqrt, insulin.sqrt,
  main="Antibody response in mice",
  ylab="Square root of nitrogen-bound BSA",
  names=c("Normal","Alloxan\n diabetic","Alloxan diabetic\n with insulin"),
  cex.axis=0.8)
par(mfrow=c(1,3))
qqnorm(normal.sqrt, main="Normal mice")
qqnorm(alloxan.sqrt, main="Diabetic mice")
qqnorm(insulin.sqrt, main="Diabetic mice with insulin")
# Check the SDs as well
c(sd(normal.sqrt), sd(alloxan.sqrt), sd(insulin.sqrt))
```





The spreads are now broadly similar, though there's one minor outlier in the insulin group. The QQ plots for the normal and diabetic mice look straight. The QQ plot for the insulin group is slightly curved, but it's a minor concern. The ANOVA assumptions may not be literally true, but the data isn't bad enough to stop us from doing the t-test

(though we might put an asterisk by our P-value at the end.)

(c) Let μ_n be the (population) mean of the square root nitrogen-bound BSA for non-diabetic mice, μ_a be the mean square root nitrogen-bound BSA for alloxan diabetic mice, and μ_i be the mean square root nitrogen-bound BSA for alloxan diabetic mice treated with insulin. The null hypothesis is that $\mu_n = \mu_a = \mu_i$, while the alternative is that at least one of the μs is different.

```
\begin{verbatim}
n1 = length(normal.sqrt)
n2 = length(alloxan.sqrt)
n3 = length(insulin.sqrt)
N = n1 + n2 + n3
grand.mean = mean(sqrt(mice))
mean.n = mean(normal.sqrt)
mean.a = mean(alloxan.sqrt)
mean.i = mean(insulin.sqrt)
SSB = n1*(mean.n-grand.mean)^2 + n2*(mean.a-grand.mean)^2 +
 n3*(mean.i-grand.mean)^2
SSW = (n1-1)*var(normal.sqrt) + (n2-1)*var(alloxan.sqrt) + (n3-1)*var(insulin.sqrt)
SST = SSB + SSW
between.ms = SSB/2
within.ms = SSW/(N-3)
F = between.ms / within.ms
1 - pf(F, df1=2, df2=N-3)
```

Variation	Sum of squares	DF	Mean square	F-statistic	P-value
Between	94.74	2	47.37	1.88	0.16
Within	1359	54	25.18		
Total	1454	57			

Table 3: ANOVA table for antibody responses in non-diabetic mice, diabetic mice, and diabetic mice treated with insulin.

The P-value is 0.16, so we do not reject the null hypothesis. There's no proof of a difference in immune responses between the three groups of mice, though we should be cautious about drawing conclusions that are too strong when the sample sizes are small (and the assumptions might be iffy.)

(d) There are several ways to approach this (the book outlines some of them), but the most straightforward thing to do is just do three Welch t-tests. Using μ_n, μ_a, and μ_i as defined above, the hypotheses of interest are:

- $H_0 : \mu_n - \mu_a = 0$ vs. $H_1 : \mu_n - \mu_a \neq 0$
- $H_0 : \mu_a - \mu_i = 0$ vs. $H_1 : \mu_a - \mu_i \neq 0$
- $H_0 : \mu_n - \mu_i = 0$ vs. $H_1 : \mu_n - \mu_i \neq 0$

Note that in each case, it doesn't matter which way around we do the subtraction. (We could also write these down as contrast vectors but that's not especially interesting here.)

```
t.test(normal.sqrt, alloxan.sqrt)
t.test(alloxan.sqrt, insulin.sqrt)
t.test(normal.sqrt, insulin.sqrt)
```

"Bonferroni" just means we should compare our P-values to $\alpha/3$ rather than α. In this case, the P-values are 0.97, 0.09, and 0.09 respectively, none of which are close to $\alpha/3$. The data is consistent with no difference among the three pairs, though as usual small samples don't prove the null.

If instead you use the method of Trosset 12.3.2:

```
Tna = (mean(normal.sqrt) - mean(alloxan.sqrt)) / sqrt((1/n1 + 1/n2) * within.ms)
2 * (1 - pt(abs(Tna), df = N-3))
Tai = (mean(alloxan.sqrt) - mean(insulin.sqrt)) / sqrt((1/n2 + 1/n3) * within.ms)
2 * (1 - pt(abs(Tai), df = N-3))
Tni = (mean(normal.sqrt) - mean(insulin.sqrt)) / sqrt((1/n1 + 1/n3) * within.ms)
2 * (1 - pt(abs(Tni), df = N-3))
```

The P-values are 0.97, 0.11, and 0.09. The difference is pretty negligible.