

Homework 9 answers

S320/520

Upload your answers as a PDF file or Word document through the Assignments tab on Canvas by 4pm, Thursday 5th November.

Note: Answers should be in your own words. Answers that are the lecturer's words will not receive credit.

This is a long homework! Start early!

1 (5 points.) Trosset chapter 11.4 problem set B part 3

1. The experimental unit is an aerobics student.
2. There's one population, aerobics students. There is one independent sample.
3. There are two measurements taken on each aerobics student. Let D_i denote student i 's watts expended after dynamics stretching and let S_i denote student i 's watts expended after static stretching.
4. $\mu = EX_i$, where $X_i = D_i - S_i$ is the amount by which the number of watts expended after dynamic stretching exceeds the number of watts expended after static stretching.
5. Irmina would like to demonstrate that $\mu > 0$, so she tests $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$.

2 (10 points.) Trosset chapter 11.4 problem set C parts 1–3

Data: <http://mypage.iu.edu/~mtrosset/StatInfer/Data/cholesterol.dat>

1. (a) The experimental unit is a person.
(b) There are two populations of heavy middle-aged men, those with Type A behavior and those with Type B behavior. This is a 2-sample problem.
(c) There is one measurement taken on each person: Cholesterol level. Let X_i denote the cholesterol level of person i in the first sample and let Y_j denote the cholesterol of person j in the second sample.
(d) $\mu_1 = EX_i, \mu_2 = EY_j, \Delta = \mu_1 - \mu_2$.
(e) The theory is that $\mu_1 > \mu_2$, i.e. $\Delta > 0$, so we test $H_0 : \Delta \leq 0$ vs. $H_1 : \Delta > 0$.

```

2. data = scan("http://mypage.iu.edu/~mtrosset/StatInfer/Data/cholesterol.dat")
   TypeA = data[1:20]
   TypeB = data[21:40]
   qqnorm(TypeA)
   qqnorm(TypeB)

```

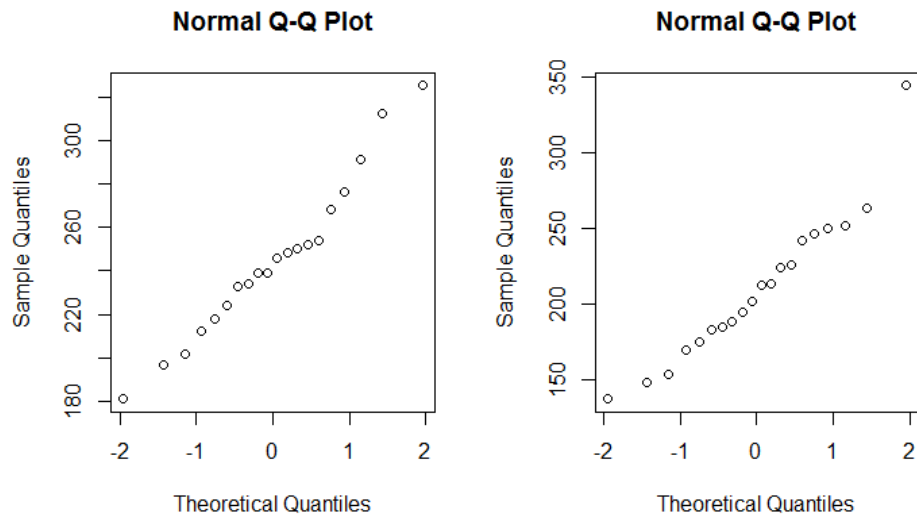


Figure 1: Normal quantile plots of cholesterol levels for Type A men (left) and Type B men (right).

As is often the case with real data, it's borderline. The normal quantile plot for Type A men is reasonably straight, though there's a slightly worrying kink at 254. The normal quantile plot for Type B men has a pretty bad outlier (344) that's hard to reconcile with a normal distribution. Both samples are probably close enough to normal to justify doing Welch's test, particularly as neither sample size is tiny. (If you were concerned about this assumption, you might instead do a Wilcoxon rank-sum test, which is resistant to outliers.)

```

3. (a) Delta.hat = mean(TypeA) - mean(TypeB)
      std.error = sqrt(var(TypeA)/20 + var(TypeB)/20)
      Tw = Delta.hat / std.error
      df = (var(TypeA)/20+var(TypeB)/20)^2 / ((var(TypeA)/20)^2/19 + (var(TypeB)/20)^2/19)
      1 - pt(Tw, df=df)

```

We compare a test statistic of 2.56 to a t -distribution with 35.4 degrees of freedom, and get a P -value of 0.007. This is smaller than α , so we reject the null hypothesis. At least for individuals like those in the study, Type A men do, on average, have higher cholesterol than Type B men.

```

(b) Delta.hat - qt(0.95, df=df) * std.error
     Delta.hat + qt(0.95, df=df) * std.error

```

The 90% confidence interval is (11.8, 57.7). That is, we're 90% confident that Type A men have higher average cholesterol by between 12 and 60 units.

3 (10 points.) Trosset chapter 11.4 problem set D.

Data: <http://mypage.iu.edu/~mtrosset/StatInfeR/Data/globulin.dat>

```
1. data = scan("http://mypage.iu.edu/~mtrosset/StatInfeR/Data/globulin.dat")
   normal = data[1:10]
   diabetic = data[11:20]
   par(mfrow=c(2,2))
   hist(normal, xlab="Thromboglobulin")
   hist(diabetic, xlab="Thromboglobulin")
   plot(density(normal, adjust=0.5), main="Density plot for controls",
        xlab="Thromboglobulin")
   plot(density(diabetic, adjust=0.5), main="Density plot for diabetics",
        xlab="Thromboglobulin")
```

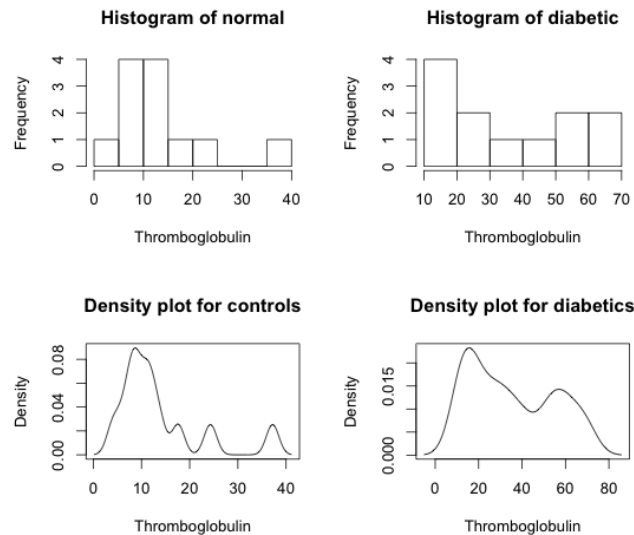


Figure 2: Problem Set D, Q1: Histograms and density plots to assess symmetry.

As usual, we can't be entirely sure from small samples, but there's some evidence of right skew from histograms and density plots.

```
2. plot(density(log(normal), adjust=0.5), main="Density plot for controls",
       xlab="Log of thromboglobulin")
   plot(density(log(diabetic), adjust=0.5), main="Density plot for diabetics",
       xlab="Log of thromboglobulin")
   plot(density(sqrt(normal), adjust=0.5), main="Density plot for controls",
       xlab="Sqrt of thromboglobulin")
   plot(density(sqrt(diabetic), adjust=0.5), main="Density plot for diabetics",
       xlab="Sqrt of thromboglobulin")
```

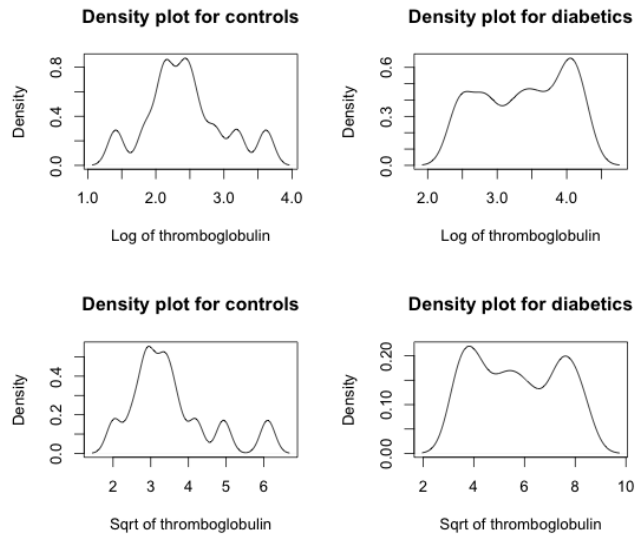


Figure 3: Problem Set D, Q2: Density plots to assess symmetry of transformed data.

It's a matter of judgment, but to me the log transformed data looks to be left-skewed, while the square root data is closest to symmetric. For that reason I prefer the square root transformation. (However, the log transformation would give results that are easier to interpret.)

3.

```
qqnorm(log(normal), main="Log of control data")
qqnorm(log(diabetic), main="Log of diabetic data")
qqnorm(sqrt(normal), main="Sqrt of control data")
qqnorm(sqrt(diabetic), main="Sqrt of diabetic data")
```

My answer to the question “Are the samples from normal distributions?” would be “Who knows?” Nevertheless, the normal QQ plots for the square root data look a little closer to straight lines.

4. One option is to do a Welch test on the square root data.

```
Delta = mean(sqrt(diabetic)) - mean(sqrt(normal))
s1 = sd(sqrt(diabetic))
s2 = sd(sqrt(normal))
n1 = 10
n2 = 10
std.error = sqrt(s1^2/n1 + s2^2/n2)
Tw = Delta/std.error
df = (s1^2/n1 + s2^2/n2)^2 / ((s1^2/n1)^2/(n1-1) + (s2^2/n2)^2/(n2-1))
2 * (1 - pt(Tw, df=df))
```

The (two-tailed) P -value is 0.0005, meaning that we reject the null hypothesis of no average difference. The evidence for the researchers' claim is strong. (A confidence interval isn't that useful here as it's hard to interpret a confidence interval for square root data.)

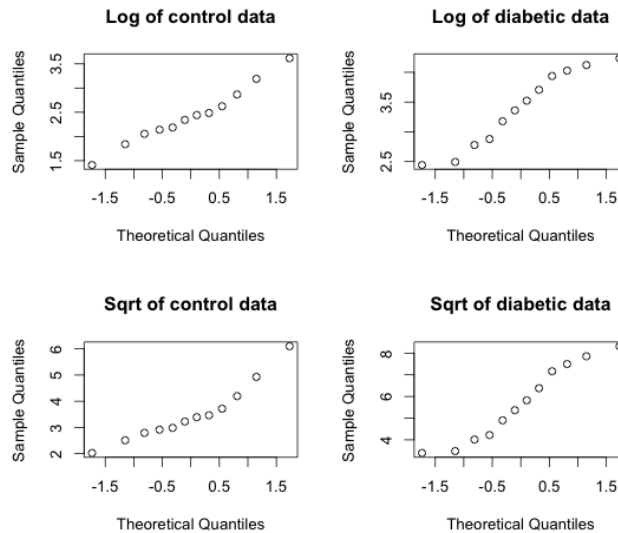


Figure 4: Problem Set D, Q3: QQ plots to assess normality of transformed data.

There are many possible choices here. One-tailed or two-tailed? It's not clear from the information given. Wilcoxon test? Sure, that works. Draw side-by-side boxplots and say it's obvious? Maybe, though the skewness means it might not completely obvious to statistical pedants.

4 (Compulsory for 5 points for S520, optional for 3 points for S320.) Trosset chapter 11.4 problem set E, part 3.

Data: <http://mypage.iu.edu/~mtrosset/StatInfer/Data/films.dat>

Note: For this question, you may use `t.test()` and similar high-level functions. You will be graded primarily on accurate statement of assumptions, explanation, and interpretation.

```
data = scan("http://mypage.iu.edu/~mtrosset/StatInfer/Data/films.dat")
movies1956 = data[1:14]
movies1996 = data[15:28]
par(mfrow=c(2,2))
hist(movies1956, main="1956 movie times", xlab="Length (minutes)")
hist(movies1996, main="1996 movie times", xlab="Length (minutes)")
qqnorm(movies1956, main="QQ plot for 1956 movies")
qqnorm(movies1996, main="QQ plot for 1996 movies")
```

It's unlikely to be true that either 1956 movie times or 1996 movie times follow the normal curve. In 1956 few movies were longer than two hours (aside from the “epics,” and they could go on for hours and hours). By the 1990s, there was an unwillingness to make movies shorter than about 85 minutes (the one outlier in the data set is an unfinished movie made on a budget of, you guessed it,

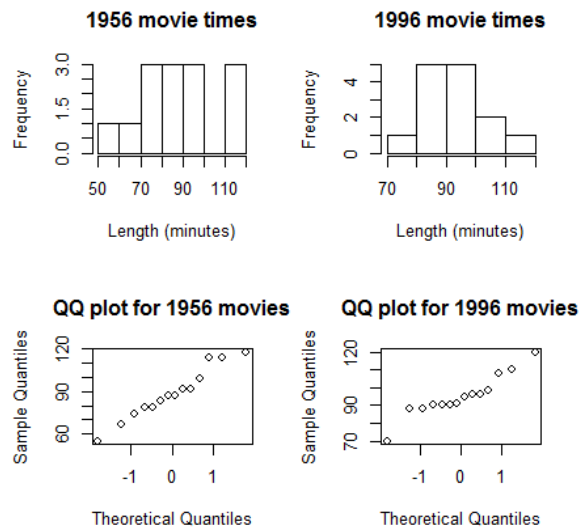


Figure 5: Problem Set E, Q3: Histograms and normal QQ plots of samples of 1956 and 1996 movie lengths.

\$40,000). As the two populations have different shapes, transformations are unlikely to work. The assumptions of Welch's two-sample t -test (two independent samples from normal populations) are unlikely to be satisfied. Nevertheless, since this isn't exactly a life-or-death issue, let's do Welch's two-sample t -test.

```
> t.test(movies1996, movies1956)

Welch Two Sample t-test

data:  movies1996 and movies1956
t = 1.105, df = 22.395, p-value = 0.2809
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.623821 18.480963
sample estimates:
mean of x mean of y
 95.00000  88.57143
```

The two-tailed P -value is 0.28, which means a one-tailed P -value would be 0.14. We wouldn't reject a null hypothesis that average movie lengths were the same (or shorter) in 1996 as in 1956. However, note the confidence interval is very wide (because the samples are small): the average runtime of 1996 movies could be anywhere from 6 minutes shorter to 18 minutes longer than the average runtime of 1956 movies. So really we don't have enough data to say.

As a sanity check, we can compare this to the result of a Wilcoxon rank-sum test:

```
> wilcox.test(movies1996, movies1956, conf.int=TRUE)
```

Wilcoxon rank sum test with continuity correction

```
data: movies1996 and movies1956
W = 126, p-value = 0.2058
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -4.000047 17.999975
sample estimates:
difference in location
      7.000011
```

It tells us much the same thing. We can't reject the null, but that might just be because our sample sizes are small.