Answers (Problem set 10)

Online S520

1 Trosset chapter 11.4 problem set C parts 1–3

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. 20 units were drawn from each population. 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)
```

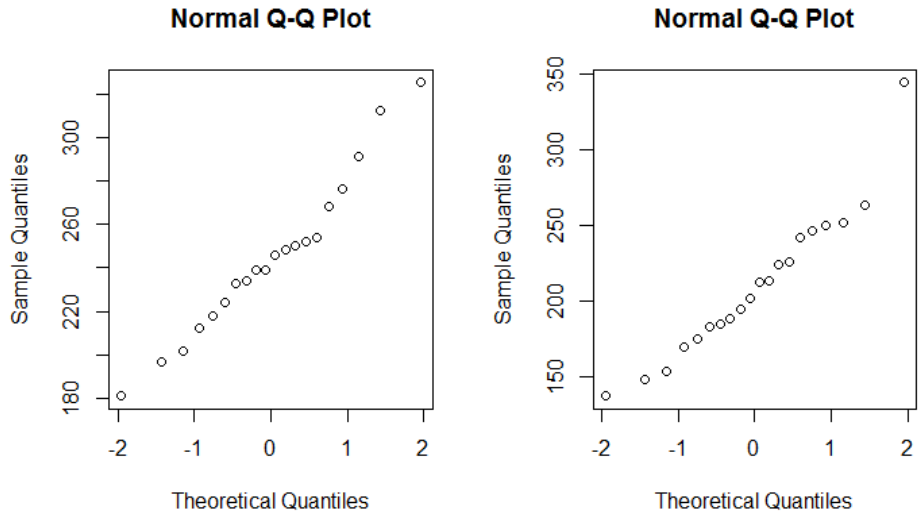

Figure 1: Normal quantile plots of cholesterol levels for Type A men (left) and Type B men (right).

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 an outlier (344) that's hard to reconcile with a normal distribution,

but isn't bad enough to totally wreck. Both samples are probably close enough to normal to justify doing Welch's test, though a skeptic would be happier with more data. (If you were very concerned about the normality assumption, you might take logs, but that would make things harder to interpret and is probably overkill here.)

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)
# or just do
# t.test(TypeA, TypeB, alt="greater")
```
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
```
or just do
```
t.test(TypeA, TypeB, conf.level =0.9)
```
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 58 units.

2 Trosset chapter 11.4 problem set D parts 1–4

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

1.
```
data = scan("http://mypage.iu.edu/~mtrosset/StatInfeR/Data/globulin.dat")
normal = data[1:12]
diabetic = data[13:24]
par(mfrow=c(2,2))
hist(normal, xlab="Thromboglobulin")
hist(diabetic, xlab="Thromboglobulin")
plot(density(normal), main="Density plot for controls",
xlab="Thromboglobulin")
plot(density(diabetic), main="Density plot for diabetics",
xlab="Thromboglobulin")
```
As usual, we can't be entirely sure from small samples, but from the histograms and density plots there's some evidence of right skew in the control group especially (even though it's mostly due to one outlier.)

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

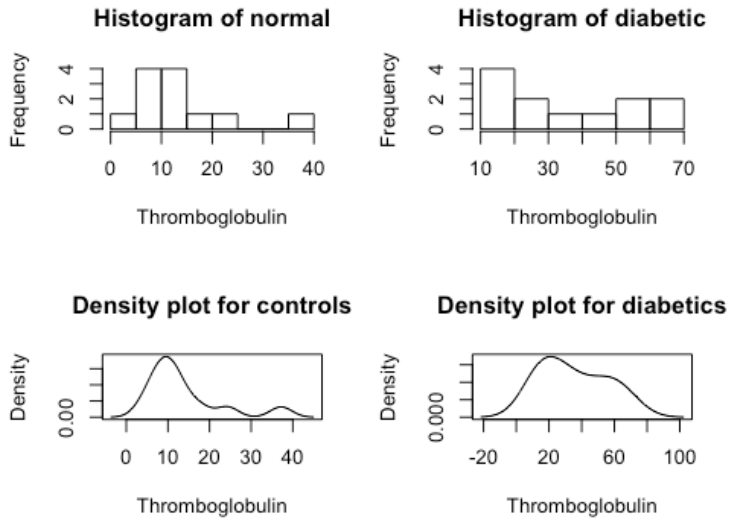

Figure 2: Problem Set D, Q1: Histograms and density plots to assess symmetry.

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

It's a matter of judgment, but to me the log transformed data looks close to symmetric, while the square root of the control data still looks right-skewed. This is a point in favor of a log transformation.

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")
```

   To the question "Are the samples from normal distributions?" I would be inclined to answer "Who knows?" Nevertheless, the normal QQ plots for the logged data look as straight as we could reasonably expect, while the QQ plot for the square root of the control data looks like it's bending upward a little more than I would like.

4. One option is to do a one-sided Welch test on the logged data (though you could also justify using the square root data.)

   $H_0 : \Delta \leq 0, H_1 : \Delta > 0$ where $\Delta = \mu_{diabetic} - \mu_{normal}$

```
Delta.hat = mean(log(diabetic)) - mean(log(normal))
s1 = sd(log(diabetic))
s2 = sd(log(normal))
```

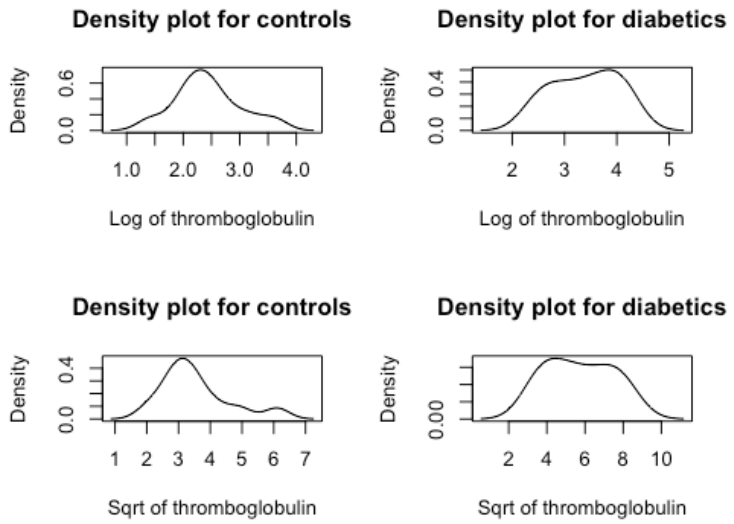

Figure 3: Problem Set D, Q2: Density plots to assess symmetry of transformed data.

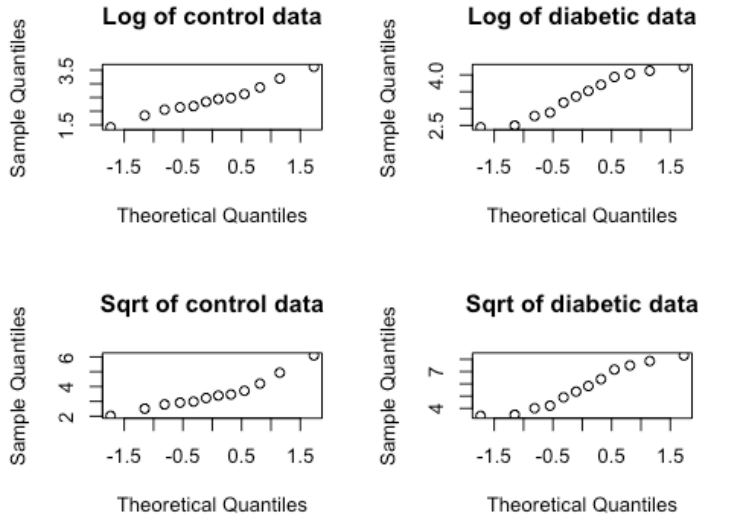

Figure 4: Problem Set D, Q3: QQ plots to assess normality of transformed data.

```
n1 = 12
n2 = 12
std.error = sqrt(s1^2/n1 + s2^2/n2)
Tw = Delta.hat / std.error
df = (s1^2/n1 + s2^2/n2)^2 / ((s1^2/n1)^2/(n1-1) + (s2^2/n2)^2/(n2-1))
1 - pt(Tw, df=df)
```

Or if you're lazy:

```
> t.test(log(diabetic), log(normal), alt="greater")

Welch Two Sample t-test

data:  log(diabetic) and log(normal)
t = 3.8041, df = 21.9, p-value =
0.0004888
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
0.5250797       Inf
sample estimates:
mean of x mean of y
3.390628  2.433349

> t.test(log(diabetic), log(normal))$conf.int
[1] 0.4352589 1.4792986
attr(,"conf.level")
[1] 0.95
```

The P-value is 0.0005, meaning that there's strong evidence that the (population) mean of the logged diabetic thromboglobulin exceeds the mean of the logged diabetic thromboglobulin. On the original scale, this means that median thromboglobulin is higher for diabetics than for others (assuming the logged populations are normal.) A log scale 95% confidence interval is 0.44 to 1.48; back-transforming gives 1.5 to 4.4. That is, we can be confident the median thromboglobulin for diabetics is 1.5 to 4.4 times higher than for others.

3 Trosset chapter 11.4 problem set E, part 3.

Data: `http://mypage.iu.edu/~mtrosset/StatInfeR/Data/films.dat`

```
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")
```

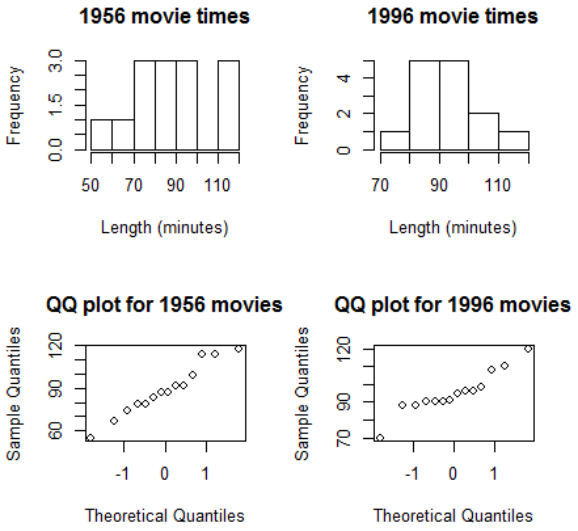

Figure 5: Problem Set E, Q3: Histograms and normal QQ plots of samples of 1956 and 1996 movie lengths.

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, $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 try Welch's two-sample t-test.

$H_0 : \Delta \leq 0, H_1 : \Delta > 0$ where $\Delta = \mu_{1996} - \mu_{1956}$

```
> t.test(movies1996, movies1956, alt="greater")
```

```
Welch Two Sample t-test

data:  movies1996 and movies1956
t = 1.105, df = 22.395, p-value = 0.1404
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
-3.553149       Inf
sample estimates:
mean of x mean of y
95.00000  88.57143
```

The one-tailed P-value is 0.14. The data is consistent with the null hypothesis that average movie lengths were the same (or shorter) in 1996 as in 1956, but it seems like that's just because our sample sizes are too small.

Besides, the 95% confidence interval, (-5.62, 18.48), 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.

The best solution would be to (and you should know what I'm going to say by now) get more data.