Two sample inference

S520

October 25, 2016

Reminder: Questions to ask

1. What is the experimental unit? (The experimental units must be independent.)
2. From how many populations were the experimental units sampled? (Remember that the units within each population must be identically distributed.) What are the populations?
3. How many measurements were taken on each experimental unit? What are the measurements?
4. What are the parameters of interest for this problem?

For one-sample location problems, first define the random variable X_i in terms of the measurements taken on unit i. The parameter is either the population mean μ or the population median θ.

For two-sample location problems, define X_i in terms of the measurements taken on unit i in the first sample and Y_j in terms of the measurements taken on unit j in the second sample. The parameter of interest is usually the *difference* in population means:

$$\Delta = \mu_1 - \mu_2$$

where $\mu_1 = EX_i$ and $\mu_2 = EY_j$. Note that it doesn't which sample you call the X's and which you call the Y's, as long as you're consistent throughout the analysis.

5. Do you need to do a significance test? If so, what are appropriate null and alternative hypotheses? In a one-sample location problem, then the hypotheses should be statements about μ (or θ.) In a two-sample location problem, then the hypotheses should be statements about Δ.

Paired data

Is typing speed on an ergnomic keyboard higher than on a standard keyboard? It's unlikely that the answer will be the same for everyone. A more specific research question we could ask is: "Is typing speed on an ergnomic keyboard higher than on a standard keyboard?" Suppose we take a sample of ten typists from some larger population of typists, and measure their typing speed on the two keyboards in words per minute. (Hopefully the order in which each typist used the keyboards was randomized.)

```r
ergonomic = c(69, 80, 60, 71, 73, 64, 63, 70, 63, 74)
standard = c(70, 68, 54, 56, 58, 64, 62, 51, 64, 53)
```

The experimental unit is a typist. We have one sample of typists from some larger population. We presume all their results are independent of each other. But it seems very likely that their standard and ergonomic typing speeds will *not* be independent: If you know someone types fast on a standard keyboard, you'd guess they'd be fast on the ergonomic keyboard as well. So we have **paired** data: two dependent measurements for each individual – typing speeds on the standard and ergonomic keyboards. With paired data, we want to reduce the data to a one-sample problem, since we only have one truly independent sample. We get this by taking *differences*: subtracting one measurement from the other. I choose to do ergonomic minus standard (but you could do standard minus ergonomic if you really wanted to):

```
differences = ergonomic - standard
differences
```

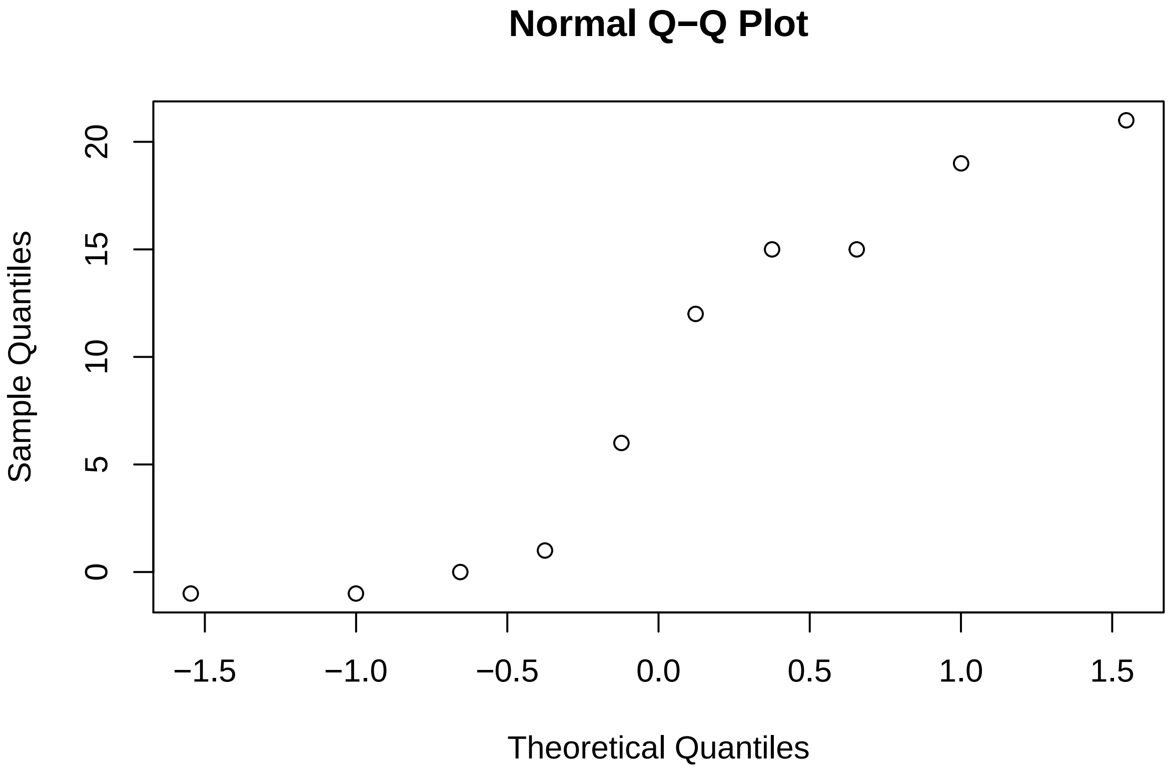
```
##  [1] -1 12  6 15 15  0  1 19 -1 21
```

More formally, let X_i be a random variable representing the ith typist's speed on the ergonomic keyboard minus their speed on the standard keyboard. Let $\mu = EX_i$. The research question had a direction, so we'll do a one-tailed test of the hypotheses. We would like to show "improvement" on average, so our alternative will be that the average difference (ergonomic minus standard) is positive.

$$H_0 : \mu \leq 0$$
$$H_1 : \mu > 0$$

Now, we have one big problem, which is that our sample size is tiny. To do inference on the mean from a small sample requires strong assumptions. Do the observed differences look normal?

```
qqnorm(differences)
```



As usual, it's difficult to judge if the normal QQ plot truly follows a straight line when the sample size is only ten. All we can really say is that there are no really bad outliers in the data (and it seems unlikely that many typists would do, say, 40 words per minute better on one keyboard than the other.) So it is not entirely unjustifiable to make a leap of faith and assume that the differences come from an approximately normal population. This lets us do the one-sample t-test (with $10 - 1 = 9$ degrees of freedom):

```
n = length(differences)
mu0 = 0
x.bar = mean(differences)
```

```
s = sd(differences)
t.statistic = (x.bar - mu0) / (s / sqrt(n))
1 - pt(t.statistic, df = n-1)
```

```
## [1] 0.005687367
```

The P-value is 0.006, which is small. The data appears to be incompatible with our null hypothesis model. (But note that someone who works for the Standard Keyboard Industrial Complex might say that it's not the null hypothesis itself that's wrong, it's some other part of the model – for example, the normal distribution assumption. And you couldn't definitively prove them wrong.)

A confidence interval for the average difference is advisable:

```
x.bar - qt(.975, df=n-1) * s / sqrt(n)
```

```
## [1] 2.490618
```

```
x.bar + qt(.975, df=n-1) * s / sqrt(n)
```

```
## [1] 14.90938
```

With 95% confidence, the average difference in speeds is 2.5 to 15 words per minute in favor of the ergonomic keyboard. Since our sample size is small, we get a pretty wide interval. Most importantly, this is an inference about the *average*. It does not tell us the chance that an *individual* will have their typing speed improved by the ergonomic keyboard.

What about a test of θ, the population *median* difference? You might do such a test because you think the median is more interesting than the mean, or because the normal distribution assumption above made you uneasy. We can use a one-tailed sign test to examine the hypotheses:

$$H_0 : \theta \leq 0$$
$$H_1 : \theta > 0$$

We want to reject the null if our observations are incompatible with the null and compatible with the alternative. If we saw zero out of ten positive observations (i.e. all the typists were *slower* on the ergonomic keyboards,) this would be completely compatible with the null. If we saw ten out of ten positive observations, this would be the result least compatible with the null. We want this least compatible result to have the smallest P-value, so we'll use the right tail: `1 - pbinom(y, 10, 0.5)`.

Now, there's one catch – in our observed differences, we found 7 positive values, 2 negative values, but also one value exactly equal to zero. A convention in statistics is to be *conservative*: that is, to do whatever gives you the *biggest* P-value, to ensure the chance of a Type I error is no bigger than α. Remember, in our test, the more positives there are, the *smaller* the P-value, so seven positive values will give you a larger P-value than eight.

```
y = sum(differences > 0)
1 - pbinom(y-1, n, 0.5)
```

```
## [1] 0.171875
```

The P-value is 0.17. This is not a small P-value (if someone rolls a die, says “I’m going to get a high number,” and rolls a six, you wouldn’t conclude the die was biased or that they had leet die-rolling skills.)

A confidence interval for the median might be more interesting. Potential confidence intervals run from the kth smallest observation to the k largest. Some possible confidence levels we could achieve are:

```
1 - 2 * pbinom(0, 10, 0.5)
```

```
## [1] 0.9980469
```

```
1 - 2 * pbinom(1, 10, 0.5)
```

```
## [1] 0.9785156
```

```
1 - 2 * pbinom(2, 10, 0.5)
```

```
## [1] 0.890625
```

The second of these gives us a confidence level of 98%, which is a nice, safe choice. So a 98% confidence interval for the median goes from the second smallest to the second largest observation (it’s the second because $k - 1$ goes into the `pbinom()` function.)

```
sort(differences)
```

```
##  [1] -1 -1  0  1  6 12 15 15 19 21
```

```
sort(differences)[c(2,9)]
```

```
## [1] -1 19
```

With 98% confidence, the median difference in speeds is between 1 word faster on the standard keyboard and 19 words faster on the ergonomic keyboard. Yet again, a small sample size leads to a wide interval.

Conclusion: If you had to pick one keyboard now, you’d guess the ergonomic keyboard was faster on average. But there’s still a lot of uncertainty, so if you wanted to be sure, take a bigger sample.

Etruscan skulls

Were the skull sizes of ancient Etruscans different from the skull sizes of modern Italians? Trosset decribes the problem on pp. 290-294. Let’s load some data from his webpage. In the data set as posted, the first 84 numbers are the breadths (in mm) of skulls of ancient Etrsucan men, while the remaining 70 numbers are breadths of a sample of skulls of ancient Italian men. (I don’t know if the samples are random – in particular, it’s hard to imagine how one would take a truly random sample of skulls of long-dead Etruscans – but we’ll assume there were no systematic biases in the data collection.)

```
data = scan("http://mypage.iu.edu/~mtrosset/StatInfeR/Data/skulls.dat")
etruscan = data[1:84]
italian = data[85:154]
```

Have a look at the numerical summaries:

```
summary(etruscan)
```

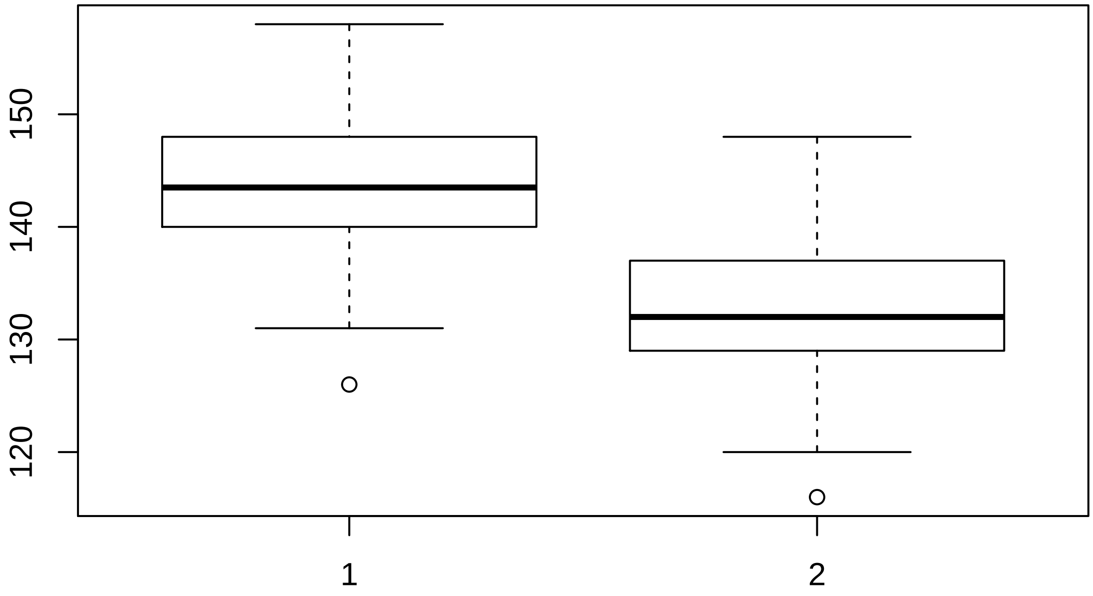
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   126.0   140.0   143.5   143.8   148.0   158.0
```

```
summary(italian)
```

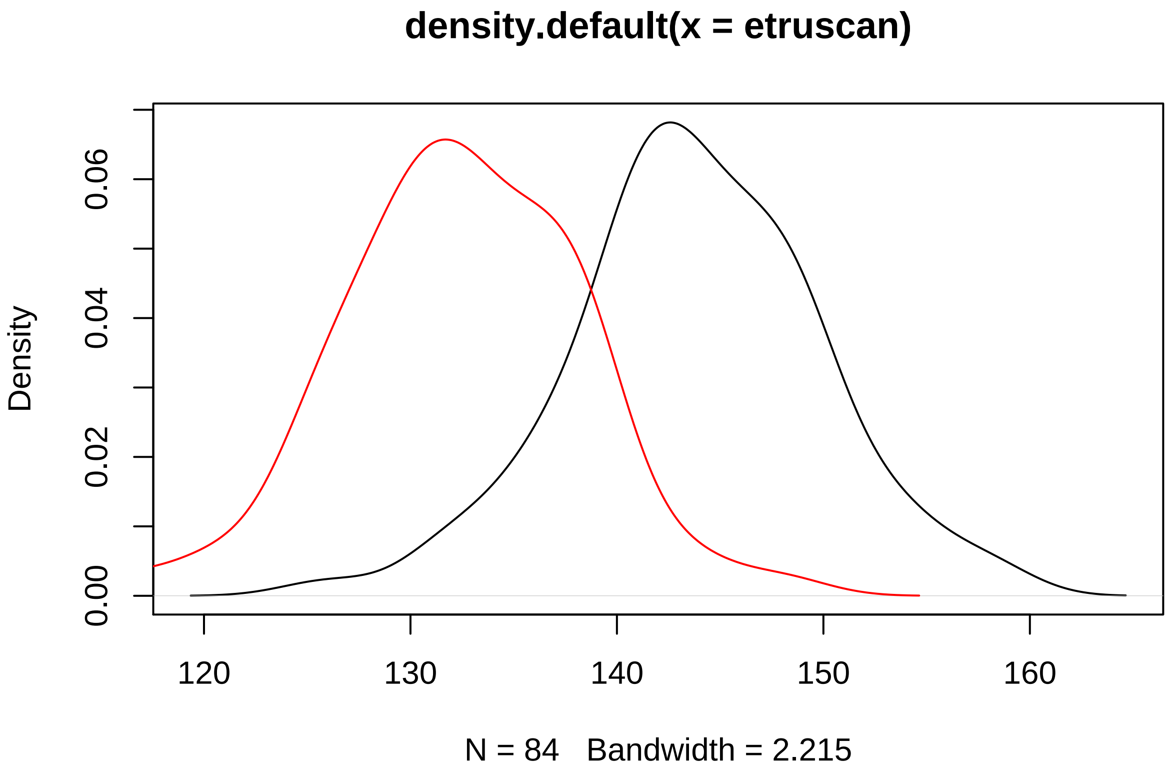
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   116.0   129.0   132.0   132.4   136.8   148.0
```

And draw some pictures:

```
boxplot(etruscan, italian)
```



```
plot(density(etruscan))
lines(density(italian), col="red")
```



It very much looks like the Italian (red) distribution is shifted to the left compared to the Etruscan (black) distribution – and the sample sizes are reasonable. Still, there might still some some doubt in your mind as to whether a difference of this size could be explained by chance, so let's do a significance test.

Answering our basic questions: The experimental unit is a skull. The skulls are sampled from two populations: ancient Etruscans and modern Italians. One measurement is taken on each skull – the breadth, in millimeters. Let X_i be the breadth of the ith Etruscan skull, and Y_j be the breadth of the jth Italian skull. Let the population mean Etruscan skull breadth be μ_1 and the population mean Italian skull breadth be μ_2. Let $\Delta = \mu_1 - \mu_2$. There was no direction to the test before looking at the data, so we'll do a two-tailed test of the hypotheses

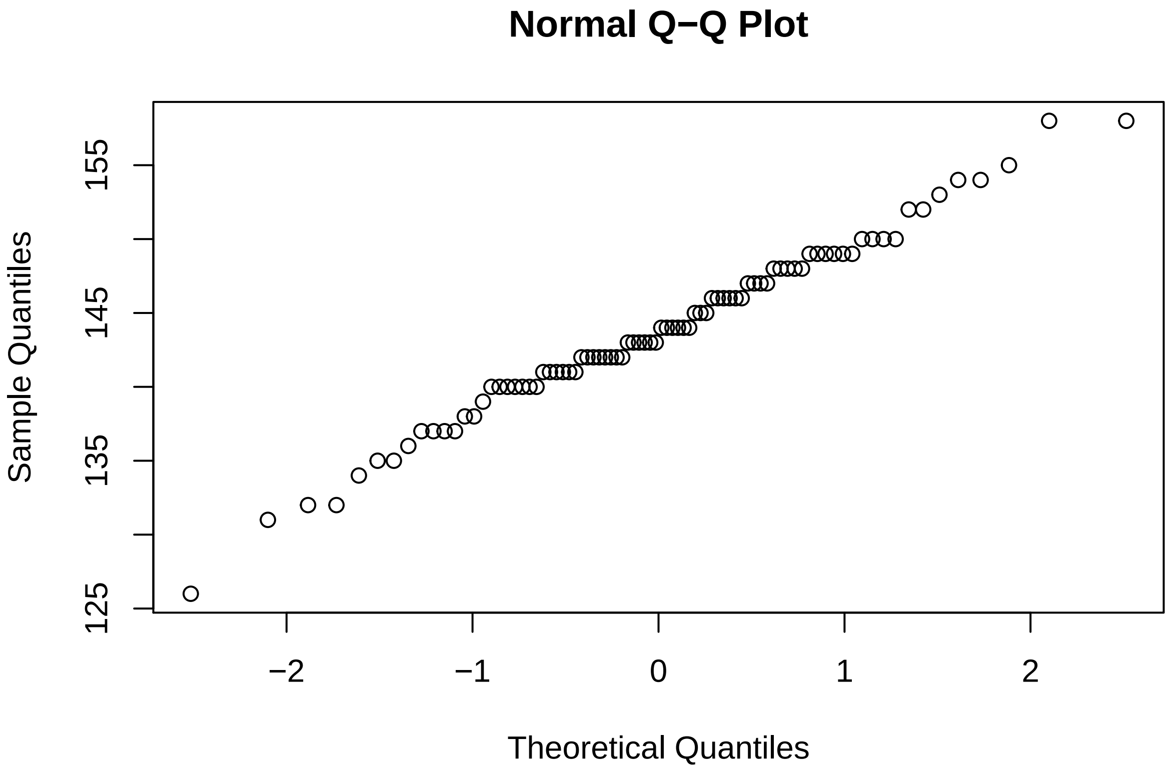
$$H_0 : \Delta = 0$$
$$H_1 : \Delta \neq 0$$

Note this is the same as testing

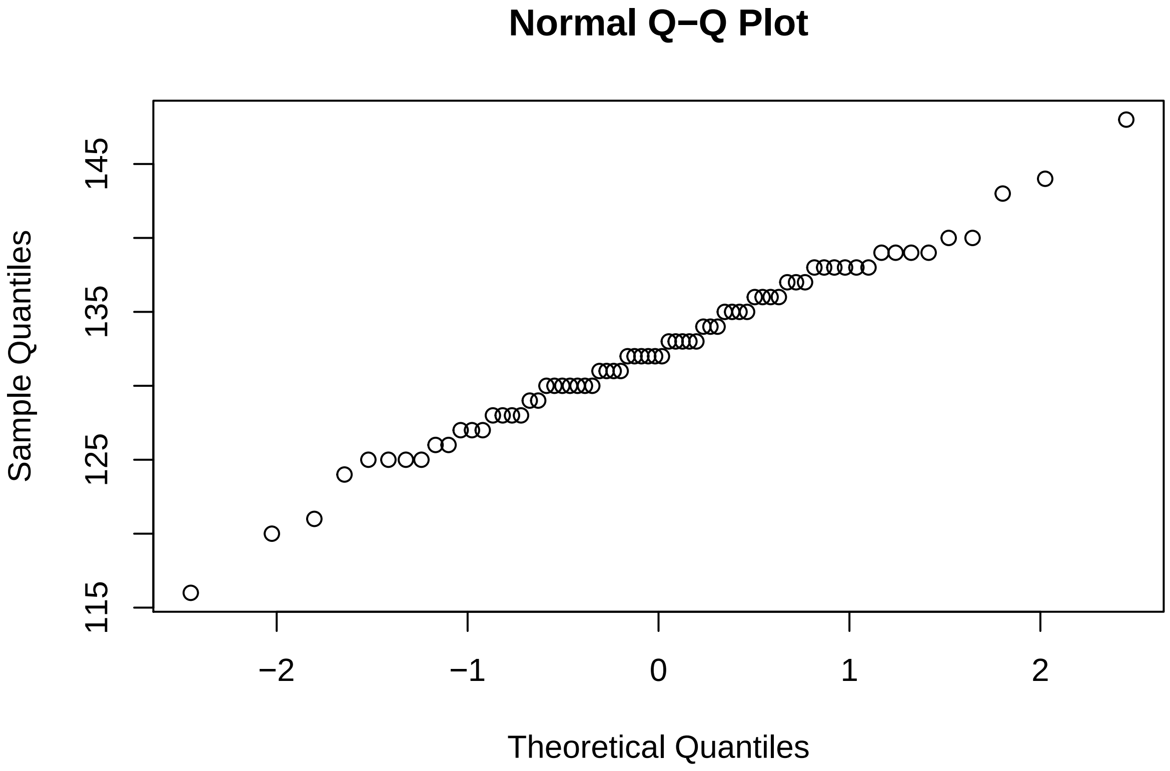
$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

It'll help if we can justify a normal distribution assumption. Draw some mornal QQ plots:

```r
qqnorm(etruscan)
```



```
qqnorm(italian)
```



They look like straight lines. When we have two samples from approximately normal populations, there are two options for significance tests concerning the difference in means:

- When we have two IID samples from two independent normal populations, we can do **Welch's two-sample t-test.**
- When we have two IID samples from two independent normal populations *with the same variance*, we can do **Student's two-sample t-test.**

Which of the two should we choose? Simulations show that:

- If the population variances are equal, both Welch's and Student's tests give similar results.
- If the population variances are not equal, Welch's test gives good results but Student's test sometimes gives very bad results.

So Welch's test is the safer bet, unless you're sure that the population variances are equal (regardless of whether the null is true or false.) This is rarely or never the case.

Welch's two-sample t-test is carefully constructed on pp. 278-280 of Trosset. We need to find a point estimate and a standard error, turn that into a t-statistic, do a really annoying calculation to get the degrees of freedom, then find a P-value. Here we go:

```
Delta = mean(etruscan) - mean(italian)
se = sqrt(var(etruscan)/84 + var(italian)/70)
T.Welch = Delta/se
nu = (var(etruscan)/84+var(italian)/70)^2/
  ((var(etruscan)/84)^2/83+(var(italian)/70)^2/69)
P.value = 2*(1-pt(abs(T.Welch), df=nu))
P.value
```

```
## [1] 0
```

The P-value is basically zero. This tells us there is a difference between Etruscan and Italian skulls, but not what the difference is. So let's do a confidence interval:

```
q = qt(0.975, df=nu)
lower = Delta - q*se
upper = Delta + q*se
lower
```

```
## [1] 9.459782
```

```
upper
```

```
## [1] 13.20212
```

We conclude that the data is not compatible with the hypothesis that ancient Etrsucan and modern Italian skulls have the same average breadth. We can be confident that the average ancient Etruscan skull was 9-13 mm broader than the average modern Italian skull.

Finally, here's the easy way to do a t-test when you have all the data:

```
t.test(etruscan, italian)
```

```
##
##  Welch Two Sample t-test
##
## data:  etruscan and italian
## t = 11.966, df = 148.82, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   9.459782 13.202123
## sample estimates:
## mean of x mean of y
##  143.7738  132.4429
```

If you must do Student's two-sample *t*-test (but you probably shouldn't):

```r
t.test(etruscan, italian, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  etruscan and italian
## t = 11.925, df = 152, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   9.45365 13.20825
## sample estimates:
## mean of x mean of y
##  143.7738  132.4429
```

Stereogram fusion times

We return to the data from the stereogram randomized experiment we looked at in class a few weeks ago. Does visual information affect the times it takes to see a "fused" stereogram image?

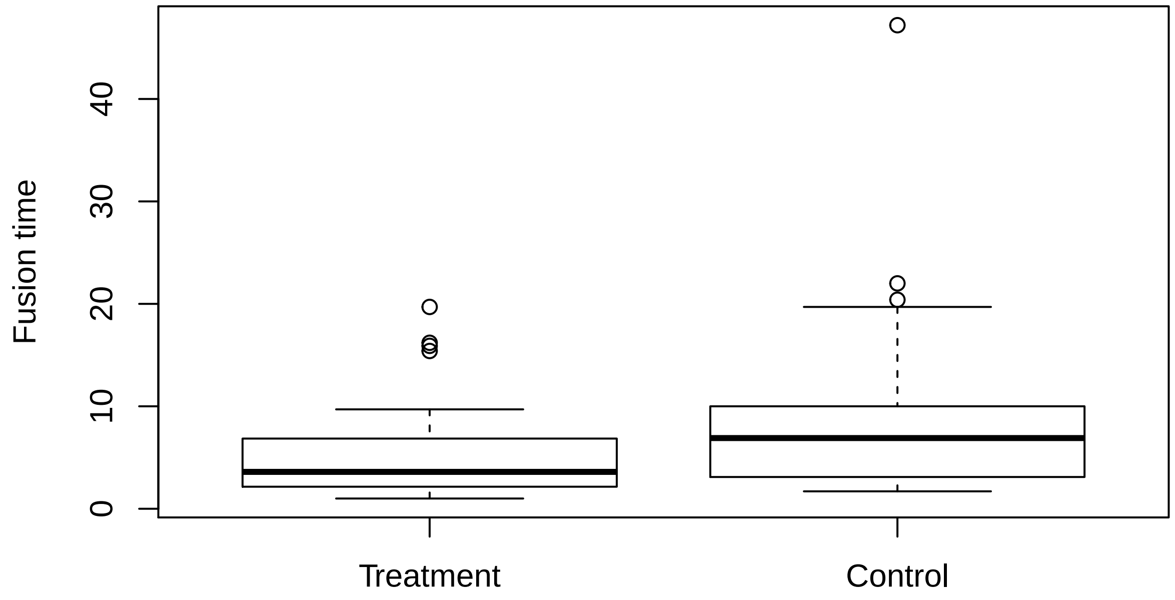
The data in `stereograms.txt` contains two variables. The variable `time` give the time (in second) taken to see the image. The variable `group` is 1 for the control group (no visual information) and 2 for the treatment group (visual information.) The experimental unit is a person looking at the stereogram. Because it's a randomized experiment, we treat it as a two-population, two-sample problem: a (hypothetical) population of people who could get the treatment and a (hypothetical) population of people who could get the control. One measurement – the fusion time – is taken on each individual. We'll leave careful definitions of the parameters and hypotheses until later.

For now, let's load the data and separate out the two groups:

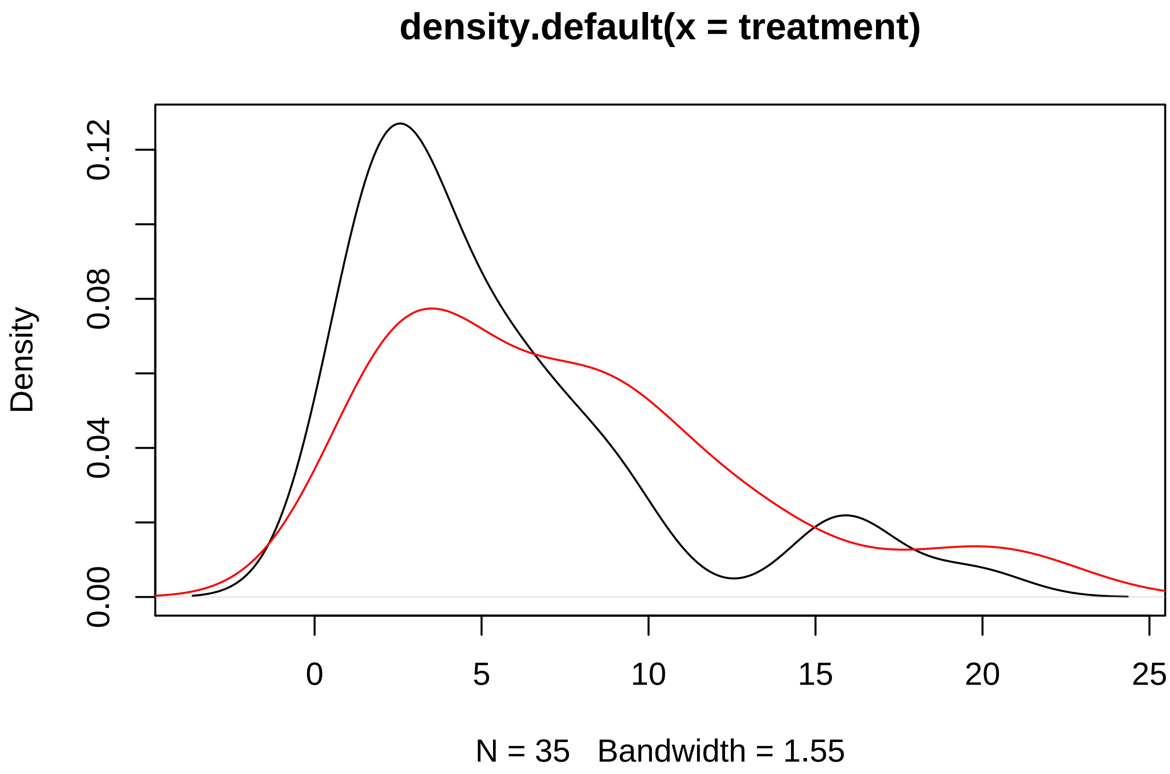
```r
stereograms = read.table("stereograms.txt", header=TRUE)
treatment = stereograms$time[stereograms$group==2]
control = stereograms$time[stereograms$group==1]
```

Look at the data:

```r
boxplot(treatment, control,
  names=c("Treatment","Control"),
  ylab="Fusion time")
```

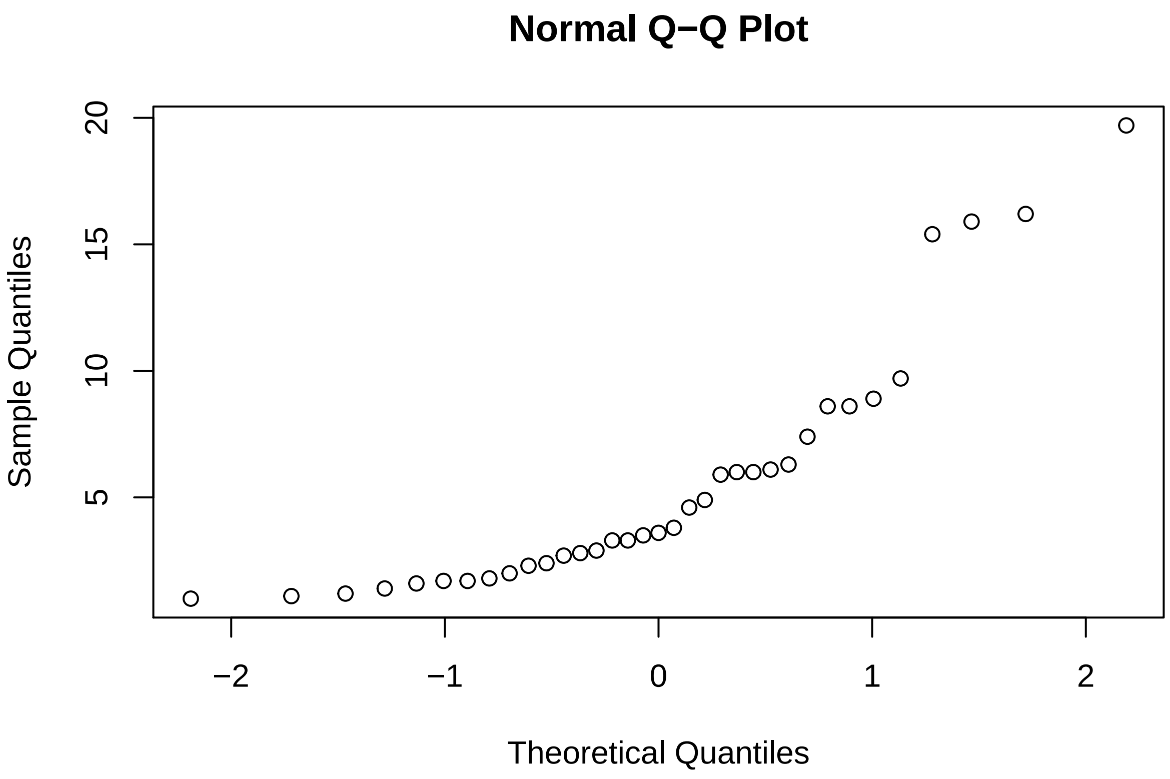


```
plot(density(treatment))
lines(density(control), col="red")
```

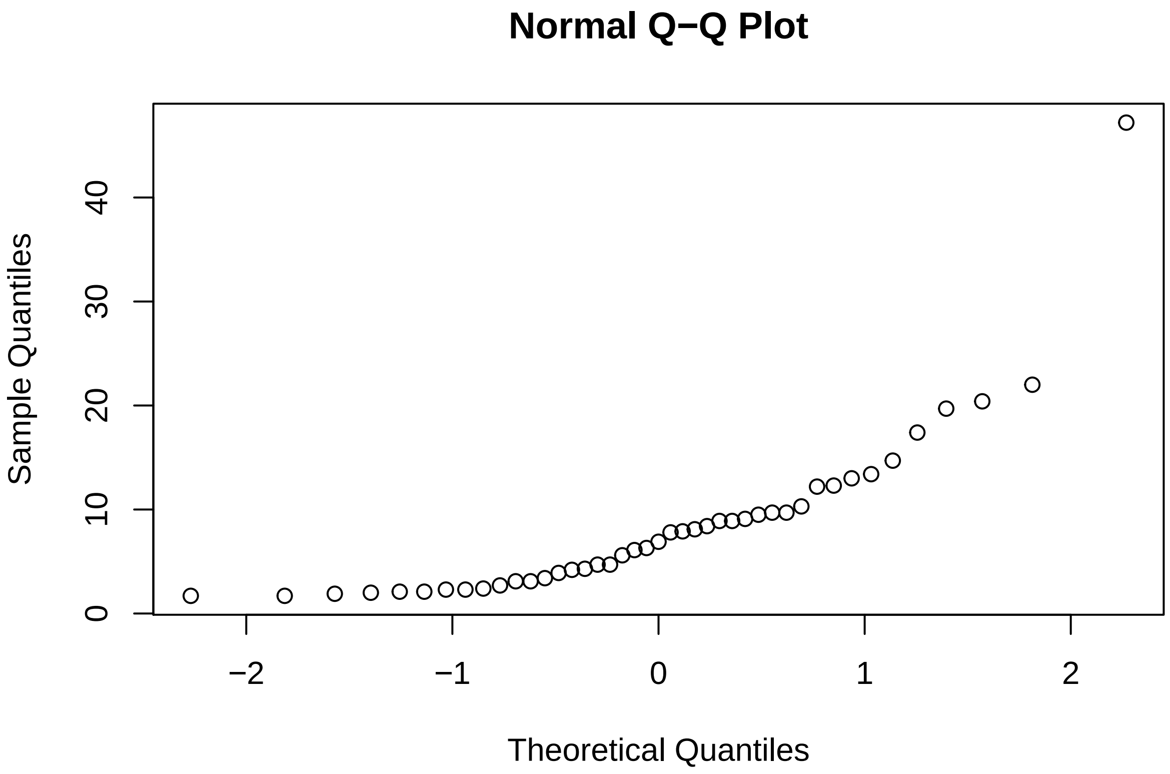


From the boxplots, it looks very much like the treatment group tends to have lower fusion times than the control group. Suppose we wish to show this more formally. Can we use a *t*-test?

```
qqnorm(treatment)
```



```
qqnorm(control)
```



Neither sample looks like it comes from a normal distribution, so it's not a good idea to do either version of the *t*-test. But let's see what happens if we do the the *t*-test anyway:

```r
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
## 95 percent confidence interval:
##  -5.95314090 -0.06493219
## sample estimates:
## mean of x mean of y
##  5.551429  8.560465
```

```r
t.test(treatment, control, var.equal=TRUE)
```

```
##
##  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
## 95 percent confidence interval:
##  -6.09901044  0.08093735
## sample estimates:
## mean of x mean of y
##  5.551429  8.560465
```

Welch's t-test gives a P-value of 0.045, while Student's t-test gives a P-value of 0.056. Once again, this is a bad idea to always use a fixed 0.05 threshold and reduce the problem to "reject" or "do not reject" – different tests can give different results. In this case, *neither* test is good because the data isn't normal, but Student's test is worse, because the variances clearly aren't equal (the control group is more spread out.)

What can we do instead? One possibility is to do **nonparametric statistics**: methods that don't assume a parametric distribution for the data. An appropriate two-sample nonparametric method is the **Wilcoxon rank-sum test**, which uses the *ranks* of the data rather than the raw values. We don't have time to go into this method in detail (see pp. 282-287 of Trosset if you're interested) but it's easy to do in R:

```r
wilcox.test(treatment, control)
```

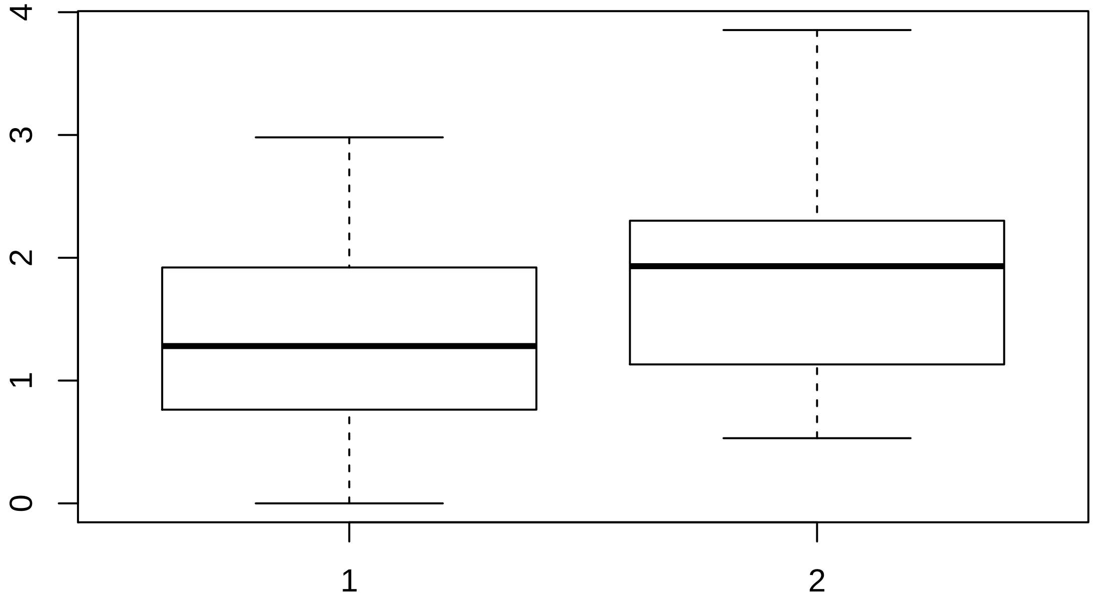
```
## Warning in wilcox.test.default(treatment, control): cannot compute exact p-
## value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  treatment and control
## W = 532, p-value = 0.02706
## alternative hypothesis: true location shift is not equal to 0
```

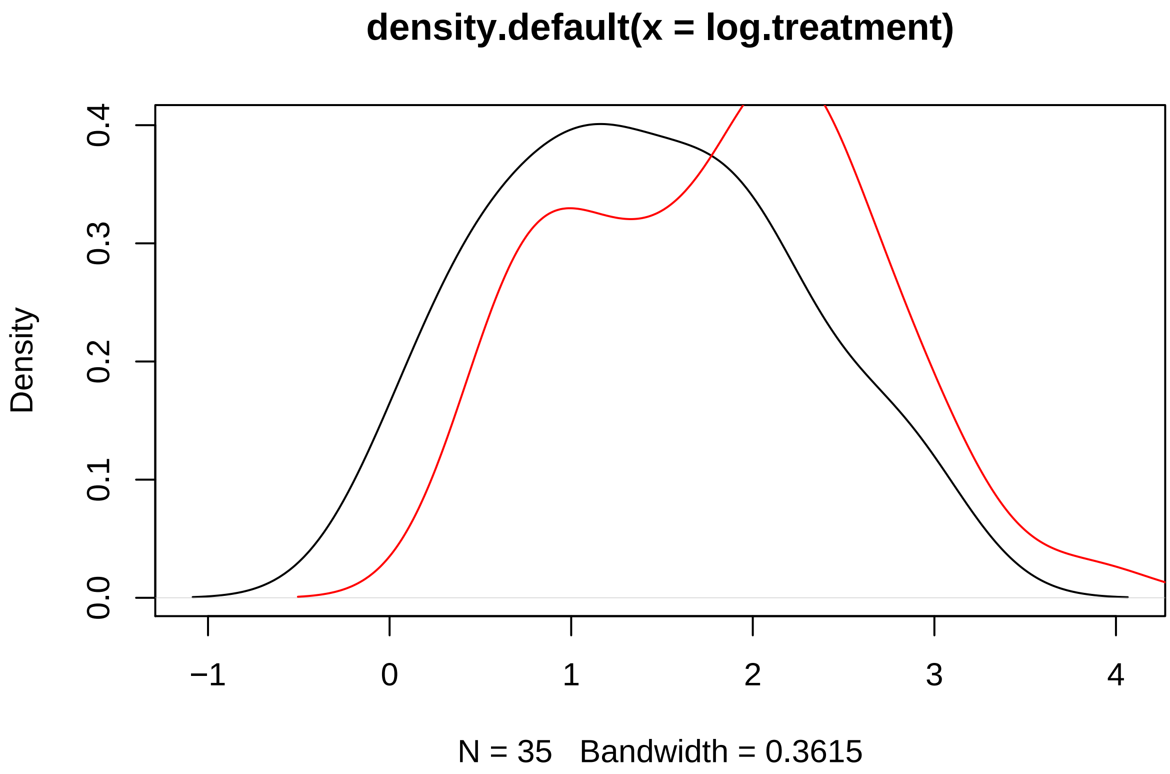
The small P-value means (roughly speaking) there's some evidence that the treatment and control times don't come from the same distribution.

Here's an arguably better solution. Remember that taking logs sometimes magically makes things normal:

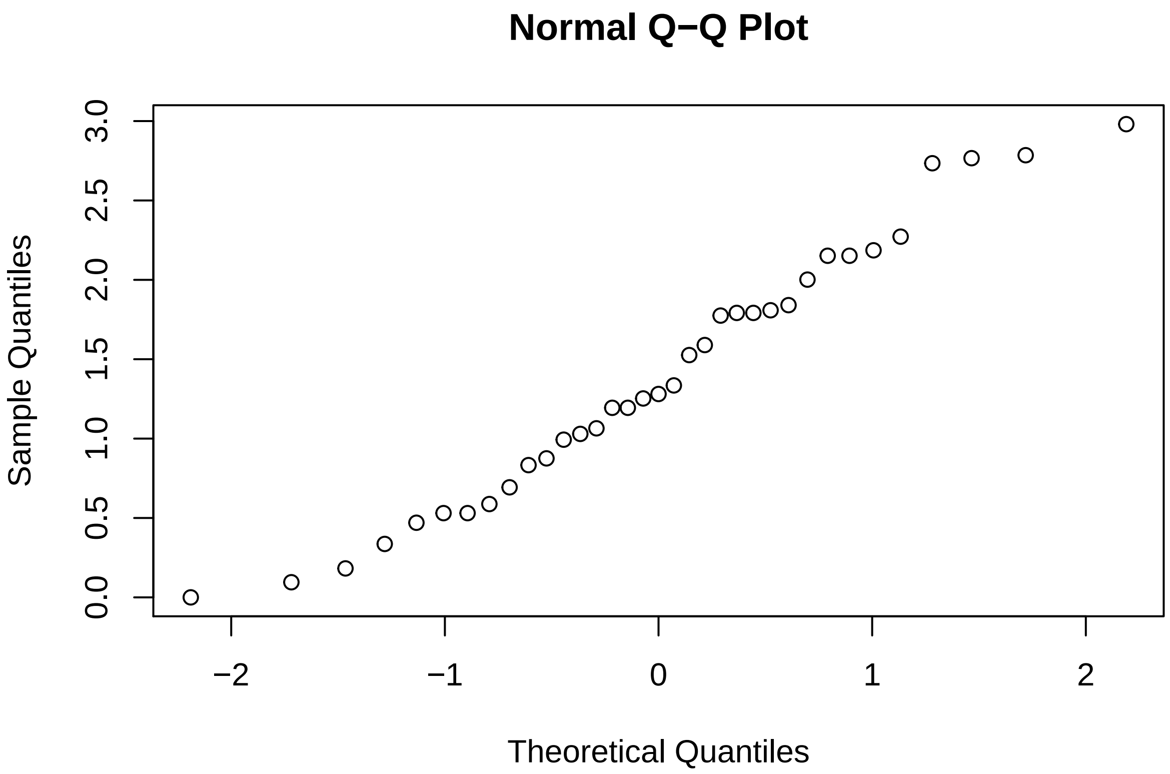
```r
log.treatment = log(treatment)
log.control = log(control)
boxplot(log.treatment, log.control)
```



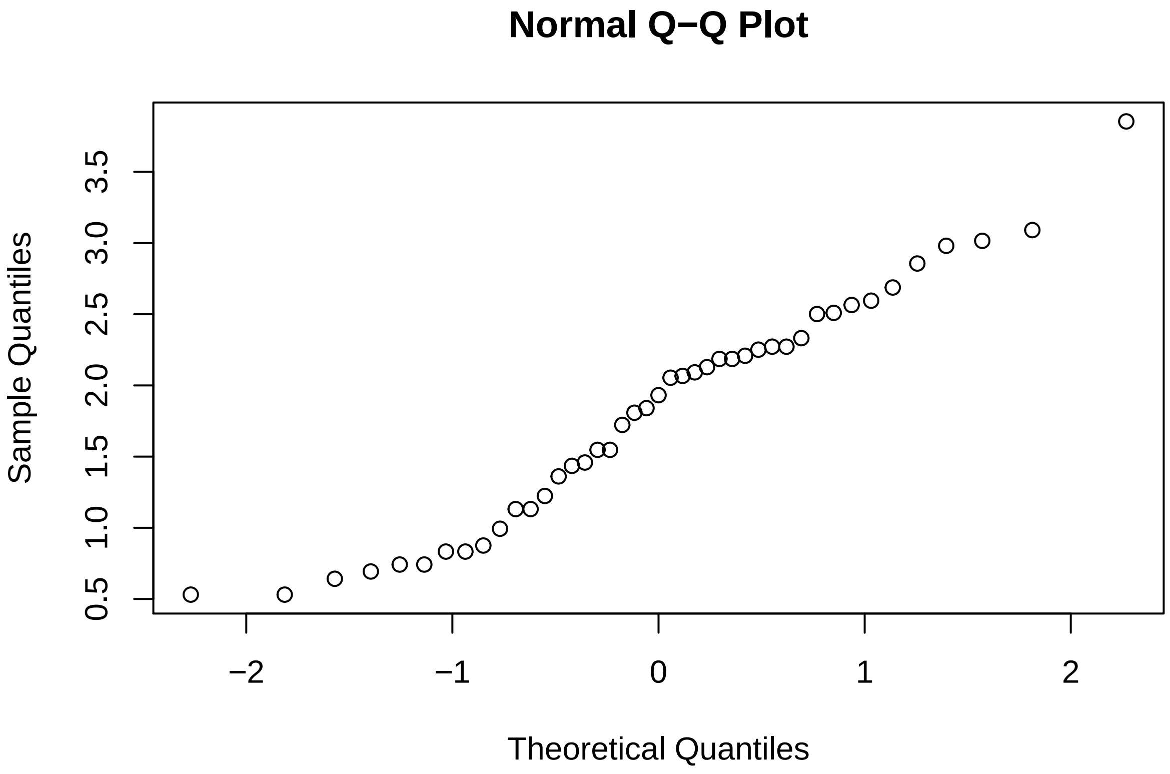
```r
plot(density(log.treatment))
lines(density(log.control), col="red")
```



```r
qqnorm(log.treatment)
```



```r
qqnorm(log.control)
```



The logged samples look much closer to normal. Since there's no a priori reason to think that the population variances will be the same, we prefer Welch's two-sample t-test (but since the sample spreads are similar, Student's t-test will give pretty much the same result.)

Now let's write down the parameters and hypotheses carefully. Let μ_1 be the population mean of *log* fusion times under the treatment. Let μ_2 be the population mean of *log* fusion times under the control. Let $\Delta = \mu_1 - \mu_2$. It's not clear whether we should do a one-tailed or two-tailed test; let's just do a two-tailed test.

$$H_0 : \Delta = 0$$
$$H_1 : \Delta \neq 0$$

```
t.test(log.treatment, log.control)
```

```
##
##  Welch Two Sample t-test
##
## data:  log.treatment and log.control
## t = -2.3178, df = 72.673, p-value = 0.02328
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.80080773 -0.06030565
## sample estimates:
## mean of x mean of y
##  1.389454  1.820011
```

The two-tailed P-value is 0.023. (The one-tailed P-value would be half this.) There's evidence that the mean log treatment time and the mean log control time are *not* the same.

The confidence interval takes a little more effort to interpret because of the transformation. We're confident that the difference in mean *log* fusion times is between -0.8 and -0.06. What does this mean in terms of actual fusion times? First, back transform by taking the exponential of the confidence interval:

```
exp(c(-0.80, -0.06))
```

```
## [1] 0.4493290 0.9417645
```

Effects that are additive on the log scale are multiplicative on the original scale. If you assume that the treatment has the same multiplicative effect on everyone, the multiplier is between 0.45 and 0.94 – that is, the treatment reduces everyone's fusion time between 6% and 55%. If you don't think the treatment has the same effect on everyone, then (assuming your normal assumptions were right on the log scale) you can say you're confident the population median for the treatment times is between 0.45 times and 0.94 times the population median for the control times. (We have to switch to medians because it's the logs that are normal, not the populations.)

Finally, let's check that Student's t-test gives more or less the same results:

```
t.test(log.treatment, log.control, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  log.treatment and log.control
## t = -2.319, df = 76, p-value = 0.02308
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.80034143 -0.06077195
## sample estimates:
## mean of x mean of y
##  1.389454  1.820011
```