Chapter 11: Two sample inference

S520

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.

Example: 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; the data is on his webpage. Before we answer the five basic questions, we'll take a look at the data.

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)
```

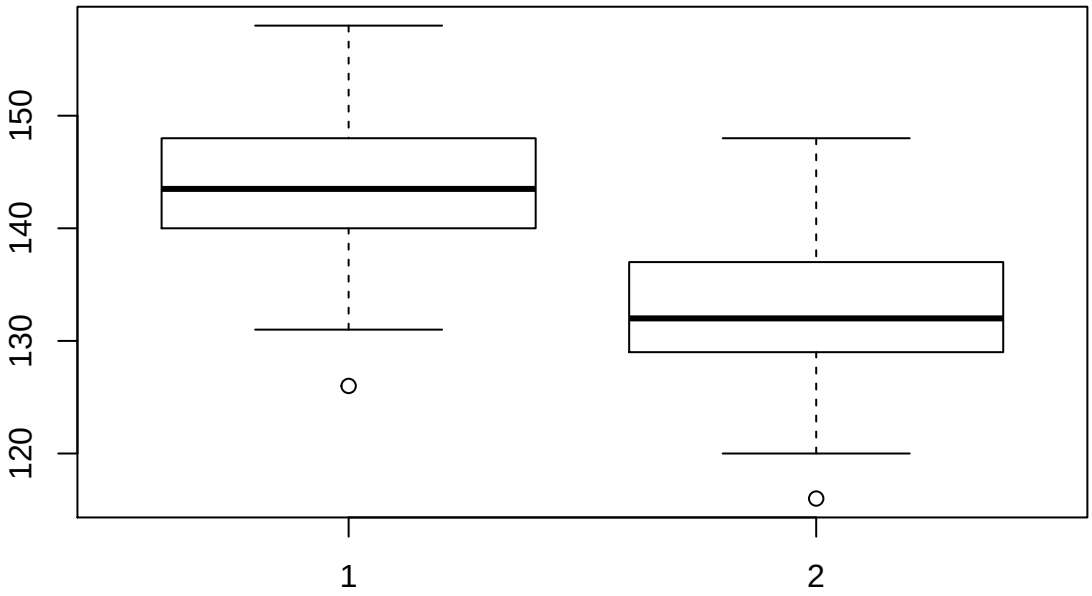
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   126.0   140.0   143.5   143.8   148.0   158.0
```

```
summary(italian)
```

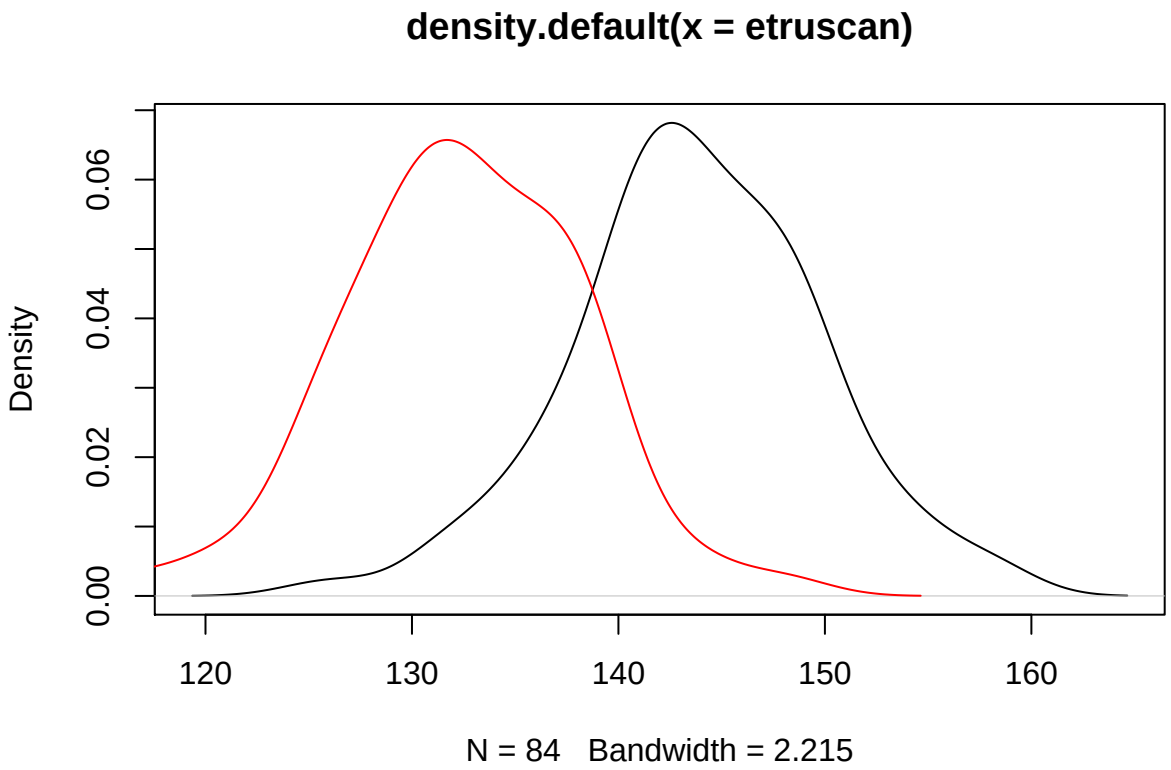
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   116.0   129.0   132.0   132.4   136.8   148.0
```

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.

Now let's answer 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

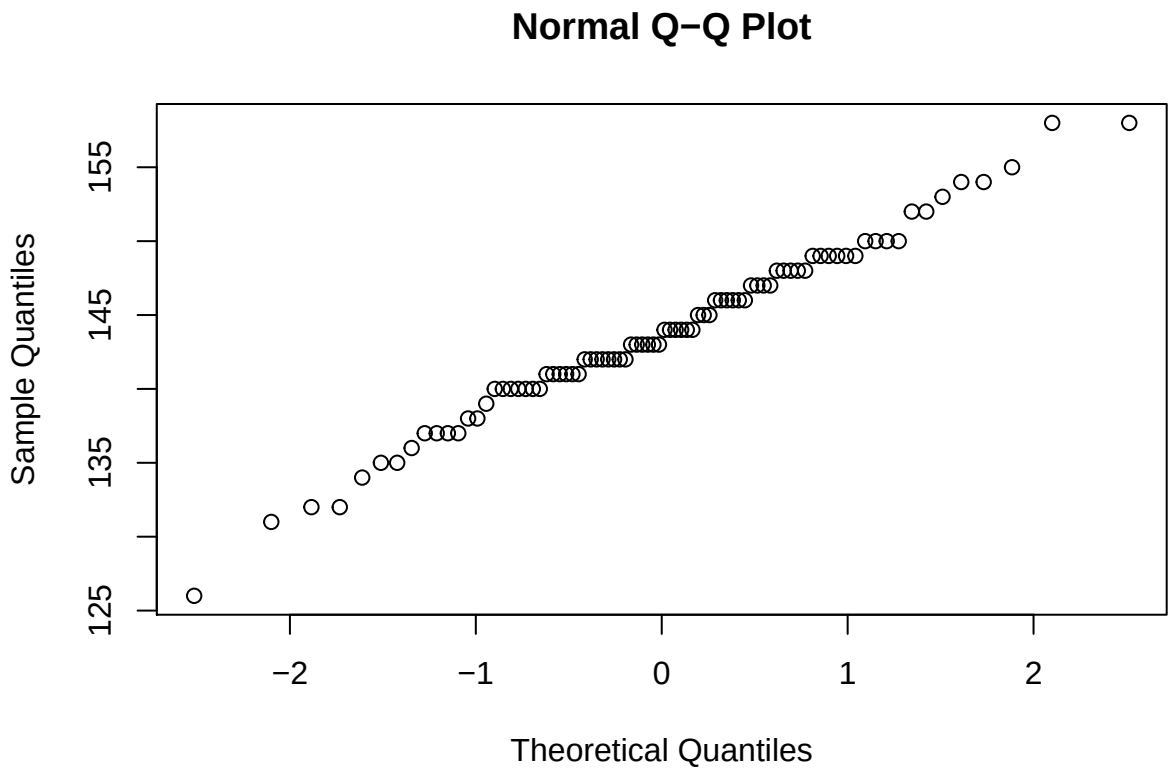
$$H_0 : \Delta = 0$$
$$H_1 : \Delta \neq 0$$

Note this is the same as testing

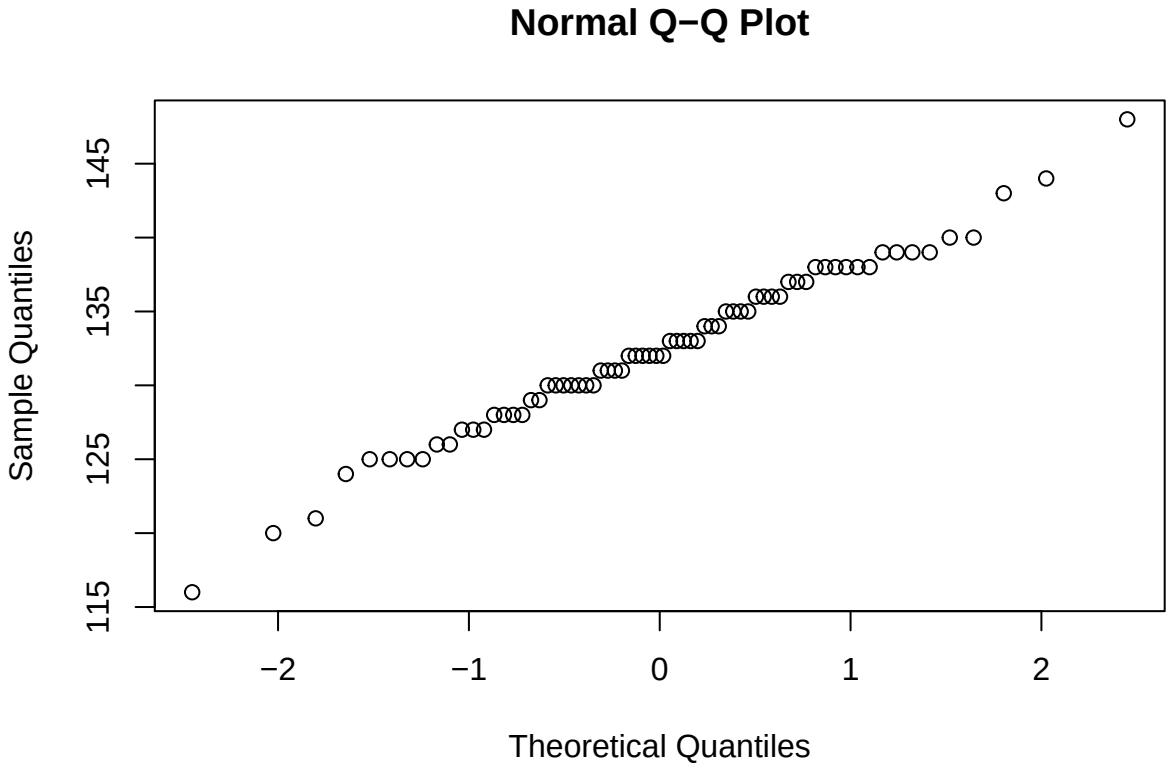
$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

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

```
qqnorm(etruscan)
```




```
qqnorm(italian)
```



They look like straight lines.

Our point estimate of Δ is just the difference in sample means:

```
mean(etruscan) - mean(italian)
```

```
## [1] 11.33095
```

We call this $\hat{\Delta}$ ("Delta hat"; remember that we put a hat on things to indicate we're dealing with an estimate rather than the true population value.)

Welch's and Student's two-sample t-tests

When we have two samples from approximately normal populations, there are three options for significance tests concerning the difference in means. The first option is to use the normal distribution for $\hat{\Delta}$, which is justifiable if you know both population standard deviations or you can estimate them very accurately. In practice, this happens in the special case where you're dealing with two populations of 0's and 1's and you're estimating the difference population proportions. We'll set this case aside.

That leaves two more choices:

- 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.

The point estimate is $\hat{\Delta}$, defined as above as the difference in sample means.

```
Delta.hat = mean(etruscan) - mean(italian)
```

The estimated standard deviation of $\hat{\Delta}$, often known as the **standard error** of $\hat{\Delta}$, is:

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

```
se = sqrt(var(etruscan)/84 + var(italian)/70)
```

Assuming a null hypothesis of no difference, the Welch t-statistic is $\hat{\Delta}$ divided by its standard deviation:

```
t.Welch = Delta.hat/se
print(t.Welch)
```

```
## [1] 11.96595
```

The approximate degrees of freedom is (you're not going to like this):

$$\hat{\nu} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

```
nu = (var(etruscan)/84 + var(italian)/70)^2/((var(etruscan)/84)^2/83 + (var(italian)/70)^2/69)
print(nu)
```

```
## [1] 148.8193
```

This isn't a whole number but that's fine, R can handle decimal degrees of freedom. (If you're stuck with software from the 1980s, just round down to the next whole number.)

If the null hypothesis is true, then Welch's t-statistic has an approximate t-distribution with $\hat{\nu}$ degrees of freedom. For a two-tailed test, the P-value is thus:

```
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 95% confidence interval. We start from $\hat{\Delta}$ and go up and down by q standard errors (where the standard error is the same one we found earlier), where q is determined using a t-distribution with the appropriate number of degrees of freedom.

```
q = qt(0.975, df = nu)
lower = Delta.hat - q * se
upper = Delta.hat + 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 ancient Etruscan average skull breadth was 9–13 mm more than the modern Italian skull average skull breadth. (Remember, this is an interval for a *difference in means*; it doesn't say anything about 95% of Etruscan skulls or 95% of Italian skulls or 95% of differences. If in doubt, state the interval in terms of Δ.)

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):

```
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
```

Assumptions of Welch's t-test

Welch's t-test assumes two independent samples from normally distributed populations. However, the test is fairly *robust* to minor violations of its assumptions: that is, nothing horrible will happen (in terms of Type I and II error probabilities) if the underlying distributions are not quite normal. In practice, you can usually get away with Welch's test as long as your samples are not tiny and reasonably symmetric with no gross outliers. On the other hand, if you have strongly skewed data or bad outliers, you should consider a transformation or a nonparametric test (see below.)

As usual, the larger the samples the better, both in terms of robustness to assumptions and (usually more importantly) in terms of power. In practical work, samples of size 50 per group should be a bare minimum if failures to reject are to be meaningful, though 1000 per group would be better.

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?

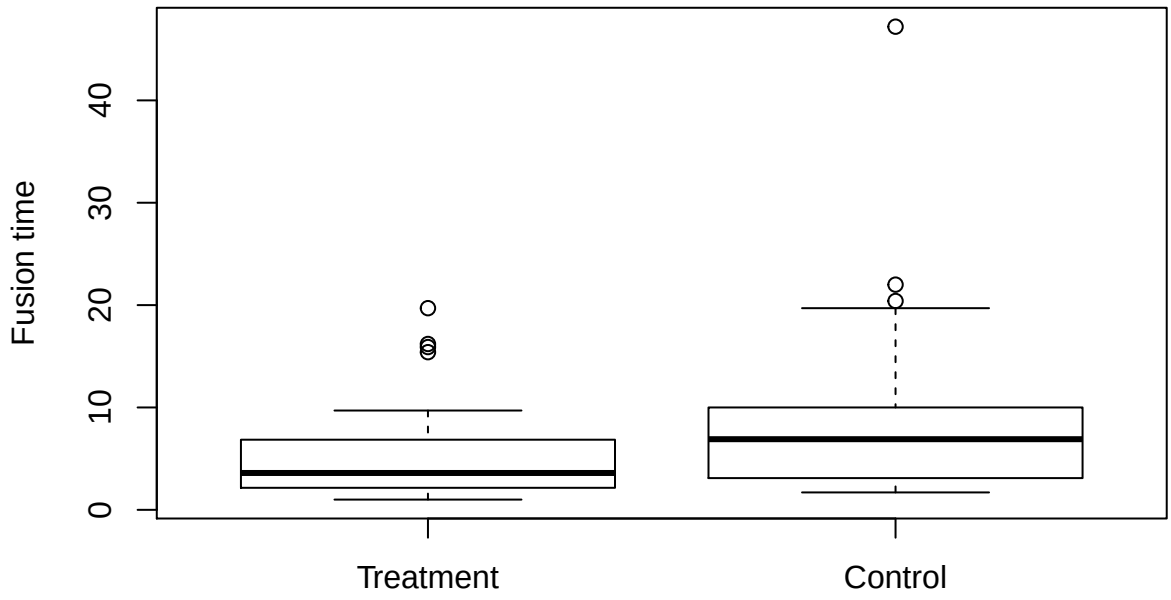
The data in `stereograms.txt` contains two variables. The variable `time` give the time (in second) taken to see the image. The variable `group` is 1 for the control group (no visual information) and 2 for the treatment group (visual information.) The experimental unit is a person looking at the stereogram. 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:

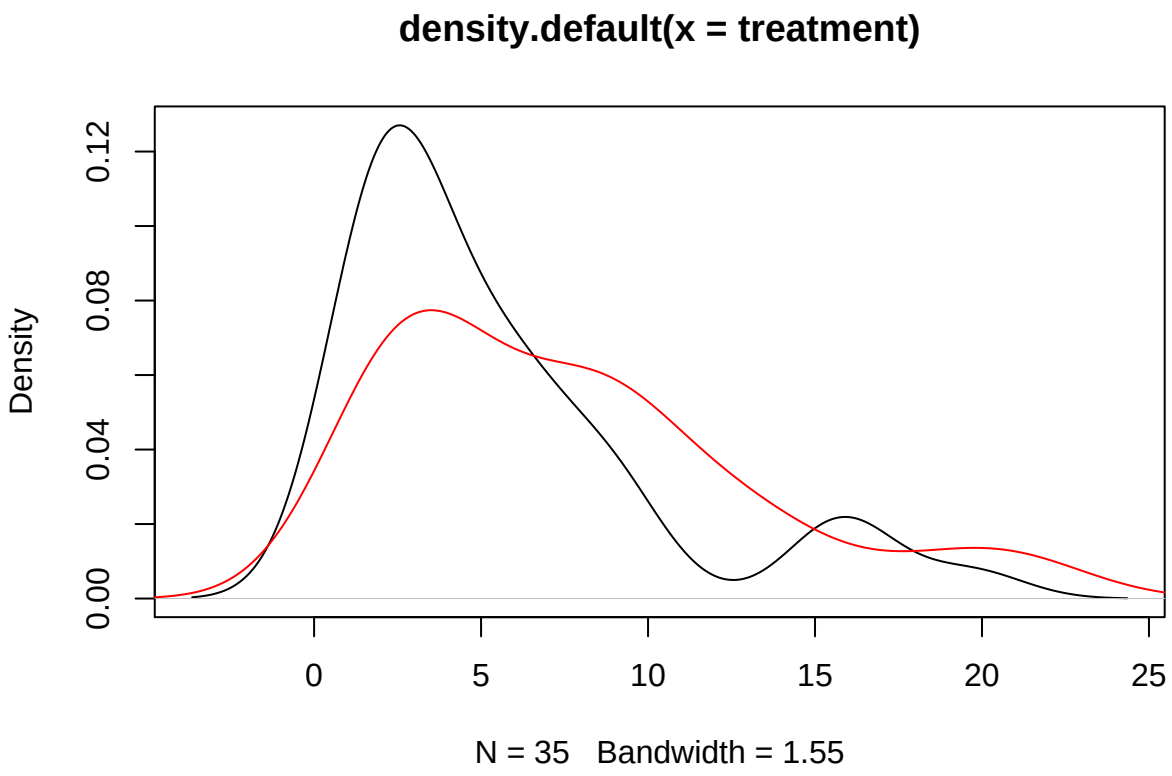
```
stereograms = read.table("stereograms.txt", header = TRUE)
treatment = stereograms$time[stereograms$group == 2]
control = stereograms$time[stereograms$group == 1]
```

Look at the data:

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

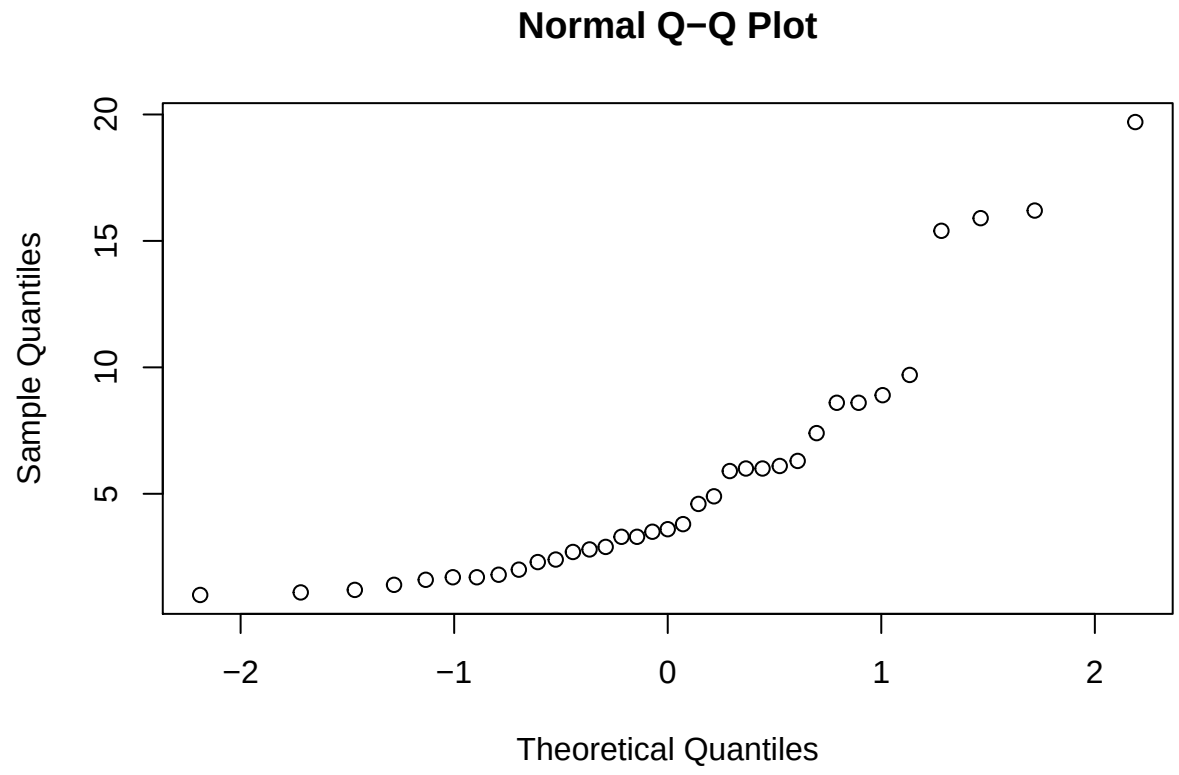



```
plot(density(treatment))
lines(density(control), col = "red")
```

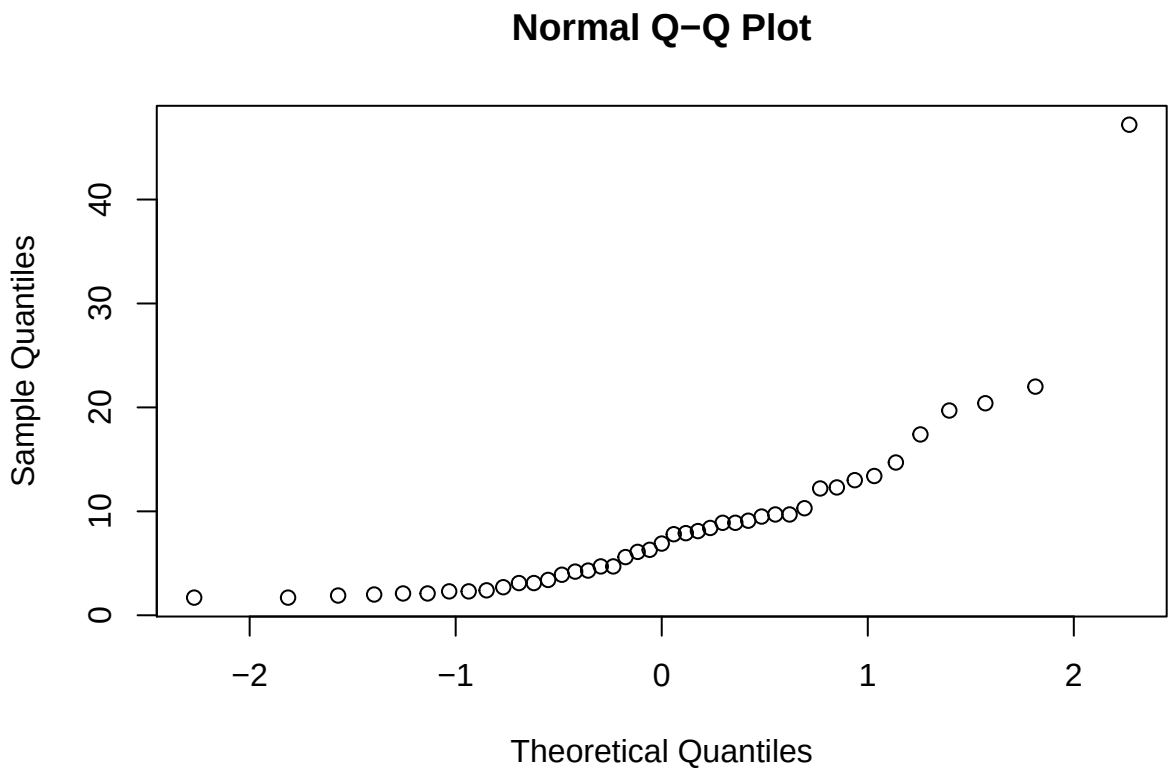


From the boxplots, it looks very much like the treatment group tends to have lower fusion times than the control group. Suppose we wish to show this more formally. Can we use a t-test?

```
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 in the interest of seeing what might go wrong, let's do the the t-test anyway:

```
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
```

```
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
```

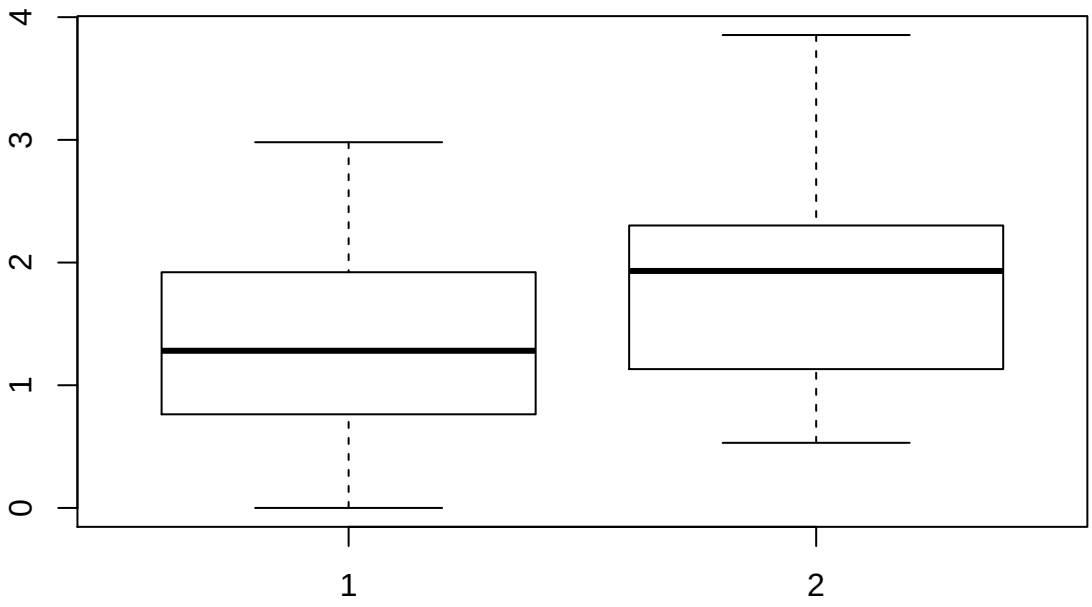
```
##  5.551429  8.560465
```

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, it's a bad idea to always use a fixed 0.05 threshold and reduce the problem to "reject" or "do not reject" – for one thing, 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.)

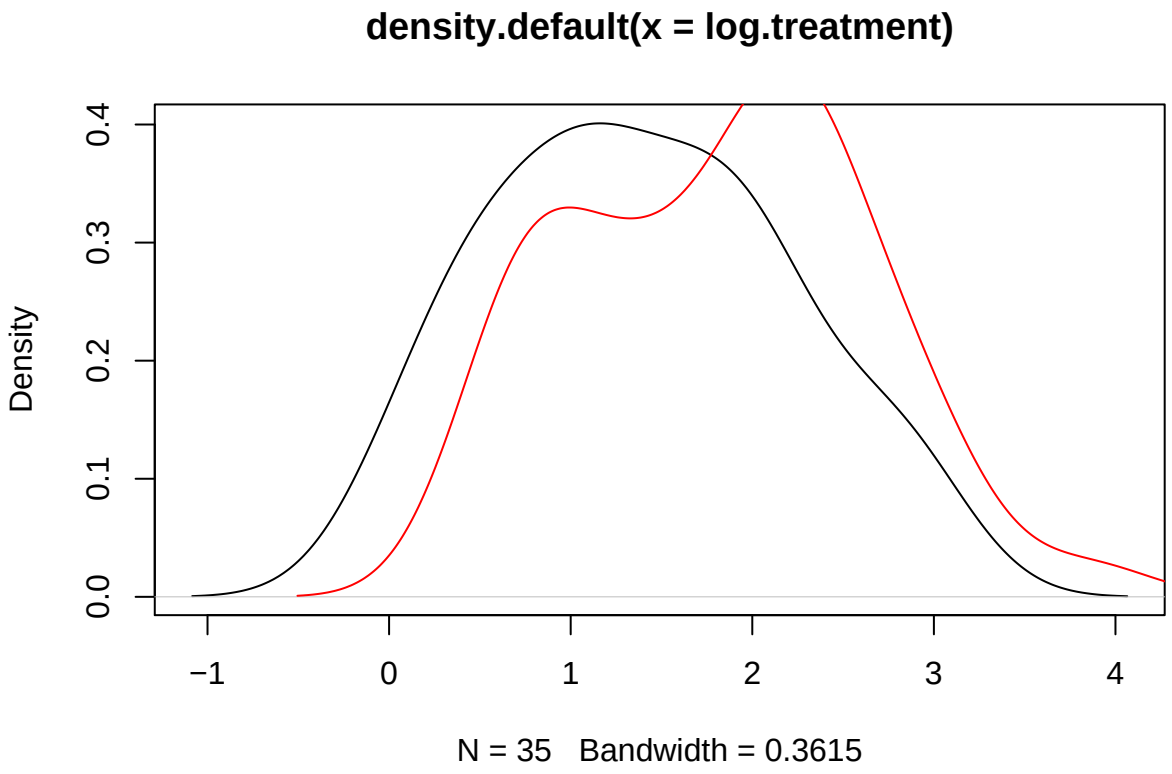
What can we do instead? One possibility is to do **nonparametric statistics**: use a method that doesn't assume a parametric distribution (such as normal) for the data. An appropriate two-sample nonparametric method is the **Wilcoxon rank-sum test**, which we'll briefly discuss later.

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

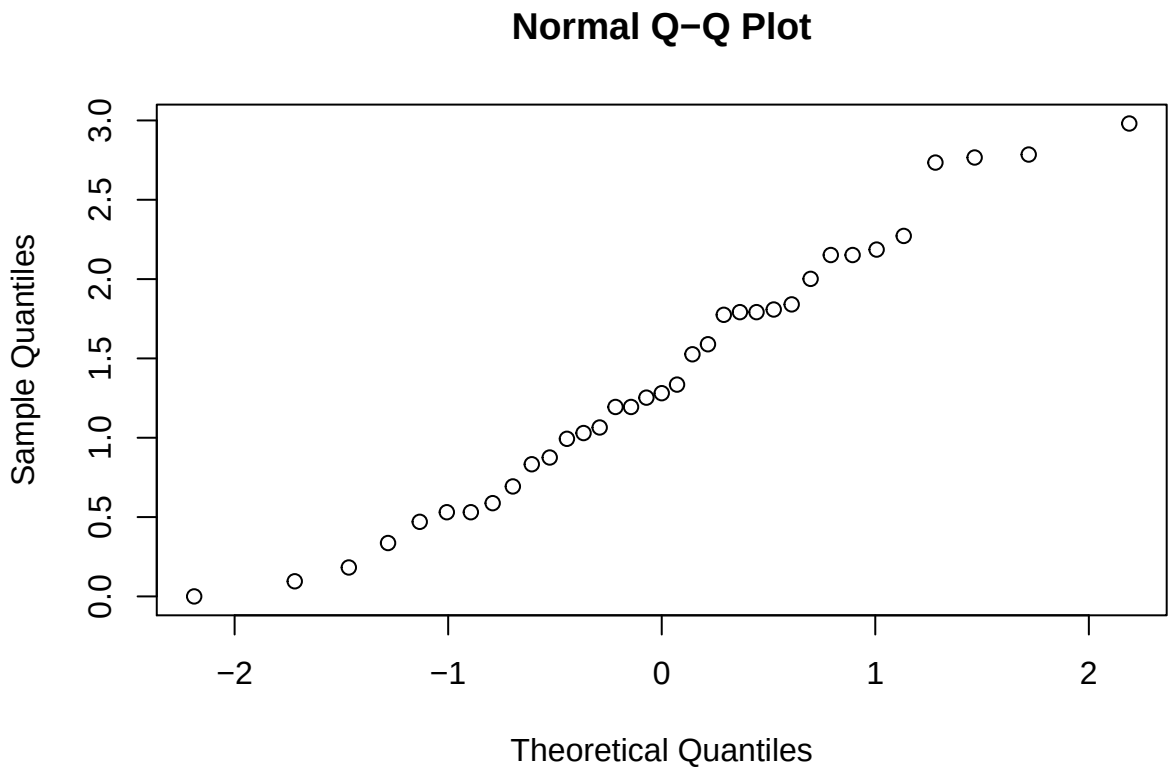
```
log.treatment = log(treatment)
log.control = log(control)
boxplot(log.treatment, log.control)
```



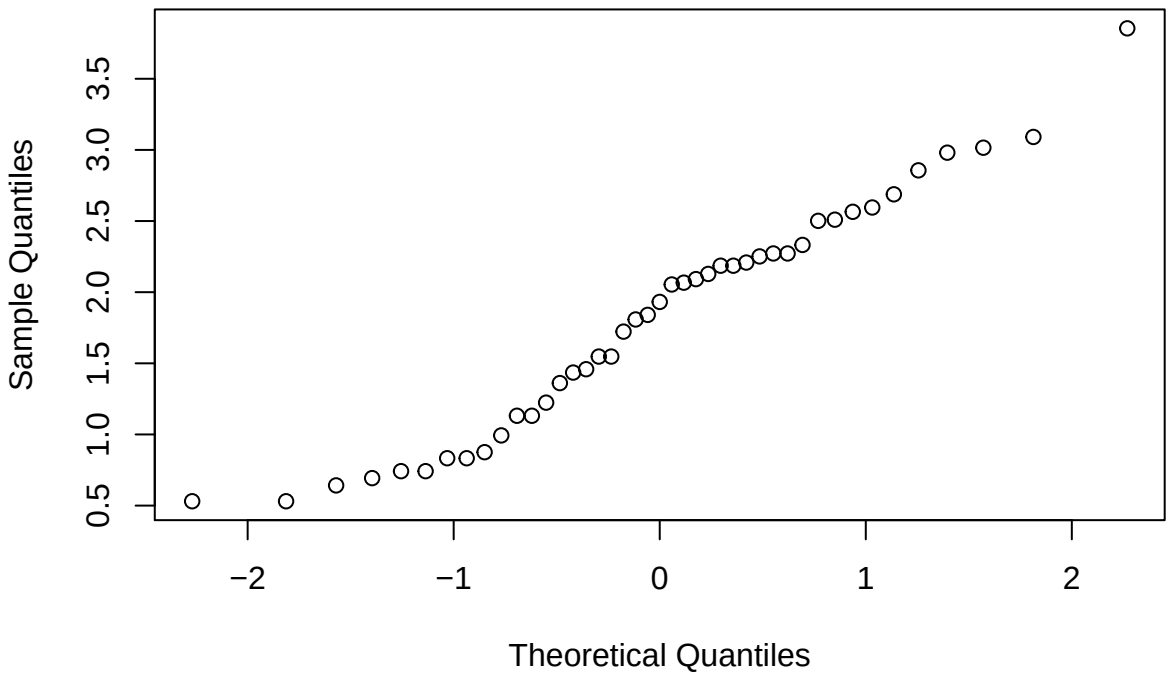

```
plot(density(log.treatment))
lines(density(log.control), col = "red")
```




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




```
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.) This is getting toward the small side, so there's some 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(-0.8, -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
```

Briefly: The Wilcoxon rank-sum test

The **Wilcoxon rank-sum test**, which is confusingly also called the **Mann-Whitney test** because as you know by now everything in statistics has multiple names, uses the *ranks* of the data rather than the raw values. See pp. 282-287 of Trosset for the full details. The cartoon idea for a two-tailed test is:

1. Rank all the data in both sample together from smallest to largest;
2. Find the average rank in each sample;
3. See if the average rank for the two samples is about the same (big P-value) or quite different (small P-value.)

The null hypothesis is that the observations are "exchangeable." This is easiest to understand (and justify) in the context of a randomized controlled experiment: then every assignment of the subjects to treatment and control was equally likely. If it made no difference at all whether a participant got the treatment or control, then every permutation of ranks would be equally likely and with decent sample sizes, the mean rank should be about the same for treatment and control.

The basic version of the test is easy to do in R. For the stereograms experiment:

```
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
```

(There's a slightly better function `wilcox_test()` in the `coin` library, but unless a large proportion of your data takes the same value it doesn't make too much difference.)

Note that we avoid having to choose and interpret a transformation. The small P-value suggests (roughly speaking) the stereogram fusion times aren't just completely randomly distributed regardless of whether a subject got the treatment or the control. A cursory inspection of the data once again suggests the most plausible conclusion is that the control reduces fusion times.

The main advantage of using Wilcoxon's rank-sum test rather than some transformation is that like other rank-based methods, it eliminates the problem of outliers (since you can never have an outlying rank.) On the other hand, if the problem with the data is skewness and a log transformation is possible, then inference on a log scale is usually easier to interpret than inference on the ranks.

An alternative way to justify the Wilcoxon rank-sum test, preferred by Trosset, is if you have two independent samples whose distributions differ only by a shift. This assumption has the advantage of allowing you to estimate the shift parameter and calculate a meaningful confidence interval for that shift. See Trosset for the details.