Bootstrap Hypothesis Test

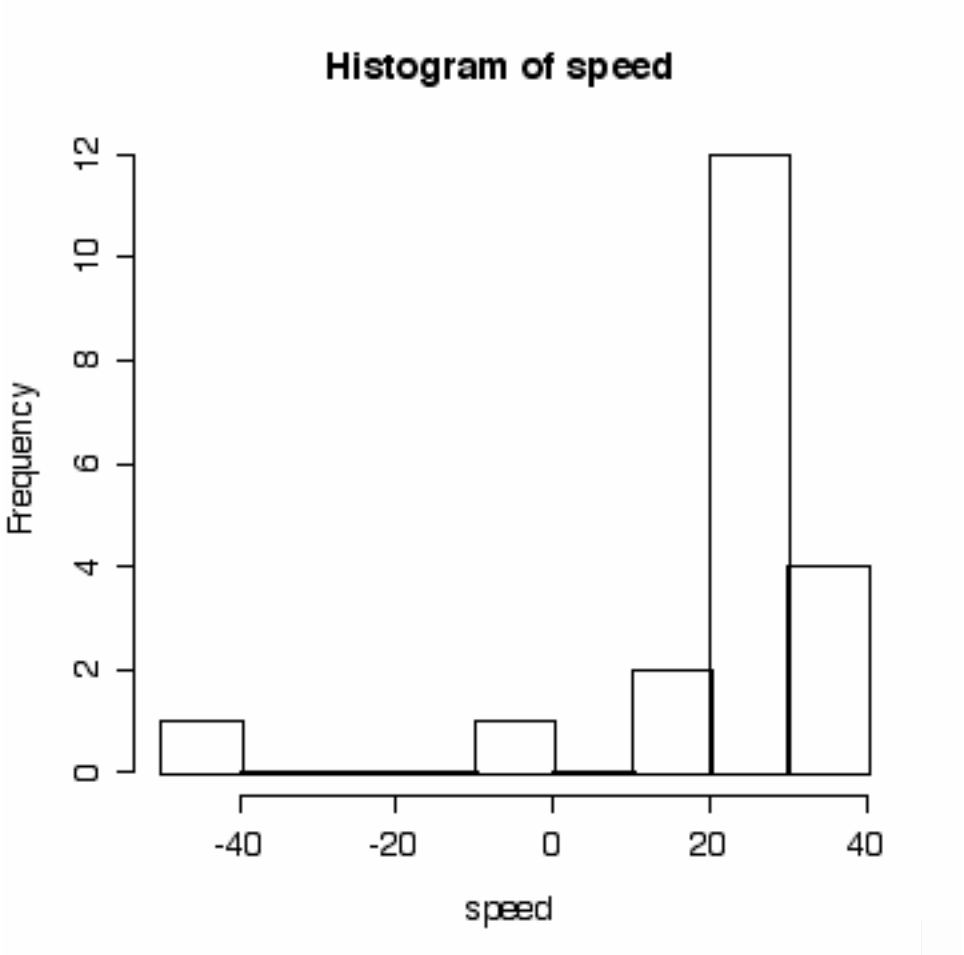
In 1882 Simon Newcomb performed an experiment to measure the speed of light. The numbers below represent the measured time it took for light to travel from Fort Myer on the west bank of the Potomac River to a fixed mirror at the foot of the Washington monument 3721 meters away. In the units in which the data are given, the currently accepted "true" speed of light is 33.02. (To convert these units to time in the millionths of a second, multiply by 10^{-3} and add 24.8.)

28 -44 29 30 26 27 22 23 33 16 24 40 21 31 34 -2 25 19

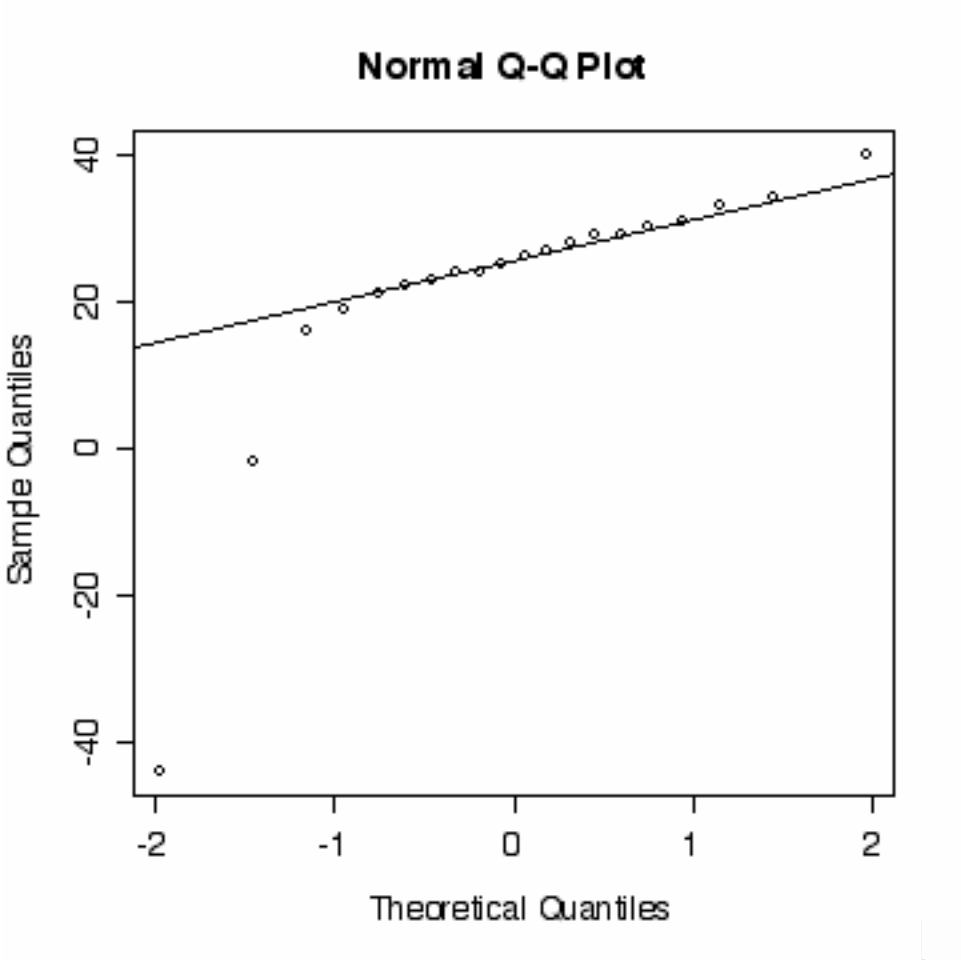
Does the data support the current accepted speed of 33.02?

First, note that the data are not normal:

```
> speed <- c(28, -44, 29, 30, 26, 27, 22, 23, 33, 16, 24, 29, 24,
40 , 21, 31, 34, -2, 25, 19)
> hist(speed)
```



```
> qqnorm(speed)
> qqline(speed)
```



To do the t-test we must assume the population of measurements is normally distributed. If this is not true, at best our tests will be approximations. But with this small sample size, and with such a severe departure from normality, we can't be gurarantee a good approximation.

The bootstrap offers one approach.

Step 1: State null and alternative hypotheses:
H0: mean = 33.02
Ha: mean <> 33.02
Step 2: Set the significance level . We'll choose 5%.
Step 3: Choose a test statistic. We wish to estimate the mean speed, and therefore we'll use the sample average.
Step 4: Find the observed value of the test statistic:

```
> mean(speed)
[1] 21.75
```

We now need the p-value, but to do this we need to know the sampling distribution of our test statistic when the null hypothesis is true. We know it is approximately normal, but also that the approximation might not be very good here.

So our approach instead is to perform a simulation under conditions in which we know the null hypothesis is true.

What we'll do is use our data to represent the population, but first we shift it over so that the mean realy is 33.02:

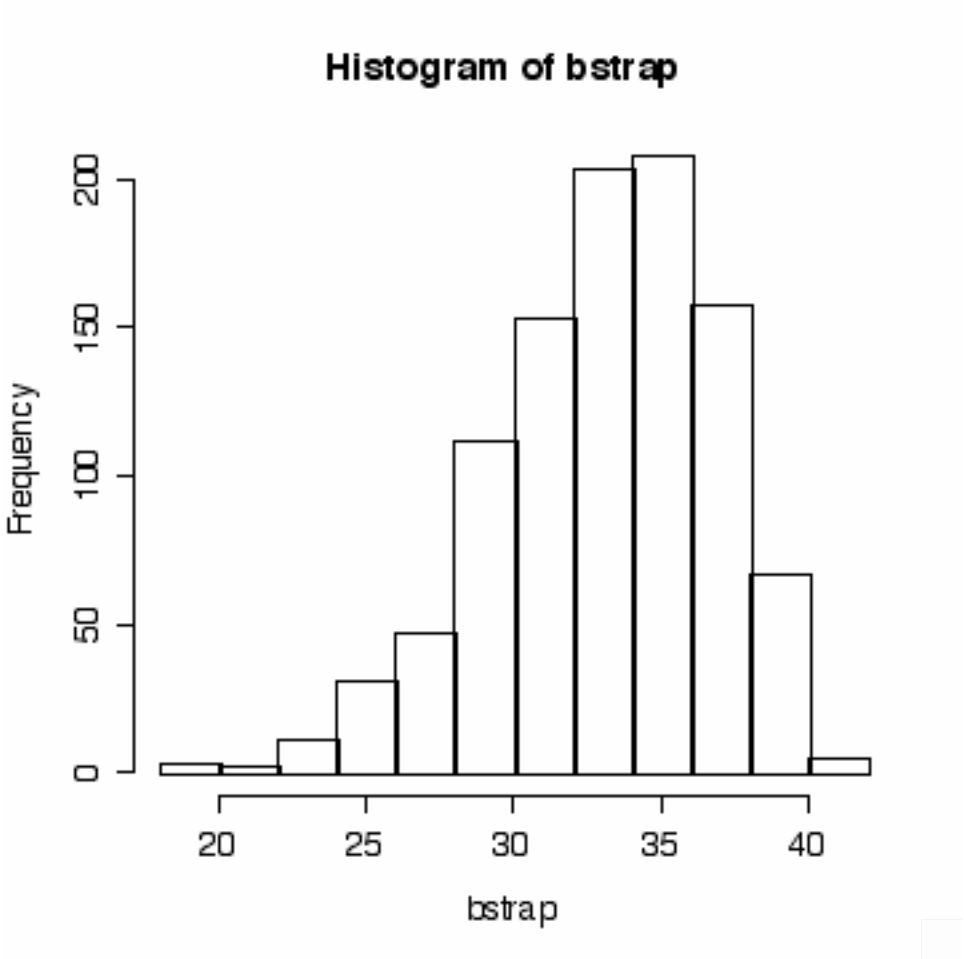
```
> newspeed <- speed - mean(speed) + 33.02
> mean(newspeed)
[1] 33.02
```

The histogram of newspeed will have exactly the same shape as speed, but it will be shifted over so that it is centered at 33.02 rather than at 21.75.

Now we reach into our fake population and take out 20 observations at random, with replacement. (We take out 20 because that's the size of our initial sample). We calculate the average. It will be close to 33.02, because that is the mean of the population. We'll save it, and then repeat this process again. After many repetitions, we'll have a good idea of what sort of sample averages we should expect when the true mean is 33.02. We will then compare these to our observed sample average (21.75) to see if it's consistent with our simulated averages. If so, then perhaps the mean really is 33.02. If not, then perhaps Newcomb had some flaws in his experimental design and was measuring incorrectly.

Here's some code:

```
> bstrap <- c()
> for (i in 1:1000){
+ newsample <- sample(newspeed, 20, replace=T)
+ bstrap <- c(bstrap, mean(newsample))}
> hist(bstrap)
```



This distribution doesn't look normal, which means that we did the right thing. With a larger sample size it would have looked normal, but 20 apparently isn't large enough. We can't, therefore, trust the normal approximation, and our bootstrap approach will be stronger.

As you can see, it's not impossible for the sample average to be 21.75 even when the true mean is 33.02. But it's not all that common, either.

The p-value is the probability of getting something more extreme than what we observed. 21.75 is 33.02 - 21.75 = 11.27 units away from the null hypothesis. So our p-value is the probability of being more than 11.27 units away from 33.02. This is P(Test Stat < 21.75) + P(Test STat > 44.29). We don't know the sampling distribution of our test statistic, but our bootstrap sample lets us estimate this probability:

```
> (sum(bstrap < 21.75) + sum(bstrap > 44.29))/1000
[1] 0.004
```

Therefore, we estimate the p-value to be 0.004. (In other words, in 4 times out of 1000, we had a sample average this extreme.)

Because our significance level is 5% and .4% < 5%, we reject the null hypothesis and conclude that Newcomb's measurements were not consistent with the currently accepted figure.

We did not have to use the sample average as our test statistic. We could have used a t-statistic, and calculated the observed value:

```
> (mean(speed) - 33.02)/(sd(speed)/sqrt(20))
[1] -2.859247
```

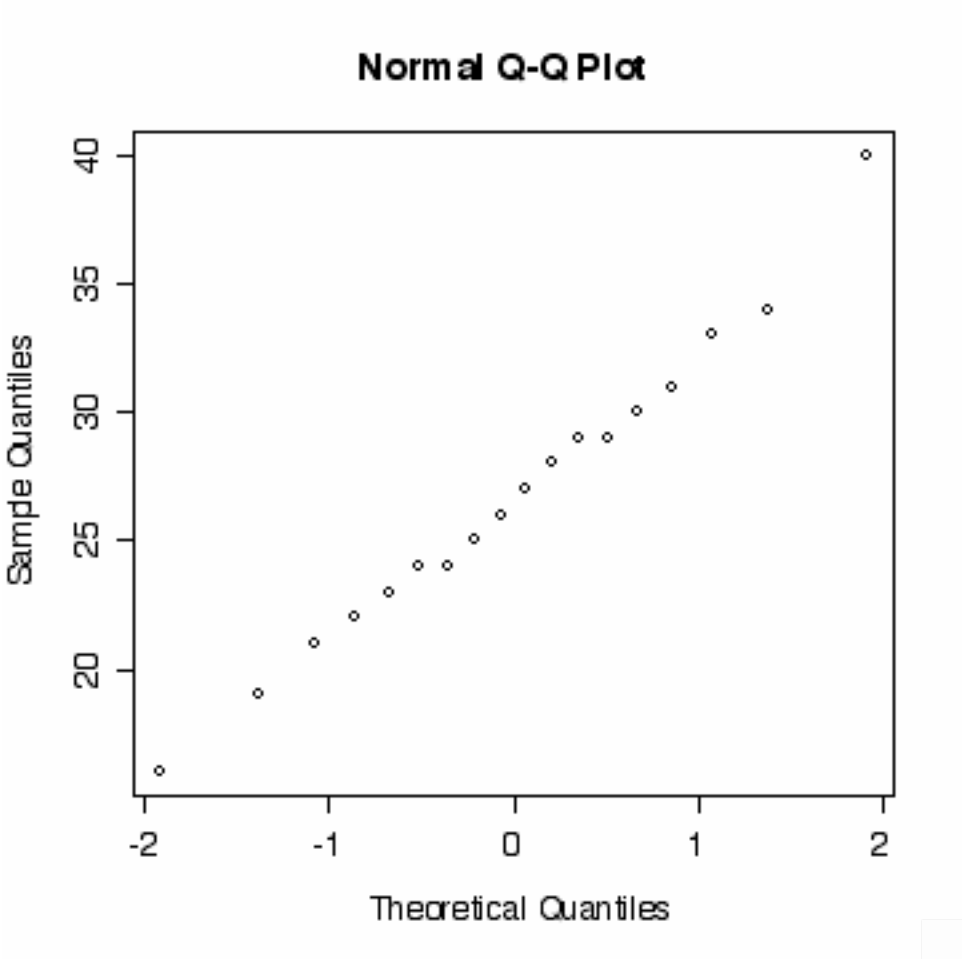
Then we would repeat as before, but instead of simply calculating the mean of the re-sample in our bootstrap code, we would instead calculate this t-statistic (replacing "newspeed" wherever "speed" appears in the formula above.)

Note that there were two extreme observations: -44 and -2. The t-test, because it depends on the sample average, is notoriously susceptible to being influenced by extreme observations. Let's take those values out and see what happens:

```
> speed
 [1]  28 -44  29  30  26  27  22  23  33  16  24  29  24  40  21
31  34  -2
[19]  25  19
> betterspeed <- speed[-c(2,18)]
> betterspeed
 [1] 28 29 30 26 27 22 23 33 16 24 29 24 40 21 31 34 25 19
```

The -44 and -2 occur in the 2nd and 18th positions of speed. The speed[-c(2,18)] part of the command refers to every item in the vector speed, with the 2nd and 18th removed.

```
> qqnorm(betterspeed)
```



Note that things look much more normal. We can still do our bootstrap test:

```
> newspeed <- betterspeed - mean(betterspeed) + 33.02
> mean(betterspeed)
[1] 26.72222
> bstrap <- c()
> for (i in 1:1000){
+ bstrap <- c(bstrap, mean(sample(newspeed,20,replace=T)))}
```

Now you'll see that the observed value of our test statistic is 26.7222.

Our pvalue is now the probability of seeing something more than (33.02 - 26.722) = 6.298 units away from 33.02.

We calculate this as:

```
> (sum(bstrap < 26.722) + sum(bstrap > 39.218))/1000
[1] 0
```

It's so extreme, that in 1000 repetitions, we never saw numbers that extreme.

What if we used the t-test? Since the data now look normal, there's no reason not to.

```
> t.test(betterspeed,alternative="two.sided",mu=33.02)

     One Sample t-test

data:  betterspeed
t = -4.6078, df = 17, p-value = 0.0002508
alternative hypothesis: true mean is not equal to 33.02
95 percent confidence interval:
 23.83863 29.60582
sample estimates:
mean of x
 26.72222
```

As you can see, the p-value is not exactly 0, but is quite small.