Chapter 10: One sample location problems

S520

These notes are written to accompany Trosset chapter 10. We focus on section 10.1 in detail and skim over the rest.

Questions to ask

Before rushing into doing a significance test or finding a confidence interval, here are some question you should ask. The answers will help you work out what method you should use.

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. We'll learn how to do inference about Δ in chapter 11.

Remember our previous rules about tails: If the test has no direction, do a two-tailed test with the null $H_0 : \mu = \ldots$ or $H_0 : \Delta = \ldots$. If the test has a direction, do a one-tailed test where the alternative hypothesis is what you want to be able to show beyond reasonable doubt.

Finally, if you want to make a binary decision at the end, you need to choose a significance level α, where α will be the maximum probability of rejecting the null hypothesis when it's true. Then at the end, you'll reject the null hypothesis in favor of the alternative if the P-value is $\leq \alpha$; otherwise, you won't reject the null.

See Trosset pp. 234–236 for examples of setting up hypothesis tests.

Normal populations

In chapter 9, we did inference based on an assumption of an approximately normal distribution for the error $\bar{X} - \mu$, with a standard error of about $s/\sqrt{n}$. This requires a large sample for two reasons:

- The approximate normal distribution for the error is justified by the Central Limit Theorem.
- The real standard error is $\sigma/\sqrt{n}$, but with a large sample, s will be close to σ.

What if we don't have so large a sample? Then we need to make further assumptions. For example, if we have a distribution with extreme skewness or bad outliers, we might not be able to say anything accurate about the population mean from a small sample.

Example. Consider the population consisting of a STAT S-520 class plus Bill Gates. Suppose we wish to estimate the population mean wealth by taking the sample mean with $n = 10$. Well, if our sample of ten includes Bill Gates, we'll horrendously overestimate the average wealth, and if our sample of ten doesn't include Bill Gates, we'll horrendously underestimate the average wealth. A simple random sample of size n just isn't big enough for this problem.

So for now let's deal with situations where we know the form of the population distribution. Suppose we take a sample of size n from a Normal distribution with mean μ and variance σ^2. Then the sample mean has distribution

$$\bar{X} \sim \text{Normal}(\mu, \sigma^2/n).$$

What if we don't know μ – what if μ is what we're trying to find out? Then

$$\text{Error} = \bar{X} - \mu \sim \text{Normal}(0, \sigma^2/n).$$

This will be easier to deal with mathematically if we rescale to get a standard normal distribution. So divide through by $\sigma/\sqrt{n}$:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1).$$

So if we know σ, we can do inference using `pnorm()` and `qnorm()`, even with small samples. However, most of the time we don't know σ. Instead we have to use s, the sample standard deviation instead. We've previously asserted that for large samples this is fine. But what happens for moderate or small n? Let's define the t-statistic:

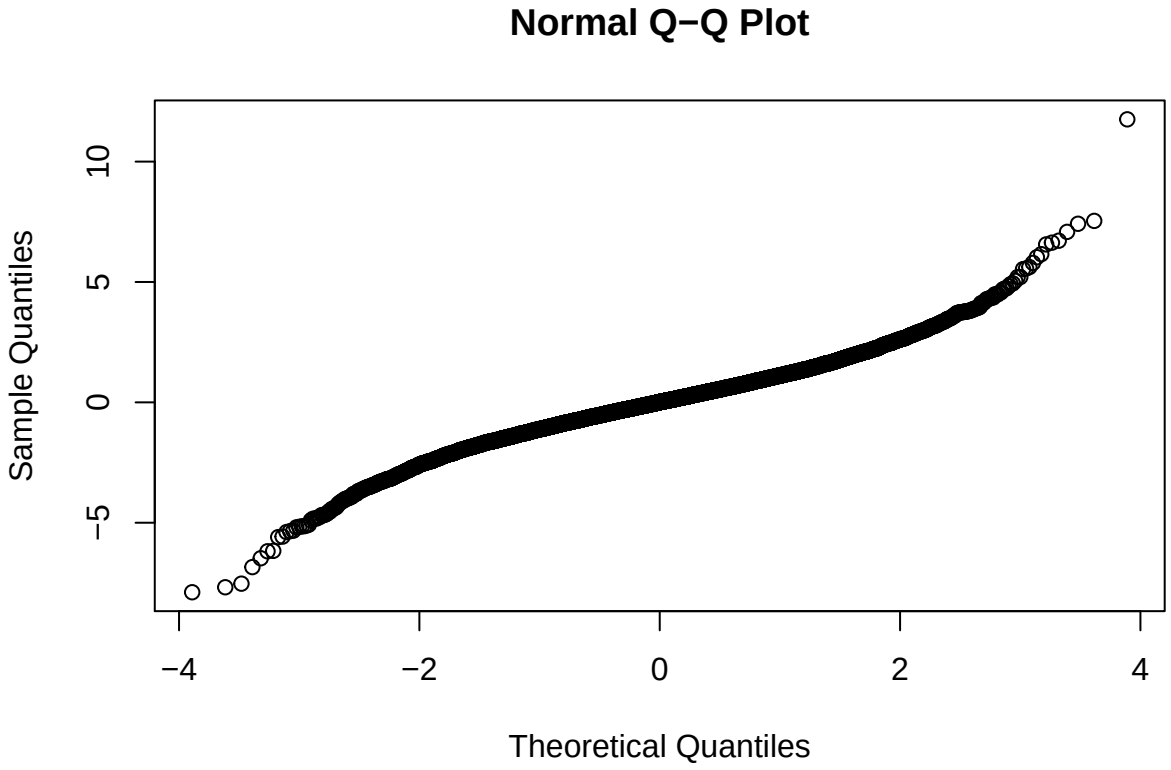
$$t_n = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Trosset p. 240 goes through some of the theory, but I'll proceed by simulation. Consider the case where we take a sample from a normal population and calculate this statistic t_n.

```
t.statistic = function(n, mu = 0, sigma = 1) {
    my.sample = rnorm(n, mu, sigma)
    x.bar = mean(my.sample)
    s = sd(my.sample)
    t = (mean(my.sample) - mu)/(s/sqrt(n))
    return(t)
}
```

Now replicate this lots of times with $n = 6$. Is the distribution normal, or something else?

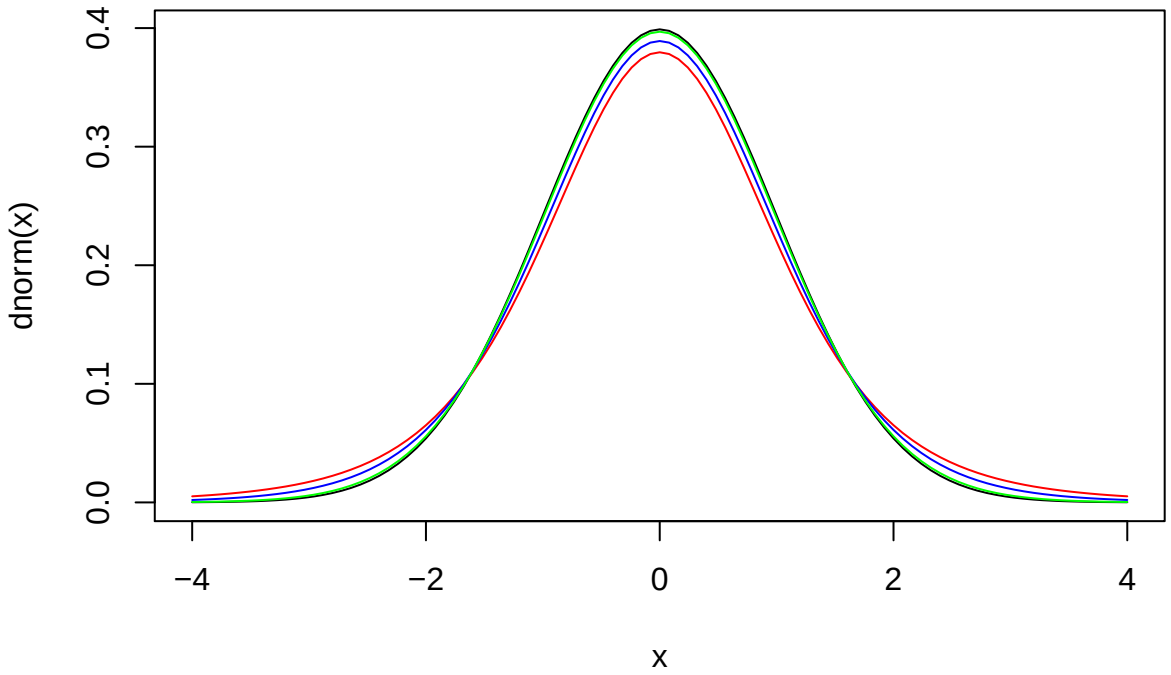
```
t.list = replicate(10000, t.statistic(n = 6))
qqnorm(t.list)
```



Nope! Instead the distribution is something called a t-distribution, and here in particular it's a t-distribution with 5 degrees of freedom. Why 5? <handwave> The sample size is 6, so if the mean is given, then 5 observations can vary freely; the sixth is determined by the other 5.</handwave> In general, if the sample size (from a Normal population) is n, the t-statistic follows a t-distribution with $n-1$ degrees of freedom.

We can get a better idea of what this is by plotting some PDFs:

```
curve(dnorm, from = -4, to = 4)
curve(dt(x, df = 5), col = "red", add = TRUE)
curve(dt(x, df = 10), col = "blue", add = TRUE)
curve(dt(x, df = 50), col = "green", add = TRUE)
```



The PDF of the t-distribution with 5 degrees of freedom is given by the red curve. Like the normal, it's symmetric. Unlike the normal, it doesn't die away so quickly and you get further away from the center. The greater the "degrees of freedom", the closer you get to the Normal curve. We can see the difference by finding some tail probabilities:

```
pnorm(-1.75)
```

```
## [1] 0.04005916
```

```
pt(-1.75, df = 5)
```

```
## [1] 0.07026118
```

```
pt(-1.75, df = 10)
```

```
## [1] 0.05534047
```

```
pt(-1.75, df = 50)
```

```
## [1] 0.04312663
```

To summarize: When the population is Normal, the variable $T_n = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a t-distribution with $n-1$ degrees of freedom. Thus the tail probabilities of the t-distribution can be used to construct a P-value for a test of the hypothesis that the population mean takes a certain value.

pnorm or pt?

Should you use the normal or the t-distribution for tests and confidence intervals for μ?

- If you know the true value of σ, there's no reason to use the t-distribution. This is rarely this case, however.
- If you have a large sample size, then `pnorm(t)` and `pt(t, df = n - 1)` will give almost the same probability, so it hardly matters. By the Central Limit Theorem, both of them will give you an approximately right answer, and this is true regardless of whether the underlying population is normal or not. If you really are using s instead of σ, then it's a bit more honest to use the t-distribution instead of the standard normal, because that's the case the t-test was designed for.

- If you have a moderate sample size and you're using s instead of σ, the t-test is better because the normal distribution will underestimate the tail probabilities. You still need to draw a QQ plot and check there's no big, systematic bend – if there is, you may need a transformation or another method.
- If you have a small sample size, then if at all possible, get more data. With very small samples (less than 20), a normal population might give a straight QQ plot or a curved QQ plot, while a skewed population might give a straight QQ plot or a curved plot – so the plot isn't of much use. Even with samples of 30 or 40, it can be hard to be sure whether the population is symmetric or skewed. If you can't get more data, then using the t-test requires a leap of faith that the population is close to normal.

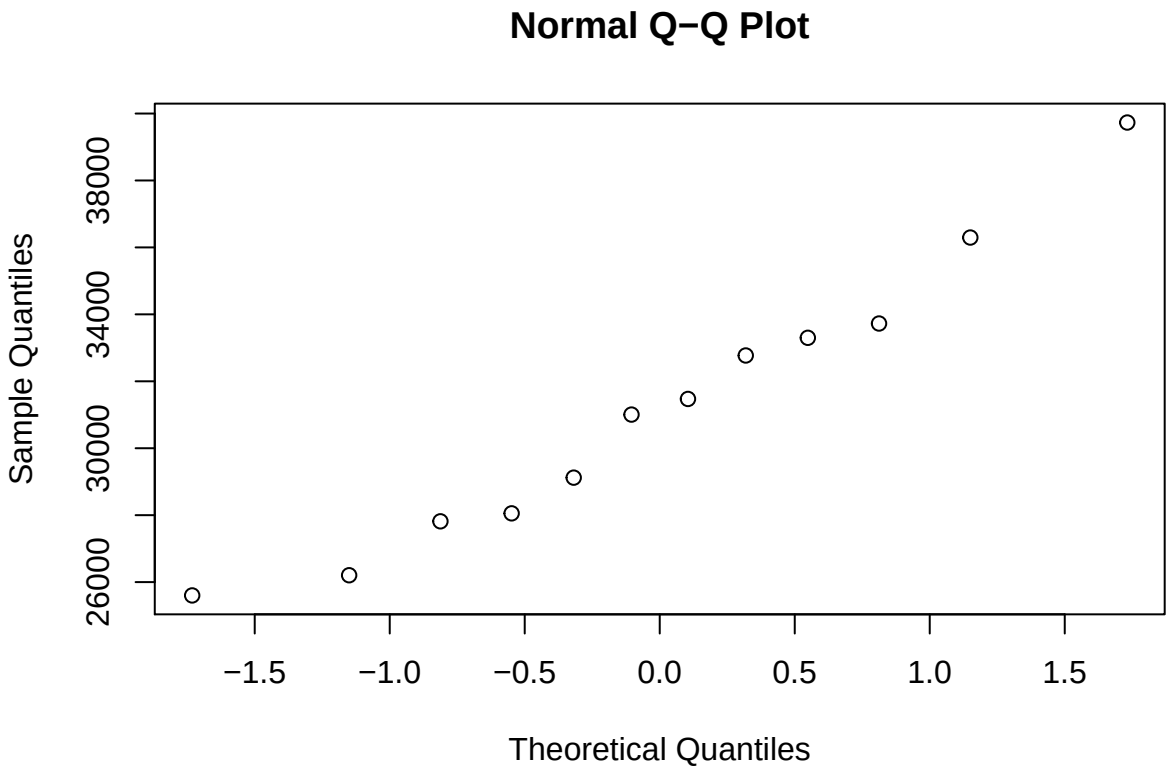
Example: Country club fees

Here's an example from a business statistics textbook I used to use. In 2008, the average country club fee was \$31,912. We have a sample of twelve country club fees from 2009. Has the average fee changed? Here's the 2009 data:

```
fees = c(29121, 31472, 28054, 31005, 36295, 32771, 26205, 33299, 25602, 33726,
    39731, 27816)
```

Are the fees normal?

```
qqnorm(fees)
```



The QQ plot is fairly close to a straight line. Unofortunately, there's no way of being sure that the population is approximately normal just from a sample of size 12. There could be huge outliers outside of our sample – this seems quite possible when it comes to country club fees. Nevertheless, we'll make a leap of faith, and assume the population is approximately normal.

```
mean(fees)
```

```
## [1] 31258.08
```

```r
sd(fees)
```

```
## [1] 4199.802
```

```r
t = (mean(fees) - 31912)/(sd(fees)/sqrt(12))
# Two-tailed P-value
2 * pt(t, df = 11)
```

```
## [1] 0.6003808
```

```r
# or
2 * (1 - pt(abs(t), df = 11))
```

```
## [1] 0.6003808
```

```r
# Find 95% CI
mean(fees) - qt(0.975, df = 11) * sd(fees)/sqrt(12)
```

```
## [1] 28589.66
```

```r
mean(fees) + qt(0.975, df = 11) * sd(fees)/sqrt(12)
```

```
## [1] 33926.51
```

What can we conclude?

- The data is compatible with the hypothesis that the average country club fee hasn't changed from 2008 to 2009.
- But the confidence interval is fairly wide. It could be that the average fee has dropped $3000, or it could be that it's risen $2000.
- The most glaring problem is the small sample size. With a sample of size 12, then you would be unlikely to detect a change of, say, a couple of thousand dollars.
- And this is with a suspiciously small standard deviation. Surely real country club fees vary a lot more than this!

Conclusions: (1) Get more data. (2) Get better data. (3) Get a better business statistics textbook.

We tested the hypothesis that the population mean was $31,912. Since the sample was small, we couldn't rely on the Central Limit Theorem, so instead we needed to assume the data came from an approximately normal population. Although the QQ plot looked straightish, this was still a leap of faith, because it's hard to show something's normal from a small sample – what if there are large outliers that we didn't happen to get in our particular sample? This is particularly easy to believe for something like country club fees, where there might be club that charge **much** more than most others. A priori, we might expect the distribution of fees to be right-skewed.

So what if you don't want to assume normality? There are a couple of options.

1. Get more data. This is the most honest approach, but may not be practically feasible.
2. *Transform* the data to have a closer-to-normal distribution. This is particularly useful with positive, right-skewed data, where a log transformation often makes the data a lot closer to normal (and is able to be interpreted.) However, with small samples, we still face the problem that it's hard to know if it's normal whether or not you do a transformation.
3. Do a **nonparametric test** that makes weaker assumptions. For example, instead of studying the mean, you might study the **median**. It turns out that you can study the median without making any assumption about kind of distribution the population, even with smallish sample (see Trosset 10.2 and below.) In addition, maybe the median really is more inherently interesting than the mean, which is often the case with skewed distributions.

What does a "significant" difference mean?

A P-value smaller than α is sometimes called "evidence of a statistically significant difference between the data and the null hypothesis," or just "statistically significant." What does this phrase really mean?

Mathematically, it means this. "Before we did our test, we picked a significance level α. The data that we have is so extreme, that if the null were true, we would see data this extreme less than α of the time. Therefore, using a rule that rejects the null hypothesis when it's true a maximum proportion α of the time, we can reject the null hypothesis."

Intuitively, it means something like this. "We have a null hypothesis model. The data we have would be unusual under this null hypothesis model. Therefore, we reject our null hypothesis."

Notice that nothing we have said here tells us *how different* our data is from what we'd expect from the null hypothesis. Perhaps our null hypothesis is that $\mu = 0$. A small P-value means there's strong evidence against our null hypothesis, but it doesn't tell us what μ is if it's not zero. That's what a point estimate or confidence interval is for, and that's why we need them: it could be that our null hypothesis isn't literally true, but it's only wrong by a tiny amount. *A statistically significant difference is one that is unusual, not one that is big and important.* One moral is that when your hypotheses involve a parameter you can estimate (such as μ,) you should almost always include a confidence interval for this parameter.

Secondly, unusual things happen sometimes. In particular, with enough repetitions, unusual things become likely. The probability that you win the lottery is tiny, but if hundreds of millions of tickets are sold, the probability that *someone* wins the lottery may be large. So if a researcher works in a field where the standard is to reject the null if the P-value is less than $\alpha = 0.05$, then if the researcher does a hundred tests, they can expect five significant results *even if all the nulls are true.* It follows that using a fixed significance level like 0.05 if you're doing many tests is at best dubious and at worst cheating. This issue is discussed further in chapter 12.4.

Finally, it's important to remember that statistical significance might not have anything to do with cause and effect. Remember back to the first week of the course. If you want to measure cause and effect, you need a randomized experiment, or something very close to it. If there's confounding, then an observational study will not accurately describe cause and effect. This is true whether or not the result is significant.

What does a not-significant difference mean?

Suppose we wish to test the null hypothesis that the average height of adult men is six foot nothing. Suppose we wish to do this by taking IID samples of size 3 from the population of size 3. What could possibly go wrong?

We do the test using a R function I wrote called `six.foot.test()` (which I'm hiding because the syntax isn't relevant.) Let's run the test and see what happens.

```
six.foot.test(n = 3)
```

```
## P-value is 0.85
```

Hmmn, we fail to reject at any reasonable significance level. Maybe we just got a bad sample? Let's try again:

```
six.foot.test(n = 3)
```

```
## P-value is 0.66
```

Maybe we got two bad samples. Try again:

```
six.foot.test(n = 3)
```

```
## P-value is 0.11
```

We failed to reject three times. Should we conclude that the average height of adult men is exactly six feet?

No. The real reason we couldn't reject was that our sample sizes were much, much too small. With small samples, we may not be able to reject even hypotheses that are obviously false.

The traditional way of think about errors in hypothesis testing is by talking about two kinds of errors:

- **Type I error:** Rejecting the null hypothesis when in fact, the null is true.
- **Type II error:** Failing to reject the null hypothesis when in fact, the alternative is true and the null should be rejected.

Instead of Type II error, we sometimes talk about **power**: the probability of correctly rejecting the null hypothesis when the alternative is true. (This is one minus the probability of a Type II error.) We leave aside the calculation of power and error probabilities as beyond the scope of this course, but we'll think about the concepts.

With small samples, you have little power. That is, your chance of rejecting the null even when you should is very low; the probability of a Type II error is very high. A larger sample size means higher power. This is good.

When your null hypothesis is very wrong, it might not take a large sample to ensure high power. For example, you should safely be able to the reject the obviously wrong hypothesis that the average height of adult men is 6'0" from a random sample of fifteen men.

Of course, most of the time we don't want to test hypotheses that are obviously wrong. Then it might take a much larger sample to get sufficient power. If you're going to do a test with any real-world importance, you should consult with a statistician *before* you collect your data to ensure that your design has enough power that you're not wasting your time and money. If you don't have a statistician on hand, a rule of thumb is that a sample size of at least 50 is necessary to give you a fighting chance of detecting a moderate-sized difference. (The numbers behind the rule: A sample size of 50 gives you 93% power for a two-tailed test if the true mean is half an SD bigger than the null hypothesis mean.)

Another sanity check is, once again, to calculate a confidence interval. If the confidence interval is wide (in real-world terms), then that's a sign your sample size isn't large enough to accurately know what the true value of μ is. So if your confidence interval for the average height of adult men is 5'1" to 6'4", that tells you that you don't have any useful idea what the true population mean is and a significance test is unlikely to be illuminating. Get more data.

tl;dr: A failure to reject the null doesn't prove the null is true. It might be that the null is false, but you don't have enough data to show it.

Transformations

The t-test assume samples come from a normal distribution. Now, hardly anything is really from a normal distribution, so approximately normal is fine. In practice, that means no strong skewness and no bad outliers, as shown by a normal QQ plot.

If there's skewness, a transformation may help. As before, our first choice with positive data is the log transformation. We hope that after taking logs, we get a normal QQ plot with no systematic bend. This works surprisingly often!

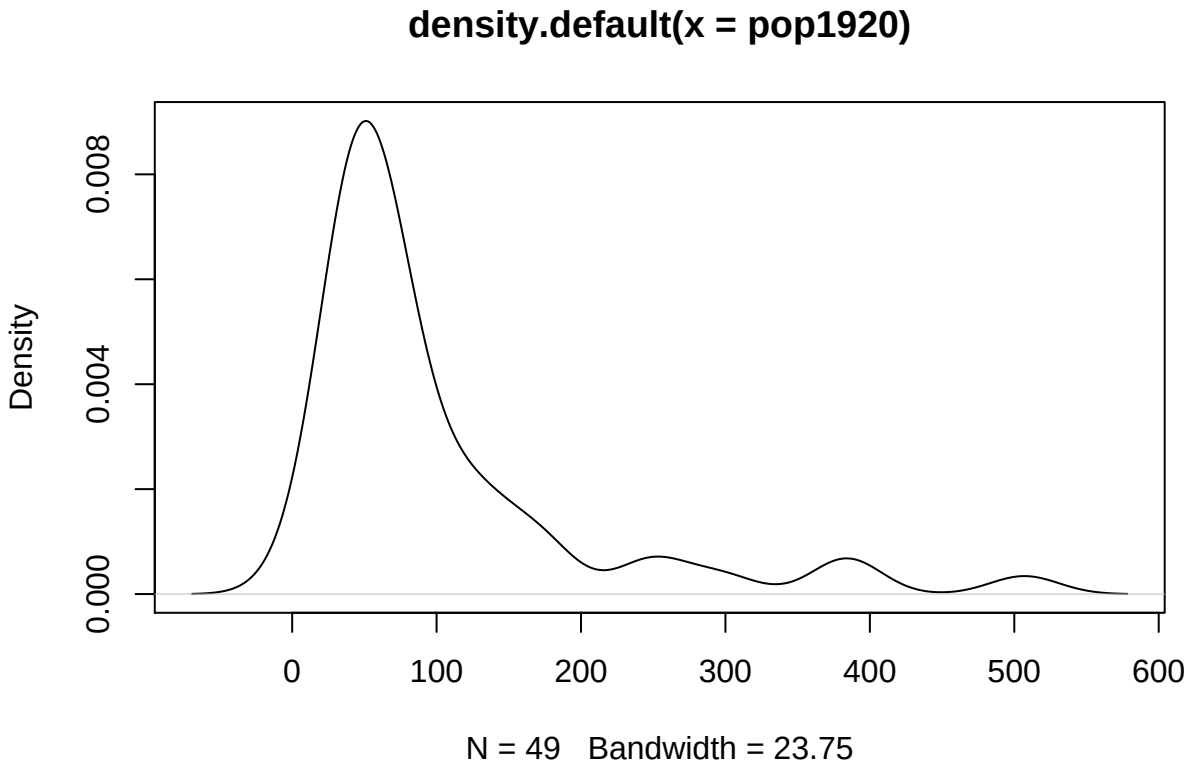
The one thing you have to be careful of is the meaning of your parameter μ. Specifically, if you take (base e) logs and then do a one-sample t-test, the μ in your hypothesis is now the mean of the (base e) logs of your population. (This is *not* the same thing as the log of the population mean.)

The mean of the logs is not something entirely intuitive. So after you find a confidence interval, you'll usually want to **back-transform** back to the original scale by taking the exponential of both ends of the interval.

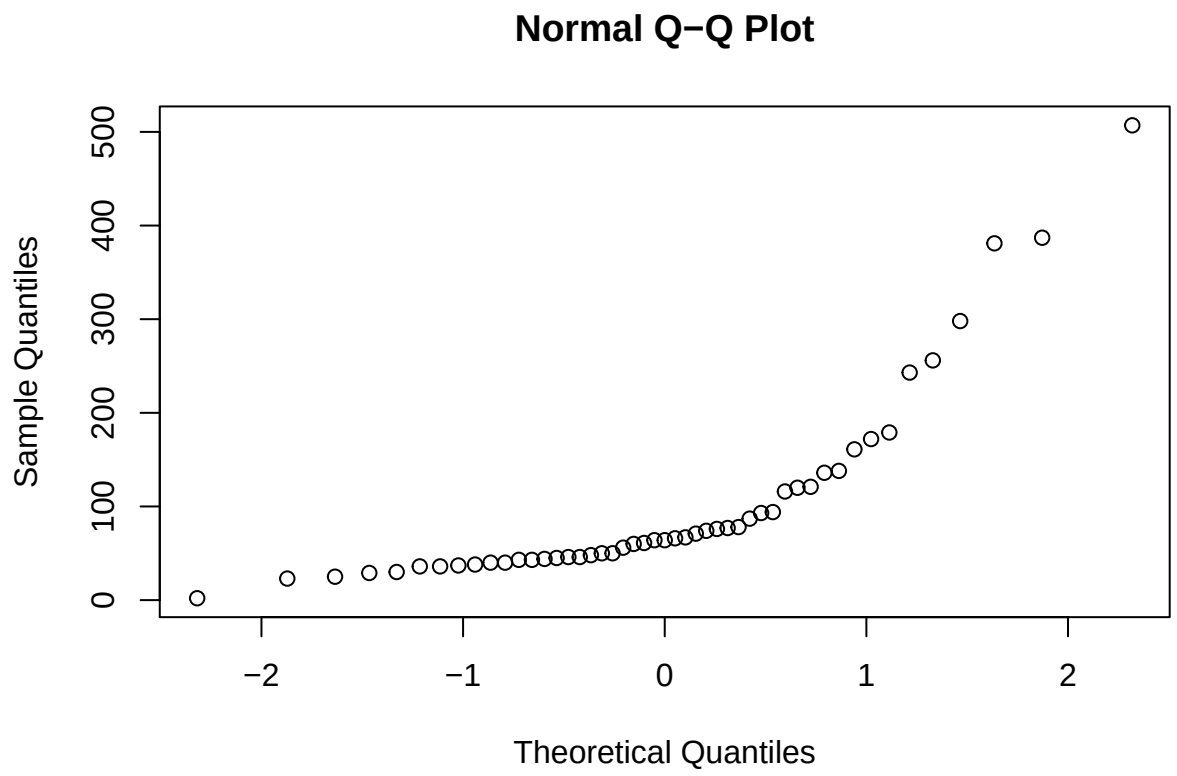
Example. The data set `bigcity` within the `boot` library contains the population in thousands of a random sample of 49 of the 196 "big cities" in the U.S. in 1920 (variable `u`) and 1930 (variable `x`.) What can we say about the typical size of a U.S. big city in 1920?

We don't have any specific hypothesis to test here, so we'll just find some kind of 95% confidence interval. First, draw some pictures:

```
library(boot)
pop1920 = bigcity$u
plot(density(pop1920))
```

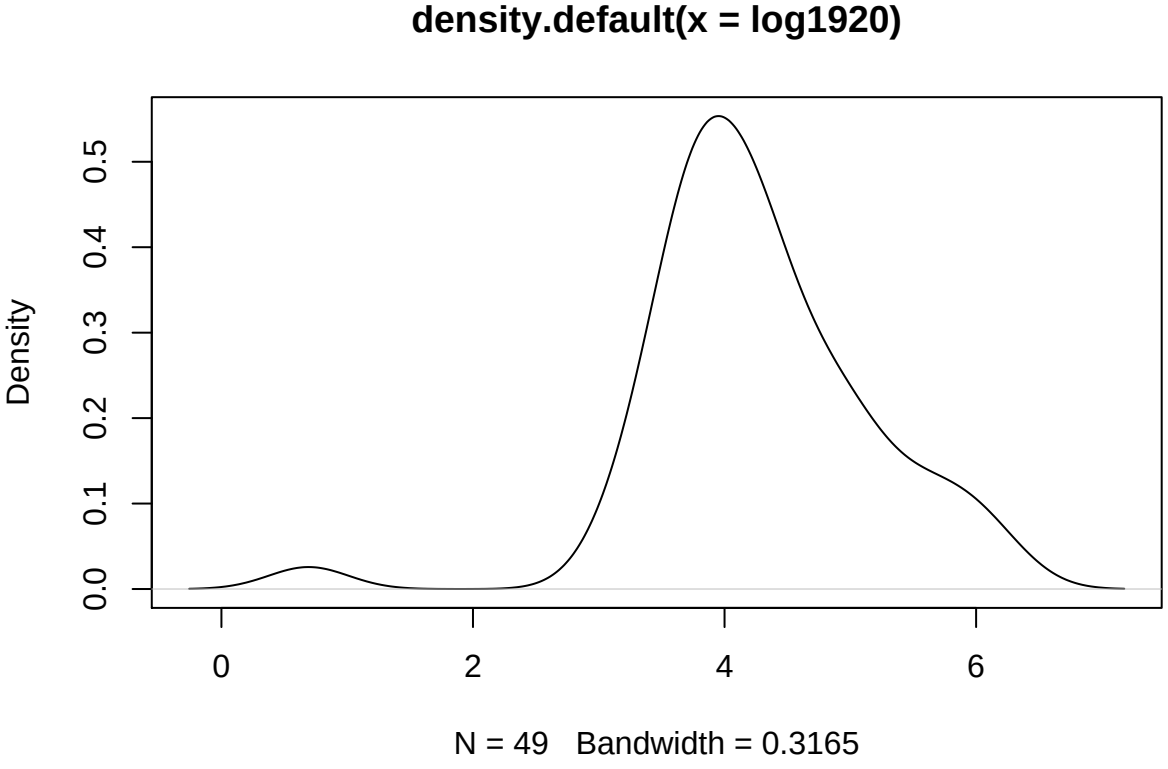



```
qqnorm(pop1920)
```

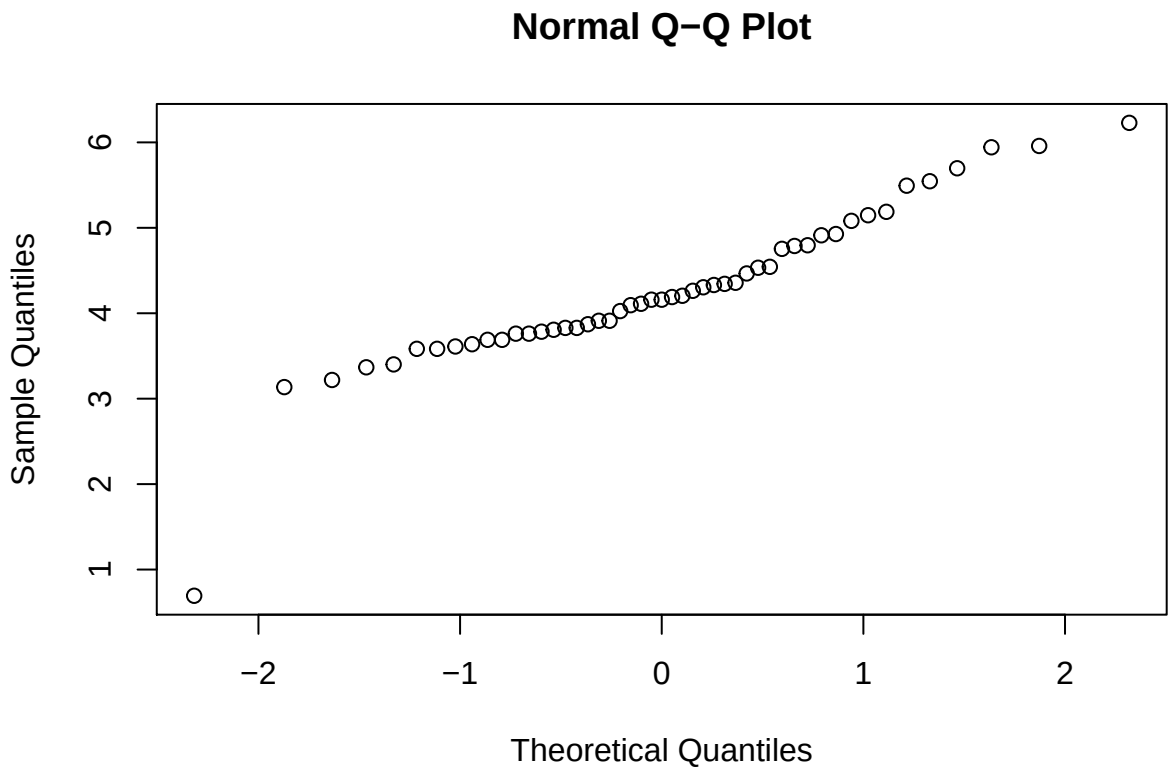


We have a skewed, non-normal sample. The sample isn't quite big enough that we can rely on the Central Limit Theorem to get us out of trouble, so we'll take logs and see if things improve.

```
log1920 = log(pop1920)
plot(density(log1920))
```



```r
qqnorm(log1920)
```



We reduce but do not entirely remove the skewness. In addition, we introduce a new problem: an outlier on the left. Somehow, a city with 2,000 people was counted as a "big city," and it really sticks out once you take logs.

Since both the raw data and the logged data show some non-normality, what do we do? The best option is to learn more statistics and skip to the sign test section below. However, with a decent sample size, the logged data isn't so bad that the good old Central Limit Theorem won't get us out of trouble, so let's try a 95% t-interval on the logged data.

```r
n = length(log1920)
x.bar = mean(log1920)
s = sd(log1920)
# Lower bound
lower = x.bar - qt(0.975, df = n - 1) * s/sqrt(n)
# Upper bound
upper = x.bar + qt(0.975, df = n - 1) * s/sqrt(n)
c(lower, upper)
```

```
## [1] 3.995711 4.519215
```

We're 95% confident that the interval 4.0 to 4.5 contains the mean log of the population in thousands.

To better understand this, we back-transform.

```r
exp(c(lower, upper))
```

```
## [1] 54.36451 91.76356
```

The interval is from 54 thousand to 92 thousand. What does this mean?

If our logged data was satisfactorily close to normal, then the exponential of the mean log would be the

median of the original population. So our interval would be a 95% confidence interval for the median big city size in 1920.

Since the transformed data was still somewhat skewed, we have to fudge this a little. We're 95% confident that the "typical" big city size in 1920 was between 54,000 and 92,000. Again, the fudging here is necessary because our methods required strong assumptions we couldn't quite satisfy. We now look at ways to do inference with weaker assumptions.

Additional Exercise (Trosset chapter 10 Problem Set C, questions 1, 3.)

1. (a) The experimental unit is a pair of seedlings of the same age. The measurements taken on each unit are the final height of the cross-fertilized plant and the final height of the self-fertilized plant, both in inches.

   (b) Let X_i be the difference in heights (cross minus self) for each pair $1, \ldots, n$.

   (c)

$$\begin{aligned} H_0 &: \theta \leq 0 \\ H_1 &: \theta > 0 \end{aligned}$$

   (It'll be the other way around if we did the subtraction in part (b) the other way around.)

3.
```
seedlings = read.table("http://mypage.iu.edu/~mtrosset/StatInfeR/Data/seedlings.dat")
diffs = seedlings[,1] - seedlings[,2]
n = length(diffs)
t.stat = mean(diffs) / (sd(diffs) / sqrt(n))
P.value = 1 - pt(t.stat, df = n-1)
lower = mean(diffs) - qt(0.95, df=n-1) * sd(diffs) / sqrt(n)
upper = mean(diffs) + qt(0.95, df=n-1) * sd(diffs) / sqrt(n)
```

   The P-value (significance probability) is 0.025, less than the specified level of α, so we would reject the null hypothesis. A 90% confidence interval for the mean difference in heights is 0.5 to 4.7 inches.

Briefly: The sign test

Trosset 10.2

What happens when you a do a test when your assumptions are wrong? There are two bad things that might happen:

(1) Your probability of a Type I error might be much more than α.
(2) You might have low power compared to a more appropriate test.

While (1) is intellectually worse, (2) is often the bigger issue in practice.

The **sign test** for the population median is based on the idea that with a continuous distribution, every observation has a 50-50 chance of either being above the population median or below the population median. The sign test has extremely weak assumptions: you have an IID random sample. That's it. (The math is much more convenient if the data is from a continuous distribution, but even this is not strictly necessary if you're careful.) As such, it's often a last resort: if all else fails, you can probably do a sign test. We leave the mathematical details to Trosset and jump straight to the implementation.

The implementation of the sign test is in the `BSDA` package. This isn't installed by default, so you'll need to install it before you can use. (You only need to do this once.)

```
install.packages("BSDA")
```

Once you've installed a package, you can load it with `library` just as you would the packages that come with R by default. The relevant function is `SIGN.test()`:

```
library(BSDA)
SIGN.test(pop1920)
```

```
##
##  One-sample Sign-Test
##
## data:  pop1920
## s = 49, p-value = 3.331e-15
## alternative hypothesis: true median is not equal to 0
## 95 percent confidence interval:
##  48.27362 77.86319
## sample estimates:
## median of x
##          64
```

```
##                   Conf.Level  L.E.pt  U.E.pt
## Lower Achieved CI     0.9146 50.0000 77.0000
## Interpolated CI       0.9500 48.2736 77.8632
## Upper Achieved CI     0.9556 48.0000 78.0000
```

That's a lot of stuff. Let's pick out two things to focus on.

- The P-value (3.331×10^{-15}) is for a two-tailed test of the null hypothesis that the median is zero. Well, we probably should have already known that the median big city population wasn't zero, even in 1920, so that isn't very interesting.
- There are three different confidence intervals. To be safe, we take the last and widest one ("Upper Achieved CI.") A 95% confidence interval for the population median big city size in 1920 is 48,000 to 78,000.

The disadvantage of the sign test is that if you really do have a near-normal distribution and the median and the mean are the same, it's somewhat wasteful in the sense that the intervals will generally be wider than the normal intervals. More generally, even though the sign test interval works with small samples, the intervals will be quite wide – as with any other technique, bigger samples are better.

Briefly: The signed-rank test

Trosset 10.3

Another nonparametric test is the **Wilcoxon signed rank test**. The idea behind the test is to *rank* the data, usually by distance from the null hypothesis center. By using rank, we avoid having to know the exact form of the population distribution. This test assumes that the distribution of the data is symmetric when the null hypothesis is true. (For a signed rank confidence interval, you also require the distribution to be symmetric under the alternative hypothesis, but I'll just focus on the test here.)

The main situation where this is a reasonable assumption is when you take two measurements on the same individuals, and to avoid dependence, you take *differences*. The main advantage of this test over the one-sample t for non-normal data is that it isn't as sensitive to outliers.

Example: Ergonomic keyboards. Suppose you measure the typing speed (in words per minute) of ten typists on both ergonomic and standard keyboards.

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

Suppose that our null hypothesis is that the type of keyboard makes no difference to typing speed. Mathematically, let X_i be a random variable representing typist i's ergonomist speed minus their standard speed. Let $\mu = E(X_i)$. Assuming a two-tailed test, the null is then $H_0 : \mu = 0$ and the alternative is $H_1 : \mu \neq 0$. Now, if there really is no difference, whether a typist is faster on the ergonomic or on the standard keyboard is just luck. Furthermore, under the null, a typist who was 10 wpm faster on the ergonomic would've just as likely to have been 10 wpm faster on the standard. This symmetry is the basis of the signed rank test.

Calculate the differences, and summarize numerically and graphically.

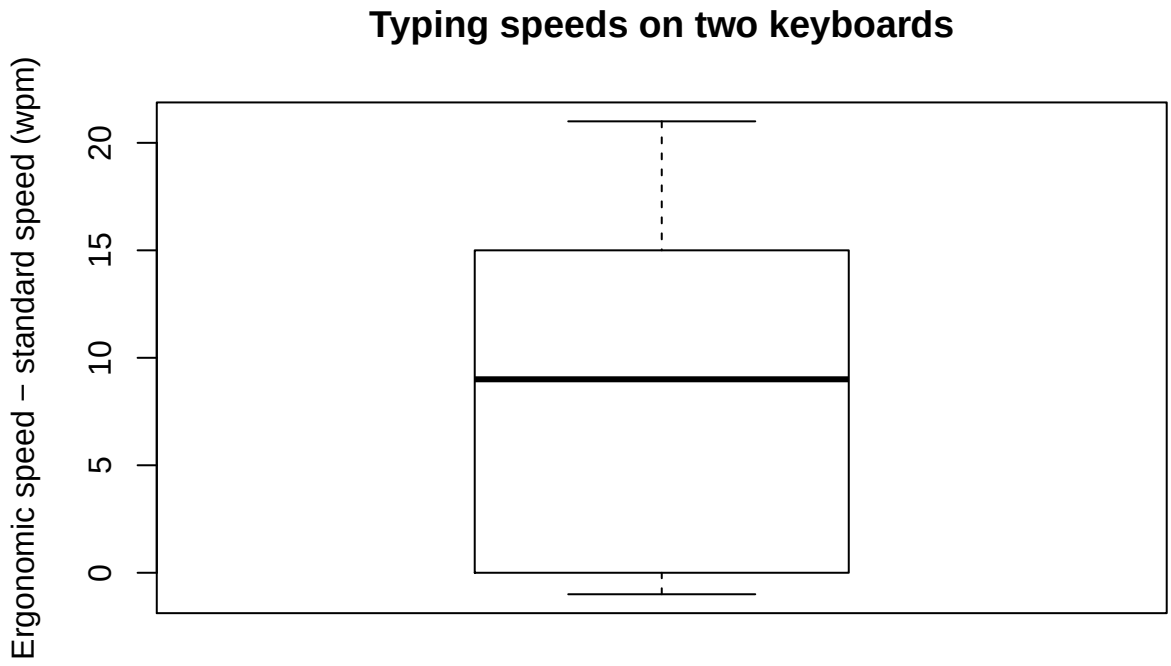
```
differences = ergonomic - standard
differences
```

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

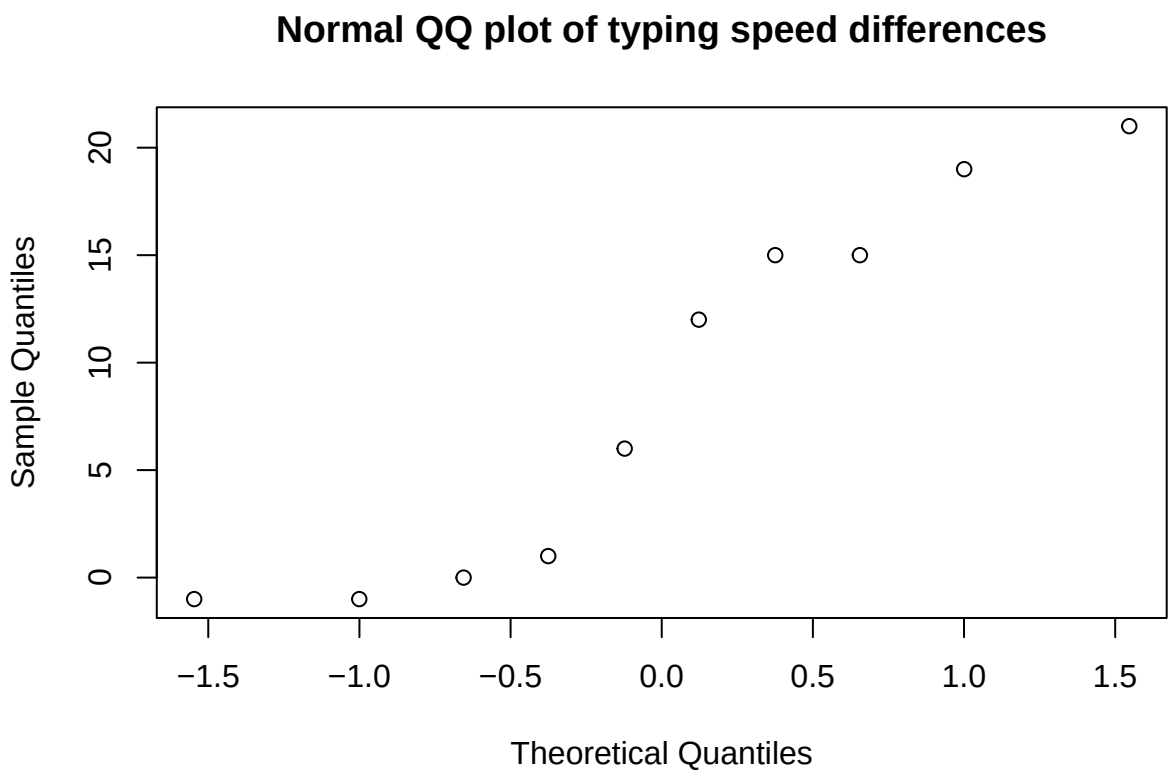
```
summary(differences)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -1.00    0.25    9.00    8.70   15.00   21.00
```

```
boxplot(differences, ylab = "Ergonomic speed - standard speed (wpm)", main = "Typing speeds on two keyb
```




```
qqnorm(differences, main = "Normal QQ plot of typing speed differences")
```



Most of the typists are quite a bit faster on the ergonomic keyboard (positive differences.) The normal QQ plot isn't a straight line, but recall that with small samples this doesn't tell us very much. It's more a matter of philosophy whether you'd do a t-test or a signed rank test on this data: are you willing to make an assumption that you can't be sure is justified, or would you prefer to make weaker assumptions at the possible cost of some power? (We note that as usual, having much more data would make this decision easier.) We'll try it out both ways and see if there's any real difference in the results. First, the t-test:

```
n = length(differences)
x.bar = mean(differences)
s = sd(differences)
t.statistic = (x.bar - 0)/(s/sqrt(n))
# P-value
2 * (1 - pt(abs(t.statistic), df = n - 1))
```

```
## [1] 0.01137473
```

The most accurate version of the signed rank test is in the `coin` package. Again, if you've never used the package before, you'll need to install it:

```
install.packages("coin")
```

Otherwise, you can skip to loading the package and doing the test. The calculation of the P-value is quite complex, and the `wilcoxsign_test()` function will be default only find an approximation. However, since the sample here is tiny, we can calculate it exactly:

```
library(coin)
wilcoxsign_test(ergonomic ~ standard, distribution = "exact")
```

```
##
##  Exact Wilcoxon-Pratt Signed-Rank Test
##
## data:  y by x (pos, neg)
##   stratified by block
## Z = 2.1503, p-value = 0.03516
## alternative hypothesis: true mu is not equal to 0
```

The t-test gives a P-value of 0.011, while the signed rank test's is 0.035. In both cases, the P-value are small but not tiny (though it's hard to get a really tiny P-value from a small sample.) There's some evidence that ergonomic keyboards result in faster typing, at least for these typists.

It's instructive to compare these results to those of the sign test:

```
# library(BSDA) # if you haven't loaded this already
SIGN.test(differences)
```

```
##
##  One-sample Sign-Test
##
## data:  differences
## s = 7, p-value = 0.1797
## alternative hypothesis: true median is not equal to 0
## 95 percent confidence interval:
##  -0.6755556 17.7022222
## sample estimates:
## median of x
##           9
```

```
##                   Conf.Level  L.E.pt  U.E.pt
## Lower Achieved CI     0.8906  0.0000 15.0000
## Interpolated CI       0.9500 -0.6756 17.7022
## Upper Achieved CI     0.9785 -1.0000 19.0000
```

The two-tailed sign test P-value is 0.18, meaning no real evidence against the null hypothesis. Really?

The problem is not that we violated any assumptions: it's that the sign test isn't very powerful. For this reason the sign test is best thought of as a last resort. If other tests are possible, do them.