Assumptions

T tests and other hypothesis tests all make assumptions about the data. For the T test we discussed the assumption of $\sigma_1^2 = \sigma_2^2$ vs. $\sigma_1^2 \neq \sigma_2^2$, and noted that this is important. In general, we shouldn't do hypothesis tests without verifying that the assumptions of these tests are actually okay. Let's take a look at some of these assumptions of the two sample T test, starting with the most important.

The data are random. This assumption is central to all of hypothesis testing. If we pick and choose the data we want to analyze, then nothing we do is really valid. It is trivial to show that women are taller than men. We simply pick 10 women that are taller than 6 feet, and then pick 10 men that are shorter than 6 feet tall. If we analyze these data with a two sample t test, it'll most likely show that women are taller than men.

Obviously, this sample is *biased.* We didn't pick 10 men and 10 women at random. If your data are biased like this, you can pretty much say anything you want with your hypothesis test.

The data between samples are independent. This is really a continuation of the first assumption (random data), but is mentioned separately as we can violate this if we're very careful. But let's figure out what we mean.

For a two sample test, we need to make sure that the data in the first sample are collected independently from the second sample. Just because a data point in the first sample is in the first row, shouldn't influence a data point in the second sample being in the first row.

It's probably best to do a simple example. Instead of biology, let's look at political affiliation for just a moment. Suppose we want to determine if there's a difference between men and women and political party affiliation (e.g., are more men Republican than women?). To do this correctly, we randomly pick, say, 25 men, and 25 women, and ask each of them if they are Republican or Democrat.

But suppose we get lazy. Let's pretend we knock on a door in a neighborhood and the husband answers the door. We ask him if he's a Democrat or Republican. He says "Republican". Now we ask for his wife to come to the door and ask her. What will she most likely say?? Republican, of course. Most people marry people with similar political and religious views to themselves.

In this little example we have violated the assumption of independence between our samples. We will most likely find just as many Democratic men as Democratic women, and will not find a difference between men and women.

Finally, we should mention that this example can not be analyzed with a two sample t test - we're really interested in comparing proportions (e.g., the proportion of Democratic (or Republican) women vs. the proportion of Democratic men).

The data in each sample have a distribution that is approximately normal. This is where we'll spend most the chapter. Tests such as the t test are called *parametric* tests because the depend on assumptions about the distribution of the data. There are other tests (*non-parametric* or *distribution free*) that do not have assumptions like this. We'll learn about one of them in the next chapter.

Two sample t tests assume that the data in *each* sample have a normal distribution. There are numerous ways of checking this, from statistical tests to graphical methods. We will briefly discuss statistical tests, and then move on to graphical methods, in particular QQ plots.

Statistical tests for normality. Hypothesis tests attempt to test the following hypotheses:

$H_0 : D \sim N$

$H_1 : D \nsim N$

So if you fail to reject, you decide in favor of H_0 and decide to assume the data are normal. The problem with hypothesis tests (as mentioned!) is that we can *not* prove the null hypothesis. Even if we fail to reject, we do not know if H_0 is true, so we can't be certain our data are really normally distributed.

For this reason many statisticians do not like tests for normality. However, if you ever do find yourself in a situation where you need to do a test like this (maybe someone is insisting on this), then there are two tests available that can help:

Shapiro-Wilks

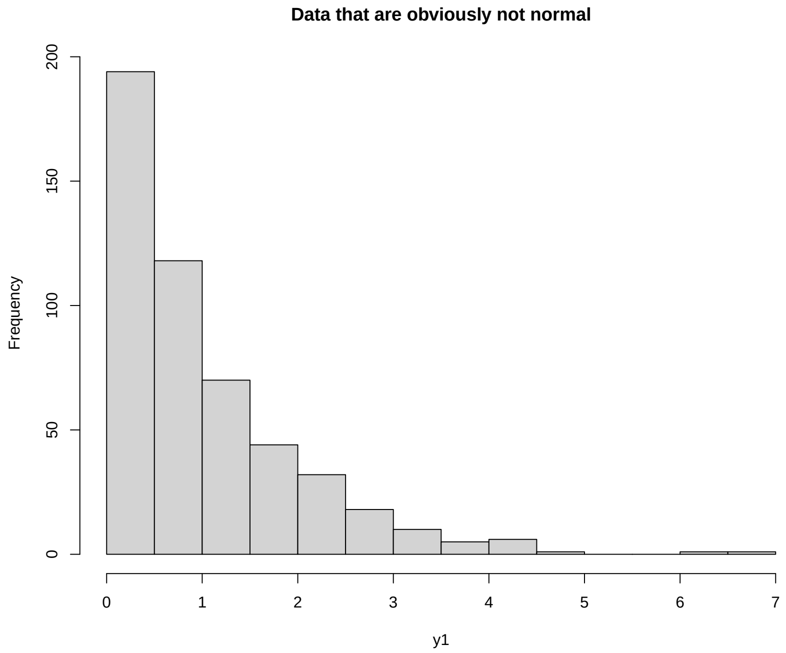
Kolmogorov

Both let you perform the above hypothesis test. For smaller samples, the Shapiro-Wilks test is generally preferred. If you want, you can check these out on Wikipedia, as usual.

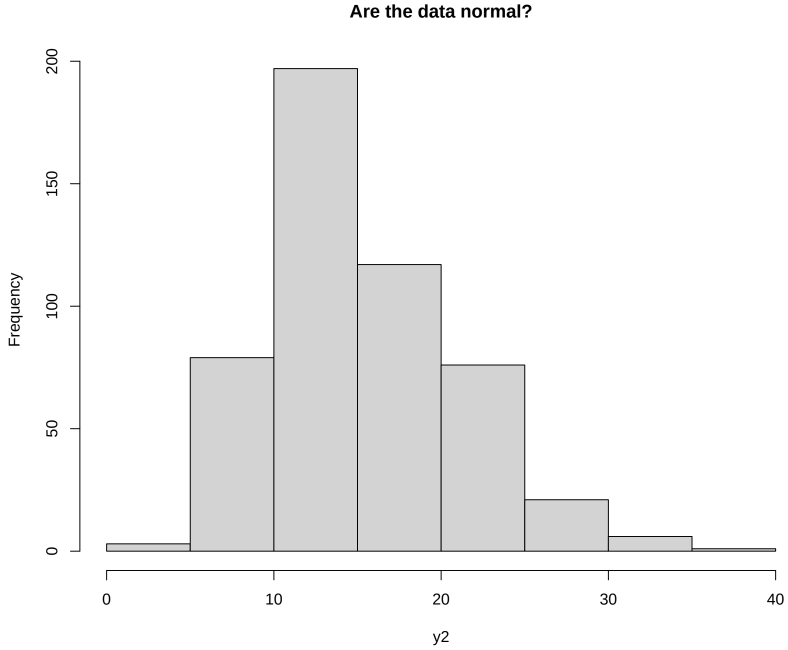
Incidentally, you should never do a *Goodness of Fit* test to evaluate normality. It is way too easy to do anything you want with this test, although it is sometimes recommended in other textboooks.

Graphical methods to check normality (introducing QQ plots). There are a variety of graphical methods you can use, but many of them suffer from a similar drawback as the statistical tests mentioned above. It is easy to see if your data are not normal, but more difficult to determine if your data are normal.

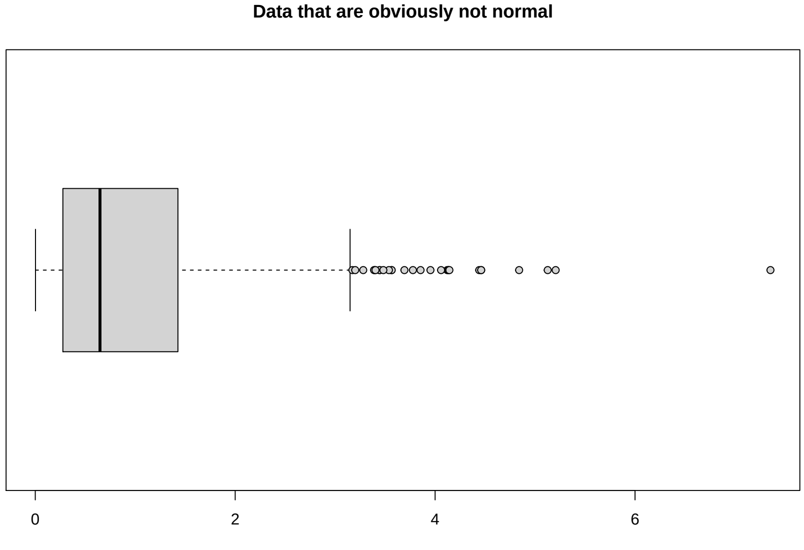
For example, here is a histogram where the data are obviously not normal:



On the other hand, here the data might be normal; it's hard to know. Drawing a normal curve over the histogram can help, but is still not ideal.



The same thing happens with boxplots. They can quickly tell you if your data are not normal, but they're not so good at telling you if your data are normal.



A much better graph is something called a *QQ* plot. *QQ* stands for ***Q***uantile-***Q***uantile (think of these as percentiles). Quantile-quantile plots are quite versatile and can be used not just to evaluate whether data have a normal distribution, but also numerous other distributions, although we're just interested in normally distributed data.

Many people refer to the plot we are learning as a *normal probability plot*, which is essentially correct, but we will call the QQ plots since that's what R calls them.

So how do we do a QQ plot? A QQ plot essentially plots the actual data on the Y-axis, and the *expected* value of each data point on the X-axis. If the actual values and expected values form a line that is more or less straight, we can conclude that our data are approximately normal.

How do we figure out what values we expect? Suppose we have the following 10 random values for IQ:

101 102 106 120 110 107 119 94 100 95

(Before we go on we should say that there a many, many problems with the IQ scale. It often does not do what it was designed to do).

Let's sort these IQ values:

94 95 100 101 102 106 107 110 119 120

Now we ask: in a sample of size 10, where do we *expect* the smallest value for IQ to be if the data have a normal distribution?

Answer: at the 10th percentile. We have 10 data points, so the smallest value should be at the 10th percentile of our normal curve. We know how to do this! Just use our normal tables for reverse lookup and find the z-score for the area closest to 0.10 (or just use R).

Then we can proceed with the 20th percentile, the 30th, and so on, all the way up to the 100th percentile. Oops - what is the z-score for the 100th percentile? How far up does the normal curve go? It goes to $+\infty$.

Before we figure out how to fix this, let's take a look at a formula for what we've been doing. It's pretty simple. For each value (where $q_i = \text{i}^{\text{th}}$ quantile) we just used:

$$q_i = \frac{i}{n}$$

Where i is the i^{th} observation (i.e, $i = 1, 2, 3, ..., n$). But because this eventually gives us a 100th percentile, we will modify this formula and use:

$$q_i = \frac{i - {}^{1}\!/_{2}}{n}$$

This will give us almost the same graph and ensure that we can get all the way to 100%.

Comment: R actually uses $q_i = \frac{i - {}^{3}\!/_{8}}{n}$ when $n \leq 10$, but for simplicity we'll stick with the formula given above. But if the QQ plot in R looks just a little different than what you do by hand, that's probably why.

So here's what we do:

1. Sort your data from smallest to largest (this isn't strictly necessary, but it'll avoid confusion).
2. Number your data points from 1 to n. Again, this isn't really necessary, particularly with small sample sizes, but it does help keep things straight.
3. For each data point, calculate:
$$q_i = \frac{i - {}^1\!/\!_2}{n}$$
Make sure you use i, not the actual data value in this equation.
4. Use q_i in your normal tables and do a reverse look up to get a z-score. You want z_{q_i}.
5. Plot your z-score on the x-axis, against the actual data value on the y-axis.

(We'll worry about interpretation a little later).

Let's continue with our IQ example before we get lost. We already sorted the data, so let's add i:

IQ:	94	95	100	101	102	106	107	110	119	120
i:	1	2	3	4	5	6	7	8	9	10

Now let's start to calculate our quantiles. For $i = 1$ we have:

$$q_1 = \frac{1 - {}^1\!/\!_2}{10} = 0.05$$

For $i = 2$ we have:

$$q_2 = \frac{2 - {}^1\!/\!_2}{10} = 0.15$$

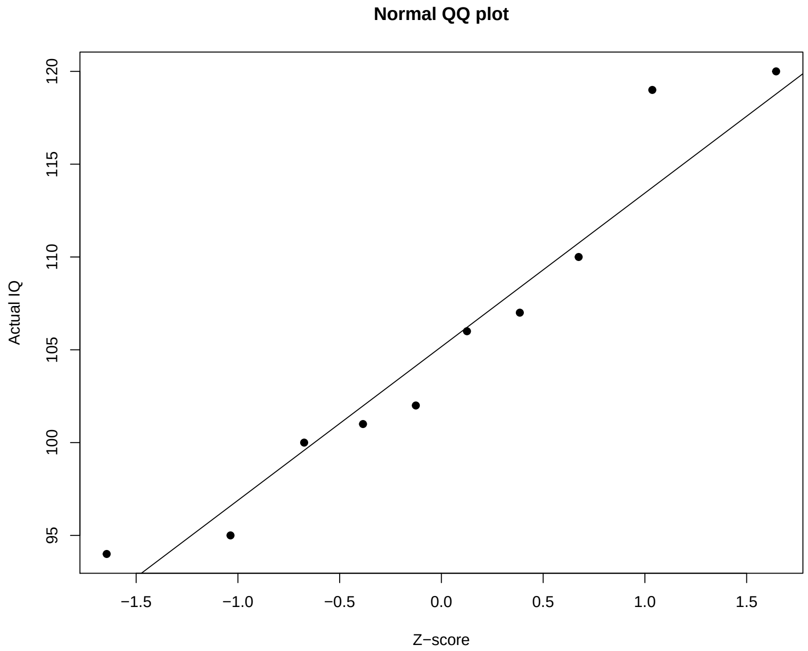
And so on for i = 3, 4, ..., 10. Let's add these values to our table:

IQ:	94	95	100	101	102	106	107	110	119	120
i:	1	2	3	4	5	6	7	8	9	10
q_i:	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95

And finally, we do a reverse look up for each of our values of q_i. For $i = 1$, we have $Z_{q_i} = Z_{0.05}$ = -1.64; for $i = 2$, we have Z_{q_i} = -1.04, and so on. If we add all these values to our table we now have:

IQ:	94	95	100	101	102	106	107	110	119	120
i:	1	2	3	4	5	6	7	8	9	10
q_i:	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
Z_{q_i}:	-1.64	-1.04	-0.67	-0.39	-0.13	0.13	0.39	0.67	1.04	1.64

Finally we are done with our calculations, and we can plot our QQ plot. We use the values in the first row on your y-axis (IQ), and the values in the last row on the x-axis (Z_{q_i}):



Why don't we use convert our Z_{q_i} values into *expected* IQ values? Because it doesn't make any difference to the plot. It will look the same whether we plot Z_{q_i} values or the expected IQ values. Let's do the plot again using expected IQ values, but first let's calculate all the expected IQ scores:

The lowest expected IQ score would be:

$$IQ_1 = -1.64 \times 8.95 + 105.4 = 90.68$$

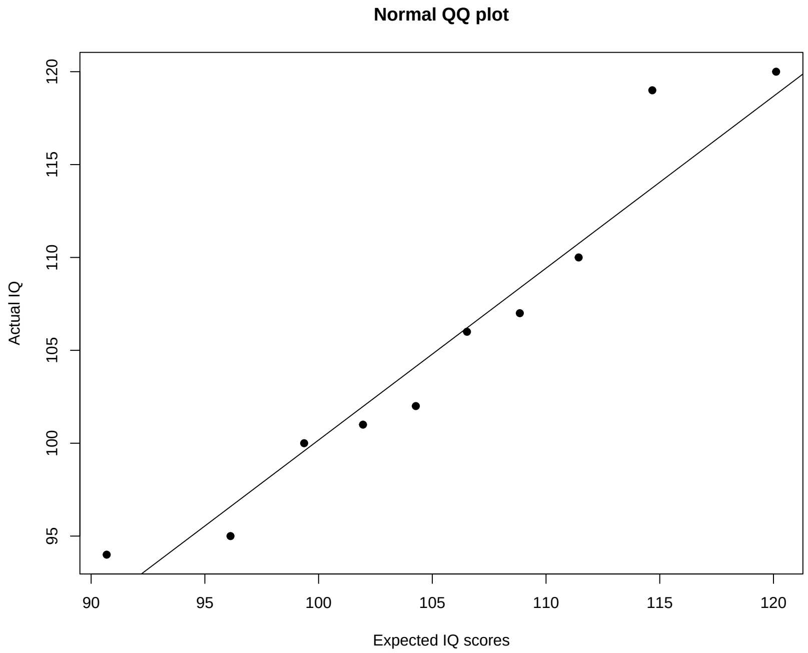
The second lowest expected IQ score would be:

$$IQ_2 = -1.04 \times 8.95 + 105.4 = 96.13$$

And so on. Notice that for this it's fine to use the sample mean ($\bar{y}$ and sample standard deviation (s) instead of μ and σ. All 10 expected IQ scores (from lowest to highest) would be:

Expected IQ:	90.7	96.1	99.4	102.0	104.3	106.5	108.8	111.4	114.7	120.1

And now we can plot the actual IQ values on the y axis, and the expected IQ values on the x axis:



Notice that the plot appears *identical* to the one above using Z_{q_i}. So we can save a step by just plotting the actual data values versus the z-scores.

Interpreting *QQ* plots. So we now know how to make a QQ plot. How do we interpret it? If the data are *perfectly* normal (no real data will ever be *perfectly* normal) then the points on the graph will form a straight line. Normally you will see some small squiggles or deviations from a straight line. Don't worry about these. Both of the graphs above (which are really identical) are good examples of data that are reasonably normal.

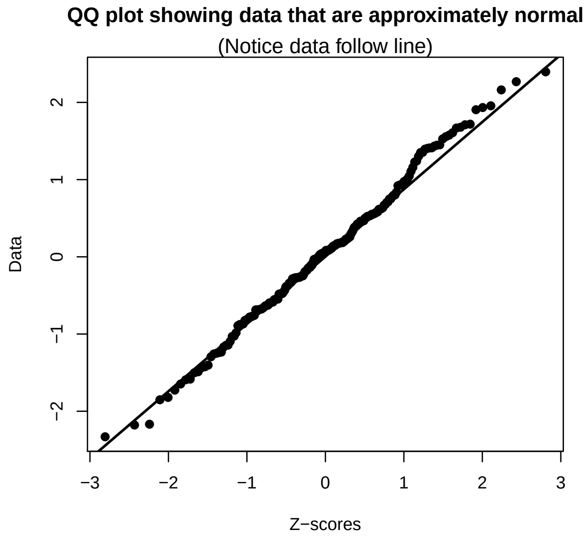
So what does a bad (non-normal) QQ plot look like? Generally, if you see points far away from the line (particularly at the ends) or if the points make a curve, that means you need to look closer at your plot. Let's list some of the problems you can encounter:

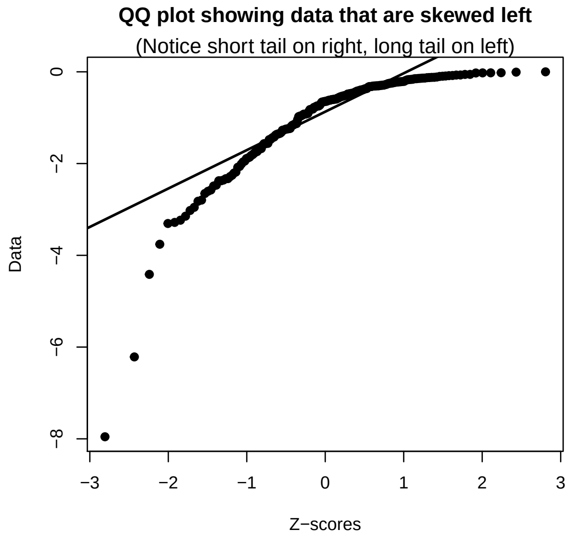
1. If you see a backwards S-shaped curve (the ends of the curve point up or down) this shows that your data have long tails. This is bad and can be difficult to work with. In addition, if some of your points are far away from the rest, this can indicate outliers, which are often a symptom of long tails.

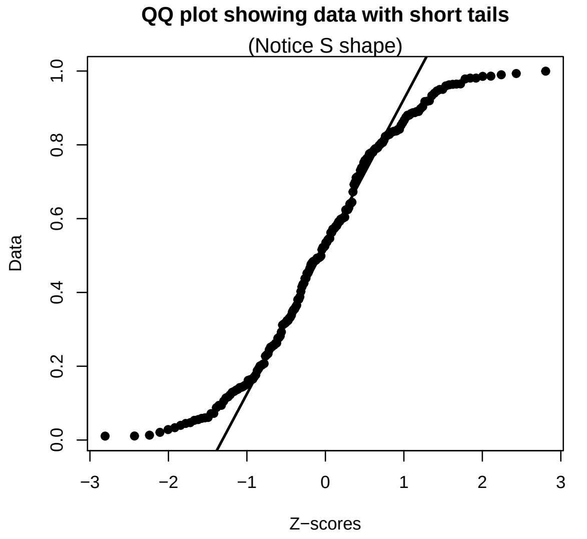
2. If you see a regular S-shaped curve (the ends of the curve point sideways) this shows data that have short tails. This isn't as bad as long tails.

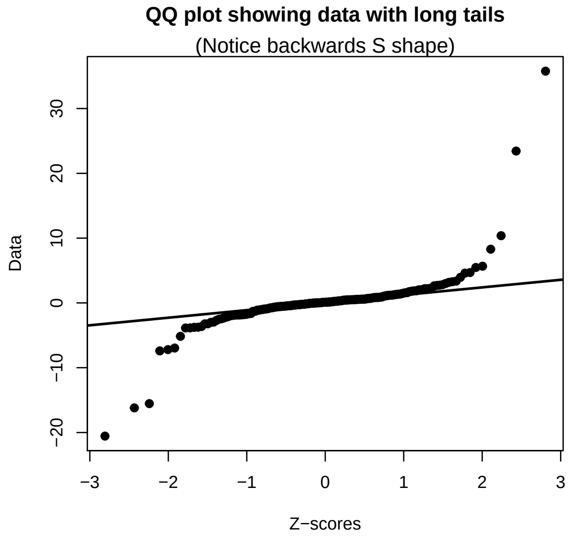
3. if you see just one curve (i.e., the points curve up or down, but don't make an S), this shows skewed data. Depending on the severity of the skewness, this can also be bad.

Examples of all three of these problems in addition to another QQ plot for normally distributed data are in the graphs:


QQ plot showing data that are approximately normal
(Notice data follow line)


QQ plot showing data that are skewed left
(Notice short tail on right, long tail on left)


QQ plot showing data with short tails
(Notice S shape)


QQ plot showing data with long tails
(Notice backwards S shape)

A word of caution. Some software (e.g., Minitab) will reverse the axes on a QQ plot. In other words, the actual data are now on the x-axis and the z-scores on the y-axis. This means that your interpretations change. For example, a regular S-shaped curve now has long tails. In this class we will never reverse the axes, but you need to know about this in case you use different software or look at someone else's QQ plots. Always check the axes before you figure out what the problems might be.

Incidentally, the normal distribution assumption applies to *each* sample. To check if your data are normal, you need to make a *QQ* plot for *each* sample. If you're comparing blood oxygen levels in men vs. women after exercise, you need to make *two QQ* plots. One for men, and another for women. If either one (or both) is not normal you need to worry about the normal distribution assumption. In more advanced classes you can learn about a short cut that does let you make one plot, but this requires a lot more explanation.

So how does all this affect the normal distribution assumption? Or, what does this mean for our *t* test?
This depends on how badly not normal our data are. Remember, the CLT will eventually take care of data that are not normal, so you can still use a *t* test. However, this might require a large sample size if the data are seriously not normal. Let's try to summarize some of this:

> If the data have long tails, this is particularly bad. The CLT may take a while before the means start to behave normally. In this case you may need a large sample size (e.g., 40 or 50, depending on how bad the problem is).
>
> If the data have short tails, that's not so bad. The CLT will start to work much quicker. A sample size of 20 or 25 might be enough for you to be able to use a *t* test.
>
> If the data are skewed, this could be bad. It depends on how badly skewed they are. If you have a really long tail on one side, you're back to needing large samples for the CLT to work (e.g., 40 or 50). If the data are only slightly skewed, you could get away with smaller sample sizes (20 or 25).

Remember that the above summary applies to *each* sample for a two sample *t* test.