Estimation

Ott & Longnecker Sections: 4.11-4.12, 4.14, 5.2-5.3, 5.8, 10.2

Key Concepts: Independence and dependence of RVs, simple random sample, independent and identically distributed (iid) RVS, estimator, statistic, estimate, bias, unbiased estimator, standard error, mean squared error, QQ plot, Central Limit Theorem (CLT), point estimate, confidence interval, t distribution, degrees of freedom, inference for population proportions.

There are three basic forms of inferential statistics: estimation, testing, and fitting. Continuing very naturally from concepts of RVs, we now discuss some basic concepts of the first of these, estimation. To begin talking about estimation though, we must extend the properties of random variables that we have learned to functions of random variables.

1 Distributions of Functions of RVs

The concepts in this section are covered in sections 4.11 and 4.12 of Ott and Longnecker.

1.1 Independence and dependence of RVs

We start with a definition:

- Two RVs are said to be **independent** if the realization of one of them does not change the probability distribution of the other, and vice versa. If two RVs are not independent, then they are **dependent**.

Examples.

1. Recall the ant farm with 20 ants, of which 5 are poisonous. You select two ants at random. Let X_1 be a 1 if the first ant is poisonous, and 0 otherwise. Let X_2 be a 1 if the second ant is poisonous, and 0 otherwise. What distribution does X_1 have? *Answer:* $X_1 \sim Ber(1/4)$. What distribution does X_2 have? *Answer: more difficult, but marginally* $X_2 \sim Ber(1/4)$ *regardless of whether* X_1 *is known and if we are sampling with or without replacement.*

   - If the two ants are selected **with replacement**, then X_1 and X_2 are independent since knowledge of whether $X_1 = 1$ (poisonous) or $X_1 = 0$ (non-poisonous) won't change the distribution of X_2 – it's an identical draw from the same population, so X_2 is still Bernoulli(1/4).
   - If the two ants are selected **without replacement**, then X_1 and X_2 are dependent. If we know $X_1 = 1$ (poisonous), then now $X_2 \sim Ber(4/19)$. If $X_1 = 0$ (not poisonous), then now $X_2 \sim Ber(5/19)$. Knowing the outcome of the first ant changed the probability distribution of the second!

2. In a Statistics class, let M be the midterm score of a randomly selected student and let F be the final score. Suppose the median of both exams is 75. Consider two ways of drawing M and F:

   - You draw M from the full roster of midterm scores. Then you draw F from the full roster of final scores. Then M and F are independent – regardless of whichever midterm score you drew, you could draw any final score with equal probability.
   - You draw a random *student* from the roster, and record M and F for that same student. Now the distribution of M and F are dependent. For instance, if you know that $M > 75$ then it is likely there is now a greater than 50% chance that $F > 75$ as well, since F would be a random draw from that reduced list of students which scored above the median on the midterm.

1.2 More properties of expectation and variance

We now continue with some properties of expectation and variance. Let X and Y be any RVs, and let c be a constant:

1. $E(c) = c$.

2. $E(c * X) = c * E(X)$.

3. $E(X + c) = E(X) + c$.

4. $E(X + Y) = E(X) + E(Y)$.

5. $VAR(c) = 0$.

6. $VAR(c * X) = c^2 VAR(X)$.

7. $VAR(X + c) = VAR(X)$.

8. If X and Y are independent, $VAR(X + Y) = VAR(X) + VAR(Y)$.

Example. We can use these properties to get more intuition about some results from the previous section. We stated that:

$$\text{If } X \sim N(\mu, \sigma^2), \text{ then } Z = \tfrac{X-\mu}{\sigma} \sim N(0, 1).$$

Recall that $E(X) = \mu$ and $VAR(X) = \sigma^2$. It's not trivial to show that Z is normal, but it is relatively easy to show that $E(Z) = 0$ and $VAR(Z) = 1$.

$$E(\tfrac{X-\mu}{\sigma}) = \tfrac{E(X)-\mu}{\sigma} = \tfrac{0}{\sigma} = 0, \text{ by properties (1), (2), and (3).}$$
$$VAR(\tfrac{X-\mu}{\sigma}) = \tfrac{VAR(X)-0}{\sigma^2} = \tfrac{\sigma^2}{\sigma^2} = 1, \text{ by properties (5), (6), and (7).}$$

Example. When discussing the binomial, we stated that:

$$\text{If } B \sim Bin(n, \pi), \text{ then } E(B) = n\pi, \text{ and } VAR(B) = n\pi(1-\pi).$$

Recall that a binomial is a sum of n iid Bernoulli RVs, call them X_1, X_2, ..., X_n. Then $B = \sum_{i=1}^{n} X_i$. Since each of these Bernoulli RVs has expectation π and variance $\pi(1-\pi)$, we can use our rules to show that:

$$E(B) = E(\sum_{i=1}^{n} X_i) = n\pi, \text{ by repeated use of property (4).}$$
$$VAR(B) = VAR(\sum_{i=1}^{n} X_i) = n\pi(1-\pi), \text{ by repeated use of property (8).}$$

These match what we stated previously.

1.3 Different kinds of samples

We now initiate discussion of what happens when we collect more than one random draw from a population into a random *sample.* We mention two types of random samples in this class:

- A random sample of size n from a population is called a **simple random sample** if every possible sample of size n is equally likely to be drawn. Unless otherwise specified, all samples in this class are simple random samples.

  - **Note:** A simple random sample of size n can be collected by randomly drawing n samples from a population **without replacement.**
  - **Example:** Dealing a hand of cards from a well-shuffled deck constitutes a simple random sample. Suppose you are playing five card stud, a kind of poker game where each player is dealt 5 cards. If the deck is shuffled well, as a player in the game, you are equally likely to get any of the five cards. The best poker players know this and can quickly calculate probabilities in their heads (alongside being able to read their opponents!)

- A random sample of n RVs X_1, X_2, ..., X_n are said to be **independent and identically distributed**, or **iid**, if: (1) the RVs are all independent of one another, that is, the realization of any one of them does not change the probability distribution of any other one; (2) they all have exactly the same probability distribution.

  - **Note:** An iid sample of size n can be generated by randomly drawing n samples from a population **with replacement.**
  - **Example:** the results of repeated flips of a coin, or rolls of a die, are i.i.d. The outcome of a single flip (roll) doesn't affect the probabilities of the outcomes of any other, and it's the same coin (die) so the distribution in each trial is the same.

In this course random samples will be iid unless stated otherwise. But, simple random samples aren't iid, being sampled without replacement! Moreover, we have often encountered finite populations from which sampling with replacement doesn't make much sense (why would you put a sample back in the population only to sample it again? You run the risk of collecting redundant information! There are some more sophisticated mathematical arguments to back this up, but we won't pursue them here.) How can we reconcile this? It turns out, if the population size is large enough relative to the sample size, a sample without replacement closely approximates a sample with replacement, because the populations change very little as elements are removed.

Example. Let's return once more to the ant farm. Suppose you choose two ants at random. As we have seen, when sampling with replacement, the RVs are independent; but if the ants are sampled without replacement, there will be some dependence. The degree of dependence, however, depends on the size of the ant farm. Consider two scenarios:

- Start with the previously worked example, where there are 20 ants in the farm, and 5 are poisonous. The probability that the first ant selected is poisonous is $5/20 = 0.2500$. If a poisonous ant is selected first, the probability that the second ant is poisonous is $4/19 = 0.2105$. This is a fairly large change.
- Suppose now that there are 1000 ants in the farm, and 250 of them are poisonous. The probability that the first ant selected is poisonous is $250/1000 = 0.2500$, just as above. If a poisonous ant is selected first, the probability that the second ant is poisonous is $249/999 = 0.2492$. This probability, while certainly not exactly 0.2500, is so close that for practical purposes it could be considered the same. So these two RVs could be considered nearly independent.

Note: Be aware that for any population of finite size, as the sample size increases, the probability distributions change more and more. So 2 draws from a population of size 1000 could be considered nearly independent, but 500 draws out of the same population of 1000 probably could not be. Yet a sample of 500 out of a population of size one million are probably close to independent. Practically speaking, so long as the sample size is small compared to the population size, without replacement sampling can be considered approximately the same as with replacement sampling.

2 Estimation

The concepts in this section are covered in sections 4.12, 4.14, and 5.2 of Ott and Longnecker

We still haven't gotten to any estimation, but we finally do so now. We start with a very common situation, that of estimating the mean of a population.

Example. A car manufacturer uses an automatic device to apply paint to engine blocks. Since engine blocks get very hot, the paint must be heat-resistant, and it is important that the amount applied is of a certain minimum thickness. A warehouse contains thousands of blocks painted by the automatic device. The manufacturer wants to know the average amount of paint applied by the device, so 16 blocks are selected at random, and the paint thickness is measured in mm. The results are below:

1.29, 1.12, 0.88, 1.65, 1.48, 1.59, 1.04, 0.83, 1.76, 1.31, 0.88, 1.71, 1.83, 1.09, 1.62, 1.49

How might we go about using this sample to make inferences about the population mean paint thickness of the entire population of blocks, which we call μ? As it happens, our intuition serves us well here, and, though it will take some work to thoroughly show this, the sample mean of these observations, which we define below, will be a good estimate of the population mean, μ:

$$\text{Sample mean: } \hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Note that it is customary in statistics to put a hat on a parameter symbol to indicate an estimator of that parameter. To apply our theory to this situation, start by thinking of each of the numbers listed above as the realization of an RV. In particular, they are the realizations of 16 iid RVs, that we might call X_1, X_2, ..., X_{16}. For now, we don't care what the exact distribution of these RVs is, but let's call the expectation $E(X_i) = \mu$, and variance $VAR(X_i) = \sigma^2$. Note that since these RVs are iid, the expectations and variances are the same for every one. We make one quick definition before moving on:

- The formula that describes how data from a sample would be used to compute a guess about a population parameter is called an **estimatOR**, or a **statistic**. The sample mean formula given above is an example of an estimator. The numerical value computed using the estimator once the data is collected is called an **estimATE**. The sample mean of the 16 data points given above, which we compute shortly, is an example of an estimate. An estimator is a RV, and an estimate is a realization of that RV.

It is **very important** to understand that the sample mean, being an estimator, is itself an RV - it is an RV constructed as a function of other RVs. In order to get one realization of the RV that represents the sample mean, we would randomly sample n elements from the population in an iid fashion, and calculate their mean.

There are many ways that we could define an estimator of any given population parameter. But some estimators are definitely better than others. Since estimators are RVs, they have probability distributions (this is why we studied the distributions of functions of RVs, since that's exactly what estimators are!). We can evaluate the performance of estimators by analyzing the properties of their probability distributions. In particular, their expectations and variances turn out to be helpful in determining the relative merit of estimators. The expectation of an estimator determines where the estimates tend to be located on average, whereas the variance determines how spread out the estimates are from sample to sample. To formalize this, let θ be any population parameter, and let $\hat{\theta}$ be a possible estimator of that parameter:

- The **bias** in an estimator $\hat{\theta}$ is defined as:

  $$bias(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

  If the bias is equal to zero, the estimator $\hat{\theta}$ is called **unbiased** for θ. All other things being equal, smaller bias is better.

- The variance of an estimator $\hat{\theta}$ is defined as $VAR(\hat{\theta})$. All other things being equal, smaller variance is better. The square root of the variance is usually called the standard deviation or SD. However, when we are talking about estimating a parameter, we instead use the term **standard error** or **SE**, to remind us that this is the amount of error in estimation. Thus the square root of the variance of an estimator will be denoted $SE(\hat{\theta})$.

- The **mean squared error**, or **MSE**, of an estimator $\hat{\theta}$ can be calculated as:

  $$MSE(\hat{\theta}) = VAR(\hat{\theta}) + \Big(bias(\hat{\theta})\Big)^2.$$

  All other things being equal, smaller MSE is better. Note how MSE incorporates information about both bias and variance.

Example. (Conceptual) Suppose you are playing darts and aiming at the bullseye on the dart board. Think of the position of the bullseye as the population parameter we are interested in. With each throw of the dart, its ultimate position could be thought of as a random draw from the dart throw "population" (or perhaps more aptly, dart throw process).

If you are pulling the darts to the right (this happens to me!) this could be thought of as a bias: the average position of your darts differs from the bullseye. Of course, each dart throw is subject to small uncontrolled variations that could be considered random, causing slight variations around the average dart position. Then the average squared distance of a dart to the bullseye is the squared difference from the expected position plus the average squared distance around the expected position.

Example. Let's return to the sample mean. It turns out it is fairly easy to work with. We can use our rules of expectation and variance to derive the expectation and variance of the sample mean:

$$E(\bar{X}) = E(\tfrac{X_1+X_2+...+X_n}{n}) = \tfrac{\mu+\mu+...+\mu}{n} = \mu.$$
$$VAR(\bar{X}) = VAR(\tfrac{X_1+X_2+...+X_n}{n}) = \tfrac{\sigma^2+\sigma^2+...+\sigma^2}{n^2} = \tfrac{\sigma^2}{n}.$$
$$SE(\bar{X}) = \sqrt{VAR(\bar{X})} = \tfrac{\sigma}{\sqrt{n}}.$$

The expectation uses properties (2) and (4). The variance uses properties (6) and (8), with the required independence in (8) being valid since the X_i were iid. We can see immediately that $\bar{X}$ is unbiased for μ, since $E(\bar{X}) = \mu$. And though it is somewhat hard to show, it can be proven that among all unbiased estimators of μ, $\bar{X}$ has the smallest possible variance provided the distribution of the individual observations is continuous. So it is a very good estimator of μ.

Example. To give a somewhat silly example of an alternative estimator of μ that does not perform as well as $\bar{X}$, consider taking a sample of size $n = 2$. Then, $\bar{X} = \frac{X_1+X_2}{2}$, so $E(\bar{X}) = \mu$ and $VAR(\bar{X}) = \frac{\sigma^2}{2}$. Consider now an estimator that we might call simply $\hat{\mu} = \frac{2X_1+X_2}{3}$. Using our rules, we find $E(\hat{\mu}) = \mu$, so $\hat{\mu}$ is unbiased for μ. But our rules also tell us that $VAR(\hat{\mu}) = \frac{5\sigma^2}{9}$. Since $5/9 > 1/2$, the estimator $\hat{\mu}$ has a larger variance than $\bar{X}$. So, while both estimators will hit the correct answer on average, $\bar{X}$ has a smaller variance and thus will also usually be closer to correct than $\hat{\mu}$. This should make sense: if we have two identical random draws from a population, what sense is there to weight one observation twice as much as the other? (There are situations where we do want to weight some data points more than others, because we suspect they have higher quality and are more reliable. But such situations certainly aren't iid!)

Computing the sample mean of the paint data, we find $\bar{x} = 1.348$ mm. (Note that $\bar{x}$ with a lower case x is an estimate.) This is our best guess, but it tells us nothing about how sure we are of the value we got. We need some measure of accuracy. Either the variance, or, more commonly, the standard error of the estimator helps us with this. The formula above says that the standard error of the sample mean is the standard deviation of any single observation divided by the square root of number of observations, or $\frac{\sigma}{\sqrt{n}}$. As n gets larger, $\frac{\sigma}{\sqrt{n}}$ gets smaller, and this is intuitive, since as we take a larger sample, we should do a better job of estimating. Unfortunately, in most cases we don't know the value of σ^2, and therefore, while σ is not the population parameter we are interested in, we still need to estimate it. We can estimate it from a sample using:

$$\text{Sample variance of } X\text{: } \hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Note that this, like the sample mean, is an estimator, and thus an RV. In this case though, S^2 is an estimator of σ^2. Here we can pause to explain something that previously I asked you to simply believe: the reason we use $n-1$ in the denominator is that this makes the estimator S^2 unbiased for σ^2. We can also estimate σ using:

$$\text{Sample standard deviation of } X\text{: } \hat{\sigma} = S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Strangely enough, even though this is just the square root of the sample variance, it is a slightly biased estimator of σ. However, the bias is quite small, and smaller than it would be if n was used in the denominator. If we plug in the estimate of σ into the formula for the standard error, we find:

$$\text{Estimated standard error of } \bar{X}\text{: } \widehat{SE(\bar{X})} = \tfrac{\hat{\sigma}}{\sqrt{n}} = \tfrac{S}{\sqrt{n}}.$$

Again note the difference between the terms standard deviation and standard error. The standard deviation is a property of the distribution of the X_i, whereas the standard error is a property of the estimator $\bar{X}$. Also, note that the estimated standard error is an estimator of the standard error. Trust me, I know how confusing this is!!

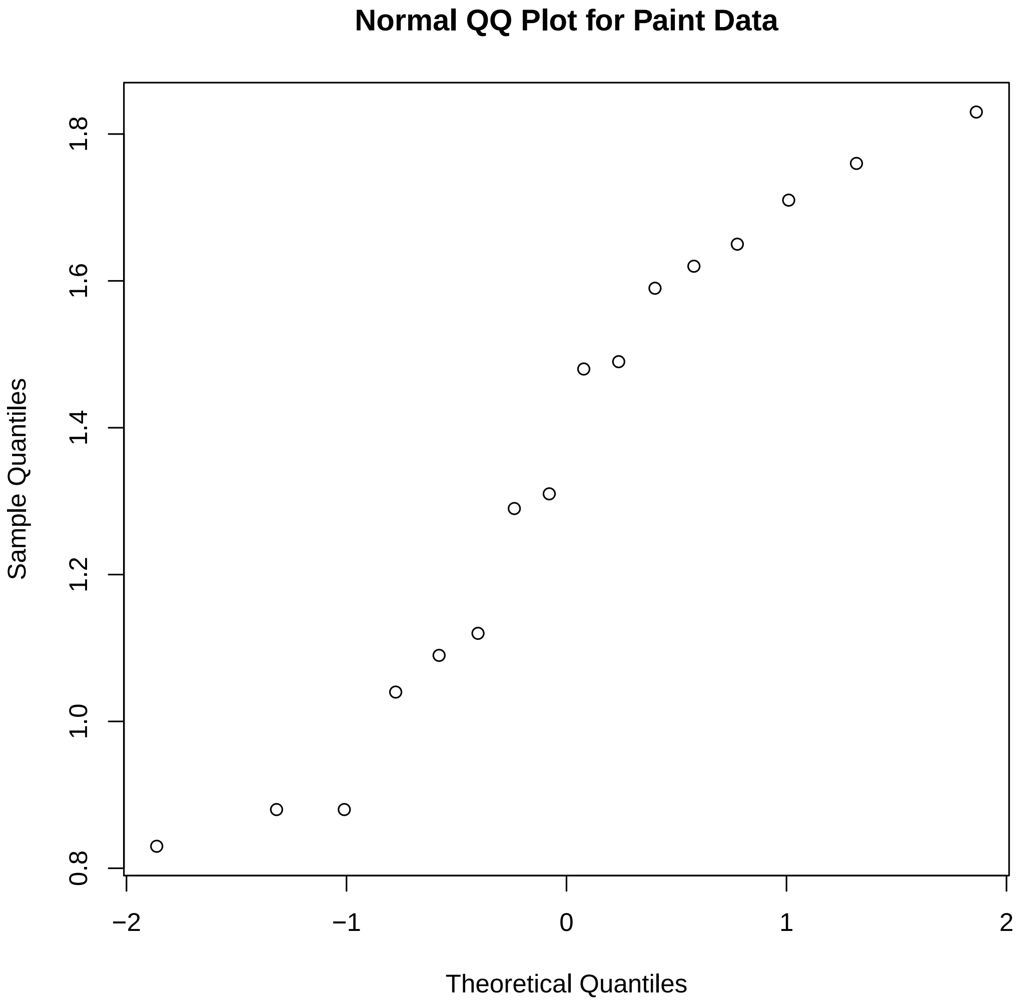
For the paint data, we find $s = 0.3385$ mm, so $\widehat{SE(\bar{X})} = \frac{0.3385}{\sqrt{16}} = 0.085$ mm. In a somewhat non-technical way, we might say that our estimate of the population average paint thickness is 1.348 mm, give or take 0.085 mm. To be a little more specific, the estimated SE tells us about how far away the estimate will usually be from the actual value of the parameter. When we say ‘usually’, we’re thinking about theoretically taking many samples of the same size from the population and making an estimate based on each sample.

Sometimes this give or take number is not sufficient, and we desire to make a stronger statement. To do this, though, we need to make an additional assumption about the underlying distribution of the X_i. In many common situations, it is a fairly reasonable assumption that the X_i are distributed as iid normal. If we still let $E(X_i) = \mu$ and $VAR(X_i) = \sigma^2$, but additionally assume normality, so that $X_i \sim N(\mu, \sigma^2)$, then it turns out that the sample mean is also normally distributed: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. This statement is a result of the following definition and fact:

- Let $X_1, X_2, ..., X_n$ be any collection of RVs. If we let $a_1, a_2, ..., a_n$ be any collection of constants, then $\sum_{i=1}^{n} a_i X_i$ is called a **linear combination** of the X_i.
- Any linear combination of independent normal RVs is itself distributed as a normal RV.

Since the sample mean is a linear combination of the X_i (set all of the $a_i = 1/n$), it follows that the sample mean must be normally distributed. We already computed the expectation and variance so the result follows. We can use this additional information to create what is called a **confidence interval**, or **CI**, which conveys even more valuable information about the value of μ than just $\bar{X}$ and $\widehat{SE(\bar{X})}$.

But before we move on to CIs, we must address the issue of normality. This is an *assumption* we are making about the population, and if that assumption is wildly incorrect, our inferences can be inaccurate. We need some method to evaluate the validity of this assumption. One easy way is with a **normal quantile-quantile plot** or **normal QQ plot**. The details are somewhat complicated, but essentially the plot is constructed in such a way that if the data is normal, the points in the plot will fall roughly on a straight line, up to sampling error. Here is the plot for the paint data:

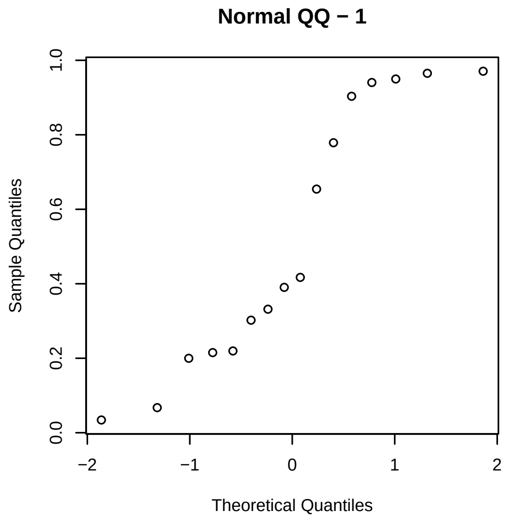


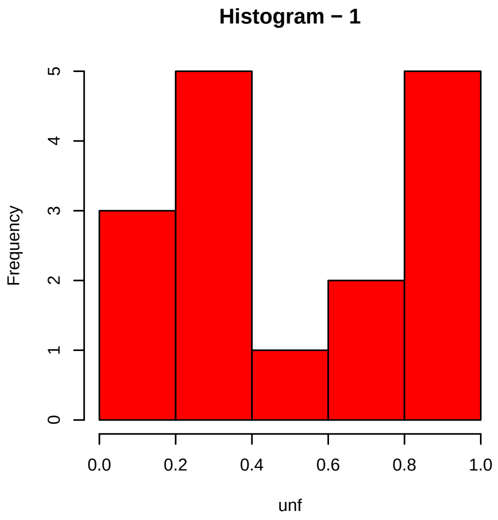
The plot is not perfectly straight, but it is pretty good. You can get a sense for the natural variation that occurs in these plots, even with truly normal data, by using R to simulate draws from a normal:

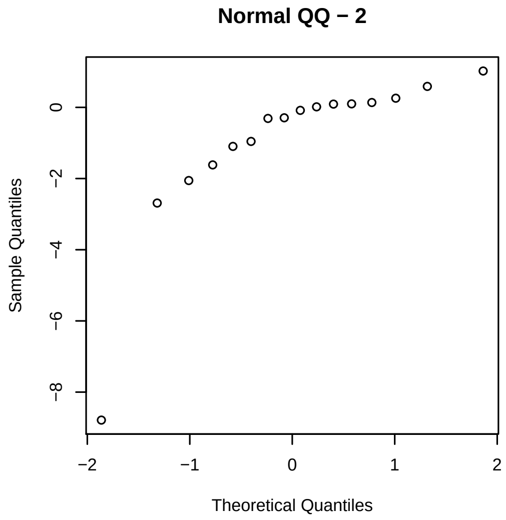
Go to QQ plot R code.

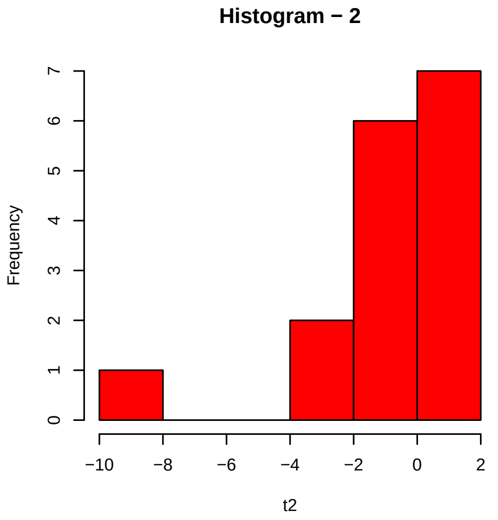
Some of these might look a little questionable, but they are all from true normal data. You shouldn't be too picky when using such plots.

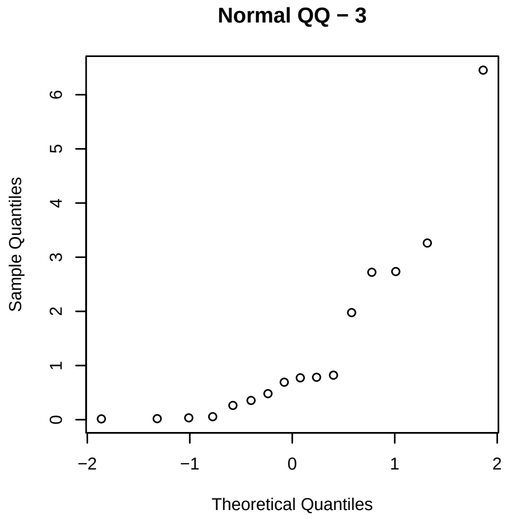
Here are some examples of QQ plots made with non-normal data, along with their corresponding histograms:

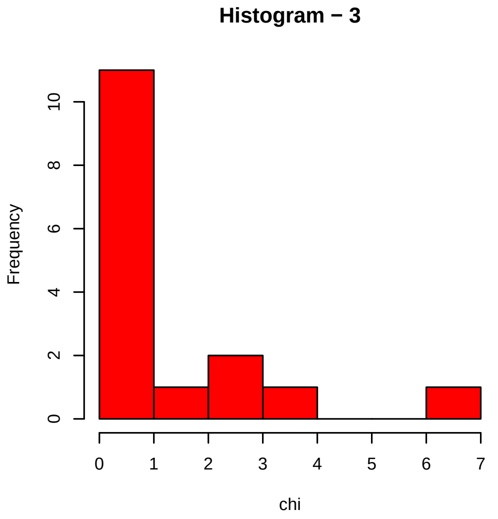

Normal QQ − 1


Histogram − 1


Normal QQ − 2


Histogram − 2


Normal QQ − 3


Histogram − 3

Some of these QQ plots exhibit outliers or show strong curves. Certainly none of the histograms look much like bell shapes. As we will find out shortly, however, there is some very strong theory that can help us, even when the plots don't look very normal. The key is to have enough samples.

3 The Central Limit Theorem

The concepts in this section are in section 4.12 of Ott and Longnecker.

There is a very important theorem that, in many cases, formalizes the idea that 'pretty good is good enough.' It is called the **Central Limit Theorem**, or **CLT**. The statement is as follows:

- Let X_1, X_2, ..., X_n be a collection of iid RVs with $E(X_i) = \mu$ and $VAR(X_i) = \sigma^2$. For large enough n, the distribution of $\bar{X}$ will be approximately normal with $E(\bar{X}) = \mu$ and $VAR(\bar{X}) = \frac{\sigma^2}{n}$. That is, $\bar{X} \dot{\sim} N(\mu, \frac{\sigma^2}{n})$. The required size for n depends on the nature of the population distribution of X_i.

This is a pretty amazing result. It says that if you take an iid sample from any population, as long as the sample is large enough, the sample mean will be approximately normal. Practically, this means we don't need to stress out too much about evaluating normality for our sample, so long as the sample is large enough and the QQ plot isn't too bad. How large is "large enough" depends entirely on what the distribution of the X_i looks like. For reasonably symmetric distributions with no outliers, $n = 5$ could be sufficient. For distributions with extreme skew or heavy tails/outliers, you may need upwards of $n = 100$ or more. But for much real-world data, $n = 30$ is a relatively safe cut-off, and this sample size is what is typically prescribed to use the CLT.

Go to R code simulating the CLT.

4 Confidence Intervals

The concepts in this section are in section 5.2 of Ott and Longnecker.

WARNING: this section is very conceptual

We now return to CIs. To back up a bit, an estimate as we've described it should really be called a **point estimate**. It is our single best guess at the value of the population parameter.

Point estimates are almost always wrong, but if the estimated SE and bias are small, the estimate will usually be close to correct. On the other hand, CIs are what we call **interval estimates**. An interval estimate is an interval of values that represents collection of good guesses.

To make a CI, we must compute the values of the lower and upper limits of the interval. Call them l and u. Note that these are written as lower-case letters. This is on purpose, and reflects the fact that the limits of a CI are just realizations of the random variables L and U. The interval is random, and once we take a sample and realize L and U into l and u, we get one realization of a random interval. We would like to be able to say that the interval is realized from a procedure with some (hopefully large) probability of containing or "covering" the population parameter we are estimating. When making a confidence interval about the population parameter μ, this would, for example, take the form:

$$P(L \leq \mu \leq U) = 1 - \alpha$$

where we hope that α is small. Our goal is to find formulas for L and U that are functions of the data we collected. In order to do this, we start in what might seem a strange place, by making a statement about the standard normal:

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha.$$

This will need explanation, with pictures. Draw a standard normal curve. Start with the right-hand side. Draw a tick mark at $z_{\alpha/2}$, and remind them that this is in critical value notation – it is the place where the area to the right is $\alpha/2$. Then do the same thing on the left-hand side, reminding them why the symmetry of the curve implies that the correct position of this tick mark is at $-z_{\alpha/2}$, and that the area to the left of this place will be $\alpha/2$. Finally, since we have $\alpha/2$ on either side, and the total area must be 1, then we must have $1 - \alpha$ between them.

How does this help us? Well, if we make the assumption that $X_i \sim N(\mu, \sigma^2)$, then we have shown that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. (Or, if n is large enough that the CLT kicks in, this is at least approximately true.) Then, by standardization, this means that:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1),$$

and thus:

$$P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha/2}) = 1 - \alpha$$

Then we move terms around in the inequality until we get μ by itself in the middle. After a few steps, we find:

$$P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha,$$

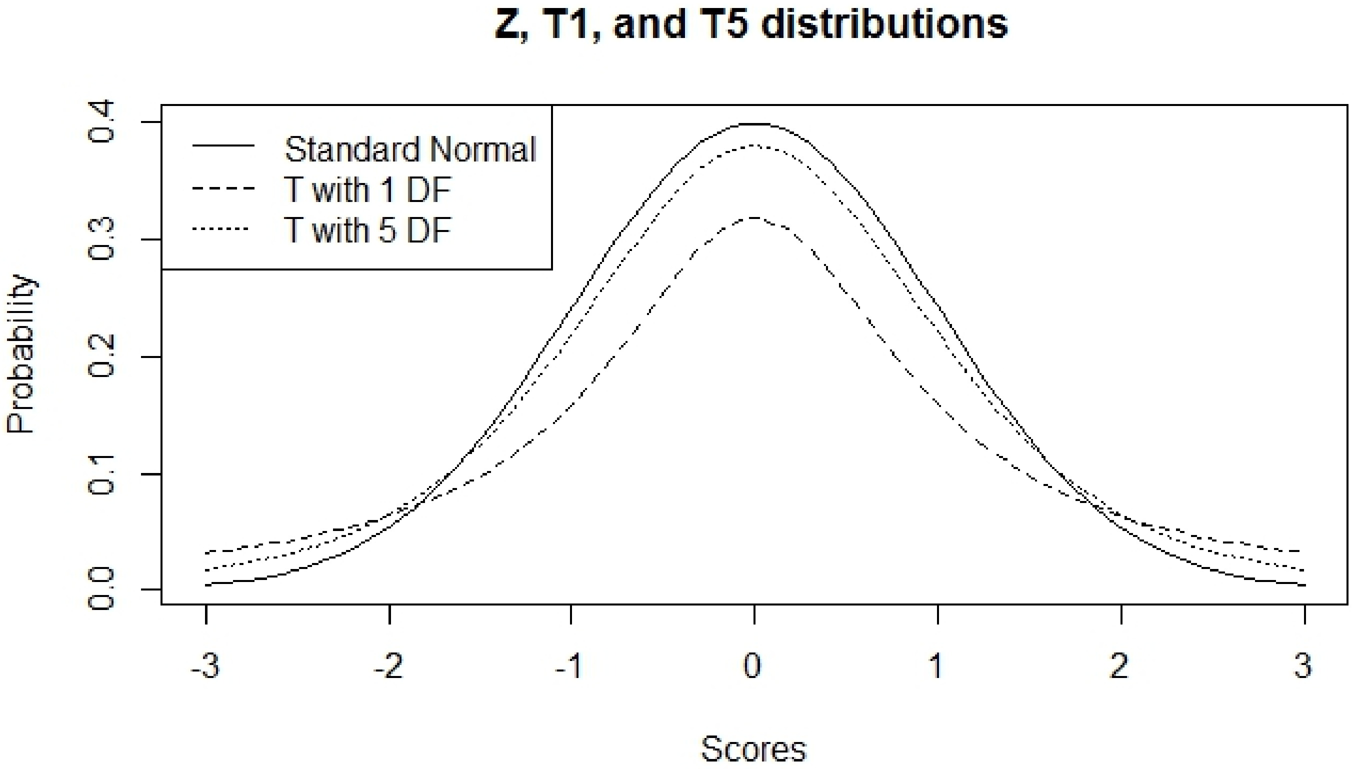
which accomplishes our first goal with $L = \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and $U = \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. The interval (L, U) is the CI. Since the interval is symmetric, we can also write it as:

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

where the right-hand-side of this expression is called the **half-width**, because it is the width of half of the interval. Now you can see why L and U are random - it is because $\bar{X}$ is random. To realize this random interval, simply replace $\bar{X}$ with the estimate $\bar{x}$.

What does this interval mean? The best interpretation is pretty wordy. If you had theoretically taken many samples from this population, and created a different interval by this method for each sample, $100(1 - \alpha)\%$ of them would cover the true value of the parameter, μ. This is usually shortened to saying we have $100(1 - \alpha)\%$ **confidence** that the interval covers μ. Note that this doesn't say anything about any particular interval we calculate. A CI once it is realized either covers μ, or it doesn't. But it does say that the *procedure* has a $100(1 - \alpha)\%$ coverage rate.

Before we actually compute the interval for the paint data, we must make a somewhat embarrassing statement. The interval we just calculated has the stated coverage probability only when σ is known. The reason is that $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$ only when σ is known exactly. If we estimate σ using the sample standard deviation S, then the true distribution of $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ is called the t **distribution on** $n - 1$ **degrees of freedom**. It looks very similar to a standard normal: it is symmetric and bell-shaped, but it is a little more spread out. The amount of additional spread decreases as the the quality of the estimate of σ improves, that is, as the degrees of freedom (i.e. the sample size) increase.



The t-distribution was first discovered in 1908 by W. S. Gosset (1876-1937), who worked at the time at Guinness Brewing Company, mostly on barley experiments. Since the brewery was worried about company secrets aiding the competition, he was forced to publish his work under the pseudonym 'Student.'

We write:

$$\frac{\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}}{} \sim t_{\nu},$$

where $\nu = n-1$ is called the **degrees of freedom**, and determines the amount of additional spread. As ν (or equivalently, n) gets larger, the extra spread diminishes. When $\nu = \infty$, the t distribution is exactly the same as the standard normal. Actually, when the sample size is large, usually somewhere above 30, the t is so close to the normal that using the normal is a very good approximation. But when σ is unknown and we estimate it with S, the exact probability statement about a CI for μ is:

$$P(\bar{X} - t_{(n-1,\alpha/2)}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{(n-1,\alpha/2)}\frac{S}{\sqrt{n}}) = 1-\alpha,$$

where $t_{(n-1,\alpha/2)}$ denotes the $\alpha/2$ critical value of the t-distribution on $n-1$ degrees of freedom. *DRAW THIS!* We realize this random interval by plugging in our realizations $\bar{x}$ and s. Probabilities for the t distribution can be looked up in tables. *Show the T-table and explain it!*

Example. Now we can create an interval for the paint data. For this data, $\bar{x} = 1.348$ mm, $s = 0.3385$ mm, and $n = 16$. Suppose we want 95% confidence, so that $\alpha = 0.05$. From the table we see $t_{(15,0.05/2)} = 2.13$, so the interval is:

$$(1.348 - 2.13(0.3385/\sqrt{16}), 1.348 + 2.13(0.3385/\sqrt{16}))$$

i.e.

$$(1.168, 1.528).$$

Notice that $\frac{S}{\sqrt{n}} = \widehat{SE(\bar{X})}$. It turns out that many confidence intervals have a very general form for their realization that looks something like:

$$\text{estimate} \pm \text{multiplier} * \text{estimated SE(estimator)}$$

In our case, the estimate is $\bar{x}$, the multiplier is $t_{(n-1,\alpha/2)}$, and the estimated SE of the estimator is $\frac{S}{\sqrt{n}}$. These values will change depending on what is being estimated, and how it is estimated.

Let's now consider the form of the confidence interval based on the t-distribution:

$$\bar{X} \pm t_{(n-1,\alpha/2)} \frac{S}{\sqrt{n}}.$$

Note that not only is the center of this interval estimator random, but the length will be random since S is also a random variable. We now illustrate this phenomenon with an **example.** Suppose the true mean paint thickness is $\mu = 1.25$ mm with a population standard deviation $\sigma = 0.5$mm. *Go to computer demonstration of CIs.*

5 Determining Sample Size

The concepts in this section are covered in section 5.3 of Ott and Longnecker

Of course, it is always desirable to have the confidence as high as possible, and the confidence interval as narrow as possible, because these would be indications of a very accurate estimate. *Here you could ask the class, "Is it possible to have a 100% CI?" Well, yes, but only if it was infinitely long, or if it encompassed all possible values of the parameter. For example, in estimating a proportion, we are always 100% confident that the value is between 0 and 1, but this is trivial and completely uninformative.* For any given confidence level, we can in principle control the width of the interval by increasing the sample size. We will show how through an example.

Example: Consider the paint example again. The foreman at the car manufacturer is not satisfied with the precision of interval we just calculated, $(1.168, 1.528)$: it's too long! He wants a 95% CI for μ that is shorter. He desires the half-width to be no larger than 0.1mm. About many blocks should be sampled in a new experiment in order to achieve this?

Now if σ, the true population standard deviation of paint thickness, were known, this problem would be very simple. Because in this case the multiplier would come from the normal distribution, which, unlike the T-distribution, doesn't depend on n. Since $z_{\alpha/2} = 1.96$ (*refer to the normal table!!*), we would just need to solve the equation

$$0.1 = 1.96 * \frac{\sigma}{\sqrt{n}},$$

which just involves some simple algebra (remember, σ is known in this hypothetical!).

Unfortunately, in the paint example, σ is unknown, meaning we have to use intervals based on the T-distribution. In this case, solving the equation

$$0.1 = t_{(n-1,\alpha/2)}\frac{s}{\sqrt{n}}$$

for n is challenging because the multiplier also depends on n and, further, s would need to be calculated from the new data that we haven't yet sampled! We need to make a few simplifying assumptions to solve for n and yield an educated guess at the requisite sample size. First, let's assume $s = 0.3385$ mm is a pretty good estimate of the population SD so we can use it. Second, let's assume that the value of n required to hit 95% confidence with such a short window is large enough that the t-distribution and the normal distribution are close, so the multipliers would be about the same. (Remember, the T and normal distributions are close around 30 to 40 degrees of freedom.) Then an educated guess for n can be obtained by solving the equation:

$$0.1 = 1.96(0.3385/\sqrt{n}),$$

which gives:

$$n = \tfrac{(1.96^2)(0.3385^2)}{0.1^2} = 44.01, \text{ which we round up to 45.}$$

Note that the final value for n is in fact large enough that the normal and t-distribution for this sample size would be very close. If this were not the case, a conservative approach to sample size calculation would use a t-multiplier with small degrees of freedom rather than a normal multiplier. E.g. in the paint example we could keep $t_{15,0.025} = 2.13$. This would give us a larger value for n than the normal approach. Otherwise trial and error would be needed to find the right n. And, there is still the issue of using an estimate for σ before the new data are even collected!

6 Estimation and Inference for Population Proportions

The concepts in this section are in section 10.2 of Ott and Longnecker.

We just spent a lot of time talking about the estimation of population means. We'd now like to discuss estimation of population proportions. We once again begin with an example.

Example. An accounting firm has a large list of clients (the population), and each client has a file with information about that client. The firm has noticed errors in some of these files, and has decided that it would be worthwhile to know the proportion of files that contain an error. Call the population proportion of files in error π. It was decided to take a simple random sample of size $n = 50$, and use the results of the sample to estimate π. Then they will decide whether it is worth the cost of examining and rectifying all the files. Each selected file was thoroughly reviewed, and classified as either containing an error (call this 1), or not (call this 0). The results are as follows:

Files with an error: 10; Files without any errors, 40.

We should begin by finding an estimator of π. To do that, observe that the procedure by which the files were selected is a binomial process. To see this, let the random variable Y_i be the *indicator* that the ith file sampled had errors: that is, Y_i is 1 if the file contains an error and 0 otherwise. The pmf of Y_i for all i is:

y_i	$p(y_i)$
0	$1-\pi$
1	π

Then the random variable $B = Y_1 + Y_2 + ... + Y_n = \sum_{i=1}^{n} Y_i \sim Bin(n, \pi)$. Note that the RVs being iid and each one having two possible outcomes satisfies the criteria of a binomial process. Also, observe that B just counts the number of files with errors. (In the example, we happened to realize $b = 10$ errors out of $n = 50$ files sampled.) So what is a natural estimator of the true proportion of files with errors? The sample proportion. That is, the proportion of successes in the sample, which is given by the formula:

$$\text{Sample proportion: } \hat{\pi} = P = \frac{\sum_{i=1}^{n} Y_i}{n}.$$

We can determine the expectation and variance by again using the E and Var of a Bernoulli RV: recall $E(Y_i) = \pi$ and $VAR(Y_i) = \pi(1-\pi)$. Hence:

$$E(P) = \pi,\ VAR(P) = \frac{\pi(1-\pi)}{n}, SE(P) = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

This tells us that the estimator P is unbiased for π, and additionally gives us a theoretical measure of accuracy of the estimator P (that is, the standard error of P). Note that if $\pi = 0$ or 1 the standard error is 0 – this should make sense! As in the discussion of $\bar{X}$, we can get the estimated standard error of P by plugging in our estimator of π:

$$\text{Estimated standard error of P: } \widehat{SE(P)} = \sqrt{\tfrac{P(1-P)}{n}}.$$

If we want to make a CI for π, we must know the distribution of P. The exact distribution of P is related to a binomial, but it turns out that making an exact CI based on this fact, while possible, is mathematically challenging and difficult, so we will not do that. But if n is large the CLT gives us a way around this problem. So long as the sample size is large enough, all the conditions of the CLT are met, because the Y_i are iid, and P is really just a sample mean of a bunch of zeros and ones. Thus, for large samples, P is approximately distributed as a normal:

$$P \dot{\sim} N(\pi, \tfrac{\pi(1-\pi)}{n}).$$

This means that an approximate $100(1-\alpha)\%$ CI for π would be of the form:

$$P \pm z_{\alpha/2}\sqrt{\tfrac{P(1-P)}{n}}.$$

The necessary sample size for the CLT to provide a good approximation in this case depends on the value of π. The closer π is to 0.5, the smaller n is necessary. Generally, a rule of thumb is that if $n\pi > 5$ and $n(1-\pi) > 5$, the approximation will be good. In this expression π can be approximated by p as estimated by the sample. The rule then becomes, you should have observed at least 5 successes and at least 5 failures.

Returning to the audit data, our estimate would be $p = 10/50 = 0.2$, with estimated standard error $\sqrt{(0.2 * 0.8)/50} = 0.057$. The CLT should be a good approximation since we have 10 successes and 40 failures, more than 5 each. Thus an approximate 95% CI for π would be $0.2 \pm 1.96 * 0.057$, or $(0.088, 0.312)$.