Chapter 7, Part 2

1 Sign Test

The concepts in this section are in section 5.9 of Ott and Longnecker. The example is from Ott and Longnecker pg. 268, with a few changes.

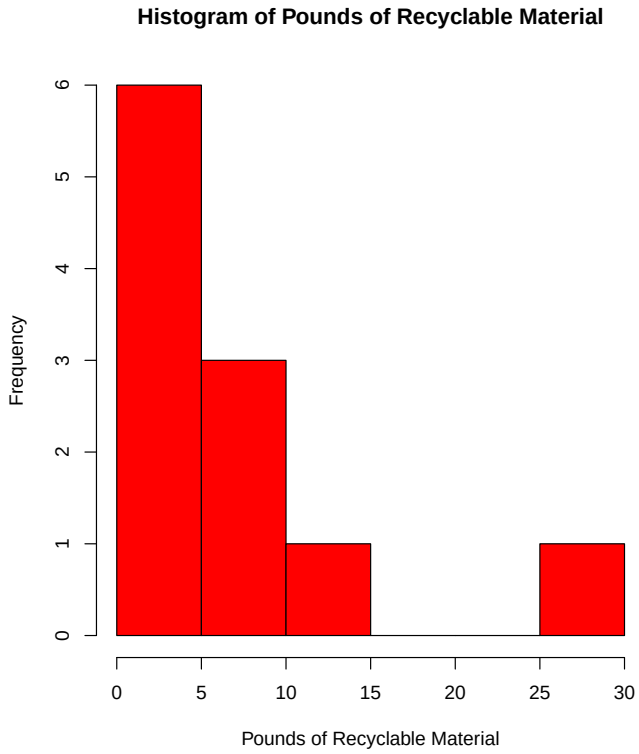
Another option–if the data does not appear to be drawn from a normal distribution and the sample size is small–is the sign test. The sign test in its most general form is technically a test concerning the value of the population *median.* However, if the sample data suggests that the population has a roughly symmetric distribution, it is equivalent to a test for the population mean.

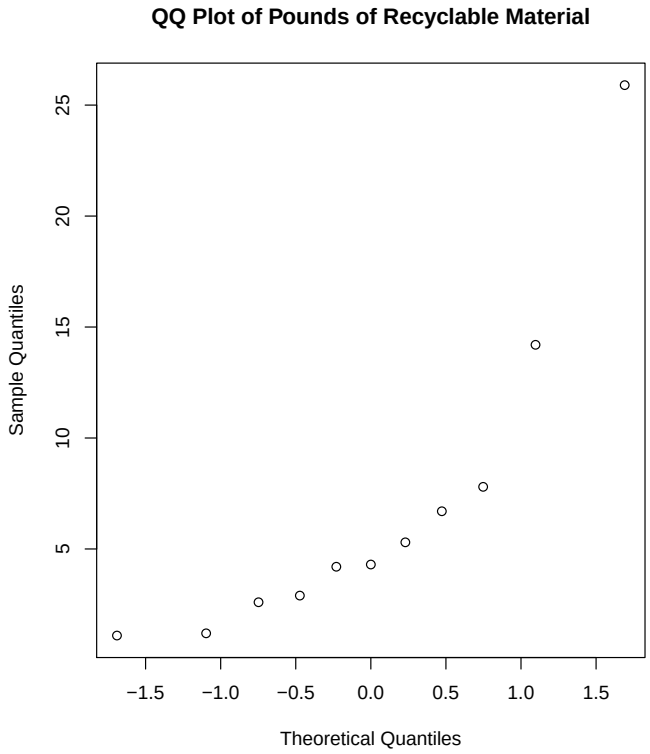
1.1 Example

The sanitation department in a large city is considering separating recyclable material out of the trash to save on landfill space and make money selling the recyclables. Based on data from other cities, it is determined that if at least half of the households in the city have 4.6 lbs or more of reclaimable recyclable material, then the separation will be profitable for the city. A random sample of 11 households yields the following data on pounds of recyclable material found in the trash:

14.2, 5.3, 2.9, 4.2, 1.2, 4.3, 1.1, 2.6, 6.7, 7.8, 25.9

Note that the median of the sample is 4.3 lbs. We are interested in whether the median of the whole city is 4.6lbs. We should always begin by plotting the data to get a sense of its shape, and check population normality. A QQ plot and histogram are good places to start:





The histogram looks neither normal nor symmetric, and the QQ plot also does not support normality. Since we only have $n = 11$, the CLT is risky, and in any event the nature of our scientific question is really concerned about the median weight of recyclable material. So, letting m be the population median, we wish to test:

$$H_0 : m = 4.6$$
$$H_A : m > 4.6.$$

To do the test we need to find a test statistic. If the null hypothesis is correct, then by the definition of the median, the sample should have about half of the observations greater than 4.6 and half less than 4.6. Said another way, the probability of observing a value greater than 4.6 in the sample should be 0.5. Thus a natural choice of test statistic is the number of observations in the sample that are greater than (or, with no loss of generality, less than) 4.6, which we might call B. If the null hypothesis is true, $B \sim Bin(11, 0.5)$. (**Technical note:** our binomial test statistic is on $n = 11$ trials since none of the 11 observed data values are exactly equal to the hypothetical median 4.6. In general, if some of the observations are tied with the null median value, we remove them from the data set and proceed only using the reduced data set.)

Finding rejection regions for such a test ends up being a little complicated, so we will not pursue it, but finding p-values is comparatively straightforward. Our observed test statistic is $b = 5$. Remember that we are counting the number of data values larger than a hypothetical

median and our alternative hypothesis is that the true median is greater than 4.6. Therefore $b = 6$ would be more contrary than what we observed, as would $b = 7$, and so forth. Hence, the p-value is $P(B \geq 5) = 0.726$. This was found by adding Binomial$(11, 0.5)$ probabilities for $b = 5$, $b = 6$, ..., $b = 11$ when $n = 11$ and $p = 0.5$. Therefore there is not sufficient evidence to reject the null. There is not strong evidence that separation of the recyclables would be profitable in this city.

Note that correct computation of the p-value does take a little thought and care. It depends on the fact that B counts the number of observations greater than a hypothetical median and the direction of H_A. If H_A contains a '>' sign, then you get the p-value by counting the probabilities greater than the observed value of $B = b$. To get a feel for why this is true, look at some extreme counterfactuals. Suppose we retain $H_0 : m = 4.6$ vs. $H_A : m > 4.6$, but now suppose all the data values are greater than 4.6, so $b = 11$. The data align in the direction of H_A and strongly suggest H_0 is false, so the p-value is small. And indeed, $P(B \geq 11) = P(B = 11) = 0.5^{11} =$ tiny. Now, suppose we had a counterfactual data set where wall the data values are less than 4.6, so $b = 0$. Such data support H_0 far more than H_A, and so correspondingly, we should get a large p-value; and indeed, $P(B \geq 0) = 1$.

Now, suppose we keep the 11 original data points and change the alternative hypothesis to $H_A : m < 4.6$. We still have $b = 5$ but what constitutes 'more extreme' or 'more contrary' under this H_A? Note that $b = 6, 7, \ldots, 11$ support H_0 more than H_A because this suggests the true median is more than 4.6. So in this case, we compute the p-value by calculating $P(B \leq b)$ for $B \sim$ Binomial$(11, 0.5)$. That is we calculate binomial probabilities *less than* what we observed. For this H_A, the p-value would be $P(B = 5) + P(B = 4) + \ldots + P(B = 1) + P(B = 0) = 0.5$. So the data do not suggest the true median is less than 4.6 either.

What if the test was two-sided? In this situation you would compute $P(B \geq b)$ and $P(B \leq b)$ and take twice the smaller of these two values as the p-value. **Example:** in the garbage example, suppose $H_A : m \neq 4.6$. Since $P(B \leq 5) = 0.5$ is the smaller of the two-values, the p-value against this alternative is $2 * 0.5 = 1$. This makes sense since neither one-direction test yielded compelling evidence against H_0.

1.2 Recap of sign test

- **Assumption.** $X_1, \ldots, X_n$ are an i.i.d. sample from some population.
- **Hypothesis.** Let m be the population *median*. We wish to test $H_0 : m = m_0$ (m_0 is some hypothetical median of interest to us).
- **Test statistic.** $B =$ # of data values greater than m_0. (Ignore values tied with m_0.)

  Note that if H_0 is true, $B \sim$ Binomial$(n^*, 0.5)$, where n^* is the number of data points not equal to m_0 (remember, these values are removed from the data set).
- **P-value**. Let b be the observed number of data points greater than m_0. If:

- $H_A : m > m_0$: calculate $P(B \geq b) = P(B = b) + P(B = b+1) + \ldots + P(B = n^*)$.
- $H_A : m < m_0$: calculate $P(B \leq b) = P(B = b) + P(B = b-1) + \ldots + P(B = 1) + P(B = 0)$.
- $H_A : m \neq m_0$: calculate $P(B \geq b)$ and $P(B \leq b)$. Then, take the smaller of these two values, and double that.

2 Test for Population Proportions

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

2.1 Example

Recall the accounting example. An accounting firm has a large list of clients, and each client has a file with information. Call the population proportion of files with any errors π. The company CEO decides that if π is greater than 0.1, then it will be worthwhile to review and fix every file. A simple random sample of size $n = 50$ is taken, and as previously, the results are as follows:

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

We wish to test:

$$H_0 : \pi = 0.1$$
$$H_A : \pi > 0.1.$$

We would like to use the CLT to simplify calculations. We check that indeed $0.1(50) = 5$ and $0.9(50) = 45 > 5$, so we should be able to use the CLT (we need the CLT to hold *under the null hypothesis*, that's why we use $\pi = 0.1$ instead of the observed proportion). Thus, if H_0 were true:

$$P \dot{\sim} N(0.1, \tfrac{0.1(1-0.1)}{50}),$$

which means that:

$$Z = \frac{P-0.1}{\sqrt{\frac{0.1(1-0.1)}{50}}} \dot{\sim} N(0,1).$$

Thus to compute the p-value, we compute the observed value of the right hand side of the above display, and compare that to a normal distribution. Our observed statistic is $z_{obs} = \frac{0.2-0.1}{\sqrt{\frac{0.1(1-0.1)}{50}}} = 2.357$, so the p-value is $P(Z > 2.357) = 0.009$. At the 5% level, we would reject the null, and conclude that too high of a proportion of files are in error. All files should be checked and fixed.

2.2 Recap of test for population proportion

- **Assumption.** $X_1, \ldots, X_n$ are i.i.d. Ber(π) and n is large.
- **Hypotheses.** $H_0 : \pi = \pi_0$.
- **Test statistic.** Let $P = \frac{\sum_{i=1}^{n} X_i}{n}$ be the proportion of successes. Check that $n\pi_0 > 5$ and $n(1-\pi_0) > 5$ so that the CLT holds under H_0. Then the test statistic is a Z-statistic:

$$Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}.$$

- **P-value.** Let $P = p$ be the observed proportion of successes in the data. If:
  - $H_A : \pi > \pi_0$: calculate $z_{obs} = \frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$ and calculate $P(Z > z_{obs})$ using a standard normal table.
  - $H_A : \pi < \pi_0$: calculate $P(Z < z_{obs})$.
  - $H_A : \pi \neq \pi_0$: calculate $2 * P(Z > |z_{obs}|)$.

3 Next lecture

In the next section, we will discuss how to test whether two independent populations have the same location. This is where things start to get really interesting and applicable to real world situations.