Chapter 4 Hypothesis Tests

4.1 Hypothesis Testing

Source: MATH 221 Textbook

Whenever sample data is used to infer a characteristic of a population, it is called making an inference. **Inferential statistics** represents a collection of methods that can be used to make inferences about a population.

This foundational assumption is called the null hypothesis. The **null hypothesis** is a statement about the population that represents the status quo, conventional wisdom, or what is generally accepted as true. Using the made-up Toyota Camry example from the last chapter, the null hypothesis is:

H_0: 1% of all Toyota Camrys have a defect in the braking system.

The purpose of a statistical study or experiment is to see if there is sufficient evidence against the null hypothesis. If there is sufficient evidence, we reject the null hypothesis. If the null hypothesis is rejected, it is rejected in favor of another statement about the population: the **alternative hypothesis**. In our example, let's assume Toyota wants to know if more than 1% of the population has a deffective braking system. The alternative hypothesis would then be:

H_a: More than 1% of all Toyota Camrys have a defect in the braking system

Notice that both the null and alternative hypotheses are statements about the population, not just the sample being tested. Again, the goal is to determine if what is observed in the sample can be assumed to be true in the population.

There is a formal procedure for testing the null and alternative hypotheses, called a **hypothesis test**. In a hypothesis test, the null hypothesis is always assumed to be true. If there is sufficient evidence against the null hypothesis, it is rejected. The evidence against the null hypothesis is assessed using a number called the P-value. The **P-value** is the probability of obtaining a result (called a test statistic) at least as extreme as the one you calculated, assuming the null hypothesis is true. We reject the null hypothesis if the P-value is small, say less than 0.05. If we assume that only 1% of Camrys have the defective braking system, the P-value is the probability of observing a number of defective cars that is as large or larger than that which was observed in the test sample.

For this example (again, which is completely made up), the P-value was determined to be 0.68. Assuming the null hypothesis is true, the probability of observing defects in the braking system at least as often as was observed in the test sample was 0.68. This is a very large value. So, it is not surprising to have observed the number of the defects in the sample in this case. The probability that these differences could occur due to chance is very high. The conclusion is that the null hypothesis should not be rejected.

If the P-value is low, the null hypothesis is rejected. If this probability is large, the null hypothesis is not rejected.

Hypothesis tests sometimes lead accidentally to incorrect conclusions because we use data from samples (as opposed to data from entire populations). When random samples are selected, some of the samples will contain disproportionately few or many cars with defects, just by chance.

Think about drawing marbles from a container in which most of the marbles are white and a few are red. Each marble represents a Toyota Camry, and the red marbles represent Camrys with defective brakes.

If you choose a random sample of the marbles in the jar, you might get all the red marbles in your sample, just by chance. This might lead you to conclude that there are many red marbles in the container, which is false. This is like Toyota rejecting the null hypothesis when it is true, because their sample contains more Camrys with bad brake systems than it should—just due to chance.

Likewise, when drawing marbles from your container, you might select none of the red marbles. This may lead you to conclude that there are no red marbles in the container, or very few, which is false. This is like Toyota failing to reject the null hypothesis when it is false, because their sample contains fewer cars with bad brakes than it should—again, just due to chance.

Notice that if you draw only one marble, it will be either white or red, and you will be in one of the situations discussed in the previous two paragraphs. On the other hand, if you draw a larger sample, say 40, the chances are you will get a pretty good idea of the proportion of red to white marbles. Certainly better than if you only draw one marble. This emphasizes the roll of sample size in making inference; in general, the larger the sample size the better we understand the population.

Such errors are no one’s fault; they are an inherent part of hypothesis testing. They make it impossible for us to be certain of the conclusions we draw using the statistical process. The thing to remember is that if we carry out the process correctly, our results are correct often enough to be very useful.

For more information about hypothesis testing, look at the Making Inference section of the Statistics Notebook.

This course will focus on Chi-Square testing, but there are several other types of hypothesis tests available to statisticians depending on the circumstances of their test and data. More information about some of them can be found under the “Making Inference” tab in the Statistics Notebook. The course *Intermediate Statistics* (MATH 325) covers these different tests in more depth.

4.2 Test for Two Proportions

Source: MATH 221 Textbook

The ability to taste the chemical Phenylthiocarbamide (PTC) is hereditary. Some people can taste it, while others cannot. Even though the ability to taste PTC was observed in all age, race, and sex groups, this does not address the issue about whether men or women are more likely to be able taste PTC.

Further exploration of the PTC data allows us to investigate if there is a difference in the proportion of men and women who can taste PTC. The following contingency table summarizes Elise Johnson’s results:

Can Taste PTC?	Female	Male	Total
No	15	14	29
Yes	51	38	89
Total	66	52	118

Researchers want to know if the ability to taste PTC is a sex-linked trait. This can be summarized in the following research question: **Is there a difference in the proportion of men and the proportion of women who can taste PTC?** The hypothesis is that there is no difference in the

the true proportion of men who can taste PTC compared to the true proportion of women who can taste PTC.

$$H_0 : \text{There is no difference in the proportion of PTC tasters for men and women}$$

$$H_1 : \text{There is a difference in the proportion of PTC tasters for men and women}$$

A sample of 66 females and 52 males were provided with PTC strips and asked to indicate if they could taste the chemical or not. (This research was approved by the BYU-Idaho Institutional Review Board.)

When working with categorical data, it is natural to summarize the data by computing proportions. If someone has the ability to taste PTC, we will call this a success. The sample proportion is defined as the number of successes observed divided by the total number of observations. For the females, the proportion of the sample who could taste the PTC was:

$$\overbrace{\hat{p}_1}^{\text{Proportion of females PTC tasters}} = \frac{x_1}{n_1} = \frac{51}{66}$$

This is approximately 77.3% of the people who were surveyed. For the males, the proportion who could taste PTC was:

$$\overbrace{\hat{p}_2}^{\text{Proportion of male PTC tasters}} = \frac{x_2}{n_2} = \frac{38}{52}$$

This works out to be about 73.1%.

Recall the hypothesis described earlier, in simpler terms, it can be read as

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

If the null hypothesis is true, then the proportion of females who can taste PTC is the same as the proportion of males who can taste PTC.

The test statistic is a z, and is given by:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p})(\frac{1}{n_1} + \frac{1}{n_2)})}}$$

In our case:

$$z = \frac{(\frac{51}{66} - \frac{38}{118}) - (0)}{\sqrt{\frac{89}{118}(1 - \frac{89}{118})(\frac{1}{66} + \frac{1}{52)})}}$$

Assume we use $\alpha = 0.05$ to help us make our decision of whether to reject or accept the null hypothesis.

After plugging in our values, we end up with $z = 0.526$. When converted to a P-value, we get $\text{P-value} = 0.599 > 0.05 = \alpha$. **We fail to reject the null hypothesis**. In English we say, there is insufficient evidence to suggest that the true proportion of males who can taste PTC is different from the true proportion of females who can taste PTC.

Men and women appear to be able to taste PTC in equal proportions. There is not enough evidence to say that one gender is able to taste PTC more than the other. It appears that the ability to taste PTC is not a sex-linked trait.

4.3 Chi-Squared Test of Independence

People often wonder whether two things influence each other. For example, people seek chiropractic care for different reasons. We may want to know if those reasons are different for Europeans than for Americans or Australians. This question can be expressed as “Do reasons for seeking chiropractic care depend on the location in which one lives?”

This question has only two possible answers: “yes” and “no.” The answer “no” can be written as “Motivations for seeking chiropractic care and one’s location are independent.” (The statistical meaning of “independent” is too technical to give here. However, for now, you can think of it as meaning that the two variables are not associated in any way. For example, neither variable depends on the other.) Writing the answer “no” this way allows us to use it as the null hypothesis of a test. We can write the alternative hypothesis by expressing the answer “yes” as “Motivations for seeking chirporactic care and one’s location are not independent.” (Reasons for wording it this way will be given after you’ve been through the entire hypothesis test.)

When we have our observed counts in hand, software will calculate the counts we should expect to see, if the null hypothesis is true. We call these the “expected counts.” The software will then subtract the observed counts from the expected counts and combine these differences to create a single number that we can use to get a P-value. That single number is called the χ2 test statistic. (Note that χ is a Greek letter, and its name is “ki”, as in “kite”. The symbol χ2 should be pronounced “ki squared,” but many people pronounce it “ki-square.”)

4.3.1 Assumptions

The following requirements must be met in order to conduct a χ2 test of independence:

- You must use simple random sampling to obtain a sample from a single population.
- Each expected count must be greater than or equal to 5. Let’s walk through the rest of the chiropractic care example.

A study was conducted to determine why patients seek chiropractic care. Patients were classified based on their location and their motivation for seeking treatment. Using descriptions developed by Green and Krueter, patients were asked which of the five reasons led them to seek chiropractic care :

- Wellness: defined as optimizing health among the self-identified healthy
- Preventive health: defined as preventing illness among the self-identified healthy
- At risk: defined as preventing illness among the currently healthy who are at heightened risk to develop a specific condition
- Sick role: defined as getting well among those self-perceived as ill with an emphasis on therapist-directed treatment
- Self care: defined as getting well among those self-perceived as ill favoring the use of self vs. therapist directed strategies The data from the study are summarized in the following contingency table :

Location	Wellness	Preventive Health	At Risk	Sick Role	Self Care	Total
Europe	23	28	59	77	95	282
Australia	71	59	83	68	188	469
United States	90	76	65	82	252	565
Total	184	163	207	227	535	1316

The research question was whether people’s motivation for seeking chiropractic care was independent of their location: Europe, Australia, or the United States. The hypothesis test used to address this question was the chi-squared (χ2) test of independence. (Recall that the Greek letter χ is pronounced, “ki” as in “kite.”)

The null and alternative hypotheses for this chi-squared test of independence are:

$$H_0 : \text{The location and the motivation for seeking treatment are independent}$$

$$H_1 : \text{The location and the motivation for seeking treatment are not independent}$$

When the Test statistic ($\chi^2 = 49.743$) is calculated, we get a p-value that is essentially 0, which is lower than our $\alpha = 0.05$ and thus we **reject the null hypothesis**.

4.3.2 Interpretation

If the null hypothesis is true, then the interpretation is simple, the two variables are independent. End of story. However, when the null hypothesis is rejected and the alternative is concluded, it becomes interesting to interpret the results because all we know now is that the two variables are somehow associated.

One way to interpret the results is to consider the individual values of

$$\frac{(O_i - E_i)^2}{E_i}$$

which, when square-rooted are sometimes called the Pearson residuals.

$$\sqrt{\frac{(O_i - E_i)^2}{E_i}} = \frac{(O_i - E_i)}{E_i}$$

The Pearson residuals allow a quick understanding of which observed counts are responsible for the χ2 statistic being large. They also show the direction in which the observed counts differ from the expected counts.

4.4 Chi-Square Goodness of Fit Test

The chi-square goodness of fit test is shown below.

$$\sum \frac{(O_i - E_i)^2}{E_i}$$

The purpose of this test is to show whether a group of observations follows an expected distribution.

4.4.1 Hypothesis

For a chi-square goodness of fit test, the hypotheses take the following form.

- Ho: The data are consistent with a specified distribution.
- Ha: The data are not consistent with a specified distribution.

These are then to be compared to an α (alpha) of your choosing (0.01, 0.05, 0.1)

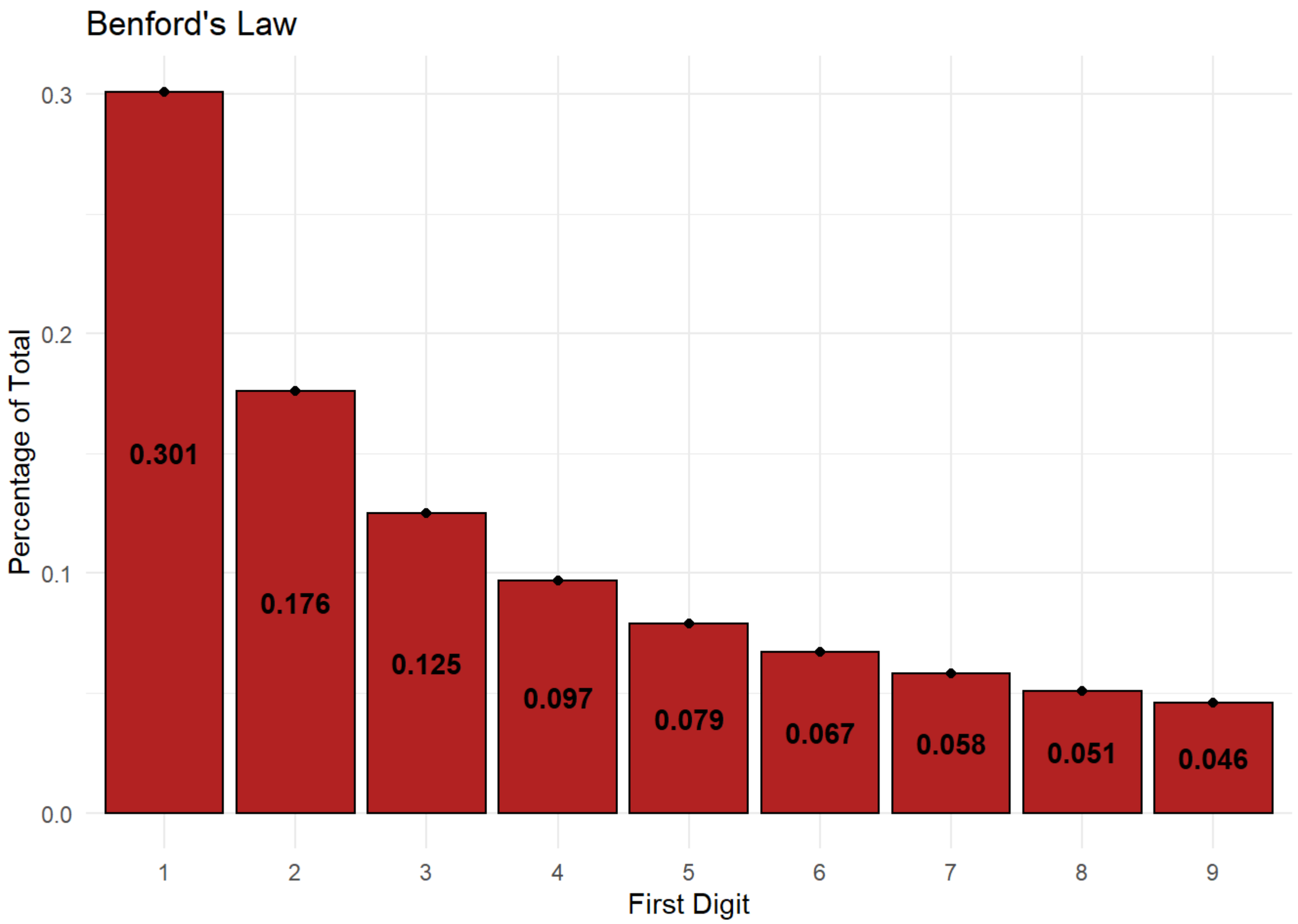
4.4.2 Assumptions

The chi-square goodness of fit test is appropriate when the following conditions are met:

- The data are from a simple random sample
- The groups that are being looked at is categorical, i.e. Qualitative
- There are at least five expected observations per group.

4.4.3 Benford’s Law

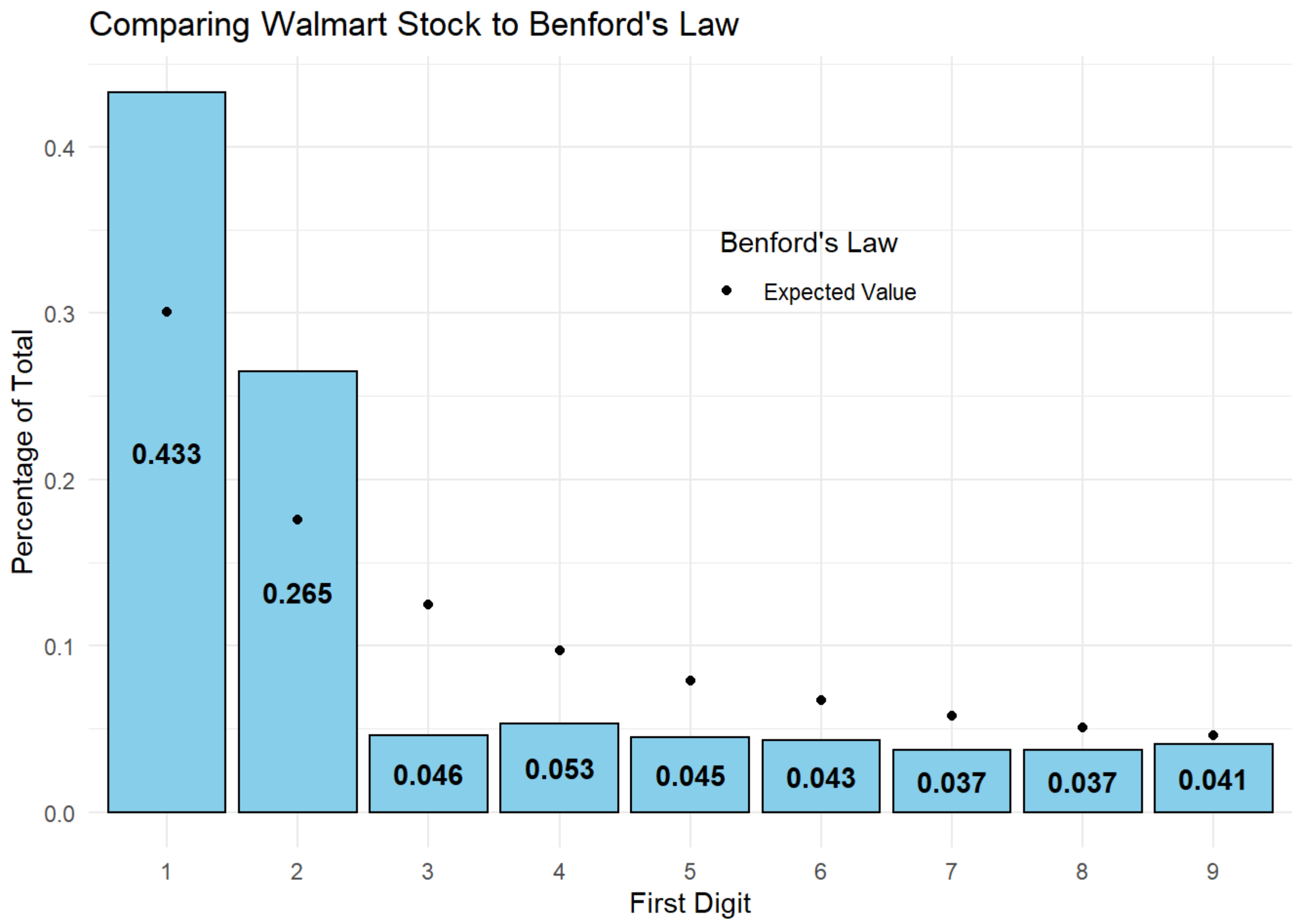
Let’s take a look at Benford’s law as an example of how to use the chi-square goodness of fit test. Benford’s law states that the first digits of a random group of numbers always follows a certain pattern.



We can assume that if I were to take a set of numbers and pull their first digits from that set then we should see this same distribution or something close to it. This is the purpose of the chi-square goodness of fit test.

Example

Let’s take a look at the distribution of the first digits of the Walmart stock. This is a good test set because there are a lot of records of what the stock has been over many years, and we also know that it’s truly random. From there we will randomly subset all of walmart data and we will perform a goodness of fit test.



As we can se it looks pretty similar to Benford’s Law, but lets take a look at the actual p-value calculated from the goodness of fit test. The first thing we need to do is find the expected counts of what Benford’s Law says the distribution should be. To do this, we multiply Benford’s distribution percentages by the total.

301, *176*, *125*, *97*, *79*, *67*, *58*, *51* and *46*

This should be the expected count, now let’s calculate the chi-squared value based on the formula at the beginning of this page. In layman’s - The observed values minus the expected values all squared / divided by the expected values, all summed.

$$\sum \frac{(O_i - E_i)^2}{E_i}$$

The chi-square value is *208*

We can now calculate the the p-value by using a chi-square test. The p-value that is produced is: *0.0000000000000000000000000000000000000347*

As wee can see the p-value is less than our α of .05, therefore we fail to reject and we can say that the random sample of numbers from the Walmart stock follows Benford's Law.