Hypothesis Testing

In the previous topic, we introduced statistical inference. That is, the ideas of making descriptive statements about the population parameters from information in a sample from that population. Those statements took the form of confidence intervals. These intervals gave a range of plausible values for the population parameter. In this topic, we will focus on another aspect of statistical inference, hypothesis testing.

In this section we will be making decisions about a specific values for a population parameter, specifically whether or not that particular value is viable given the data. It is important to remember that when we are making a decision (in effect choosing between two possible outcomes) based upon a sample, that we can make mistakes. As before, the reason for this is variability in the sampling process. We can make the decision that is best given the information at hand, but still be wrong. Statistics gives us a way to quantify the chance that we will make a mistake. Consider the following: we have to decide between two possible choices, A and B. One of these is the correct answer. There are two possible errors we can make. We can decide A when B is the correct answer and we can decide B when A is the correct answer. Regardless of how seemingly strong the information may be in one direction or the other it is important to remember that there is always a probability of making one of these errors. These errors are the result of variability from sample to sample. Another way to view this is to think of the problem as the following picture.

<table>
<thead>
<tr>
<th>Truth</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Correct</td>
<td>Error</td>
</tr>
<tr>
<td>B</td>
<td>Error</td>
<td>Correct</td>
</tr>
</tbody>
</table>

The rows A, B represent the decision that you make, while the columns represent the Truth which we don’t know when we make our decision. As a consequence, regardless of the decision that we make it could be the wrong one.

Definitions

We will be testing whether or not a particular value or particular attribute is reasonable. We might be interested in whether the mean of a population is 500 or whether there is a difference between the mean score of a group given Drug A and the mean of a group given Drug B. Or we might want to know whether or not two categorical random variables are independent. We can test all of these ideas using a hypothesis test.

One important facet of this topic is that we will be making decisions about the population parameters. We begin this process with some definitions that help us define a hypothesis test.

*Definition: A null hypothesis is the “hypothesized value” for the population parameter that is being tested.*

The null hypothesis is often denoted with $H_0$. The null hypothesis is the concept being tested. This is the theoretical value that we want to show is not true. For example the null hypothesis might be that $H_0: \mu = 450$ or $H_0: p = 0.45$.

*Definition: A alternative hypothesis is a collection of alternatives to the null hypothesis.*

The alternative hypothesis is denoted by either $H_1$ or $H_A$. The alternative hypothesis often will take one of three forms. These are:

not equal to (e.g., $H_A: p \neq 0.45$, $H_A: \mu \neq 450$),
less than (e.g., $H_A$: $p < 0.45$, $H_A$: $\mu < 450$),
greater than (e.g., $H_A$: $p > 0.45$, $H_A$: $\mu > 450$).
These are collections of values for the population parameter that differ from the “hypothesized value” of
the null hypothesis.

**Definition:** A test statistic is a quantity whose value we will use to determine whether or not we reject the null hypothesis.

The test statistic is often abbreviated by T.S., which we will use in this topic. Each parameter that we will
test will have it’s own test statistic. The test statistic measures the discrepancy between the observed data
and the hypothesized value of the parameter.

**Definition:** The significance level for a hypothesis test is the probability that you reject the null hypothesis
when the null hypothesis is true.

Chosen by the experimenter (always before the experiment takes place), it determines how different is
“different enough” to reject the null hypothesis. It is the probability that you reject the null hypothesis
when the null hypothesis is true. $\alpha = P(\text{reject null hypothesis} | \text{null hypothesis is true})$. Recall that we are
making a decision and that decision may not be correct.

**Definition:** The standard error is the estimated standard deviation for a statistic.

Often denoted by S.E., the standard error is the estimated variability that exists for the quantity that we’re
using to estimate a parameter. For the sample mean, $\bar{X}$, if we know $\sigma_x$ then the standard deviation of $\bar{X}$
is $\frac{\sigma_x}{\sqrt{n}}$. However, we usually do not know $\sigma_x$. As a consequence, we estimate $\sigma_x$ with $s_x$. So the estimated
standard deviation of $\bar{X}$ is $\frac{s_x}{\sqrt{n}}$. This is then called the standard error of $\bar{X}$.

**Definition:** A decision rule is a rule that determines which of two decisions will be made.

In the case of hypothesis testing, our decision will be between “rejecting” and “not rejecting” a null hypothesis.

A comparison is often made between hypothesis testing and the U.S. legal system. In the U.S. legal system
there is a presumption of innocence and the onus is on the prosecution to present it’s case beyond a
reasonable doubt. The decision that a jury has in criminal cases after hearing the evidence is either “guilty”
or “not guilty”. The jury never finds anyone “innocent”. In hypothesis testing, we will presume that the
null hypothesis is true (like a presumption of innocence). The decision at the end of each statistical test is
to “reject” or “not reject” the null hypothesis. We never “accept” a hypothesis as true, we only say there is
not enough evidence to reject. This is analogous to never finding someone innocent, only saying there is
not enough evidence to convict.

The basis for hypothesis test is the following. We will assume (for the purposes of the test) that the
hypothesized value (found in the null hypothesis) is the true, but unknown, population parameter. Then if
the observed value falls far away from the hypothesized value, we reject the hypothesized values. How far
away is “far”, is determined by $\alpha$. $\alpha$ determines the percentile of the distribution of the test statistic beyond
which we will reject the null hypothesis.

**One-sided Hypothesis Tests**

It is often the case that we suspect that the observed data will be in a particular direction (larger or smaller)
relative to the null value of the parameter. For example, it may be that Ma and Pa, who run Ma and Pa’s
grocery, suspect that they sell out of eggs more than 60% (prior to data collection). In that case you only
reject if the observed value is significantly more than the hypothesized value. For a hypothesis test with a one-sided alternative, the alternative tells us to only look in one direction to reject the null hypothesis. So if the one-sided alternative hypothesis were greater than the hypothesized value, we would only reject the null hypothesis if the observed test statistic were significantly larger than the hypothesized value. Likewise, if the one-sided alternative hypothesis were less than the hypothesized value, then we would only reject the null hypothesis if the observed test statistic were significantly smaller than the hypothesized value.

For a one-sided hypothesis test, we will only reject if that value is at the one tail where we might anticipate the observed test statistic to be. In the picture above, we reject the null hypothesis only if the test statistics is large (in the right tail of the distribution). This would correspond to an alternative hypothesis of greater than the hypothesized parameter value. For an alternative hypothesis that involves less than the hypothesized parameter value, a (symmetric) picture would exist with the area for rejection on the left tail of the sampling distribution.

For the hypothesis tests that follow, both forms of the one-sided alternative hypothesis are given. For each \( H_A \), one alternative is given and another is placed in brackets [ ] after it. Likewise, the corresponding decision rule is given in part 6, with the corresponding decision rule for the other alternative is given in brackets [ ] as well.

**Hypothesis test for a population mean:**

If

1) \( \sigma \) is unknown AND \( n>2 \) and the data is normal

OR

2) \( n \geq 30 \)

then we can use the following to test a particular value for the population mean.

1. \( H_0: \mu = \mu_0 \)
2. \( H_A: \mu < \mu_0 \) [or \( \mu > \mu_0 \)]
3. Significance level is \( \alpha \).
4. The test statistics is \( t = \frac{\bar{X} - \mu_0}{s_x/\sqrt{n}} \)
5. The table value will be \( t_{(n-1),1-\alpha} \)
6. Decision Rule: Reject \( H_0 \) if T.S. < -tabled value
   Do not reject \( H_0 \) if T.S. > -tabled value

[Decision Rule: Reject \( H_0 \) if T.S. > tabled value
   Do not reject \( H_0 \) if T.S. < tabled value]
Example:

It is thought that the mean time for mice to finish the maze that Prof. Statley has in her laboratory is 65 sec. Based upon some recent literature, a new genetically modified mouse may be able to complete the maze in less time. Prof. Statley orders 24 of these genetically modified mice and runs them through the maze. The average time for these mice to complete the maze was 62.48 sec with a standard deviation 6.25 sec. Test the null hypothesis against a one-sided alternative hypothesis.

First, we must check the conditions under which we can use the test given above.

Since we need either $n > 2$ and a distribution that is approximately Normal OR $n > 30$, we cannot use the above methods to test this hypothesis.

Example:

It is thought that the mean time for mice to finish the maze that Prof. Statley has in her laboratory is 65 sec. Based upon some recent literature, a new genetically modified mouse may be able to complete the maze in less time. Prof. Statley orders 24 of these genetically modified mice and runs them through the maze. The average time for these mice to complete the maze was 62.48 sec with a standard deviation 5.27 sec. Test the null hypothesis against a one-sided alternative hypothesis using $\alpha = 0.05$. Assume that the population of times that it takes a mouse to complete the maze has a Normal distribution.

In this example, since $n > 2$ and we are the data (maze completion times) come from a Normal distribution, then we can use the following to test that hypothesis.

1. $H_0: \mu = 65$
2. $H_1: \mu < 65$
3. Significance level is $\alpha = 0.05$
4. The test statistics is $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{62.48 - 65}{5.27/\sqrt{24}} = -2.342$
5. The table value will be $t_{(n-1, 1-\alpha)} = t_{(23, 0.99)} = 2.500$
6. Decision Rule: Reject $H_0$ if $T.S. < -$tabled value (If Calc $t < -$Tab $t$ in the texts lingo.)
   Do not reject $H_0$ if $T.S. > -$tabled value (if calc $t >$- tab $t$ in the texts lingo).

Since $T.S.(-2.342) < -$tabled value (-2.500) we do not reject the null hypothesis.

We conclude at the 5% significance level that the mean completion time for all genetically altered mice is not significantly less than 65 sec.

In this example, the one-sided alternative is less than 65. Since we believe that the genetically altered mice may complete the maze in less than this time. Thus, we only want to reject the null hypothesis, if there is enough evidence from the data that the mean completion time is significantly less than the hypothesized completion time of 65 sec.

Two-sided Hypothesis Tests

We will differentiate between two types of hypothesis tests: two-sided and one-sided hypothesis tests. This section will focus on two-sided hypothesis testing. For a two-sided hypothesis the decision rule is based on the following idea. If the observed value for the test statistic is at the edges of the sampling distribution when the null hypothesis is true then we reject the null hypothesis. The edge(s) are defined by $\alpha$, the significance level of a hypothesis test. For a two-sided hypothesis test, we reject if the test statistic falls
outside the inner central \((1-\alpha)\times100\%\) of the distribution. So if the distribution below is the distribution for the test statistic when the null hypothesis is true, then we reject if the test statistic falls in the upper \(\alpha/2\)th percent of the distribution or in the lower \(\alpha/2\)th percent of the distribution. The rational for this is that if the hypothesized population parameter (the value specified in \(H_0\)) is incorrect then the test statistic will be more likely to be at the extremes of the distribution. However, if the hypothesized population parameter is correct, then we incorrectly reject that value with probability \(\alpha = (\alpha/2 + \alpha/2)\), if the test statistic falls beyond the double-bar lines below.

In the following tests, we will refer to a tabled value. For a two-sided hypothesis, the tabled value represents the double lines in the graph above. The exact location of the tabled value on the x-axis is determined by the value of \(\alpha\).

**Hypothesis test for a population mean:**
If
1) \(\sigma\) is unknown AND \(n>2\) and the data is normal
OR
2) \(n \geq 30\)
then we can use the following to test a particular value for the population mean.

1. \(H_0: \mu = \mu_0\)
2. \(H_1: \mu \neq \mu_0\)
3. Significance level is \(\alpha\).
4. The test statistic is \(t = \frac{\bar{X} - \mu_0}{s_x/\sqrt{n}}\)
5. The tabled value will be \(t_{(n-1,1-\alpha/2)}\).
6. Decision Rule: Reject \(H_0\) if T.S. > tabled value or T.S. < -tabled value
   Otherwise Do not reject \(H_0\)

**Example:**

Consider a two-sided hypothesis test of \(\mu = 50\) with \(\alpha = 0.05\). Suppose that \(n = 45\), \(\bar{X} = 53.42\), \(s_x = 8.45\).

Since \(n>30\), we can use the hypothesis test, described above.

The hypothesis test goes as follows:
1. $H_0$: $\mu = 50$
2. $H_1$: $\mu \neq 50$
3. Significance level is $\alpha = 0.05$
4. The test statistic is $t = \frac{\bar{X} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{53.42 - 50.00}{8.45 / \sqrt{45}} = 2.715$
5. The tabled value will be $t_{(44,0.975)} = 1.960$
6. Decision Rule: Reject $H_0$ if T.S. $> \text{tabled value}$ or $T.S. < -\text{tabled value}$

Since the T.S.(2.715) $> \text{tabled value}$ (1.960) then we reject the null hypothesis.

At the 5% significance level we conclude that the population mean is not 50.

**Example:**

In recent years, there has been an increasing amount of information available (research) on how manual laborers conceptualize and understand their work. One aspect of this is how manual workers deal with stress. Psychology researchers are interested in the average scores of mine workers on a test of stress management. Previous studies have suggested that manual laborers score have a mean score of 22 on this test. These studies also suggest that data from this test is approximately Normally distributed. The psychologists obtain a random sample of 80 mine workers from a mine in southwestern Ohio. These 80 workers each take the test of stress management. Their average score is 21.4 with a standard deviation of 2.43. Test using $\alpha = 0.05$, whether or not we can conclude that mine workers have a mean of 22 on this test.

Since $n>30$ we can use the following to test the null hypothesis.

1. $H_0$: $\mu = 22$
2. $H_1$: $\mu \neq 22$
3. Significance level is $\alpha = 0.05$
4. The test statistic is $t = \frac{\bar{X} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{21.4 - 22}{2.43 / \sqrt{80}} = 2.208$
5. The tabled value will be $t_{(79,0.975)} = 1.960$
6. Decision Rule: Reject $H_0$ if T.S. $> \text{tabled value}$ or $T.S. < -\text{tabled value}$

Since the T.S.(2.208) $> \text{tabled value}$ (1.960) then we reject the null hypothesis.

We conclude that the mean score on this stress test for the population of mine workers is significantly different from 22 at the 0.05 significance level.

Before we move on to another facet of hypothesis testing, it is important to update this notation of making a decision. In the introduction to this topic we discussed choosing (or making a decision) between $A$ and $B$. More explicitly, for hypothesis testing that decision is between $H_0$ and $H_1$, the null and alternative hypotheses. And $\alpha$, the significance level equals $P(\text{reject } H_0 \mid H_0 \text{ is true})$. We can create another table like we had at the beginning of the topic. However, we replace $A$ and $B$ with $H_0$ and $H_1$. 

6 of 8
In hypothesis testing (and the field of Statistics, in general), we differentiate between the types of errors that result from this decision making process. We call the mistake of rejecting $H_0$ when $H_0$ is true a Type I error. The probability of making a Type I error is the significance level, $\alpha$. We call the mistake of not rejecting $H_0$ when $H_1$ is true a Type II error. Two facets of hypothesis tests are noteworthy. First is the significance of the test, $\alpha$. The second is the power of the test that is defined as the probability of rejecting $H_0$ when $H_1$ is true. As with confidence intervals, we don’t get to know the “Truth.” As a consequence we make errors, both Type I and Type II errors. The difference between statistical tests and other methodologies is that with statistical tests we can know the probability of making those errors.

Another aspect of hypothesis testing that has gone unmentioned is statistically significant. When we reject a hypothesis as we did in the first example in this section (an example for hypothesis testing for $\mu$), we can claim that the observed value (the sample mean) is statistically significant at the $\alpha$ significance level. More specifically, the observed value is statistically significantly different from the hypothesized value, $\mu_0$, which was 50. Continuing with that example, the observed value of 53.42 is statistically different from 50 at the 0.05 level. This notion is especially useful when we are talking about testing whether the means of two populations are the same. When we reject the null hypothesis of equally for two population means (either paired or independent populations), then we say that the means of the two populations are statistically significant at the $\alpha$ level.

Finally, we mention that the list presented here is only a fraction of the possible hypothesis tests that exist. There are hypothesis tests (and confidence intervals) for the population standard deviation, for comparing proportions from two populations, for comparing the means of three or more populations, for testing whether or not the standard deviations of two populations are the same, etc. It is important to recognize when you have one of the preceding situations and when you do not have one of those scenarios.

**p-values**

There is another way (the modern method) to determine whether or not a hypothesis should be rejected or not rejected. That way is through the idea of p-values. For hypothesis testing, our logic is that if the observed test statistic is at the edge of the distribution, then we reject the null hypothesis. To do this we compare the test statistic to some tabulated value that corresponds to a percentile of the distribution. The other way is to compare the area beyond the test statistic to the area defined by $\alpha$. If the area beyond the test statistic is smaller than $\alpha$, then you reject the null hypothesis. If the area beyond the test statistic is larger than $\alpha$, then you do not reject the null hypothesis. Consider the following graph.

![Sampling distribution of the test statistic.](image)

This area = p-value.

Observed value of the test statistic.
The observed test statistic will be somewhere along the x-axis of this distribution. The area beyond that value will be called the *p*-value. If the *p*-value is small, then we reject since a small *p*-value implies a smaller area beyond the observed value of the test statistic than is beyond the “tabled value”. If the *p*-value is larger than \( \alpha \), then we do not reject since a large *p*-value implies an area beyond the observed value of the test statistic that is larger than the area beyond the “tabled value.” The upside to *p*-values is that you need only compare the *p*-value to \( \alpha \).

The rules for using *p*-values.

1. For any hypothesis test, if the *p*-value is more than the significance level, \( \alpha \), then you do not reject the null hypothesis.
2. For any hypothesis test, if the *p*-value is less than \( \alpha \), then you reject the null hypothesis.

Another facet of using *p*-values which makes them appealing is they can be used regardless of the value of \( \alpha \). That is, if you simply report the *p*-value then regardless of the value of \( \alpha \) that the reader wants to use, they can make their own decision. For this reason presenting the *p*-value for a hypothesis test is quite common.

For \( H_A: \mu < \mu_0 \) the *p*-value = \( P(t < T.S.) \)

For \( H_A: \mu > \mu_0 \) the *p*-value = \( P(t > T.S.) \)

For \( H_A: \mu \neq \mu_0 \) the *p*-value = \( 2 \times P(t > |T.S.|) \)

**Note:** When you are using computer software to calculate test statistics and *p*-values, there are different methods for calculating *p*-values for one-sided and two-sided alternative hypotheses. It is important to know how a *p*-value is reported by the software packages. The problem arises because there are two ways to report *p*-values for a two-sided alternative hypothesis. These leads to two different decision rules.

**Example:**

If we are testing a null hypothesis using \( \alpha = 0.10 \) and we find that the *p*-value is 0.04, we reject the null hypothesis since the *p*-value is smaller than \( \alpha \).

**Example:**

A medical journal reports the results of a two-sided hypothesis test for comparing the means effects of two drugs for tuberculosis. Their null hypothesis is the two drugs have the same effect and the alternative is two-sided. They report a *p*-value for that test of 0.067. If you want to test the null hypothesis using \( \alpha = 0.05 \), you would not reject the null hypothesis since the *p*-value of 0.067 > \( \alpha = 0.05 \).