7. HYPOTHESIS TESTING

*The most exciting phrase to hear in science, the one that heralds new discoveries, is not ‘Eureka!’ but ‘That's funny...’*

— Isaac Asimov

In the last chapter, we showed how sample data can be used to construct a confidence interval within which the true population mean should fall. These confidence intervals can be used to test various hypotheses concerning the population mean. One example would be to test to see if the population mean was equal to a certain value (say zero). Another example would be to test to see if two populations have the same mean or whether the difference between two population means is zero or nonzero. These are called *hypothesis tests* and they are another example of *inferential statistics*.

Hypothesis tests are based on the concept of *proof by contradiction* ... we state our hypothesis then try to prove that it’s *not* true. Hypothesis tests are composed of the following five parts:

1. Null hypothesis, $H_0$
2. Alternative hypothesis, $H_a$ (also called the "research" hypothesis)
3. The test statistic
4. The rejection region
5. The conclusion

Suppose, for example, that we know that, statewide, the mean yield (per acre) of a particular variety of soybeans is 520 bushels per acre. Let us also suppose that 36 one-acre plots scattered around the state have been treated with a new fertilizer. The average yield from these 36 plots is 573 bushels per acre with a standard deviation of 124 bushels per acre. Can we conclude that the mean yield at the test farms is greater than the others in the state? That is, can we conclude that the difference in means is due to the new fertilizer rather than just randomness?

For this example, we would state the null hypothesis as:

$$H_0: \mu = 520$$

Similarly, we can state the alternative hypothesis as:

$$H_a: \mu > 520$$

If the null hypothesis is true, then, according to the central limit theorem, the sample means should be approximately normally distributed with mean $\mu = 520$ and variance $\sigma^2/n = 427$. The values of $\bar{X}$ that contradict the null hypothesis are located in the upper tail of Figure 1 on the next page.

If the observed value of $\bar{X}$ falls in the rejection region of the figure, we would reject the null hypothesis that the population mean is 520 bushels per acre in favor of the alternative hypothesis that yield is more than 520 bushels per acre. The line marking the rejection region is determined by the probability $(1-\alpha)$ that we would normally use in establishing a confidence interval.
One must understand from the outset that rejection of the null hypothesis does not mean that we are assured
that the alternative hypothesis is true. Likewise, failure to reject the null hypothesis does not mean that we are
assured that the alternative hypothesis is false. Proper interpretations of hypothesis testing results are:

a. **reject H₀** – the observed data provides evidence that H₀ is not true

b. **do not reject H₀** – there is insufficient evidence in the data to indicate that H₀ is not true

Even if H₀ is true, it is possible for the results of the experiment to lead to the rejection of H₀. This is
undesirable, but it cannot be avoided when you’re using a small sample to draw conclusions about a large
population. On the other hand, it is possible that the results of an experiment may not lead to the rejection of
H₀ when in fact, H₀ is false. Proper experimental design, sample selection, etc. can minimize this possibility.

We say that a **type I error** has occurred if we reject the null hypothesis when it is true and a **type II error** has
occurred if we fail to reject the null hypothesis when it is false. The probability of a type I error is denoted by
the symbol α and the probability of a type II error is denoted by the symbol β.

The simple table below illustrates the four possible decisions for the outcome of an experiment:

<table>
<thead>
<tr>
<th>Null Hypothesis</th>
<th>Decision</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept H₀</td>
<td>Correct</td>
<td>β</td>
<td></td>
</tr>
<tr>
<td>Reject H₀</td>
<td>α</td>
<td>Correct</td>
<td></td>
</tr>
</tbody>
</table>

The usual approach in hypothesis testing is to set an upper bound on the chance one is willing to take of
making a type I error (rejecting a “true” H₀) and to minimize the chance of making a type II error (not
rejecting a “false” H₀) through proper experimental design, use of a large enough sample size, and choosing
the correct test statistic. The probability of making a type I error, α, is called the **significance level** of the test.
Traditional levels of significance are 0.10, 0.05, and 0.01, though the level in any particular problem will
depend on the seriousness of a type I error.
SINGLE SAMPLE HYPOTHESIS TESTS ON $\mu$

1. Normal Population, $\sigma$ Known

If the population being sampled is known to be normally distributed or the sample size is large enough that the Central Limit Theorem holds (typically $n \geq 30$), a one-sample hypothesis test on the population mean is as follows:

<table>
<thead>
<tr>
<th>HYPOTHESIS TEST ON A MEAN ($\sigma$ KNOWN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$: $\mu = \mu_0$</td>
</tr>
<tr>
<td>$H_a$: $\mu \neq \mu_0$</td>
</tr>
<tr>
<td>$H_a$: $\mu &gt; \mu_0$</td>
</tr>
<tr>
<td>$H_a$: $\mu &lt; \mu_0$</td>
</tr>
<tr>
<td>$TS$: $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$ uses known $\sigma$</td>
</tr>
</tbody>
</table>

Example

A company that produces bias-ply tires is considering a modification in the tread design. An economic feasibility study indicates that the modification can be justified only if the true average tire life under standard test condition exceeds 40,000 miles. A random sample of $n = 16$ prototype tires is manufactured and tested, resulting in a sample average tire life of $\bar{x} = 40,758$ miles. Suppose the standard deviation for the current version of the tire is $\sigma = 1500$ miles and is not expected to change. Do the data suggest that the modification meets the condition required for changeover? Test the appropriate hypothesis using significance level $\alpha = 0.01$. 
2. Normal Population, \(\sigma\) Unknown, \(n \geq 30\)

If the population being sampled is known to be normally distributed and the sample size is large (typically \(n \geq 30\)), the sample standard deviation can be used as a surrogate for an unknown population standard deviation with little loss of accuracy:

\[
\begin{align*}
\text{HYPOTHESIS TEST ON A MEAN (\(\sigma\) UNKNOWN, \(n\) LARGE)} \\
H_0: & \quad \mu = \mu_0 \\
\text{TS:} & \quad z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad \leftarrow \text{uses } s \text{ instead of } \sigma \\
H_a: & \quad \mu \neq \mu_0 \quad H_a: \quad \mu > \mu_0 \quad H_a: \quad \mu < \mu_0 \\
\text{RR:} & \quad |z| > z_{\alpha/2} \quad \text{RR:} \quad z > z_\alpha \quad \text{RR:} \quad z < -z_\alpha
\end{align*}
\]

Example

A certain type of brick is being considered for use in a particular construction project. The brick will be used unless sample evidence strongly suggests that the true average compressive strength is below 3200 psi. A random sample of 36 bricks is selected and each is tested to failure. The sample average compressive strength is 3109 psi with a standard deviation of 156 psi. At a level of significance of \(\alpha = 0.05\), should the brick be used?
3. Normal Population, σ Unknown, n < 30

If the population being sampled is known to be normally distributed but the sample size is small (typically n < 30), the sample standard deviation can still substitute for the population standard deviation, but the test statistic follows a t distribution instead of a z distribution:

**HYPOTHESIS TEST ON A MEAN (σ UNKNOWN, n SMALL)**

Ho:  μ = μ₀

TS:  \[ t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \] ← uses t instead of z

Ha:  μ ≠ μ₀  Ha:  μ > μ₀  Ha:  μ < μ₀

RR:  |t| > t_{α/2,n-1}  RR:  t > t_{α,n-1}  RR:  t < -t_{α,n-1}

**Example**

In order to test gasoline mileage performance for a new version of one of its compact cars, an automobile manufacturer selected six nonprofessional drivers to drive test cars from Phoenix to Los Angeles. At the conclusion of the trip, the resulting gas mileage numbers for the six cars were:

32.2  29.3  31.5  28.7  30.2  30.0

The manufacturer wishes to advertise that this car gets 30 mpg or better on the highway. Do the sample data support the claim that the manufacturer would like to make? Assume α = 0.05.
TWO SAMPLE HYPOTHESIS TESTS ON \( \mu \)

1. **Comparison of Two Means, \( \sigma_1 \) and \( \sigma_2 \) Known**

   If the populations being sampled are known to be normally distributed or the sample sizes are large enough that the Central Limit Theorem holds (typically \( n \geq 30 \)), a two-sample hypothesis test on the difference between two population means is as follows:

   \[
   \text{DIFFERENCE BETWEEN MEANS (}\sigma\text{ KNOWN)}
   \]

   \[
   H_0: \quad \mu_1 - \mu_2 = \Delta_0 \\
   \text{TS:} \quad z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad \text{← uses known } \sigma
   \]

   \[
   \begin{align*}
   H_a: \quad & \mu_1 - \mu_2 \neq \Delta_0 & H_a: \quad & \mu_1 - \mu_2 > \Delta_0 & H_a: \quad & \mu_1 - \mu_2 < \Delta_0 \\
   \text{RR:} \quad & |z| > z_{\alpha/2} & \text{RR:} \quad & z > z_{\alpha} & \text{RR:} \quad & z < -z_{\alpha}
   \end{align*}
   \]

**Example**

A random sample of 20 specimens of cold-rolled steel had an average yield strength of 29.8 ksi. A second random sample of 25 galvanized steel specimens gave an average yield strength of 34.7 ksi. Assuming that the two yield strength distributions are normal with \( \sigma_1 = 4.0 \) and \( \sigma_2 = 5.0 \), do the data indicate that the true average yield strengths, \( \mu_1 \) and \( \mu_2 \), are different? Assume \( \alpha = 0.01 \).
2. **Comparison of Two Means, $\sigma_1$ and $\sigma_2$ Unknown**

If the populations being sampled are known to be normally distributed and the sample size is large (typically $n \geq 30$), the sample standard deviation can be substituted for unknown population standard deviations with little loss of accuracy:

\[
\text{DIFFERENCE BETWEEN MEANS (}$\sigma$\text{ UNKNOWN, } n \text{ LARGE)}
\]

\[
H_0: \quad \mu_1 - \mu_2 = \Delta_0
\]

\[
\text{TS: } z = \left( \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \right) - \Delta_0 \quad \leftarrow \text{uses } s \text{ instead of } \sigma
\]

\[
H_a: \quad \mu_1 - \mu_2 \neq \Delta_0 \quad H_a: \quad \mu_1 - \mu_2 > \Delta_0 \quad H_a: \quad \mu_1 - \mu_2 < \Delta_0
\]

\[
\text{RR: } |z| > z_{\alpha/2} \quad \text{RR: } z > z_{\alpha} \quad \text{RR: } z < -z_{\alpha}
\]

**Example**

In a sample of 30 women who did not live near a freeway, the sample average blood lead level was 9.9 and the sample standard deviation was 4.9, while a second sample of 35 females who did live near a freeway had a sample average and sample standard deviation of 16.7 and 7.0, respectively. Does proximity to heavily traveled roads result in higher blood lead levels? Test at $\alpha = 0.01$. 
3. **Comparison of Two Means, $\sigma_1$ and $\sigma_2$ Unknown but Equal (Pooled t test)**

If the populations being sampled are known to be normally distributed but the sample sizes are small (typically $n < 30$), the sample standard deviations can still be substituted for the population standard deviations, but the test statistic follows a t distribution instead of a z distribution. If the population standard deviations can be presumed to be equal to each other, you can use a pooled t test:

**DIFFERENCE BETWEEN MEANS (\(\sigma\) UNKNOWN BUT EQUAL, \(n\) SMALL)**

$$H_0: \mu_1 - \mu_2 = \Delta_0$$

$$TS: \quad t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \leftarrow \text{uses } t \text{ instead of } z$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$H_0: \mu_1 - \mu_2 \neq \Delta_0 \quad H_a: \mu_1 - \mu_2 > \Delta_0 \quad H_a: \mu_1 - \mu_2 < \Delta_0$$

RR: $|t| > t_{\alpha/2, n_1 + n_2 - 2}$  \quad RR: $t > t_{\alpha, n_1 + n_2 - 2}$  \quad RR: $t < -t_{\alpha, n_1 + n_2 - 2}$

**Example**

A random sample of 15 ceramic insulators doped in a certain manner yielded a sample average holdoff voltage of 110 kV and a sample standard deviation of 24 kV. A random sample of 76 undoped ceramic insulators produced a sample average holdoff voltage of 101 kV with a standard deviation of 22 kV. If we can assume that the actual population standard deviations should be the same, do the data suggest that the true average holdoff voltage for doped specimens exceeds that for plain specimens by more than 5 kV at a significance level of 0.10?
4. Comparison of Two Means, $\sigma_1$ and $\sigma_2$ Unknown and Unequal

If the populations being sampled are known to be normally distributed but the standard deviations are unknown and cannot be presumed to be equal to each other, you can use the Smith-Satterthwaite Procedure, which uses a $t$ distribution with the number of degrees of freedom calculated as a sort of variance-weighted average of the sample sizes:

![DIFFERENCE BETWEEN MEANS (\(\sigma\) UNKNOWN AND UNEQUAL)]

H$_0$: $\mu_1 - \mu_2 = \Delta_0$

TS: $t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{s_2^2}{n_2}\right)^2} \Leftarrow \text{round down!}$

H$_0$: $\mu_1 - \mu_2 \neq \Delta_0$

H$_a$: $\mu_1 - \mu_2 > \Delta_0$

H$_a$: $\mu_1 - \mu_2 < \Delta_0$

RR: $|t| > t_{\alpha/2,\nu}$

RR: $t > t_{\alpha,\nu}$

RR: $t < -t_{\alpha,\nu}$

Example

Dextroamphetamine is a drug commonly used to treat hyperkinetic children. A paper in the Journal of Nervous and Mental Disorders (1968, vol. 146, pp. 136-146) reported the following data on the percentage of the drug excreted within seven hours of its administration by children having organically related disorders and children with nonorganic disorders.

1. Organic: 17.53 20.60 17.62 28.93 27.10

2. Nonorganic: 15.59 14.76 13.32 12.45 12.79

The summary values are $\bar{x}_1 = 22.36$, $\bar{x}_2 = 13.78$, $s_1^2 = 28.63$, and $s_2^2 = 1.80$. The data suggest that there is much less variability in recovery percentage for children with organically related disorders. Compare the means at a significance level of 0.01.
5. Comparison of Means from Paired Samples (Paired $t$ Test)

From the chapter on confidence intervals, we know that sometimes tests are made using paired data. In those instances, there is one set of $n$ individuals or objects, and two observations (let’s say “before” and “after”) are made on each one. Rather than compare the “before” and “after” sample means, we instead compute the differences $d_i$ in the “before” and “after” test measurements for each individual in the sample set, then test the mean value of the differences, $\mu_D$:

\[
\text{DIFFERENCE BETWEEN MEANS (PAIRED SAMPLES)}
\]

\[
\begin{align*}
H_0: \quad & \mu_D = \Delta_0 \\
\text{TS:} \quad & t_{\text{paired}} = \frac{\bar{d} - \Delta_0}{s_D/\sqrt{n}} \\
\bar{d} &= \frac{\sum_{i=1}^{n}d_i}{n} \\
\sum_{i=1}^{n}(d_i - \bar{d})^2 &= \frac{\sum_{i=1}^{n}d_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}d_i\right)^2}{n-1} \\
S_D^2 &= \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1} \\
H_a: \quad & \mu_D \neq \Delta_0 \quad \quad H_a: \quad & \mu_D > \Delta_0 \quad \quad H_a: \quad & \mu_D < \Delta_0 \\
RR: \quad & \left|t_{\text{paired}}\right| > t_{\alpha/2,n-1} \quad \quad RR: \quad & t_{\text{paired}} > t_{u,n-1} \quad \quad RR: \quad & t_{\text{paired}} < -t_{u,n-1}
\end{align*}
\]

Example

In an experiment designed to evaluate an additive to increase the strength of concrete, each of five batches of concrete was divided in half and the additive added to one half of each batch. The resulting compressive strength measurements (load in kips at failure) were shown below. Does the additive work? Test at $\alpha = 0.01$.

<table>
<thead>
<tr>
<th></th>
<th>Treated</th>
<th>Untreated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16.1</td>
<td>14.8</td>
</tr>
<tr>
<td></td>
<td>14.7</td>
<td>13.2</td>
</tr>
<tr>
<td></td>
<td>17.4</td>
<td>15.5</td>
</tr>
<tr>
<td></td>
<td>13.7</td>
<td>12.3</td>
</tr>
<tr>
<td></td>
<td>16.9</td>
<td>15.9</td>
</tr>
</tbody>
</table>
P-VALUES

So far, our work in hypothesis testing simply rejects or does not reject the null hypothesis. There are two problems with this (maybe not huge ones, but we could do better). First, the conclusion does not tell us how close the test statistic was to the rejection region boundary. Did we barely reject H₀? Did we barely fail to reject H₀? Second, the level of significance is decided by the person performing the hypothesis test. People might have different opinions concerning an appropriate confidence level.

A P-value conveys much more information concerning the strength of evidence against H₀ and allows an individual to draw a conclusion at any level of α with which he/she is comfortable. The P-value is defined as the smallest level of significance at which H₀ would be rejected. Put another way, it’s the probability of incorrectly rejecting the null hypothesis. If that probability is less than the significance level, α, that you are personally comfortable with, you should be perfectly comfortable rejecting H₀. If, on the other hand, that probability is more than you are comfortable with, you would fail to reject H₀ and conclude that the strength of the evidence is insufficient to reject H₀.

Procedures exist for determining P-values for both z tests and t tests, although (as we will see shortly) P-values are not always easy to calculate. It has, fortunately, become common for statistical software to include P-values in their output.

The P-value for a z Test

\[
P = 2\left[1 - F(|z|)\right] \quad \text{for a two-tailed test } (H₀: \mu = \mu₀; H₁: \mu \neq \mu₀)
\]

\[
P = 1 - F(z) \quad \text{for an upper-tailed test } (H₀: \mu = \mu₀; H₁: \mu > \mu₀)
\]

\[
P = F(z) \quad \text{for a lower-tailed test } (H₀: \mu = \mu₀; H₁: \mu < \mu₀)
\]

Example

A certain type of brick is being considered for use in a particular construction project. The brick will be used unless sample evidence strongly suggests that the average compressive strength is below 3200 psi. A random sample of 36 bricks is selected and each is tested to failure. The sample average compressive strength is 3109 psi with a standard deviation of 156 psi. Use the P-value to test the hypothesis that the compressive strength is below 3200 psi.
The P-value for a $t$ Test

The standard normal distribution gives critical values for many values of $z$. However, for any given number of degrees of freedom, the $t$ table contains only 10 values. Therefore, the exact P-value cannot usually be determined. The procedure (unless you're using software that gives you exact P-values) is to use upper and lower bounds for the P-value.

Example

In order to test gasoline mileage performance for a new version of one of its compact cars, an automobile manufacturer selected six nonprofessional drivers to drive test cars from Phoenix to Los Angeles. At the conclusion of the trip, the resulting gas mileage numbers for the six cars were:

32.2  29.3  31.5  28.7  30.2  30.0

The manufacturer wishes to advertise that this car gets 30 mpg or better on the highway. Use P-values to determine whether or not it would be wise for the manufacturer to make this claim.

Solution

$H_0$: $\mu = 30$

$H_a$: $\mu > 30$

TS: $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{30.32 - 30}{1.32/\sqrt{6}} = 0.59$ (from an earlier example)

The $t$-table in your textbook doesn’t have a column for upper tail areas this large. An alternative approach is to use the TDIST function in Excel:

\[ \text{TDIST}(t, \nu, \text{tails}) \]

If $\text{tails} = 1$, this function returns the upper tail area of a $t$ distribution with $\nu$ degrees of freedom corresponding to the specified $t$ value. If $\text{tails} = 2$, this function returns the combined upper and lower tail areas corresponding to $\pm t$ instead.

If you enter $t = 0.59$, $\nu = 5$, and $\text{tails} = 1$, Excel returns a value of 0.29, which means the $P$-value for this hypothesis test is 0.29. Thus, the conclusion that $\mu = 30$ is only valid at a significance level of $\alpha = 0.29$ or greater. This means that there is a 29% chance that the conclusion is wrong! This would be unacceptable in the real world.