• **Statistical Inference:**

Recall from chapter 5 that statistical inference is the use of a subset of a population (the sample) to draw conclusions about the entire population. In chapter 5 we studied one kind of inference called estimation. In this chapter, we study a second kind of inference called hypothesis testing.

The validity of inference is related to the way the data are obtained, and to the stationarity of the process producing the data.

• **The Components of a Statistical Hypothesis Testing Problem**

  1. **The Scientific Hypothesis**
  2. **The Statistical Model**
  3. **The Statistical Hypotheses**
  4. **The Test Statistic**
  5. **The P-Value**

• **Example:**

One stage of a manufacturing process involves a manually-controlled grinding operation. Management suspects that the grinding machine operators tend to grind parts slightly larger rather than slightly smaller than the target diameter of 0.75 inches while still staying within specification limits, which are $0.75 \pm 0.01$ inches. To verify their suspicions, they sample 150 within-spec parts. We will use this example to illustrate the components of a statistical hypothesis testing problem.

  1. **The Scientific Hypothesis** The scientific hypothesis is the hypothesized outcome of the experiment or study. In this example, the scientific hypothesis is that there is a tendency to grind the parts larger than the target diameter.

  2. **The Statistical Model** We will assume these data were generated by the C+E model:

$$Y = \mu + \epsilon,$$

where the random error, $\epsilon$, follows a $N(0, \sigma^2)$ distribution model.

  3. **The Statistical Hypotheses** In terms of the C+E model, management defined “a tendency to grind the parts larger than the target diameter” to be a statement about the population mean diameter, $\mu$, of the ground parts. They then defined the statistical hypotheses to be

$$H_0: \mu = 0.75$$
$$H_a: \mu > 0.75$$

Notice that $H_a$ states the scientific hypothesis.

  4. **The Test Statistic** In all one-parameter hypothesis test settings we will consider, the test statistic will be the estimator of the population parameter about which inference is being made. As you know from chapter 5, the estimator of $\mu$ is the sample mean, $\bar{Y}$, and this is also the test statistic. The observed value of $\bar{Y}$ for these data is $\bar{Y} = 0.7518$. 

1
5. **The P-Value** Think of this as the *plausibility value*. It measures the probability, given that \( H_0 \) is true, that a randomly chosen value of the test statistic will give as much or more evidence against \( H_0 \) and in favor of \( H_a \) as does the observed test statistic value.

For the grinding problem, since \( H_a \) states that \( \mu > 0.75 \), large values of \( \overline{Y} \) will provide evidence against \( H_0 \) and in favor of \( H_a \). Therefore any value of \( \overline{Y} \) as large or larger than the observed value \( \overline{Y} = 0.7518 \) will provide as much or more evidence against \( H_0 \) and in favor of \( H_a \) as does the observed test statistic value. Thus, the *p*-value is \( P_0(\overline{Y} \geq 0.7518) \), where \( P_0 \) is the probability computed under the assumption that \( H_0 \) is true: that is, \( \mu = 0.75 \).

To calculate the *p*-value, we standardize the test statistic by subtracting its mean (remember we’re assuming \( H_0 \) is true, so we take \( \mu = 0.75 \)) and dividing by its estimated standard error:

\[
\hat{\sigma}(\overline{Y}) = \frac{s}{\sqrt{n}} = \frac{0.0048}{\sqrt{150}} = 0.0004.
\]

If \( H_0 \) is true, the result will have a \( t_{n-1} = t_{149} \) distribution.

Putting this all together, the *p*-value is

\[
P_0(\overline{Y} \geq 0.7518) = P_0 \left( \frac{\overline{Y} - 0.75}{0.0004} \geq \frac{0.7518 - 0.75}{0.0004} \right) = P(t_{149} \geq 4.5) = 6.8 \times 10^{-6}.
\]

### Two-Sided Tests

In all examples we’ll look at, \( H_0 \) will be ***simple*** (i.e. will state that the parameter has a single value.) as opposed to ***compound***. Alternative hypotheses will be ***one-sided*** (that the parameter be larger the null value, or smaller than the null value) or ***two-sided*** (that the parameter not equal the null value).

In the grinding example, we had

\[
H_0 : \mu = 0.75 \text{ (simple)} \\
H_a : \mu > 0.75 \text{ (compound, one-sided)}
\]

Suppose in the grinding problem that management wanted to see if the mean diameter was off target. Then appropriate hypotheses would be:

\[
H_0 : \mu = 0.75 \text{ (simple)} \\
H_a : \mu \neq 0.75 \text{ (compound, two-sided)}
\]

In this case, evidence against \( H_0 \) and in favor of \( H_a \) is provided by both large and small values of \( \overline{Y} \).

To compute the *p*-value of the two-sided test, we first compute the standardized test statistic \( t \), and its observed value, \( t^* \):

\[
t = \frac{\overline{Y} - 0.75}{0.0004}, \quad t^* = \frac{0.7518 - 0.75}{0.0004} = 4.5.
\]

Recall that under \( H_0 \), \( t \sim t_{149} \). By the symmetry of the \( t \) distribution about 0, we compute the *p*-value as \( P(|t| \geq |t^*|) = P(|t| \geq 4.5) = 13.6 \times 10^{-6} \).
• **The Philosophy of Hypothesis Testing**

  Statistical hypothesis testing is modeled on scientific investigation. The two hypotheses represent competing scientific hypotheses.

  – The **alternative hypothesis** is the hypothesis that suggests change, difference or an aspect of a new theory.

  – The **null hypothesis** is the hypothesis that represents the accepted scientific view or that, most often, suggests no difference or effect.

  For this reason the null hypothesis is given favored treatment.

• **Other Issues**

  – Statistical significance

  – Cautions

    o Statistical vs. practical significance

    o Exploratory vs. confirmatory

    o Lotsa tests means false positives

    o Data suggesting hypotheses

    o Lack of significance \( \neq \) failure

• **One Sample Hypothesis Tests for the Mean in the C+E Model**

  Check out Appendix 6.1, p. 346, with me!

• **One Sample Hypothesis Tests for a Population Proportion**

  First, check out Appendix 6.1, p. 347, with me!

  **Example:**

  Back at the grinding operation, management has decided on another characterization of the scientific hypothesis that “there is a tendency to grind the parts larger than the target diameter.” They decide to make inference about \( p \), the population proportion of in-spec parts with diameters larger than the target value. The hypotheses are

  \[
  H_0: \quad p = 0.5 \\
  H_a: \quad p > 0.5
  \]

  The datum is \( Y \), the number of the 150 sampled parts with diameters larger than the target value. Of the 150 parts, \( y^* = 93 \) (a proportion 0.62) have diameters greater than the target value 0.75.

  We will first perform an exact test of these hypotheses. Under \( H_0 \), \( Y \sim b(150, 0.5) \), so the \( p \)-value is

  \[
  p^+ = P(b(150, 0.5) \geq 93) = 0.0021.
  \]
Now, for illustration, we will use the large-sample test. This is valid since \( np_0 \) and \( n(1 - p_0) \) both equal \( 75 > 10 \).

The observed continuity-corrected standardized test statistic is

\[
z^* = \frac{93 - (0.5)(150) - 0.5}{\sqrt{(150)(0.5)(1 - 0.5)}} = 2.858.
\]

The approximate \( p \)-value is then

\[
P(N(0, 1) \geq 2.858) = 0.0021.
\]

- **The Two Population C+E Model**

  We assume that there are \( n_1 \) measurements from population 1 generated by the C+E model

  \[
  Y_{1,i} = \mu_1 + \epsilon_{1,i}, \ i = 1, \ldots, n_1,
  \]

  and \( n_2 \) measurements from population 2 generated by the C+E model

  \[
  Y_{2,i} = \mu_2 + \epsilon_{2,i}, \ i = 1, \ldots, n_2.
  \]

  We want to compare \( \mu_1 \) and \( \mu_2 \).

- **Hypothesis Test for Paired Comparisons**

  Sometimes each observation from population 1 is paired with another observation from population 2. For example, each student may take a pre- and post-test. In this case \( n_1 = n_2 \) and by looking at the pairwise differences, \( D_i = Y_{1,i} - Y_{2,i} \), we transform the two population problem to a one population problem for C+E model \( D = \mu_D + \epsilon_D \), where \( \mu_D = \mu_1 - \mu_2 \) and \( \epsilon_D = \epsilon_1 - \epsilon_2 \). Therefore, an hypothesis test for the difference \( \mu_1 - \mu_2 \) is obtained by performing a one sample hypothesis test for \( \mu_D \) based on the differences \( D_i \).

- **Example:**

  The manufacturer of a new warmup bat wants to test its efficacy. To do so, it selects a random sample of 12 baseball players from among a larger number who volunteer to try the bat. For each player, company researchers compute \( D_i \) the difference between the player’s test year average and his previous year’s average. Assuming that these differences follow a C+E model, they want to test

  \[
  H_0 : \ \mu_D = 0 \\
  H_a : \ \mu_D > 0
  \]

  The data (found in SASDATA.BATTING) are:
<table>
<thead>
<tr>
<th>PLAYER</th>
<th>AVG92</th>
<th>AVG93</th>
<th>DIFFAVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.254</td>
<td>0.262</td>
<td>0.008</td>
</tr>
<tr>
<td>2</td>
<td>0.274</td>
<td>0.290</td>
<td>0.016</td>
</tr>
<tr>
<td>3</td>
<td>0.300</td>
<td>0.304</td>
<td>0.004</td>
</tr>
<tr>
<td>4</td>
<td>0.246</td>
<td>0.267</td>
<td>0.021</td>
</tr>
<tr>
<td>5</td>
<td>0.278</td>
<td>0.291</td>
<td>0.013</td>
</tr>
<tr>
<td>6</td>
<td>0.252</td>
<td>0.257</td>
<td>0.005</td>
</tr>
<tr>
<td>7</td>
<td>0.235</td>
<td>0.248</td>
<td>0.013</td>
</tr>
<tr>
<td>8</td>
<td>0.313</td>
<td>0.324</td>
<td>0.021</td>
</tr>
<tr>
<td>9</td>
<td>0.305</td>
<td>0.317</td>
<td>0.012</td>
</tr>
<tr>
<td>10</td>
<td>0.255</td>
<td>0.252</td>
<td>−0.003</td>
</tr>
<tr>
<td>11</td>
<td>0.244</td>
<td>0.276</td>
<td>0.032</td>
</tr>
<tr>
<td>12</td>
<td>0.322</td>
<td>0.332</td>
<td>0.010</td>
</tr>
</tbody>
</table>

An inspection of the differences shows no evidence of nonnormality or outliers, so we proceed with the test. For these data, $\bar{d} = 0.0127$, and $s_d = 0.0092$. Then $\hat{\sigma}(\bar{D}) = 0.0092/\sqrt{12} = 0.0027$, so the observed value of the standardized test statistic is

$$t^* = \frac{0.0127}{0.0027} = 4.70,$$

resulting in a $p$-value

$$P(t_{11} \geq 4.7) = 0.006.$$

**Testing Differences in Population Means of Independent Populations**

Let $\bar{Y}_1$ and $\bar{Y}_2$ denote the sample means from populations 1 and 2, $S_1^2$ and $S_2^2$ the sample variances. The point estimator of $\mu_1 - \mu_2$, is $\bar{Y}_1 - \bar{Y}_2$. We will test

$$H_0 : \mu_1 - \mu_2 = \delta_0$$

Versus one of

- $H_{a-} : \mu_1 - \mu_2 < \delta_0,$
- $H_{a+} : \mu_1 - \mu_2 < \delta_0,$
- $H_{a\pm} : \mu_1 - \mu_2 \neq \delta_0.$

**Equal Variances**

If the population variances are equal ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), then we estimate $\sigma^2$ by the pooled variance estimator

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

The estimated standard error of $\bar{Y}_1 - \bar{Y}_2$ is then given by

$$\hat{\sigma}_p(\bar{Y}_1 - \bar{Y}_2) = \sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}.$$
Then, if \( H_0 \) is true,
\[
t^{(p)} = \frac{\bar{Y}_1 - \bar{Y}_2 - \delta_0}{\hat{\sigma}_p(Y_1 - Y_2)}
\]
has a \( t_{n_1+n_2-2} \) distribution.

Suppose \( t^{(p)*} \) is the observed value of \( t^{(p)} \). Then the \( p \)-value of the test of \( H_0 \) versus \( H_{a-} \) is
\[
p_- = P(t_{n_1+n_2-2} \leq t^{(p)*}),
\]
versus \( H_{a+} \) is
\[
p^+ = P(t_{n_1+n_2-2} \geq t^{(p)*}),
\]
and versus \( H_{a\pm} \) is
\[
p^\pm = 2 \min(p_-, p^+).
\]

- **Unequal Variances**

If \( \sigma_1^2 \neq \sigma_2^2 \), then the standardized test statistic
\[
t^{(ap)} = \frac{\bar{Y}_1 - \bar{Y}_2 - \delta_0}{\hat{\sigma}(Y_1 - Y_2)}
\]
approximately follows a \( t_\nu \) distribution model, where \( \nu \) is the largest integer less than or equal to
\[
\frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1-1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2-1}},
\]
and
\[
\hat{\sigma}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{\bar{S}_1^2}{n_1} + \frac{S_2^2}{n_2}}.
\]

If \( t^{(ap)*} \) denotes the observed value of \( t^{(ap)} \), the \( p \)-values for \( H_0 \) versus \( H_{a-} \), \( H_{a+} \) and \( H_{a\pm} \), respectively, are
\[
p_- = P(t_\nu \leq t^{(ap)*}), \quad p^+ = P(t_\nu \geq t^{(ap)*}) \quad \text{and} \quad p^\pm = 2 \min(p_-, p^+).
\]

**Example:**

A company buys cutting blades used in its manufacturing process from two suppliers. In order to
decide if there is a difference in blade life, the lifetimes of 10 blades from manufacturer 1 and 13
blades from manufacturer 2 used in the same application are compared. A summary of the data
shows the following (units are hours): (The data are in SASDATA.BLADE2)

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>( n )</th>
<th>( \bar{y} )</th>
<th>( s )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>118.4</td>
<td>26.9</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>134.9</td>
<td>18.4</td>
</tr>
</tbody>
</table>

The experimenters want to test
\[
H_0 : \mu_1 - \mu_2 = 0
\]
\[
H_a : \mu_1 - \mu_2 \neq 0
\]
The experimenters generated histograms and normal quantile plots of the two data sets and found no evidence of nonnormality or outliers. The estimate of \( \mu_1 - \mu_2 \) is \( \bar{y}_1 - \bar{y}_2 = 118.4 - 134.9 = -16.52 \).

- **Pooled variance test** The pooled variance estimate is

\[
s_p^2 = \frac{(10 - 1)(26.9)^2 + (13 - 1)(18.4)^2}{10 + 13 - 2} = 503.6,
\]

So the standard error estimate of \( \bar{Y}_1 - \bar{Y}_2 \) is

\[
\hat{\sigma}_p(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{503.6}{10}} \left( \frac{1}{10} + \frac{1}{13} \right) = 9.44.
\]

Therefore, \( t^{(p)*} = -16.52/9.44 = -1.75 \), with 21 degrees of freedom. So \( p_- = P(t_{21} \leq -1.75) = 0.0473 \), \( p^+ = P(t_{21} \geq -1.75) = 0.9527 \), and the \( p \)-value for this problem is \( 2 \min(0.0473, 0.9527) = 0.0946 \).

- **Separate variance test** The standard error estimate of \( \bar{Y}_1 - \bar{Y}_2 \) is

\[
\hat{\sigma}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{(26.9)^2}{10} + \frac{(18.4)^2}{13}} = 9.92.
\]

The observed value of the standardized test statistic is \( t^{(p)*} = -16.52/9.92 = -1.67 \). The degrees of freedom \( \nu \) is computed as the greatest integer less than or equal to

\[
\frac{\left( \frac{(26.9)^2}{10} \right)^2 + \left( \frac{(18.4)^2}{13} \right)^2}{(26.9)^2/10 - 1} + \frac{\left( \frac{(18.4)^2}{13} \right)^2}{13 - 1} = 15.17,
\]

so \( \nu = 15 \).

Therefore, \( p_- = P(t_{15} \leq -1.67) = 0.0583 \), \( p^+ = P(t_{15} \geq -1.67) = 0.9417 \), and the \( p \)-value for this problem is \( 2 \min(0.0583, 0.9417) = 0.1166 \).

The results for the two \( t \)-tests are not much different.

**Comparing Two Population Proportions**

\( Y_1 \sim b(n_1, p_1) \) and \( Y_2 \sim b(n_2, p_2) \) are observations from two independent populations. The estimator of \( p_1 - p_2 \) is

\[
\hat{p}_1 - \hat{p}_2 = \frac{Y_1}{n_1} - \frac{Y_2}{n_2}.
\]

We wish to test a null hypothesis that the two population proportions differ by a known amount \( \delta_0 \),

\[
H_0 : p_1 - p_2 = \delta_0,
\]

against one of three possible alternative hypotheses:
Case 1: $\delta_0 = 0$

Suppose $H_0$ is $p_1 - p_2 = 0$. Then, let $p = p_1 = p_2$ denote the common value of the two population proportions. If $H_0$ is true, the variance of $\hat{p}_1$ equals $p(1 - p)/n_1$ and that of $\hat{p}_2$ equals $p(1 - p)/n_2$. This implies the standard error of $\hat{p}_1 - \hat{p}_2$ equals

$$\sqrt{\frac{p(1 - p)}{n_1} + \frac{p(1 - p)}{n_2}}.$$ 

Since we don’t know $p$, we estimate it using the data from both populations:

$$\hat{p} = \frac{Y_1 + Y_2}{n_1 + n_2}.$$ 

The estimated standard error of $\hat{p}_1 - \hat{p}_2$ is then

$$\hat{\sigma}_0(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n_1} + \frac{\hat{p}(1 - \hat{p})}{n_2}} = \sqrt{\hat{p}(1 - \hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}.$$ 

The standardized test statistic is then

$$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\hat{\sigma}_0(\hat{p}_1 - \hat{p}_2)}.$$ 

which has approximately a $N(0, 1)$ distribution if $H_0$ is true.

Case 2: $\delta_0 \neq 0$

If $\delta_0 \neq 0$, the (by now) standard reasoning gives the standardized test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\hat{\sigma}(\hat{p}_1 - \hat{p}_2)},$$ 

where

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

is the estimated standard error of $\hat{p}_1 - \hat{p}_2$.

**Example:**

In a recent survey on academic dishonesty 24 of the 200 female college students surveyed and 26 of the 100 male college students surveyed agreed or strongly agreed with the statement “Under some circumstances academic dishonesty is justified.” Suppose $p_f$ denotes the proportion of all female and $p_m$ the proportion of all male college students who agree or strongly agree with this statement.

o Suppose we want to Test
\[
H_0: \ p_f - p_m = 0 \\
H_a: \ p_f - p_m \neq 0
\]

Then since \(Y_f = 24, 200 - Y_f = 176, Y_m = 26, \) and \(100 - Y_m = 74\) all exceed 10, we may use the normal approximation.

The point estimate of \(p_f - p_m\) is

\[
\hat{p}_f - \hat{p}_m = \frac{24}{200} - \frac{26}{100} = -0.140,
\]

and the estimate of the common value of \(p_f\) and \(p_m\) under \(H_0\) is \(\hat{p} = (26 + 24)/(200 + 100) = 0.167\).

Thus,

\[
\hat{\sigma}_0(\hat{p}_f - \hat{p}_m) = \sqrt{(0.167)(0.833)\left(\frac{1}{200} + \frac{1}{100}\right)} = 0.046,
\]

and

\[
Z_0 = \frac{-0.140}{0.046} = 3.04.
\]

From this, we obtain \(p^+ = P(N(0, 1) \geq 3.04) = 0.0012, p_- = P(N(0, 1) \leq 3.04) = 0.9988,\) and \(p_\pm = 2\min(0.0012, 0.9988) = 0.0024,\) this last being the \(p\)-value we want.

\begin{itemize}
  \item Suppose we want to test
  \[
  H_0: \ p_f - p_m = -0.10 \\
  H_a: \ p_f - p_m < -0.10
  \]

  The estimated standard error of \(p_f - p_m\) is

  \[
  \hat{\sigma}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{0.12(1 - 0.12)}{200} + \frac{0.26(1 - 0.26)}{100}} = 0.05,
  \]

  which gives

  \[
  Z = \frac{24/200 - 26/100 - (-0.10)}{0.05} = -0.80,
  \]

  and a \(p\)-value of \(P(N(0, 1) \leq -0.80) = 0.2119.\)
\end{itemize}

**Fixed Significance Level Tests**

Steps, illustrated using grinding example:

1. **Specify hypotheses to be tested.**

   \[
   H_0: \ \mu = 0.75 \\
   H_a: \ \mu > 0.75
   \]

   (i.e. \(\mu_0 = 0.75\))

2. **Set the significance level \(\alpha\).** Usual choices are 0.01 or 0.05. We'll choose the latter.
3. **Specify the (standardized) test statistic and it’s distribution under H₀.** For simplicity, assume we know \( \sigma = 0.048 \). Then the standardized test statistic is

\[
Z = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{Y} - 0.75}{0.0048/\sqrt{150}},
\]

and under \( H₀ \) it has a \( N(0, 1) \) distribution.

4. **Find the critical region of the test.** The critical region of the test is the set of values of the (standardized) test statistic for which \( H₀ \) will be rejected in favor of \( Hₐ \). Here, \( Hₐ \) tells us that the critical region has the form

\[
[zₐ, \infty) = [z_{0.05}, \infty) = [1.645, \infty),
\]

meaning \( H₀ \) will be rejected if and only if the observed value of \( Z \) is greater than or equal to 1.645.

5. **Perform the test.** The observed value of \( Z \) is

\[
z^* = \frac{0.7518 - 0.75}{0.0048/\sqrt{150}} = 4.5,
\]

which falls in the critical region, so \( H₀ \) is rejected in favor of \( Hₐ \).

- **Power** In a fixed significance level test, power is the probability of rejecting \( H₀ \) in favor of \( Hₐ \). Power will vary for different values of the parameter being tested, so it is written as a function of that parameter.

In the grinding example, the power is

\[
\Pi(\mu) = P(Z \geq 1.645|\mu) = P(\bar{Y} \geq 0.75 + (1.645)\frac{0.0048}{\sqrt{150}}|\mu) = P(Z' \geq 1.645 + \frac{0.75 - \mu}{0.0048/\sqrt{150}}),
\]

where \( Z' = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \).

- **The relation between hypothesis tests and confidence intervals**