Statistical Inference: Recall from chapter 6 that statistical inference is the use of a subset of a population (the sample) to draw conclusions about the entire population. In chapter 6 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 stability 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:

Recall the example from Chapters 4 and 6:

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, 0.75 inches while still staying within specification limits, which are 0.75 ± 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 measurement model:

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

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

3. **The Statistical Hypotheses** In terms of the measurement model, management defined "a tendency to grind the parts larger than the target diameter" to be a statement about the population mean diameter, μ, of the ground parts. They then defined the statistical hypotheses to be

$$\begin{aligned} H_0: &\quad \mu = 0.75 \\ H_a: &\quad \mu > 0.75 \end{aligned}$$

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 6, the estimator of μ is the sample mean, $\overline{Y}$, and this is also the test statistic. The observed value of $\overline{Y}$ for these data is $\overline{y}^* = 0.7518$.

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:

$$\begin{aligned}\widehat{\sigma}(\overline{Y}) &= s/\sqrt{n} \\ &= 0.0048/\sqrt{150} \\ &= 0.0004.\end{aligned}$$

If H_0 is true, the result will have a $t_{n-1} = t_{149}$ distribution. Putting this all together, the p-value is

$$\begin{aligned}P_0(\overline{Y} \geq 0.7518) &= P_0\left(\frac{\overline{Y} - 0.75}{0.0004} \geq \frac{0.7518 - 0.75}{0.0004}\right) \\ &= P(t_{149} \geq 4.5) \\ &= 6.8 \times 10^{-6}.\end{aligned}$$

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**. Alternate 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

$$\begin{aligned} H_0: &\quad \mu = 0.75 \text{ (simple)} \\ H_a: &\quad \mu > 0.75 \text{ (compound, one-sided)} \end{aligned}$$

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

$$\begin{array}{lrcl} H_0: & \mu & = & 0.75 \text{ (simple)} \\ H_a: & \mu & \neq & 0.75 \text{ (compound, two-sided)} \end{array}$$

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^*:

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

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 **alternate 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
  - Statistical vs. practical significance
  - Exploratory vs. confirmatory
  - Lotsa tests means false positives
  - Data suggesting hypotheses
  - Lack of significance $\neq$ failure

One Sample Hypothesis Tests for the Mean in the Measurement Model

Check out the appendix with me!

One Sample Hypothesis Tests for a Population Proportion

First, check out the appendix with me!

Now, let's do an

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

$$\begin{aligned} H_0: &\quad p &= &\quad 0.5 \\ H_a: &\quad p &> &\quad 0.5 \end{aligned}$$

The datum is Y, the number of the 150 sampled parts with diameters larger than the target value.

Of the 150 parts, $y^* = 93$ of 150 (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 standardized test statistic is

$$z^* = \frac{93 - (0.5)(150)}{\sqrt{(150)(0.5)(1-0.5)}} = 2.94.$$

The approximate p-value is then

$$P(N(0,1) \geq 2.94) = 0.0016.$$

The Two Population Measurement Model

We assume that there are n_1 measurements from population 1 generated by the measurement model

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

and n_2 measurements from population 2 generated by the measurement model

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

We want to compare μ_1 and μ_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 measurement 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 μ_D based on the differences D_i.

Example:

In 1993 the National League expanded by adding the Florida and Colorado teams. Many experts predicted that this expansion would dilute the quality of pitching and inflate team batting statistics. Others pointed out that the batting level would also decline, and that the result would be little or no difference. To assess who was right, we have collected the team batting averages for 1992 and 1993 for all 12 teams that were in the league in 1992. We assume that each team's batting average each year follows a measurement model centered about an overall (and unknown) league average.

Since most personnel on a team stay the same from one year to the next, we feel that paired comparisons are appropriate. Thus, we compute the differences in 1993 and 1992 averages for each team.

Thus, for each team, we will compute D, the difference between the 1993 and 1992 team batting average. We will test the hypotheses

$$\begin{aligned} H_0: &\quad \mu_D = 0 \\ H_a: &\quad \mu_D > 0 \end{aligned}$$

The data (found in SASDATA.NLAVG923) are:

TEAM	AVG92	AVG93	DIFFAVG
ATL	0.254	0.262	0.008
CHI	0.254	0.270	0.016
CIN	0.260	0.264	0.004
HOU	0.246	0.267	0.021
LA	0.248	0.261	0.013
MON	0.252	0.257	0.005
NY	0.235	0.248	0.013
PHI	0.253	0.274	0.021
PIT	0.255	0.267	0.012
SD	0.255	0.252	−0.003
SF	0.244	0.276	0.032
STL	0.262	0.272	0.010

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

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

resulting in a p-value

$$P(t_{11} \geq 4.7) = 0.0006.$$

Testing Differences in Population Means of Independent Populations

Let $\overline{Y}_1$ and $\overline{Y}_2$ 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 $\overline{Y}_1 - \overline{Y}_2$. We will test

$$\begin{array}{lrcl} H_0: & \mu_1 - \mu_2 & = & \delta_0 \\ \text{Versus one of} & & & \\ H_{a-}: & \mu_1 - \mu_2 & < & \delta_0, \\ H_{a+}: & \mu_1 - \mu_2 & < & \delta_0, \\ H_{a\pm}: & \mu_1 - \mu_2 & \neq & \delta_0. \end{array}$$

Equal Variances

If the population variances are equal ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), then we estimate σ^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 $\overline{Y}_1 - \overline{Y}_2$ is then given by

$$\widehat{\sigma}_p(\overline{Y}_1 - \overline{Y}_2) = \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Then, if H_0 is true,

$$t^{(p)} = \frac{\overline{Y}_1 - \overline{Y}_2 - \delta_0}{\widehat{\sigma}_p(\overline{Y}_1 - \overline{Y}_2)}$$

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

$$p_- = P(t_{n_1+n_2-2} < t^{(p)*}),$$

versus H_{a^+} is

$$p^+ = P(t_{n_1+n_2-2} > t^{(p)*}),$$

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

$$t^{(ap)} = \frac{\overline{Y}_1 - \overline{Y}_2 - \delta_0}{\widehat{\sigma}(\overline{Y}_1 - \overline{Y}_2)}.$$

approximately follows a t_ν distribution model, where ν is the largest integer less than or equal to

$$\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}},$$

and

$$\widehat{\sigma}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

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 $p_- = P(t_\nu \leq t^{(ap)*})$, $p^+ = P(t_\nu \geq t^{(ap)*})$ and $p\pm = 2\min(p_-, p^+)$.

Example:

A company buys grinding wheels used in its manufacturing process from two suppliers. In order to decide if there is a difference in wheel life, the lifetimes of 10 wheels from manufacturer 1 and 13 wheels 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.GRIND2)

Manufacturer	n	$\overline{y}$	s
1	10	118.4	26.9
2	13	134.9	18.4

Test

$$\begin{aligned} H_0: &\quad \mu_1 - \mu_2 = 0 \\ H_a: &\quad \mu_1 - \mu_2 \neq 0 \end{aligned}$$

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 $\overline{y}_1 - \overline{y}_2 = 118.4 - 134.9 = -16.52$.

- Pooled variance test The pooled variance estimate is

$$s_p^2 = \frac{(10-1)(26.9)^2 + (13-1)(18.4)^2}{10+13-2} = 503.6.$$

This gives the standard error estimate of $\overline{Y}_1 - \overline{Y}_2$ as

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

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 $\overline{Y}_1 - \overline{Y}_2$ is

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

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

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

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

$$\widehat{p}_1 - \widehat{p}_2 = \frac{Y_1}{n_1} - \frac{Y_2}{n_2}.$$

We wish to test a null hypothesis that the two population proportions differ by a known amount δ_0,

$$H_0: \quad p_1 - p_2 \quad = \quad \delta_0,$$

against one of three possible alternate hypotheses:

$$\begin{array}{lrcl} H_{a+}: & p_1 - p_2 & > & \delta_0 \\ H_{a-}: & p_1 - p_2 & < & \delta_0 \\ H_{a\pm}: & p_1 - p_2 & \neq & \delta_0 \end{array}$$

Case 1: $\boldsymbol{\delta_0 = 0}$

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 $\widehat{p}_1$ equals $p(1-p)/n_1$ and that of $\widehat{p}_2$ equals $p(1-p)/n_2$. This implies the standard error of $\widehat{p}_1 - \widehat{p}_2$ 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:

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

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

$$\begin{aligned} \widehat{\sigma}_0(\widehat{p}_1 - \widehat{p}_2) &= \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n_1} + \frac{\widehat{p}(1-\widehat{p})}{n_2}} \\ &= \sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \end{aligned}$$

The standardized test statistic is then

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

which has 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.

- Test

$$\begin{aligned} H_0: \quad p_f - p_m &= 0 \\ H_a: \quad p_f - p_m &\neq 0 \end{aligned}$$

  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

$$\hat{p}_f - \hat{p}_m = 24/200 - 26/100 = -0.140,$$

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

$$\begin{aligned} \widehat{\sigma}_0(\widehat{p}_f - \widehat{p}_m) &= \sqrt{(0.167)(0.833)\left(\frac{1}{200} + \frac{1}{100}\right)} \\ &= 0.046, \end{aligned}$$

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.

- Test

$$\begin{array}{rrcl} H_0: & p_f - p_m & = & -0.10 \\ H_a: & p_f - p_m & < & -0.10 \end{array}$$

The estimated standard error of $p_f - p_m$ is

$$\begin{aligned} \widehat{\sigma}(\widehat{p}_1 - \widehat{p}_2) &= \sqrt{\frac{0.12(1-0.12)}{200} + \frac{0.26(1-0.26)}{100}} \\ &= 0.05, \end{aligned}$$

which gives

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

and a p-value of $P(N(0,1) \leq -0.80) = 0.2119$.

Other Topics

- Fixed significance level tests
- Power
- The relation between hypothesis tests and confidence intervals