- # 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 ± 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, ϵ, 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, μ, of the ground parts. They then defined the statistical hypotheses to be
     $$\begin{array}{lccc} H_0: & \mu & = & 0.75 \\ H_a: & \mu & > & 0.75 \end{array}$$
     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 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:

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

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

Putting this all together, the p-value is

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

- **What's the Conclusion?**

At this point, we have two options:

- Reject H_0 in favor of H_a.
- Do not reject H_0 in favor of H_a.

If the p-value is small enough, it indicates that, relative to H_a, the data are not consistent with the assumption that H_0 is true, so our action would be to reject H_0 in favor of H_a.

How small is "small enough" to reject H_0 in favor of H_a? That depends on a number of factors, such as the type of study, the purposes of the study, and the number of hypothesis tests being conducted. Table 1 gives guidelines for a single hypothesis test.

If the p-value is less than:	*The evidence against H_0 and in favor of H_a is:*
0.100	borderline
0.050	reasonably strong
0.025	strong
0.010	very strong

Table 1: *Interpreting the strength of evidence against H_0 and in favor of H_a provided by p-values*

- # 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

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

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

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

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, $t \sim t_{149}$.

Because the test is two-sided, we compute the p-value as $P(|t| \geq |t^*|) = P(t \leq -|t^*|) + P(t \geq |t^*|)$. By the symmetry of the t distribution about 0,

this equals $2P(t \geq |t^*|)$. For the present example, the p-value is $P(|t| \geq 4.5) = 2P(t \geq 4.5) = 13.6 \times 10^{-6}$.

Here's an easier way to compute the p-value for the two-sided test:

Let $p^+ = P(t \geq t^*)$, and let $p_- = P(t \leq t^*) = 1 - p^+$. Then the p-value for the two-sided test is $p\pm = 2 \times min\{p^+, p_-\}$. In our example, $p^+ = P(t \geq 4.5) = 6.8 \times 10^{-6}$, $p_- = P(t \leq 4.5) = 1 - p^+ = 0.9999932$, so $p\pm = 2 \times min\{6.8 \times 10^{-6}, 0.9999932\} = 13.6 \times 10^{-6}$.

- # 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

    Often, prior to conducting the study, users of hypothesis tests set a pre-specified threshold level of evidence against the null and in favor of the alternative hypothesis. In order to reject H_0 in favor of H_a, a p-value must fall below this threshold. The name given to this threshold is "signifance level", and it is often denoted α.

    If, for example, we decide to use a significance level of $\alpha = 0.05$, our action would be to reject H_0 in favor of H_a if the p-value is less than 0.05, and to not reject otherwise.

  - Statistical significance and sample size

    Statistical significance measures our ability to detect a difference. As such, it is at least partly based on the amount of data we have. For instance, recall the grinding example. There were 150 parts having mean diameter 0.7518 and standard deviation .0048. To test

    $$\begin{array}{lccc} H_0: & \mu & = & 0.75 \\ H_a: & \mu & \neq & 0.75 \end{array}$$

    we computed the p-value as

    $$P_0(\overline{Y} \geq 0.7518) =$$
    $$P_0\left(\frac{\sqrt{150}(\overline{Y} - 0.75)}{0.0048} \geq \frac{\sqrt{150}(0.7518 - 0.75)}{0.0048}\right)$$

$$= P(t_{149} \geq 4.5) = 6.8 \times 10^{-6}$$

Now suppose that we had samples of sizes 10 and 50 with the same mean and standard deviations. The corresponding p-values are:

$$P_0(\overline{Y} \geq 0.7518) =$$
$$P_0\left(\frac{\sqrt{10}(\overline{Y} - 0.75)}{0.0048} \geq \frac{\sqrt{10}(0.7518 - 0.75)}{0.0048}\right)$$
$$= P(t_9 \geq 1.2) = 0.1330$$

and

$$P_0(\overline{Y} \geq 0.7518) =$$
$$P_0\left(\frac{\sqrt{50}(\overline{Y} - 0.75)}{0.0048} \geq \frac{\sqrt{50}(0.7518 - 0.75)}{0.0048}\right)$$
$$= P(t_{49} \geq 2.7) = 0.0054$$

- Statistical vs. practical significance

  Statistical significance is used to decide if there is a difference. It says nothing about practical significance: whether that difference is important or not. In the example, we found that a mean of 0.7518 inches for the 150 sampled parts provided strong evidence that the population mean diameter was larger than the target of 0.75. This result says nothing about whether a difference on the order of 0.0018 inches makes any real difference in product performance, manufacturing cost, etc.

- Other Cautions
  - Data suggesting hypotheses (Exploratory vs. confirmatory studies)
  - Lotsa tests means false positives
  - 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:**

  Here's an example of how to do a two-sided exact test. A manufacturer of high fiber cereal claims that its product Fibermax is recommended by 2 out of 3 nutritionists. In a small (but well-conducted) survey, 3 of a random sample of 6 nutritionists recommended Fibermax.

  - **The Scientific Hypothesis** Fibermax is not recommended by 2 out of 3 nutritionists.
  - **The Statistical Model** Y, the number of the 6 nutritionists surveyed who recommend Fibermax has a $b(6, p)$ distribution. (Here p is the proportion of all nutritionists who recommend Fibermax).
  - **The Statistical Hypotheses**

$$\begin{array}{lccl} H_0: & p & = & 0.667 \\ H_a: & p & \neq & 0.667 \end{array}$$

  - **The Test Statistic** Y
  - **The P-Value** This is 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, $y^* = 3$.

Under H_0, Y, the number of a sample of 6 who recommend Fibermax, has a $b(6, 0.667)$ distribution, so its pmf is

$$p_Y(y) = \left(\begin{array}{c} 6 \\ y \end{array} \right) 0.667^y (1 - 0.667)^{6-y}, \ y = 0, 1, \ldots, 6.$$

Evaluating, we find the pmf:

y	$p_Y(y)$	y	$p_Y(y)$
0	0.001364	4	0.329218
1	0.016387	5	0.263770
2	0.082058	6	0.088055
3	0.219149		

The observed value of Y is $y^* = 3$. The p value is the sum of all $p_Y(y)$ values that are less than or equal to $p_Y(y^*) = p_Y(3) = 0.219149$: That is, $p_Y(0) + p_Y(1) + p_Y(2) + p_Y(3) + p_Y(6) = 0.4070$

You may want to compare this with how the p value would be computed for a one-sided test. If, for example, the alternative hypothesis was $p < 0.667$, the p value would be $P_0(Y \leq y^*) = P_0(Y \leq 3) = p_Y(0) + p_Y(1) + p_Y(2) + p_Y(3) = 0.3190$.

- **Example: Large Sample Test for a Proportion**

  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 **scientific hypothesis** then becomes: "The population proportion of in-spec parts with diameters larger than the target value is greater than 1/2."

  The datum is Y, the number of the 150 sampled parts with diameters larger than the target value. If we assume each part represents a Bernoulli trial (independent, two possible outcomes: diameter larger than target or not, and probability p of being larger than target), we get the **statistical model**: $Y \sim b(150, p)$.

  The **statistical hypotheses** are

  $$\begin{array}{lccc} H_0: & p & = & 0.5 \\ H_a: & p & > & 0.5 \end{array}$$

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

  Of the 150 parts, y^* (the observed value of Y) equals 93 (a proportion 0.62).

  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 continuity-corrected test statistic is

  $$z_u^* = \frac{93 - (0.5)(150) - 0.5}{\sqrt{(150)(0.5)(1-0.5)}} = 2.858.$$

  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 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 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 μ_D based on the differences D_i.

- # Example:

  The manufacturer of a new warmup bat thinks its product is effective in raising batting averages. To test if this is true, it selects a random sample of 12 baseball players from among a larger number who volunteer to try the bat, and who have never used it before. The players use the warmup bat for a season, and company researchers obtain as data the batting averages from this season and the previous (pre-bat) season.

  1. **The Scientific Hypothesis** Batting averages are higher when players use the bat.
  2. **The Statistical Model** The paired C+E model: If D_i is the difference between this season's and last season's batting average for player i, we assume $D_i = \mu_D + \epsilon_i$, where the random errors, ϵ_i, are independent and follow a $N(0, \sigma^2)$ distribution model.
  3. **The Statistical Hypotheses**

     $$\begin{array}{lccc} H_0: & \mu_D & = & 0 \\ H_a: & \mu_D & > & 0 \end{array}$$

  4. **The Test Statistic** The standardized test statistic is

     $$t = \frac{\overline{D}}{\hat{\sigma}(\overline{D})},$$

     where

     $$\hat{\sigma}(\overline{D}) = \frac{S_D}{\sqrt{n}}$$

     is the estimated standard error of $\overline{D}$, and $n = 12$.

     Under H_0, t follows a t_{11} distribution model. The data (found in SASDATA.BATTING) are:

PLAYER	AVG92	AVG93	D
1	0.254	0.262	0.008
2	0.274	0.290	0.016
3	0.300	0.304	0.004
4	0.246	0.267	0.021
5	0.278	0.291	0.013
6	0.252	0.257	0.005
7	0.235	0.248	0.013
8	0.313	0.324	0.021
9	0.305	0.317	0.012
10	0.255	0.252	−0.003
11	0.244	0.276	0.032
12	0.322	0.332	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 $\hat{\sigma}(\overline{D}) = 0.0092/\sqrt{12} = 0.0027$, so the observed value of the standardized test statistic is

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

5. **The P-Value** The p-value is

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

- # 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

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

where δ_0 is a specified value.

- # 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

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

Then, if H_0 is true,

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

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

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

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

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

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

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

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 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)

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

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 $\overline{y}_1 - \overline{y}_2 = 118.4 - 134.9 = -16.5$.

1. **The Scientific Hypothesis** There is a difference in the lifetimes of blades from the two manufacturers.
2. **The Statistical Model** The two population C+E model. For illustration, we will consider both the equal-variance and general case.
3. **The Statistical Hypotheses**

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

To calculate the test statistic and p-value, we will consider separately the two cases: equal variances and unequal variances.

• Equal Variances

4. **The Test Statistic**
   The pooled variance estimate is

$$\begin{aligned} s_p^2 &= \frac{(10-1)(26.9)^2 + (13-1)(18.4)^2}{10+13-2} \\ &= 503.6, \end{aligned}$$

   So the standard error estimate of $\overline{Y}_1 - \overline{Y}_2$ is

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

   Therefore, $t^{(p)*} = -16.5/9.44 = -1.75$, with 10+13-2=21 degrees of freedom.
5. **The p-value** $p_- = P(t_{21} \leq -1.75) = 0.0473$, $p^+ = P(t_{21} \geq -1.75) = 0.9527$, and the p-value for this problem is $2\min(0.0473, 0.9527) = 0.0946$.

• Unequal Variances

4. **The Test Statistic**
   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 $t^{(ap)*} = -16.5/9.92 = -1.67$. 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 $\nu = 15$.

5. **The P-Value**
   $p_- = P(t_{15} \leq -1.67) = 0.0583$, $p^+ = P(t_{15} \geq -1.67) = 0.9417$, 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.

• The Large Sample Case

If n_1 and n_2 are large (for most cases, 100 will qualify as large), you may base the test on the statistic used in the unequal variances case:

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

Under H_0, $t^{(ap)} \dot{\sim} N(0,1)$, so you may use the standard normal distribution to compute the p-value (which means you don't have to do that nasty degrees of freedom calculation.)

• 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 δ_0,

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

against one of three possible alternative hypotheses:

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

The tests we will present rely on the normal approximation promised by the Central Limit Theorem. Therefore, you should always check that the sample sizes are large enough to justify this approximation. $y_i \geq 10$ and $n_i - y_i \geq 10$, $i = 1, 2$, suffices as a rule of thumb.

- **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

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

and 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.

  To illustrate the calculation of the two possible test statistics, we will consider two different scientific hypotheses:

  1. **Scientific Hypothesis 1:** There is a difference in the population proportions of male and female students who agree or strongly agree with the statement.
  2. **Scientific Hypothesis 2:** The population proportion of males who agree or strongly agree with the statement is at least 0.1 greater than the population proportion of females who agree or strongly agree with the statement.

  1. **Scientific Hypothesis 1** There is a difference in the population proportions of male and female students who agree or strongly agree with the statement.
  2. **The Statistical Model** The two-population binomial.
  3. **The Statistical Hypotheses**

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

  4. **The Test Statistic** 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 $\hat{p} = (26 + 24)/(200 + 100) = 0.167$.

  Thus,

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

  and

  $$z^* = \frac{-0.140}{0.046} = -3.04.$$

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

1. **Scientific Hypothesis 2** The population proportion of males who agree or strongly agree with the statement is at least 0.1 greater than the population proportion of females who agree or strongly agree with the statement.
2. **The Statistical Model** The two-population binomial.
3. **The Statistical Hypotheses**

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

4. **The Test Statistic** The estimated standard error of $p_f - p_m$ is

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

which gives

$$\begin{aligned} z^* &= \frac{24/200 - 26/100 - (-0.10)}{0.05} \\ &= -0.80, \end{aligned}$$

5. **The P-Value** $P(N(0,1) \leq -0.80) = 0.2119$.

- # Fixed Significance Level Tests

Steps, illustrated using grinding example:

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

$$\begin{array}{lrcl} H_0: & \mu & = & 0.75 \\ H_a: & \mu & > & 0.75 \end{array}$$

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

2. **Set the significance level $\boldsymbol{\alpha}$.** Usual choices are 0.01 or 0.05. We'll choose the latter.
3. **Specify the (standardized) test statistic and it's distribution under $\mathbf{H}_0$.** For simplicity, assume we know $\sigma = 0.0048$. Then the standardized test statistic is

$$Z = \frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}} = \frac{\overline{Y} - 0.75}{0.0048/\sqrt{150}},$$

and under H_0 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 H_0 will be rejected in favor of H_a. Here, H_a tells us that the critical region has the form

$$[z_\alpha, \infty) = [z_{0.05}, \infty) = [1.645, \infty),$$

meaning H_0 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 H_0 is rejected in favor of H_a.

- **Power** In a fixed significance level test, power is the probability of rejecting H_0 in favor of H_a. 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

$$\begin{aligned}
\Pi(\mu) &= P(Z \geq 1.645|\mu) \\
&= P\left(\frac{\overline{Y} - 0.75}{0.0048/\sqrt{150}} \geq 1.645|\mu\right) \\
&= P(\overline{Y} \geq 0.75 + (1.645)\frac{0.0048}{\sqrt{150}}|\mu) \\
&= P(Z' \geq 1.645 + \frac{0.75 - \mu}{0.0048/\sqrt{150}}),
\end{aligned}$$

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

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