25. ADVANCED HYPOTHESIS TESTING

Hypothesis Testing For Proportions

Suppose we are going to observe \(X\), the number of successes in \(n\) trials of a binomial experiment. We have a hypothesis on the probability \(p\) of success on any given trial. The null hypothesis is of form \(H_0: p = p_0\), where \(p_0\) is a known number. Here are the rejection rules for various large-sample hypothesis tests of level \(\alpha\), according to the nature of the alternative hypothesis. In each case, the test statistic is the \(z\)-statistic, \(z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}\), which measures the distance between \(\hat{p}\) and \(p_0\) in units of standard errors.

\[
\begin{array}{c|c}
H_A & \text{Rule For Rejecting } H_0 \\
\hline
p \neq p_0 & |z| > z_{\alpha/2} \\
p > p_0 & z > z_{\alpha} \\
p < p_0 & z < -z_{\alpha}
\end{array}
\]

Examples:
Bush-Gore Election (Test if it’s a tie, a “statistical dead heat”).
Taste Test (Test if prob of a correct guess is \(1/2\)).
Forecasting Lab (Test if prob of a correct forecast is \(1/2\)).
Lucky Coin, which always comes up Heads. (Test if prob of Heads is \(1/2\), and compute \(p\)-value).
Essex County Elections (Next slide).

Eg: The top line on county ballots is supposed to be assigned by random drawing to either the Republican or Democratic candidate. The clerk of the county is supposed to make this random drawing. In Essex County, New Jersey, the county clerk (who incidentally is a Democrat) assigned the top line to the Democrats 40 out of 41 times. What is the probability that the Democrats would do so well in a truly random lottery?

Sol: Let’s test \(H_0: p = .5\) versus \(H_A: p > .5\) at level .01, where \(p\) is the probability that the clerk assigns the top line to a Democrat for a given election. We have \(\hat{p} = 40/41 = .976\). The \(z\)-statistic is \(z = \frac{.976-.5}{\sqrt{.5(1-.5)/41}} = 6.10\).

We reject \(H_0\) at level .01, and even at much smaller values of \(\alpha\). In fact, the \(p\)-value (one-tailed) is 1 in 52 Billion, using the exact binomial formula. This is the probability that the Democrats would do so well in a truly random lottery!

Hypothesis Testing for the difference between two means

We often have two samples and we want to decide if the population means are different.

Suppose Samples 1 and 2 are drawn from Populations 1 and 2.
The sample means are \(\bar{x}_1\) and \(\bar{x}_2\).
The population means are \(\mu_1\) and \(\mu_2\).
We want to test \(H_0: \mu_1 = \mu_2\).
This is equivalent to \(H_0: \mu_1 - \mu_2 = 0\).
Sometimes, there is a natural item in sample 2 to go with each item in sample 1. Then we say the data are in matched pairs.

If we have matched pairs, we can simply work with the differences between the observations for each pair, and use a one-sample \( t \) test.

You used this method in a HW problem on attractiveness of members of the opposite sex before and after drinking. Here, the same subject serves as both treatment and control.

Another example is the use of identical twins in a study: Give the treatment to one twin, and the placebo to the other twin.

For example, there was a 1989 Finnish study of twins where one twin smoked and the other didn’t. There were 22 pairs of twins where at least one twin had died by 1989. In 17 of these 22 cases, the smoker died first.

Use of matched pairs is extremely desirable since it cuts down on extraneous sources of variability which obscure the information in the data.

But often, we have two independent samples without matched pairs.

In this case, we use Minitab’s 2-sample \( t \), which is based on the standardized difference between the sample means, \( \bar{x}_1 - \bar{x}_2 \). We won’t worry about formulas! Instead, let’s focus on interpreting the Minitab output.

Eg: A trade break occurs whenever the buyer and seller of a stock have a misunderstanding about the price they supposedly agreed upon. The data set Break.MTP contains data on all daily NY Exchange trades that occurred in a large New York City investment house from June 1995 to May 1996. (This example and data set were provided by Professor Jeffrey Simonoff, who is not at liberty to disclose the name of the investment house.) For each day, the number of breaks, the total number of trades, and the break rate (as a percentage of all trades) are given. The investment house would like to understand what factors cause a high break rate, since resolving the breaks is costly and time consuming.

Preliminary investigation revealed that the break rate may be related to the day of the week on which the trade was made.
The boxplots indicate that the break rate seems to be higher on Mondays than on other trading days. But is there a statistically significant difference? To answer this, use Minitab’s 2-sample t:

**Two Sample T-Test and Confidence Interval**

**Two sample T for Break_Rate**

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>207</td>
<td>7.30</td>
<td>2.36</td>
<td>0.16</td>
</tr>
<tr>
<td>1</td>
<td>47</td>
<td>8.61</td>
<td>4.15</td>
<td>0.60</td>
</tr>
</tbody>
</table>

95% CI for mu (0) - mu (1): (-2.57, -0.05)
T-Test mu (0) = mu (1) (vs <): T = -2.09  P = 0.021  DF = 52

The sample mean break rate is 7.30 on non-Mondays, compared to 8.61 on Mondays. Is this more than just a chance fluctuation? The 95% confidence interval for the difference in population mean break rates (non-Monday minus Monday) includes only negative values. The one-tailed t-test for the difference in the population means yields a p-value of .021. Thus, there is statistical evidence that the mean break rate is indeed higher on Mondays than on the other days.

**Hypothesis Testing for the difference between two proportions**

Suppose we want to compare two population proportions $p_1$ and $p_2$ based on two independent binomial experiments.

Our data are $x_1$, $n_1$, $x_2$, $n_2$, the number of successes and observations in the two samples. We want to test $H_0 : p_1 - p_2 = 0$.

The test is based on the standardized difference between the sample proportions, $\hat{p}_1 - \hat{p}_2$, where $\hat{p}_1 = x_1/n_1$ and $\hat{p}_2 = x_2/n_2$.

Again, we will skip the formulas and focus on the Minitab output (2-sample proportions).

Eg: The U.S. Physicians Health Study considered the effect of aspirin on fatal heart attacks. Out of 22,071 U.S. physicians, a random sample of about half of them received one Bufferin aspirin every other day, and the other half received a placebo. The study was double-blind. After five years, there were 5 fatal heart attacks in the Aspirin group (of 11,037), but there were 18 fatal heart attacks in the Placebo group (of 11,034).

The rate of fatal heart attacks is over three times greater in the placebo group than in the Aspirin group. But is this a statistically significant difference?
The Minitab 2-Proportions output is as follows.

**Test and Confidence Interval for Two Proportions**

<table>
<thead>
<tr>
<th>Sample</th>
<th>X</th>
<th>N</th>
<th>Sample p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>11037</td>
<td>0.000453</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>11034</td>
<td>0.001631</td>
</tr>
</tbody>
</table>

Estimate for p(1) - p(2): -0.00117830
95% CI for p(1) - p(2): (-0.00202954, -0.000327055)
Test for p(1) - p(2) = 0 (vs < 0): Z = -2.71  P-Value = 0.003

If aspirin were ineffective, we would see such an improvement compared to the placebo in only 3 samples of 1,000.

The benefits of aspirin were highly statistically significant.