Chapter 21: More about tests

Example: The space shuttle Challenger suffered a catastrophic failure January 27, 1986. The solid rocket motor disintegrated after the several of o-rings failed to contain combustion gases. There were two o-rings\(^1\) on each of three joints and at least one joint was breached, implying that both the primary and secondary o-rings were breached. Some testing had been done on the reliability of the o-rings, and it was believed that the resiliency depended on ambient air temperature and pressure. The lowest temperature at launch before the January 27 event was 53° F; at the time of the launch, the temperature was forecast to be 31° F.

Information on reliability was obtained by recovering 23 boosters used in 24 of the preceding flights. None of the secondary o-rings failed in these 24 flights.\(^2\) The figure to the right shows the estimated probability of o-ring failure at each launch temperature. There had been at least one heated discussion between the engineers (who were familiar with the susceptibility of the o-rings to low temperatures) and the NASA managers about aborting the launch. If the decision to launch had been based on these data, would the Challenger have been launched?

In hindsight, a hypothesis test might have convinced the managers not to proceed with the launch. Let \(p\) denote the probability that an o-ring will fail at 32° F. For brevity, let

\[
H_0 : p = .1 \quad \text{and} \quad H_a : p > .1.
\]

If \(H_0\) is true, then the probability that two independent o-rings fail is \(.1^2 = .01\), or one in 100. Simplistically, the decision to launch might be based on whether \(H_0\) is concluded to be false and that \(H_a\) is true, or conversely, that \(H_0\) is accepted as true. If \(H_0\) is concluded to be false, the space shuttle will not be launched. If \(H_0\) is accepted as true, then space shuttle will be launched.

In this situation, the strength of evidence is important, but ultimately, a decision will be made: reject \(H_0\) or accept \(H_0\) since the shuttle will or will not be launched. This approach

---

\(^1\)Each o-ring was 37.5 feet in diameter and .28 inches thick.

is undesirable from the scientific/academic view, but in light of the fact that decisions are
made based on hypothesis tests (particularly outside academia), it’s important to discuss
hypothesis testing in this context.

The decision-making approach: a significance level \( \alpha \) is set in advance of collecting the
data to define a decision rule. \( \alpha \) represents the point at which the p-value is small enough
to reject \( H_0 \) and adopt \( H_a \). The decision rule is

- If the p-value \( \leq \alpha \), reject \( H_0 \) and conclude that \( H_a \) is true.
- If the p-value \( > \alpha \), fail to reject \( H_0 \) and conclude that there is insufficient evidence
to reject \( H_0 \).

Common choices of \( \alpha \) are .01, .05, or .10.

Recall that the p-value was defined to be the probability of obtaining a \( z \)-statistic as extreme
or more so if \( H_0 \) is true. (Extreme means contradicting \( H_0 \) and supporting \( H_a \).) Suppose
that

\[
H_0 : p = p_0 \quad \text{and} \quad H_a : p > p_0,
\]

Then

\[
p-value = P(Z \geq z|H_0),
\]

where \( z \) is the observed value of the test statistic.

Suppose that a decision rule is to be used and \( \alpha \) is set to be \( \alpha = .05 \). The data yields
a one-sample \( z \) statistic value of \( z = 1.7 \) and p-value = .0446. Consequently, \( H_0 \) is rejected
in favor of \( H_a \). The figure illustrates the situation.
The test statistic also can be used to define the decision rule since \( P(Z \geq 1.645|H_0) = 0.05 \). The rule is

\[
\begin{align*}
\text{if } Z > 1.645, & \quad \text{then } p\text{-value} = P(Z > 1.645|H_0) < 0.05 \\
& \Rightarrow \text{reject } H_0 \\
\text{if } Z < 1.645, & \quad \text{then } p\text{-value} = P(Z > 1.645|H_0) > 0.05 \\
& \Rightarrow \text{fail to reject } H_0
\end{align*}
\]

If \( Z = 1.645 \), then \( p\text{-value} = P(Z > 1.645|H_0) = 0.05 \). This discussion illustrates that \( \alpha \) is the probability of rejecting \( H_0 \) if \( H_0 \) is true, i.e.,

\[ \alpha = P(\text{reject } H_0|H_0 \text{ is true}). \]

So, \( \alpha \) partially quantifies the risk in making the decision. For example, if \( \alpha = 0.05 \) is always used for testing, then \( H_0 \) will be incorrectly rejected when it’s true 5% of the time. This is similar to confidence intervals since 95% confidence intervals fails to contain the parameter 5% of the time.

Consider a test that produced \( z = -2.014 \) and \( p\text{-value} = 0.022 \). Using a decision rule, what would be concluded if the following \( \alpha \) levels were adopted?

1. \( \alpha = 0.05? \)
   There is sufficient evidence to reject \( H_0 \) at the .05 significance level because \( p\text{-value} = 0.022 < 0.05 = \alpha \).

2. \( \alpha = 0.01? \)
   There is insufficient evidence to reject \( H_0 \) at the .01 significance level because \( \alpha = 0.01 < 0.022 = p\text{-value} \).

If \( H_0 \) cannot be rejected because the \( p\text{-value} \) is greater than \( \alpha \), then it is not concluded that \( H_0 \) is true. Instead, the conclusion is that there is insufficient evidence to reject \( H_0 \). It is possible that more data (and hence evidence) will lead to rejecting \( H_0 \). Often, researchers are not willing drop the position that \( H_a \) is true based on the outcome of a single test.

Because the \( \alpha \)-level may be set to draw any desired conclusion, \( \alpha \) always should be set in advance of data collection. Further, there is no concrete division between significance (and rejecting \( H_0 \)) and non-significance (not rejecting \( H_0 \)) as implied by the use of an \( \alpha \)-level and a decision rule.

The strength of evidence as measured by the \( p\text{-value} \) is not discrete—there is only stronger and stronger evidence against \( H_0 \) as the \( p\text{-value} \) decreases. Consequently, it’s more informative to report the \( p\text{-value} \) than the \( \alpha \)-level and the outcome of the decision rule (reject, or
A company advertises with a local radio station hoping to increase product recognition among local residents. They know that 20% of their community already recognizes their product. The radio station wants to convince the company that the advertising was effective, so they sample residents after an ad runs for one week to see if there is an increase in product recognition. Here is a hypothesis test:

1. The parameter of interest is $p = \text{the proportion of town residents that can recognize the company’s product.}$

2. The hypotheses are:

   $H_0 : p = .20 \text{ vs. } H_a : p > .20$

3. The test statistic is

   $$ Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} $$

   Large values of $\hat{p}$ contradict $H_0$ and support $H_a \implies p\text{-value} = P(Z \geq z\mid H_0 \text{ is true}),$ where $z$ is the computed value of the test statistic.

4. Compute the p-value in each of the following sample cases.

   - 100 people surveyed, of whom 21 recognize the product. Thus, $\hat{p} = .21$ and the result is nonsignificant, because a 1% improvement in product recognition is worth essentially nothing. Further,

     $$ Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{.21 - .2}{\sqrt{.2 \times .8/100}} = .25. $$

     From Table $Z$, p-value $= P(Z \geq .25\mid H_0 \text{ is true}) = .401$. Conclusion: there’s no evidence supporting $H_0$.

   - 1,000 people surveyed, of whom 210 recognize the product.

     Again, $\hat{p} = .21$ and there is no practical improvement gained by the advertising. The result is nonsignificant from a practical standpoint. The $Z$-statistic depends on $n$, so it should be recomputed with the new $n$:

     $$ Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{.21 - .2}{\sqrt{.2 \times .8/1000}} = .77. $$
From Table Z, \( p\text{-value} = P(Z \geq .77|H_0 \text{ is true}) = .245 \). Again, there’s no evidence supporting \( H_0 \).

- 10,000 people surveyed, of whom 2100 recognize the product.
  
  Thus, \( \hat{p} = .21 \) and there is no practical improvement gained by the advertising. The \( Z \)-statistic depends on \( n \), so

\[
Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{.21 - .2}{\sqrt{.2 \times .8/10000}} = 2.5,
\]

and \( p\text{-value} = P(Z \geq 2.5|H_0 \text{ is true}) = .0062 \). Now the evidence is very strong supporting \( H_a : p > .2 \) and refuting \( H_0 : p = .2 \).

What happened? The precision of the estimate \( \hat{p} = .21 \) is greater when \( n = 10,000 \). Hence, \( \sigma(\hat{p}) = .0041 \) is much smaller than when \( n = 100 \). This term is the denominator of the \( z \)-statistic, and so the \( z \)-statistic is larger than when \( n = 100 \). But \( p \) is not much larger than .2.\(^3\)

In each example, \( \hat{p} = .21 \), mathematically different from \( p = .2 \), but from a practical standpoint, the difference is not significant. In the last example, the difference was found to be statistically significant because the strength of evidence was substantial and convincing. In brief, this illustrates that statistical significance is not the same as practical significance.

**Type I and Type II errors:** if a hypothesis test is conducted with a decision rule and a significance level \( \alpha \), there are two possible decisions that can be made:

1. Reject \( H_0 \) if \( p\text{-value} \leq \alpha \).
2. Accept \( H_0 \) if \( p\text{-value} > \alpha \).

Either decision may be wrong, but the consequences are different. They are

1. If \( H_0 \) is rejected when it is true, then a Type I error has been committed.
2. If \( H_0 \) is not rejected when it is false, then a Type II error has been committed.

With two possible hypotheses and two possible decisions, there are four possible outcomes of a significance test, as summarized in the table to the right.

<table>
<thead>
<tr>
<th>Truth</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0 )</td>
<td>Correct Decision</td>
</tr>
<tr>
<td>( H_a )</td>
<td>Type I Error Correct Decision</td>
</tr>
</tbody>
</table>

\(^3\)A 95% confidence interval for \( p \) is (.202, .218).
The probability of a Type I error is the probability of rejecting $H_0$ when $H_0$ is true.

For example, suppose that $H_a: p > p_0$ and $\alpha = .05$. Consequently, the decision rule states that $H_0$ is rejected if $Z \geq 1.645$. The probability of a Type I error is

$$P(\text{Type I error}) = P(Z \geq 1.645 | H_0 \text{ true}) = .05 = \alpha.$$  

The same reasoning applies to the other two types of alternative hypotheses and different values of $\alpha$, and in general,

$$P(\text{Type I error}) = \alpha.$$  

This logic provides some guidance for choosing $\alpha$ since it is the probability of a Type I error. $\alpha$ should be the largest acceptable level of Type I error. Smaller values of $\alpha$ are undesirable because if the alternative hypothesis is true, more evidence is needed before (correctly) rejecting $H_0$.

- Which type of error (Type I or II) is more important?

It depends on the problem. However, it can be said unequivocally that if the probability of a Type II error is much larger than .25, there’s little value to the study. Why? In most cases, the researcher is attempting to reject $H_0$ and conclude that $H_a$ is true, because they believe $H_a$ is true. If the researcher does not have a good chance of succeeding, it’s not worth the effort.

**Example:** A state clean air standard requires that vehicle exhaust emissions not exceed a specified limit for a certain pollutant. State regulators check a random sample of cars that a suspect repair shop has passed as meeting the standard. The state will revoke the shop’s license if they find significant evidence that the shop is certifying vehicles that do not meet standards. (Their approach is that the shop is presumed innocent of certifying vehicles that do not meet standards, but want to assess the evidence that the shop is guilty of certifying vehicles that do not meet standards.)

In this context, what is a Type I error? ($H_0$ states that the shop is innocent.)

In this context, what is a Type II error? ($H_a$ states that the shop is guilty.)

---

$^4$Concluding the shop is certifying vehicles that do not meet standards when in truth they are certifying only vehicles that meet the standards.

$^5$Concluding the shop is not certifying vehicles that fail to meet standards when in truth they are certifying these vehicles.
Which type of error would the shop’s owner consider more serious?\textsuperscript{6}

Which type of error might an environmental group consider more serious?\textsuperscript{7}

Concluding remark: Decision rules have limited use. It’s usually better to present an estimate and standard error, and the p-value and let others form their on conclusion.

\textsuperscript{6}Type I
\textsuperscript{7}Type II