Tests of Significance

Outline:
- General Procedure for Hypothesis Testing
  - Null and Alternative Hypotheses
  - Test Statistics
  - $p$-values
- Interpretation of the Significance Level
- Tests for a Population Mean
- Interpretation of $p$-values
- Statistical vs. Practical Significance
- Confidence Intervals and Hypothesis Tests
- Potential Abuses of Tests

Testing Hypotheses

A hypothesis test is an assessment of the evidence provided by the data in favor of (or against) some claim about the population.

For example, suppose we perform a randomized experiment or take a random sample and calculate some sample statistic, say the sample mean.

We want to decide if the observed value of the sample statistic is consistent with some hypothesized value of the corresponding population parameter.

If the observed and hypothesized value differ (as they almost certainly will), is the difference due to an incorrect hypothesis or merely due to chance variation?

A confidence interval is a very useful statistical inference tool when the goal is to estimate a population parameter.

When the goal is to assess the evidence provided by the data in favor of some claim about the population, test of significance are used.

Example: Filling Coke Bottles

A machine at a Coke production plant is designed to fill bottles with 16oz of Coke. The actual amount varies slightly from bottle to bottle. From past experience, it is known that the SD 0.2oz.

A SRS of 100 bottles filled by the machine has a mean 15.94oz per bottle. Is this evidence that the machine needs to be recalibrated, or could this difference be a result of random variation?

General Procedure for Hypotheses Testing

1. Formulate the null hypothesis and the alternative hypothesis

- The null hypothesis $H_0$ is the statement being tested. Usually it states that the difference between the observed value and the hypothesized value is only due to chance variation.
  
  For example, $\mu = 16$ oz.

- The alternative hypothesis $H_a$ is the statement we will favor if we find evidence that the null hypothesis is false. It usually states that there is a real difference between the observed and hypothesized values.
  
  For example, $\mu \neq 16$, $\mu > 16$, or $\mu < 16$.

A test is called

- two-sided if $H_a$ is of the form $\mu \neq 16$.

- one-sided if $H_a$ is of the form $\mu > 16$, or $\mu < 16$. 

Example: GRE Scores

The mean score of all examinees on the Verbal and Quantitative sections of the GRE is about 1040. Suppose 50 randomly sampled UC Berkeley graduate students have a mean GRE V+Q score of 1310. We are interested in determining if a mean GRE V+Q score of 1310 gives evidence that, as a whole, Berkeley graduate students have a higher mean GRE score than the national average.

What is $H_0$? What is $H_a$?

For the Coke example, we have that the mean of the sample is 15.94 oz. The population mean specified by the null hypothesis is 16 oz. A test statistic is

$$z = \frac{15.94 - 16}{0.2/\sqrt{100}} = -3$$

(We’ll have more to say about this in a moment.)

General Procedure for Hypotheses Testing

cont...

2. Calculate the test statistic on which the test will be based.

The test statistic measures the difference between the observed data and what would be expected if the null hypothesis were true. When $H_0$ is true, we expect the estimate based on the sample to take a value near the parameter value specified by $H_0$.

Our goal is to answer the question, “How extreme is the value calculated from the sample from what we would expect under the null hypothesis?”

In many common situations the test statistic has the form

$$\frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}}$$

3. Find the $p$-value of the observed result

- The $p$-value is the probability of observing a test statistic as extreme or more extreme than actually observed, assuming the null hypothesis $H_0$ is true.
- The smaller the $p$-value, the stronger the evidence against the null hypothesis.
- If the $p$-value is as small or smaller than some number $\alpha$ (e.g. 0.01, 0.05), we say that the result is statistically significant at level $\alpha$.
- $\alpha$ is called the significance level of the test.

In the case of the Coke example, $p = 0.0013$ for a one-sided test or $p = 0.0026$ for a two-sided test. (Once again, we’ll have more to say about this in a moment.)
Interpretation of the Significance Level

To perform a test of significance level \( \alpha \), we perform the previous three steps and then reject \( H_0 \) if the \( p \)-value is less than \( \alpha \).

The following outcomes are possible when conducting a test:

<table>
<thead>
<tr>
<th>Reality</th>
<th>Our Decision</th>
<th>Type I Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0 )</td>
<td>( \checkmark )</td>
<td>Type I Error</td>
</tr>
<tr>
<td>( H_a )</td>
<td>Type II Error</td>
<td>( \checkmark )</td>
</tr>
</tbody>
</table>

Suppose \( H_0 \) is actually true. If we draw many samples, and perform a test for each one, \( \alpha \) of these tests will (incorrectly) reject \( H_0 \). In other words, \( \alpha \) is the probability that we will make a Type I error.

Type II error is related to the notion of the power of a test, which we will discuss later.

Tests for a Population Mean

In the preceding example, we were able to perform an exact Binomial test. Frequently, an exact test is impractical, but we can use the approximate normality of means to conduct an approximate test.

Suppose we want to test the hypothesis that \( \mu \) has a specific value:

\[
H_0 : \mu = \mu_0
\]

Since \( \bar{x} \) estimates \( \mu \), the test is based on \( \bar{x} \), which has a (perhaps approximately) Normal distribution. Thus,

\[
z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}
\]

is a standard normal random variable, under the null hypothesis.

\( p \)-values for different alternative hypotheses:

- \( H_a : \mu > \mu_0 \) - \( p \)-value is \( P(Z \geq z) \) (area of right-hand tail)
- \( H_a : \mu < \mu_0 \) - \( p \)-value is \( P(Z \leq z) \) (area of left-hand tail)
- \( H_a : \mu \neq \mu_0 \) - \( p \)-value is \( 2P(Z \geq |z|) \) (area of both tails)

Example: An Exact Binomial Test

In the last 51 World Series (through 2003) there have been 24 seven game series. Suppose we wish to test the hypothesis

\( H_0 \): Games within a World Series are independent, with each team having probability \( \frac{1}{2} \) of winning.

For the alternative hypothesis, let’s use the generic

\( H_a \): The model in \( H_0 \) is incorrect.

Let \( X \) denote the number of games in the World Series. Under \( H_0 \), \( X \) has the following distribution:

<table>
<thead>
<tr>
<th>( k )</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(X=k) )</td>
<td>( \frac{1}{16} )</td>
<td>( \frac{1}{8} )</td>
<td>( \frac{5}{16} )</td>
<td>( \frac{1}{16} )</td>
</tr>
</tbody>
</table>

For our test statistic, let’s just use

\( M = \# \text{seven game series} \)

What is the \( p \)-value?

We need to find \( m \) such that \( P_{H_0}(M \geq m) \approx 0.05 \). Assuming different years’ World Series are independent (i.e. that the last 51 World Series are an SRS from the population of World Series), the number of seven game series in 51 “trials” is \( B(51, \frac{1}{2}) \).

\[
P(M \geq 20) = 0.086 \quad P(M \geq 21) = 0.049
\]

We want to have a significance level of no more than a 5%, so the critical value will be 21.

Do we reject \( H_0 \) at significance level \( \alpha = 0.05 \)? This is just a matter of checking whether our observed value of \( M \) (24) exceeds the critical value (21). It does, so we reject \( H_0 \).

Example: Filling Coke Bottles (cont.)

We are interested in assessing whether or not the machine needs to be recalibrated, which will be the case if it is systematically over- or under-filling bottles. Thus, we will use the hypotheses

\[
H_0 : \mu = 16
\]

\[
H_a : \mu \neq 16
\]

Recall that \( \bar{x} = 15.94 \), \( \sigma = 0.2 \), and \( n = 100 \). Thus,

\[
z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = -3
\]

The \( p \)-value for a two-sided test is

\[
p = 2P(Z \geq 3) = 0.0026
\]

If \( \alpha = 0.01 \), we reject \( H_0 \).

If \( \alpha = 0.05 \), we reject \( H_0 \).
Example: TV Tubes

TV tubes are taken at random and the lifetime measured. \( n = 100, \sigma = 300 \) and \( \bar{x} = 1265 \) (days). Test whether the population mean is 1200, or greater than 1200.

\[
H_0 : \mu = 1200 \\
H_a : \mu > 1200
\]

Under \( H_0 \), \( \bar{x} \sim N(1200, 30) \).

\[
\therefore z = \frac{\bar{x} - 1200}{30} \sim N(0, 1) \text{ under } H_0
\]

The test statistic is \( z = \frac{1265 - 1200}{30} = 2.17 \), and the \( p \)-value is \( P(Z \geq 2.17|H_0) = 0.015 \)

This is evidence against \( H_0 \) at significance level 0.05, so we reject \( H_0 \). That is, we conclude that the average lifetime of TV tubes is greater than 1200 days.

A Rough Interpretation of \( p \)-values

<table>
<thead>
<tr>
<th>( p )-value</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p &gt; 0.10 )</td>
<td>no evidence against ( H_0 )</td>
</tr>
<tr>
<td>( 0.05 &gt; p \leq 0.10 )</td>
<td>weak evidence against ( H_0 )</td>
</tr>
<tr>
<td>( 0.01 &gt; p \leq 0.05 )</td>
<td>evidence against ( H_0 )</td>
</tr>
<tr>
<td>( p \leq 0.01 )</td>
<td>strong evidence against ( H_0 )</td>
</tr>
</tbody>
</table>

Statistical vs. Practical Significance

Saying that a result is statistically significant does not signify that it is large or necessarily important. That decision depends on the particulars of the problem. A statistically significant result only says that there is substantial evidence that \( H_0 \) is false.

Failure to reject \( H_0 \) does not imply that \( H_0 \) is correct. It only implies that we have insufficient evidence to conclude that \( H_0 \) is incorrect.

Confidence Intervals and Hypothesis Tests

A level \( \alpha \) two-sided test rejects a hypothesis \( H_0 : \mu = \mu_0 \) exactly when the value of \( \mu_0 \) falls outside a \( (1 - \alpha) \) confidence interval for \( \mu \).

For example, consider a two-sided test of the following hypotheses

\[
H_0 : \mu = \mu_0 \\
H_a : \mu \neq \mu_0
\]

at the significance level \( \alpha = 0.05 \).

- If \( \mu_0 \) is a value inside the 95% confidence interval for \( \mu \), then this test will have a \( p \)-value greater than 0.05, and therefore will not reject \( H_0 \).
- If \( \mu_0 \) is a value outside the 95% confidence interval for \( \mu \), then this test will have a \( p \)-value smaller than 0.05, and therefore will reject \( H_0 \).

Example

A particular area contains 8000 condominium units. In a survey of the occupants, a simple random sample of size 100 yields the information that there are 160 motor vehicles in the sample giving an average number of motor vehicles per unit of 1.6, with a sample standard deviation of 0.8.

Construct a confidence interval for the total number of vehicles in the area.

The city claims that there are only 11,000 vehicles in the area, so there is no need for a new garage. What do you think?
More on Constructing Hypothesis Tests

Hypothesis always refer to some population or model, not to a particular outcome. As a result, $H_0$ and $H_a$ must be expressed in terms of some population parameter or parameters.

$H_a$ typically expresses the effect that we hope to find evidence for. So $H_a$ is usually carefully thought out first. We then set up $H_0$ to be the case when the hope-for effect is not present.

It is not always clear whether $H_a$ should be one-sided or two-sided, i.e., does the parameter differ from its null hypothesis value in a specified direction.

Note: You are not allowed to look at the data first and then frame $H_a$ to fit what that data show.

Potential Abuses of Tests

In many applications, a researcher constructs a null hypotheses with the intent of discrediting it.

For example:

- $H_0$: new drug has the same effect as placebo
- $H_0$: men and women are paid equally

A small $p$-value can help a drug company can get a drug approved by the FDA. Similarly, a researcher may have an easier time publishing his results if the $p$-value is smaller than 0.05.

Because of that we have to be aware of the following potential abuses:

- Using one-sided tests to make the $p$-value one-half as big
- Conducting repeated sampling and testing and reporting only the lowest $p$-value
- Testing many hypothesis or testing the same hypothesis on many different subgroups.

In the last two, even if there is actually no effect, you will probably get at least one small $p$-value.