Chapter 11
Testing Hypotheses

11.1 Hypotheses
To test whether the daily fluctuations are equally likely to be up as down, we assume that they are, and that any apparent difference from 50% is just random fluctuation.

In other words, we begin by assuming the proportion of days on which the DJIA increases is 50%. This statement will be our null hypothesis.

11.1 Hypotheses
The null hypothesis, \( H_0 \), specifies a population model parameter and proposes a value for that parameter.

We usually write a null hypothesis about a proportion in the form \( H_0: \hat{p} = p_0 \).

For our hypothesis about the DJIA, we need to test \( H_0: \hat{p} = 0.5 \).

The alternative hypothesis, \( H_A \), contains the values of the parameter that we consider plausible if we reject the null hypothesis. Our alternative is \( H_A: \hat{p} \neq 0.5 \).

11.1 Hypotheses
What would convince you that the proportion of up days was not 50%?

Start by finding the standard deviation of the sample proportion of days on which the DJIA increased.

We've seen 51.53% up days out of 1112 trading days.

The sample size of 1112 is big enough to satisfy the Success/Failure condition.

We suspected that the daily price changes are random and independent.

Is the Dow just as likely to move higher as it is to move lower on any given day?

Out of the 1112 trading days in that period, the average increased on 573 days (sample proportion = 0.5153 or 51.53%).

That is more “up” days than “down” days.

But is it far enough from 50% to cast doubt on the assumption of equally likely up or down movement?
11.1 Hypotheses

If we assume that the Dow increases or decreases with equal likelihood, we'll need to center our Normal sampling model at a mean of 0.5.

Then, the standard deviation of the sampling model is: 

$$SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\left(\frac{0.5 \cdot 0.5}{1112}\right)} = 0.015$$

For the mean, $\mu$, we use $p = 0.50$, and for $\sigma$ we use the standard deviation of the sample proportions $SD(\hat{p}) = 0.015$.

---

11.2 A Trial as a Hypothesis Test

We started by assuming that the probability of an up day was 50%.

Then we looked at the data and concluded that we couldn't say otherwise because the proportion that we actually observed wasn't far enough from 50%.

This is the logic of jury trials. In British common law and systems derived from it (including U.S. law), the null hypothesis is that the defendant is innocent.

The evidence takes the form of facts that seem to contradict the presumption of innocence. For us, this means collecting data.
11.3 P-Values

The *P-value* is the probability of seeing the observed data (or something even less likely) *given* the null hypothesis.

A low enough P-value says that the data we have observed would be very unlikely if our null hypothesis were true. If you believe in data more than in assumptions, then when you see a low P-value you should reject the null hypothesis.

When the P-value is high (or just not low enough), data are consistent with the model from the null hypothesis, and we have no reason to reject the null hypothesis. Formally, we say that we "fail to reject" the null hypothesis.

**What to Do with an "Innocent" Defendant**

If there is insufficient evidence to convict the defendant (if the P-value is not low), the jury does not conclude that the null hypothesis is true and declare that the defendant is innocent. Juries can only fail to reject the null hypothesis and declare the defendant "not guilty."

In the same way, if the data are not particularly unlikely under the assumption that the null hypothesis is true, then the most we can do is to "fail to reject" our null hypothesis.

**Example:**

Which of the following are true? If false, explain briefly.

a. A very low P-value provides evidence against the null hypothesis.
   True.

b. A high P-value is strong evidence in favor of the null hypothesis.
   False. A high P-value indicates there is no evidence to reject the null hypothesis. We can continue with our assumption that the null hypothesis is true, but we have not proved it to be true.

c. A P-value above 0.10 shows that the null hypothesis is true.
   False. No P-value ever shows that the null hypothesis is true (or false).

d. If the null hypothesis is true, you can't get a P-value below 0.01.
   False. This will happen 1 in 100 times.

**11.4 The Reasoning of Hypothesis Testing**

We divide hypothesis testing into four distinct sections: hypotheses, model, mechanics, and conclusion.

**Hypotheses**

First, state the null hypothesis.

\[ H_0: \text{parameter} = \text{hypothesized value}. \]

The alternative hypothesis, \( H_1 \), contains the values of the parameter we consider plausible when we reject the null.
11.4 The Reasoning of Hypothesis Testing

Model

Specify the model for the sampling distribution of the statistic you will use to test the null hypothesis and the parameter of interest.

For proportions, use the Normal model for the sampling distribution.

State assumptions and check any corresponding conditions. For a test of a proportion, the assumptions and conditions are the same as for a one-proportion z-interval.

11.4 The Reasoning of Hypothesis Testing

Mechanics

Perform the actual calculation of our test statistic from the data. Usually, the mechanics are handled by a statistics program or calculator.

The goal of the calculation is to obtain a P-value.

If the P-value is small enough, we'll reject the null hypothesis.

11.4 The Reasoning of Hypothesis Testing

Conclusions and Decisions

The primary conclusion in a formal hypothesis test is only a statement stating whether we reject or fail to reject that hypothesis.

Your conclusion about the null hypothesis should never be the end of the process. You can't make a decision based solely on a P-value.

Business decisions should always take into consideration three things:

• the statistical significance of the test,
• the cost of the proposed action, and
• the effect size of the statistic they observed.
11.5 Alternative Hypotheses

In a two-sided alternative we are equally interested in deviations on either side of the null hypothesis value; the P-value is the probability of deviating in either direction from the null hypothesis value.

An alternative hypothesis that focuses on deviations from the null hypothesis value in only one direction is called a one-sided alternative.

Example:
A survey of 100 CEOs finds that 60 think the economy will improve next year. Is there evidence that the rate is higher among all CEOs than the 55% reported by the public at large?

What are the appropriate hypotheses?

What conditions and assumptions must be met to proceed with the test?

Find the \( z \)-statistic.

Determine the P-value.

Write a conclusion in the context of the problem.

Example:
A survey of 100 CEOs finds that 60 think the economy will improve next year. Is there evidence that the rate is higher among all CEOs than the 55% reported by the public at large?

Find the \( z \)-statistic.

Determine the P-value.

\[ z = \frac{\hat{p} - p_0}{\sqrt{(p_0)(1-p_0)/n}} = \frac{0.60 - 0.55}{\sqrt{(0.55)(0.55)/100}} = 1.006 \]

\[ P-value = P(z > 1.006) = 0.1572 \]
**Example:**
A survey of 100 CEOs finds that 60 think the economy will improve next year. Is there evidence that the rate is higher among all CEOs than the 55% reported by the public at large?

Write a conclusion in the context of the problem.

The *P*-value of 0.1572 is large and therefore the collected data were not unusual under the assumption that the null hypothesis is true. Fail to reject the null hypothesis. There is no evidence that the proportion of CEOs who think the economy will improve next year exceeds 55%.

**11.6 P-Values and Decisions:** What to Tell about a Hypothesis Test

Hypothesis tests are particularly useful when we need to make a decision about a population parameter. Whenever possible, it’s a good idea to report a confidence interval for the parameter of interest as well.

Common alpha-levels are 0.05 and 0.01, but there is no steadfast rule in practice. The smaller the *P*-value, the more evidence we have in the data to reject the null hypothesis. Your conclusion about any null hypothesis should always be accompanied by the *P*-value of the test. Reporting the *P*-value allows each reader to decide whether or not there is enough strength in the data to reject the null hypothesis.

**11.7 How to Think about *P*-Values**

A *P*-value is a conditional probability. It tells us the probability of getting results at least as unusual as the observed statistic, given that the null hypothesis is true.

The *P*-value is *not* the probability that the null hypothesis is true.

What to Do with a Small *P*-Value

How small the *P*-value has to be for you to reject the null hypothesis depends on a lot of things, not all of which can be precisely quantified. Your belief in the null hypothesis will influence your decision. Your trust in the data, in the experimental method if the data come from a planned experiment, in the survey protocol if the data come from a designed survey, all influence your decision.

The *P*-value should serve as a measure of the strength of the evidence against the null hypothesis, but should never serve as a hard and fast rule for decisions.
11.7 How to Think about P-Values

What to Do with a High P-Value

Big P-values just mean that what we’ve observed isn’t surprising.

That is, the results are in line with our assumption that the null hypothesis models the world, so we have no reason to reject it.

A big P-value doesn’t prove that the null hypothesis is true, but it certainly offers no evidence that it’s not true.

When we see a large P-value, all we can say is that we “don’t reject the null hypothesis.”

11.8 Alpha Levels and Significance

NOTATION ALERT
The first Greek letter, \( \alpha \), is used in Statistics for the threshold value of a hypothesis test. You’ll hear it referred to as the alpha level.

Common values are 0.10, 0.05, 0.01, and 0.001.

Conclusion
If the P-value < \( \alpha \), then reject \( H_0 \).
If the P-value \( \geq \alpha \), then fail to reject \( H_0 \).

11.8 Alpha Levels and Significance

When the P-Value is small, it tells us that our data are rare given the null hypothesis.

We can define a “rare event” arbitrarily by setting a threshold for our P-value. If our P-value falls below that point, we’ll reject the null hypothesis.

We call such results statistically significant.

The threshold is called an alpha level. Not surprisingly, it’s labeled with the Greek letter \( \alpha \).

11.8 Alpha Levels and Significance

The alpha level is also called the significance level.

You must select the alpha level before you look at the data. Otherwise, you can be accused of finagling the conclusions by tuning the alpha levels to the results after you’ve seen the data.

11.8 Alpha Levels and Significance

For large samples, even small, unimportant (“insignificant”) deviations from the null hypothesis can be statistically significant.

On the other hand, if the sample is not large enough, even large, financially or scientifically important differences may not be statistically significant.

It’s good practice to report the magnitude of the difference between the observed statistic value and the null hypothesis value (in the data units) along with the P-value.
11.8 Alpha Levels and Significance

Statistically significant results come from data that was inconsistent with the claim found in the null hypothesis.

Practically significant results come from data that would cause an individual or a business to take notice, take action, or change a course of business.

Always think carefully about the practical consequences of a hypothesis test conclusion.

11.9 Critical Values

When building a confidence interval, a critical value, \( z^* \) or \( t^* \), corresponds to a selected confidence level.

Critical values can also be used as a shortcut for running a hypothesis test.

Before computers and calculators were common, P-values were hard to find. It was easier to select a few common alpha levels and learn the corresponding critical values for the Normal model or the Student’s \( t \) model.

You’d calculate how many standard deviations your observed statistic was away from the hypothesized value and compare that value directly against the critical value.

Any \( z \)-score or \( t \)-score larger in magnitude than a particular critical value will have a P-value smaller than the corresponding alpha.

11.10 Testing Hypotheses about Means – the One-Sample \( t \)-Test

For testing a hypothesis about a mean, the test is based on the \( t \) distribution.

Is there evidence from a sample that the mean is really different from some hypothesized value calls for a one-sample \( t \)-test for the mean.

Here are the traditional \( z^* \) critical values from the Normal model.

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>1-sided</th>
<th>2-sided</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>1.645</td>
<td>1.96</td>
</tr>
<tr>
<td>0.01</td>
<td>2.33</td>
<td>2.576</td>
</tr>
<tr>
<td>0.001</td>
<td>3.69</td>
<td>3.29</td>
</tr>
</tbody>
</table>
**Example: Shopping Patterns**  
A new manager of a small convenience store randomly samples 20 purchases from yesterday’s sales. If the mean was $45.26 and the standard deviation was $20.67, is there evidence that the overall mean purchase amount is at least $40?

What are the hypotheses?
What conditions must be met?
Find the \( t \)-statistic.
What is the \( P \)-value for the test?
What can you conclude about the overall mean sales?

Find the \( t \)-statistic.

\[
t = \frac{\bar{y} - \mu}{s/\sqrt{n}} = \frac{45.26 - 40}{20.67/\sqrt{20}} = 1.138
\]

What is the \( P \)-value for the test?  
There are 19 degrees of freedom.  

\[
P-value = P(t > 1.138) = 0.1346
\]

What can you conclude about the overall mean sales?

Fail to reject the null hypothesis. There is insufficient evidence that the mean sales is greater than $40.
We know that a larger sample will almost always give better results, but more data costs money, effort, and time.

We know how to find the margin of error for the mean.

$$ME = t^*_{n-1} \times SE(\bar{y})$$

We also know how to find the standard error for the mean.

$$SE(\bar{y}) = \frac{s}{\sqrt{n}}$$

We can determine the sample size by solving this equation for \(n\).

A formula that can be used to estimate sample size when working with means is given as:

$$n \approx \left( \frac{z^* \times \text{standard deviation estimate}}{ME} \right)^2$$

Sample size calculations are never exact.

The margin of error you find after collecting the data won’t match exactly the one you used to find \(n\).

Before you collect data, it’s always a good idea to know whether the sample size is large enough to give you a good chance of being able to tell you what you want to know.

Example: Sample Size

Data from a survey of 25 randomly selected customers found a mean age of 31.84 years and the standard deviation was 9.84 years. A 95% confidence interval for the mean is (27.78, 35.90) with a margin of error of 4.06.

How large a sample is needed to cut the margin of error down to 2?
11.10 Testing Hypotheses about Means – the One-Sample t-Test

Example: Sample Size

\[ n \approx \left( z^* \times \frac{\text{standard deviation estimate}}{\text{ME}} \right)^2 \]

\[ \approx \left( 1.96 \times \frac{9.84}{2} \right)^2 = 92.99 \]

So 93 customers should be targeted for the sample to decrease margin of error down to 2 years.

11.11 Confidence Intervals and Hypothesis Tests

Although a hypothesis test can tell us whether the observed statistic differs from the hypothesized value, it doesn’t say by how much.

The corresponding confidence interval gives us more information.

11.11 Confidence Intervals and Hypothesis Tests

Confidence intervals and hypothesis tests are built from the same calculations.

Because confidence intervals are naturally two-sided, they correspond to two-sided tests.

In general, a confidence interval with a confidence level of $C\%$ corresponds to a two-sided hypothesis test with an $\alpha$ level of $100 - C\%$.

11.11 Confidence Intervals and Hypothesis Tests

You can approximate a hypothesis test by examining the confidence interval.

Just ask whether the null hypothesis value is consistent with a confidence interval for the parameter at the corresponding confidence level.

Example: Shopping Patterns

Recall the new manager of a small convenience store who randomly sampled 20 purchases from yesterday’s sales.

Given a 95% confidence interval (35.586, 54.934), is there evidence that the mean purchase amount is different from $40?
11.11 Confidence Intervals and Hypothesis Tests

Example: Shopping Patterns  Recall the new manager of a small convenience store who randomly sampled 20 purchases from yesterday's sales.

Given a 95% confidence interval (35.586, 54.934), is there evidence that the mean purchase amount is different from $40?

At $\alpha = 0.05$, fail to reject the null hypothesis. The 95% confidence interval contains $40$ as a plausible value.

There is no evidence that the mean purchase amount differs from $40$. 