Chapter 6: Hypothesis Testing

Introduction

- Recall: We discussed an example in Chapter 5 about microdrills.
- Our sample had a mean of 12.68 and standard deviation of 6.83.
- Let us assume that the main question is whether or not the population mean lifetime $\mu$ is greater than 11.
- We can address this by examining the value of the sample mean. We see that our sample mean is larger than 11, but because of uncertainty in the means, this does not guarantee that $\mu > 11$. 
Hypothesis

- We would like to know just how certain we can be that $\mu > 11$.
- A confidence interval is not quite what we need.
- The statement “$\mu > 11$” is a hypothesis about the population mean $\mu$.
- To determine just how certain we can be that a hypothesis is true, we must perform a hypothesis test.

$P$-Value

- The $P$-value measures the plausibility of $H_0$.
- The smaller the $P$-value, the stronger the evidence is against $H_0$.
- If the $P$-value is sufficiently small, we may be willing to abandon our assumption that $H_0$ is true and believe $H_1$ instead.
- This is referred to as rejecting the null hypothesis.
Steps in Performing a Hypothesis Test

1. Define $H_0$ and $H_1$.
2. Assume $H_0$ to be true.
3. Compute a **test statistic**. A test statistic is a statistic that is used to assess the strength of the evidence against $H_0$. A test that uses the $z$-score as a test statistic is called a $z$-test.
4. Compute the $P$-value of the test statistic. The $P$-value is the probability, assuming $H_0$ to be true, that the test statistic would have a value whose disagreement with $H_0$ is as great as or greater than what was actually observed. The $P$-value is also called the **observed significance level**.

Example 6.2

A scale is to be calibrated by weighing a 1000 g test weight 60 times. The 60 scale readings have mean 1000.6 g and standard deviation 2 g. Find the $P$-value for testing $H_0: \mu = 1000$ versus $H_1: \mu \neq 1000$. 

![Graph showing a normal distribution with mean 1000 and standard deviation 2, highlighting the $z$-scores of -2.32 and 2.32 with corresponding $P$-values of 0.0102.]
One and Two-Tailed Tests

- When $H_0$ specifies a single value for $\mu$, both tails contribute to the $P$-value, and the test is said to be a **two-sided** or **two-tailed** test.
- When $H_0$ specifies only that $\mu$ is greater than or equal to, or less than or equal to a value, only one tail contributes to the $P$-value, and the test is called a **one-sided** or **one-tailed** test.

Summary

- Let $X_1, \ldots, X_n$ be a large (e.g., $n \geq 30$) sample from a population with mean $\mu$ and standard deviation $\sigma$.
  To test a null hypothesis of the form $H_0: \mu \leq \mu_0$, $H_0: \mu \geq \mu_0$, or $H_0: \mu = \mu_0$.
- Compute the $z$-score:
  $$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$
- If $\sigma$ is unknown it may be approximated by $s$. 
P-Value

Compute the $P$-value. The $P$-value is an area under the normal curve, which depends on the alternate hypothesis as follows.

- If the alternative hypothesis is $H_1: \mu > \mu_0$, then the $P$-value is the area to the right of $z$.
- If the alternative hypothesis is $H_1: \mu < \mu_0$, then the $P$-value is the area to the left of $z$.
- If the alternative hypothesis is $H_1: \mu \neq \mu_0$, then the $P$-value is the sum of the areas in the tails cut off by $z$ and $-z$.

Section 6.2: Drawing Conclusions from the Results of Hypothesis Tests

- There are two conclusions that we draw when we are finished with a hypothesis test,
  - We reject $H_0$. In other words, we concluded that $H_0$ is false.
  - We do not reject $H_0$. In other words, $H_0$ is plausible.
- One can never conclude that $H_0$ is true. We can just conclude that $H_0$ might be true.
- We need to know what level of disagreement, measured with the $P$-value, is great enough to render the null hypothesis implausible.
More on the $P$-value

- The smaller the $P$-value, the more certain we can be that $H_0$ is false.
- The larger the $P$-value, the more plausible $H_0$ becomes but we can never be certain that $H_0$ is true.
- A rule of thumb suggests to reject $H_0$ whenever $P \leq 0.05$. While this rule is convenient, it has no scientific basis.

Example 6.3

A hypothesis test is performed of the null hypothesis $H_0: \mu = 0$. The $P$-value turns out to be 0.03.

Is the result statistically significant at the 10% level? The 5% level? The 1% level?

Is the null hypothesis rejected at the 10% level? The 5% level? The 1% level?
Comments

- Some people report only that a test significant at a certain level, without giving the P-value. Such as, the result is “statistically significant at the 5% level.”
- This is poor practice.
- First, it provides no way to tell whether the P-value was just barely less than 0.05, or whether it was a lot less.
- Second, reporting that a result was statistically significant at the 5% level implies that there is a big difference between a P-value just under 0.05 and one just above 0.05, when in fact there is little difference.
- Third, a report like this does not allow readers to decide for themselves whether the P-value is small enough to reject the null hypothesis.
- Reporting the P-value gives more information about the strength of the evidence against the null hypothesis and allows each reader to decide for himself or herself whether to reject the null hypothesis.

Comments on P

Let \( \alpha \) be any value between 0 and 1. Then, if \( P \leq \alpha \),

- The result of the test is said to be statistically significant at the 100\( \alpha \)% level.
- The null hypothesis is rejected at the 100\( \alpha \)% level.
- When reporting the result of the hypothesis test, report the P-value, rather than just comparing it to 5% or 1%.
Example 6.4

Specifications for steel plates to be used in the construction of a certain bridge call for the minimum yield to be greater than 345 MPa. Engineers will perform a hypothesis test to decide whether to use a certain type of steel. They will select a random sample of steel plates, measure their breaking strengths, and perform a hypothesis test. The steel will not be used unless the engineers can conclude that $\mu > 345$. Assume they test $H_0: \mu \leq 345$ versus $H_1: \mu > 345$. Will the engineers decide to use the steel if $H_0$ is rejected? What if $H_0$ is not rejected?

Significance

- When a result has a small $P$-value, we say that it is “statistically significant.”
- In common usage, the word significant means “important.”
- It is therefore tempting to think that statistically significant results must always be important.
- Sometimes statistically significant results do not have any scientific or practical importance.
Hypothesis Tests and CI’s

- Both confidence intervals and hypothesis tests are concerned with determining plausible values for a quantity such as a population mean $\mu$.
- In a hypothesis test for a population mean $\mu$, we specify a particular value of $\mu$ (the null hypothesis) and determine if that value is plausible.
- A confidence interval for a population mean $\mu$ can be thought of as a collection of all values for $\mu$ that meet a certain criterion of plausibility, specified by the confidence level $100(1-\alpha)$%.
- The values contained within a two-sided level $100(1-\alpha)$% confidence intervals are precisely those values for which the $P$-value of a two-tailed hypothesis test will be greater than $\alpha$.

Section 6.3: Tests for a Population Proportion

- A population proportion is simply a population mean for a population of 0’s and 1’s: a Bernoulli population.
- We have a sample that consists of successes and failures.
- Here we have hypothesis concerned with a population proportion, it is natural to base the test on the sample proportion.
Hypothesis Test

- Let $X$ be the number of successes in $n$ independent Bernoulli trials, each with success probability $p$; in other words, let $X \sim \text{Bin}(n, p)$.
- To test a null hypothesis of the form $H_0: p \leq p_0$, $H_0: p \geq p_0$, or $H_0: p = p_0$, assuming that both $np_0$ and $n(1 - p_0)$ are greater than 10:
  - Compute the $z$-score:
    $$z = \frac{\hat{p} - p_0}{p_0(1 - p_0)/n}$$

$P$-value

Compute the $P$-value. The $P$-value is an area under the normal curve, which depends on the alternate hypothesis as follows:
- If the alternative hypothesis is $H_1: p > p_0$, the $P$-value is the area to the right of $z$.
- If the alternative hypothesis is $H_1: p < p_0$, the $P$-value is the area to the left of $z$.
- If the alternative hypothesis is $H_1: p \neq p_0$, the $P$-value is the sum of the areas in the tails cut off by $z$ and $-z$. 
Example 6.6

An article presents a method for measuring orthometric heights above sea level. For a sample of 1225 baselines, 926 gave results that were within the class C spirit leveling tolerance limits. Can we conclude that this method produces results within the tolerance limits more than 75% of the time?

Section 6.4: Small Sample Test for a Population Mean

- When we had a large sample we used the sample standard deviation $s$ to approximate the population deviation $\sigma$.
- When the sample size is small, $s$ may not be close to $\sigma$, which invalidates this large-sample method.
- However, when the population is approximately normal, the Student’s $t$ distribution can be used.
- The only time that we don’t use the Student’s $t$ distribution for this situation is when the population standard deviation $\sigma$ is known. Then we are no longer approximating $\sigma$ and we should use the $z$-test.
Hypothesis Test

- Let $X_1, \ldots, X_n$ be a sample from a normal population with mean $\mu$ and standard deviation $\sigma$, where $\sigma$ is unknown.
- To test a null hypothesis of the form $H_0: \mu \leq \mu_0$, $H_0: \mu \geq \mu_0$, or $H_0: \mu = \mu_0$.
- Compute the test statistic
  \[ t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \]

P-value

Compute the P-value. The P-value is an area under the Student’s $t$ curve with $n - 1$ degrees of freedom, which depends on the alternate hypothesis as follows.

- If the alternative hypothesis is $H_1: \mu > \mu_0$, then the P-value is the area to the right of $t$.
- If the alternative hypothesis is $H_1: \mu < \mu_0$, then the P-value is the area to the left of $t$.
- If the alternative hypothesis is $H_1: \mu \neq \mu_0$, then the P-value is the sum of the areas in the tails cut off by $t$ and $-t$. 
Example 6.7

Before a substance can be deemed safe for landfilling, its chemical properties must be characterized. An article reports that in a sample of six replicates of sludge from a New Hampshire wastewater treatment plant, the mean pH was 6.68 with a standard deviation of 0.20. Can we conclude that the mean pH is less than 7.0?
Section 6.5: The Chi-Square Test

- A generalization of the Bernoulli trial is the **multinomial trial**, which is an experiment that can result in any one of \( k \) outcomes, where \( k \geq 2 \).
- Suppose a gambler rolls a die 600 times.
- The results obtained are called the **observed values**.
- To test the null hypothesis that \( p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6 \), we calculate the **expected values** for the given outcome.
- The idea behind the hypothesis test is that if \( H_0 \) is true, then the observed and expected values are likely to be close to each other.

---

**TABLE 6.3** Notation for observed values

<table>
<thead>
<tr>
<th>Row 1</th>
<th>Column 1</th>
<th>Column 2</th>
<th>( \ldots )</th>
<th>Column ( J )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row 1</td>
<td>( O_{11} )</td>
<td>( O_{12} )</td>
<td>( \ldots )</td>
<td>( O_{1J} )</td>
<td>( O_1 )</td>
</tr>
<tr>
<td>Row 2</td>
<td>( O_{21} )</td>
<td>( O_{22} )</td>
<td>( \ldots )</td>
<td>( O_{2J} )</td>
<td>( O_2 )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>( \vdots )</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>Row ( I )</td>
<td>( O_{I1} )</td>
<td>( O_{I2} )</td>
<td>( \ldots )</td>
<td>( O_{IJ} )</td>
<td>( O_I )</td>
</tr>
<tr>
<td>Total</td>
<td>( O_1 )</td>
<td>( O_2 )</td>
<td>( \ldots )</td>
<td>( O_J )</td>
<td>( O_{..} )</td>
</tr>
</tbody>
</table>
The Test

- Therefore we will construct a test statistic that measures the closeness of the observed to the expected values.
- The statistic is called the chi-square statistic.
- Let \( k \) be the number of possible outcomes and let \( O_i \) and \( E_i \) be the observed and expected number of trials that result in outcome \( i \).
- The chi-square statistic is
  \[
  \chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}
  \]

Decision for Test

- The larger the value of \( \chi^2 \), the stronger the evidence against \( H_0 \).
- To determine the \( P \)-value for the test, we must know the null distribution of this test statistic.
- When the expected values are all sufficiently large, a good approximation is available. It is called the chi-square distribution with \( k - 1 \) degrees of freedom.
- Use of the chi-square distribution is appropriate whenever all the expected values are greater than or equal to 5.
- A table for the chi-square distribution is provided in Appendix A, Table A.5.
Chi-Square Tests

- Sometimes several multinomial trials are conducted, each with the same set of possible outcomes.
- The null hypothesis is that the probabilities of the outcomes are the same for each experiment.
- There is a chi-squared statistic for testing for homogeneity.
- There is also a chi-square test for independence between rows and columns in a contingency table.

Example 6.9

Four machines manufacture cylindrical steel pins. The pins are subject to a diameter specification. A pin may meet the specification, or it may be too thin or too thick. Pins are sampled from each machine, and the number of pins in each category is counted. Test the null hypothesis that the proportions of pins that are too thin, OK, or too thick are the same for all machines.

<table>
<thead>
<tr>
<th></th>
<th>Too Thin</th>
<th>OK</th>
<th>Too Thick</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine 1</td>
<td>10</td>
<td>102</td>
<td>8</td>
<td>120</td>
</tr>
<tr>
<td>Machine 2</td>
<td>34</td>
<td>161</td>
<td>5</td>
<td>200</td>
</tr>
<tr>
<td>Machine 3</td>
<td>12</td>
<td>79</td>
<td>9</td>
<td>100</td>
</tr>
<tr>
<td>Machine 4</td>
<td>10</td>
<td>60</td>
<td>10</td>
<td>80</td>
</tr>
<tr>
<td>Total</td>
<td>66</td>
<td>402</td>
<td>32</td>
<td>500</td>
</tr>
</tbody>
</table>
Section 6.6: Fixed-Level Testing

- A hypothesis test measures the plausibility of the null hypothesis by producing a $P$-value.
- The smaller the $P$-value, the less plausible the null.
- We have pointed out that there is no scientifically valid dividing line between plausibility and implausibility, so it is impossible to specify a “correct” $P$-value below which we should reject $H_0$.
- If a decision is going to be made on the basis of a hypothesis test, there is no choice but to pick a cut-off point for the $P$-value.
- When this is done, the test is referred to as a fixed-level test.

Conducting the Test

To conduct a fixed-level test:
- Choose a number $\alpha$, where $0 < \alpha < 1$. This is called the significance level, or the level, of the test.
- Compute the $P$-value in the usual way.
- If $P \leq \alpha$, reject $H_0$. If $P > \alpha$, do not reject $H_0$. 
Comments

- In a fixed-level test, a **critical point** is a value of the test statistic that produces a $P$-value exactly equal to $\alpha$.
- A critical point is a dividing line for the test statistic just as the significance level is a dividing line for the $P$-value.
- If the test statistic is on one side of the critical point, the $P$-value will be less than $\alpha$, and $H_0$ will be rejected.
- If the test statistic is on the other side of the critical point, the $P$-value will be more than $\alpha$, and $H_0$ will not be rejected.
- The region on the side of the critical point that leads to rejection is called the **rejection region**.
- The critical point itself is also in the rejection region.

Example 6.11

A new concrete mix is being evaluated. The plan is to sample 100 concrete blocks made with the new mix, compute the sample mean compressive strength ($\bar{X}$), and then test $H_0: \mu \leq 1350$ versus $H_1: \mu > 1350$, where the units are MPa. It is assumed that previous tests of this sort that the population standard deviation $\sigma$ will be close to 70 MPa. Find the critical point and the rejection region if the test will be conducted at a significance level of 5%.
Example 6.11

Errors

When conducting a fixed-level test at significance level $\alpha$, there are two types of errors that can be made. These are

- Type I error: Reject $H_0$ when it is true.
- Type II error: Fail to reject $H_0$ when it is false.

The probability of Type I error is never greater than $\alpha$. 
Section 6.7: Power

- A hypothesis test results in Type II error if $H_0$ is not rejected when it is false.
- The power of the test is the probability of rejecting $H_0$ when it is false. Therefore,
  \[ \text{Power} = 1 - P(\text{Type II error}). \]
- To be useful, a test must have reasonable small probabilities of both type I and type II errors.

More on Power

- The type I error is kept small by choosing a small value of $\alpha$ as the significance level.
- If the power is large, then the probability of type II error is small as well, and the test is a useful one.
- The purpose of a power calculation is to determine whether or not a hypothesis test, when performed, is likely to reject $H_0$ in the event that $H_0$ is false.
Computing the Power

This involves two steps:
1. Compute the rejection region.
2. Compute the probability that the test statistic falls in the rejection region if the alternate hypothesis is true. This is power.

When power is not large enough, it can be increased by increasing the sample size.

Example 6.12

Find the power of the 5% level test of $H_0: \mu \leq 80$ versus $H_1: \mu > 80$ for the mean yield of the new process under the alternative $\mu = 82$, assuming $n = 50$ and $\sigma = 5$. 

$z_0 = 1.645$
$z_1 = -1.19$

Power = 0.8830
Section 6.8: Multiple Tests

- Sometimes a situation occurs in which it is necessary to perform many hypothesis tests.
- The basic rule governing this situation is that as more tests are performed, the confidence that we can place in our results decreases.
- The Bonferroni method provides a way to adjust $P$-values upward when several hypothesis tests are performed.
- If a $P$-value remains small after the adjustment, the null hypothesis may be rejected.
- To make the Bonferroni adjustment, simply multiply the $P$-value by the number of test performed.

Example 6.22

Four different coating formulations are tested to see if they reduce the wear on cam gears to a value below 100 $\mu$m. The null hypothesis $H_0: \mu \geq 100 \, \mu$m is tested for each formulation and the results are

- Formulation A: $P = 0.37$
- Formulation B: $P = 0.41$
- Formulation C: $P = 0.005$
- Formulation D: $P = 0.21$

The operator suspects that formulation C may be effective, but he knows that the $P$-value of 0.005 is unreliable, because several tests have been performed. Use the Bonferroni adjustment to produce a reliable $P$-value.
Summary

We learned about:
- Large sample tests for a population mean.
- Drawing conclusions from the results of hypothesis tests.
- Tests for a population proportion
- Small sample tests for a population mean.
- Chi-Square test
- Fixed level testing
- Power
- Multiple Tests