

Hypothesis Testing and Tests of Significance (Chapter 26.2-26.6)

A. Overview

Remember the basic idea about hypothesis testing: we make either make assumptions about the parameters or the parameters are known, and then we test to see if those parameters could have led to the sample outcome we observed. We then use a probability calculation to express the strength of our conclusions.

What we are doing in Chapters 26 is learning about a "method" for making decisions. This involves identifying the parameter and statistic, constructing a test, and then interpreting the results of the test to determine whether any differences between what was expected and the actual outcome could have happened by chance or is the difference meaningful?

B. An example

The Statistical Abstract of the United States reported that the net average earnings of all M.D.'s was \$155,800 per year with a standard deviation of \$23,400.

A random sample of income tax returns of 9 M.D.'s practicing in rural communities showed that their net earnings to be:

93,700 110,500 173,600 123,300 136,800 142,700 129,900 153,400 140,200

Assuming that the earnings of these M.D.'s follow a distribution that is approximately normal, use a 1% level of significance to test the claim that the mean earnings of all rural M.D.'s is less than the national average.

1) What is the population? What are the parameters?

All MDs in the US. Average income is \$155,800 and the SD is \$23,400

2) What is the sample? What are the statistics?

9 MDs. Sample Average= 133,789, Sample SD = \$22,045

3) What is the null hypothesis? (See definitions, next section)

Rural average earnings are not different from US average earnings, that is \$155,800

4) What is the alternative hypothesis? (See definitions, next section)

Rural average earnings are less than US average earnings.

5) What is the appropriate test statistic? (See definitions, next section)

Ask yourself...what do I know about the population?

I'll say, although the sample size is very small (9) but the population standard deviation is known (Chapter 26.6 deals with situations when population standard deviation is unknown) so a Z-test is appropriate here.

$z = (\text{observed value} - \text{expected value}) / \text{appropriate measure of spread}$

$(133789 - 155800) / \text{SE of the average.}$

Why the SE of the average?

$$\text{SE of the average} = \frac{\sqrt{9} * 23,400}{9} = 7,800 \text{ dollars .}$$

So how far is 133789 from 155800? about a Z of -2.85.

6) What is the interpretation of the test statistic given the stated significance level of 1% (or p-value of 1%)?

First, what is being suggested is that the person asking the question will only accept p-values of 1% or less. In other words, returning to concepts in Chapters 17, 19, and 20 if there is less than a 1% chance of getting an average this low if we were expecting 155,800 I will reject the null hypothesis in favor of the alternative.

The p-value associated with a Z of -2.85 is about .22%. How did I get that? The probability of getting a Z between +2.85 and -2.85 is 99.56% (from table A 105). So the chance of being outside of that is (100 - 99.56) or .44%. Since the question stated the alternative as a one-tailed (or one-sided) we're only interested in the left side of the curve.

C. Definitions Again (26.2)

1. The NULL HYPOTHESIS is that any differences between the parameter (the expected value) and the observed (or actual) outcome was due to chance alone. In this case, the null hypothesis is a statement about a parameter: the population average is \$155,800
2. The ALTERNATIVE HYPOTHESIS is that the observed results are due to more than just chance. It implies that the NULL is not correct and any observed difference is real, not luck.

Usually, the ALTERNATIVE is what we're setting out to prove. The NULL is like a "straw man" that we might knock over with the alternative. As Freedman says, it's unfortunate that these rather confusing names are so standard.

3. The TEST STATISTIC measures how different the observed results are from what we would expect to get if the null hypothesis were true.

where $z = (\text{observed value} - \text{expected value}) / \text{some measure of spread}$

All a Z does is it tells you how many SEs away the observed value is from the expected value when the expected value is calculated by using the NULL HYPOTHESIS.

4. The SIGNIFICANCE LEVEL (or P-VALUE). This is the chance of getting results as or more extreme than what we got, IF the null hypothesis were true. P-VALUE could also be called "probability value" and it is simply the area associated with the calculated Z.

p-values are always "if-then" statements:

"If the null hypothesis were true, then there would be a p% chance to get these kind of results."

5. If the p-value is less than 5%, we say the results are STATISTICALLY SIGNIFICANT; if $p < 1\%$, the results are HIGHLY STATISTICALLY SIGNIFICANT. A "significant" result means that it would be unlikely to get such extreme observed values by chance alone.

D. Hypothesis Testing Summarized

1. Clearly identify the parameter and the outcome.
2. State the null hypothesis. This is what is being tested. A test of significance assesses the strength of evidence (outcomes) against the null hypothesis. Usually the null hypothesis is a statement of "no-effect" In the tax law example, the senator thinks the law is neutral.
3. State the alternative hypothesis. This is a counter to the null. The counter comes from the IRS who believes that the law is not neutral and hurts the government.
4. Calculate the test statistic.

The significance test assesses evidence by examining how far the test statistic falls from the proposed null. The test is Z.

5. The probability that you observe is called a P-VALUE. The smaller the p-value the stronger is the evidence against the null hypothesis.

6. On significance levels. Sometimes prior to calculating a score and finding its p-value, we state in advance what we believe to be a decisive value of p . This is the significance level. 5% and 1% significance levels are most commonly used. If your p-value is as small or smaller than the significance level you have chosen then you would say that "the data is statistically significant at the ---- (e.g. 1% or 5% or some other) level."

E. One more example

A lawyer who knows a little bit about statistics decides to find out exactly how many sheets of designer toilet paper are in the rolls he usually buys at his local "Why-Pay-Less?" store. Although the package advertises that there are 1000 sheets per roll, he thinks the rolls run out too quickly and therefore must have less than 1000 sheets. He calls the manufacturer and learns that the rolls are normally distributed with a mean of 1000 sheets per roll and a standard deviation of 12 sheets.

He goes out and buys a package of 9 rolls. Treat this like a random sample. He counts each sheet in the 9 rolls and comes up with the following:

998, 999, 1001, 1000, 921, 999, 1001, 998, 997

1. Does the lawyer have enough evidence to sue the manufacturer for false advertising? Use a 5% level of significance as your rule of thumb.

2. Lawsuits are expensive and time-consuming and this lawyer is cautious. He hires you to analyze his data with this thought in mind: given what you have learned about means, medians, standard deviations, extreme observations, sample sizes, perhaps this data set has an unusual observation. If you think it does, go ahead and remove it from the sample and perform a second significance test. Would you advise him to pursue his lawsuit against the manufacturer?

If you think it does not, explain why not.

F. Some Notes (optional 26.5)

Statistical Significance. We use the term "significant" in the statistical sense to mean that an observed difference is likely a real difference and not explainable in terms of sampling error. The level of likelihood used in our analysis is 95%. In other words, when we say a difference is significant we are saying that there is a 95% chance that it is a real difference and not the result of sampling error. This does not mean that the particular result is important in a managerial or any other sense.

If a particular difference is large enough to be unlikely to have occurred due to chance or sampling error, then the difference is statistically significant.

Mathematical differences. By definition, if numbers are not exactly the same, they are different. This does not however suggest that the difference is either important or statistically significant.

Managerially important differences. If results or numbers are different to the extent that the difference would matter from a managerial perspective, we can argue that the difference is important. For example, the difference in consumer response to two different packages in a test market might be statistically significant but yet so small as to have little practical or managerial significance.

G. The t-test (optional chapter 26.6)

When you are in small sample situations (usually less than size 25) and the population SD is unknown, the z-test must be modified. So we have something called a t-test.

It is like the z-scores you learned in Chapters 21 and 23 except the difference is in the calculation of the "SD of the box". You really aren't allowed to just substitute the SD of the box with the sample SD when your samples are small (below 25). And you can no longer use the normal curve on page A-105 but instead you must use the t-table on page A-106 to read off areas and get p-values. Furthermore, you need to find something called "the degrees of freedom" to help you locate the correct t-values. The degrees of freedom are calculated as the sample size minus 1.

The new SD is simply SD+ as described in Chapter 4.7. Or for some of you, it's the SD your calculator gives automatically.

An example.

Suppose from the population of UCLA male undergraduates, a random sample of size 16 is picked and each male's height is measured. Suppose for this random sample, the average height is 68 inches with a standard deviation of 2 inches. The Chancellor claims that UCLA men are as tall as USC men (who average 69 inches in height). Is his claim correct based on the evidence from this sample?

The corrected SD = 2.0656 (this is SD+ in your text)

Test is t now and t looks a lot like Z:

$$t = \frac{\text{observed average} - \text{hypothesized average}}{\text{Standard Error of the average}} = \frac{68 - 69}{\frac{\sqrt{16 * 2.0656}}{16}} = -1.9465$$

with 16-1=15 degrees of freedom. This falls between 5% and 2.5% on the t table. If your rule of thumb was a 5% level of significance, we'd say that the chancellor's claim is not correct. We would reject the null hypothesis so the difference we are seeing is not due to chance. There is evidence that UCLA men are shorter.

If your rule of thumb was 1%, then we'd say would not reject the null so the difference we are seeing is due to chance and there is no evidence that UCLA men are shorter.

Generally, you are asked to either state a level of significance or are given a level of significance. Perhaps you can see why here...if we didn't like the results at the 5% level...we could say "oh, we don't use a 5% level, we use a 1% level"