A. Chapter 26.1-26.4: Overview

The basic idea on chapter 26.1-26.4: we make assumptions about the value of the parameters, and then test to see if those assumptions could have led to the outcomes (statistics) we observed. We then use a probability calculation (a Z score and areas) to express the strength of our conclusions.

The basic question: Suppose you have a sample outcome (sample statistic) which is different from the expected value (population parameter) was the difference in the sample outcome we observed due to chance error or something else?

Let's walk through the example at the beginning of Chapter 26 together.

> A senator introduces a bill to simplify the tax code. His claim is the bill is revenue-neutral. Basically it won't change the amount of taxes the government collects, it just simplifies the law.
>
> A law can be evaluated. The IRS could SAMPLE from the POPULATION of all tax returns, figure out the effect the bill would have on these revenues, and then check to see if the bill is really revenue-neutral.
>
> In the example, the IRS samples 100 forms. The sample average comes out to -$219 which means that the government would have collected 219 fewer dollars from taxpayers. The sample standard deviation is $725.
>
> The senator's argument (issued through an aide) is that the SD is so large, $725, that an average of -219 is inconsequential.
>
> The IRS's argument is what you want to learn. To understand the -219 and the 725, you need to convert the sample SD to an SE for the sample average.
>
> Remember what the SE is, it is variation association with sample statistics. The SD is the variation in a given sample (e.g. the 725 here) or the variation in a population, or in a list (see Chapter 4).
>
> The IRS goes on to say, the senator may think/argue that the population parameter is $0, but the IRS thinks it's not zero and in fact it is also negative.

How do they figure that?

First, they calculate an SE for the sample average

$$\frac{\sqrt{100} * 725}{100} \approx \$72$$

Second, they set up a "test" and use the Z score as the "test statistic"

$$\frac{-219 - 0}{72} \approx -3.0 = Z$$

This test statistic Z is interpreted in this manner in chapter 26: that if the true parameter was zero dollars and the samples of size 100 have a variation (standard error) of $72 then the chance (probability) that you

could have picked a sample of size 100 with a mean of -219 is about 0.1 of 1% which is the area to the left of -3 under the normal curve.

In Chapter 26, when using the Z score, you are always interested in the area in the "tail" which is as extreme or more extreme than your statistic. So in this case, we are looking at areas beyond –3 Z scores.

In previous chapters, you learned to work with the formula for a Z score and the normal curve. In Chapter 26, it all comes together.

The basic question restated: one side thinks any observed difference between what you expect and what you get is REAL (so perhaps something is wrong with the expected value or with the sample process). The other side thinks the difference is just random chance error operating.

If your observed value is too many STANDARD ERRORS away from the expected value, this is hard to explain by chance alone. For example, here, we are seeing a chance of .1% which is very small, this means you could have gotten this outcome randomly only once in 1000 samples. The number of standard errors away is called a Z score and the method you use to arrive at this score is a "test of significance"

B. Vocabulary

The **NULL HYPOTHESIS (26.2)** is that the observed results are purely due to chance alone. That is, any differences between the parameter (the expected value) and the observed (or actual) outcome are due to chance only. In this case, the null hypothesis is a statement about a parameter: the population average is 0 for the IRS example in 26.1.

The **ALTERNATIVE HYPOTHESIS (26.2)** 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 wish to knock down.

The **TEST STATISTIC (26.3)** measures how different the observed results are from what we would expect to get if the null hypothesis were true. When using the normal curve, the test statistic is z,

$$Z = \frac{observed\_statistic - hypothesized\_value}{standard\_error}$$

All a Z does in Chapter 26 is tell you how many SEs away the observed statistic is from the expected (hypothesized) value when the expected (hypothesized) value is generated from the **NULL HYPOTHESIS**.

The **SIGNIFICANCE LEVEL (or P-VALUE) (26.3).** 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."

So in our case: if the bill is truly revenue-neutral, there would be less than 0.1% chance to get a result of –219 from a sample of 100 returns.

If the p-value is less than 5%, we say the results are **STATISTICALLY SIGNIFICANT (26.4)**; 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.

C. Hypothesis Testing Summarized

1. Clearly identify the parameter (expected value) and the observed outcome (sample statistic).

2. State the null hypothesis. This is what is being tested. A test of significance assesses the strength of evidence (observed outcomes) against the null hypothesis (expected value). Usually the null hypothesis is a statement of "no-effect" or "no difference"

3. The alternative hypothesis is the claim about the population that we are trying to find evidence in favor of. In the tax law example, you are seeking evidence that the law is not neutral. The null hypothesis would say that the average return will not change (i.e. 0) , the alternative would say it is negative . Note this is a ONE-SIDED alternative because you are only interested in deviations in one direction. (Advanced: A two-sided situation occurs when you do not know the direction, you just think the evidence suggests something different from the null)

4. The test statistic. It is the statistic that estimates the parameter of interest. In the above example, the parameter is the population average and the outcome is the sample average and the test-statistic is Z. (Advanced: other test statistics exist, the method is the same, see Chapter 26.6 if you plan to take more courses involving statistics)

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

To answer these types of questions, you find the probability of getting an outcome as extreme or MORE than you actually observed. So to test the outcome of -219, you would ask "what is the chance of getting a -$219 or lower number?"

5. The probability that you calculate from your Z score is called a P-VALUE. The smaller the p-value the stronger is the evidence against the null hypothesis. Suppose instead you had gotten a sample average of -$100 (instead of –219) with the same SE, the senator may well be right. (A Z of about -1.4 has about 8% of the normal, so here, there was an 8% chance of getting a sample with an average of -$100 or lower).

6. On significance levels. Sometimes prior to calculating a score and finding it's 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 level --- "

NOTE: Significant is not the same as important. All it means is that the outcome you observed probably did not happen by chance.

D. One more example

The following letter appeared in the "Dear Abby" column in the 1970s:

Dear Abby,

You wrote in your column that a woman is pregnant for 266 days. Who said so? I carried my baby for 10 months and 5 days (Prof's note: that's 310 days) and there is no doubt about it because I know the exact date my baby was conceived. My husband is in the Navy and it couldn't have been conceived at any other time because I saw him only once for an hour, and I didn't see him again until the day before the baby was born.

I don't drink or run around, and there is no way this baby isn't his, so please print a retraction about the 266 day carrying time because otherwise I am in a lot of trouble.

San Diego Reader

OK....suppose it is known that pregnancy durations are normally distributed with a mean of 266.0 days and a standard deviation of 16.0 days.

A chapter 23-like question might be: what is the chance of observing a single pregnancy duration of 310 days or more?

$$Z = \frac{310 - 266}{\frac{\sqrt{1} * 16}{1}} = +2.75$$ and that translates to a 0.3% chance in the tail area

Remember, this is like a sample of size one and you know the population parameters. A pregnancy as long as 310 days can happen, and the chance is about 3 in 1000 pregnancies.

In chapter 26, when we are calculating chances, we're thinking more along the lines of larger samples and of trying to make a decision -- choosing between hypotheses.

A chapter 26-like question: In a recent study of 100 pregnancy durations selected at random, the average pregnancy duration was 270 days with standard deviation of 20 days. Does this study provide evidence that pregnancy durations have increased since the 1970s? Perform a test of significance and state the p-value.

The null hypothesis is 266 days. (we expect the average woman to have a pregnancy like this)
The alternative is something longer than 266 days (because we're not exactly sure, we just think it's longer)
This is a one-side test, we're only interested if the durations are longer now
The test statistic is

$$Z = \frac{270 - 266}{\frac{\sqrt{100} * 16}{100}} = +2.50$$ and that translates into a 0.62% in the tail area

This would suggest that the probability of getting a sample average of 270 if the true average is 266 is about 6 times in 1000. This is evidence for the alternative, that is, that durations are getting longer.

Things to note: why did we use Standard Deviation of 16 and not 20?