STA 111: Probability & Statistical Inference

Lecture Fourteen – Hypotheses Testing
D.S. Sections 9.1, 9.2, 9.4 & 9.6

Instructor: Olanrewaju Michael Akande

Department of Statistical Science, Duke University

Outline

- Questions from Last Lecture
- Introduction
- Hypotheses Tests
- Examples
- Recap

Introduction

– So far in the second half of this class, we have been talking about making inference on a population using observed data.

– In line with that, we have talked about building confidence intervals for population parameters by relying on central limit theorem and bootstrap (in the lab).

– Today we will extend that discussion to what is called hypotheses testing. Loosely speaking, hypotheses testing is asks the question: how strongly does the data support my preconceived hypothesis about a population parameter?

Introduction

– This is the same as asking: if my preconceived hypothesis about a population parameter is true, how extreme is the data I just observed? As we will see, there is a duality between confidence intervals and hypothesis testing.

– We will then look at some examples to understand how hypotheses testing works formally.

Hypotheses Tests

A hypothesis test (significance test) is a way to decide whether the data strongly support one point of view or another.

There are many kinds of significance tests, but all involve:

- a null and alternative hypothesis
- a test statistic
- a significance probability (P-value).

The following gives an overview of most of the different kinds of significance tests.

Hypotheses Tests

Step 1: Pick the null and alternative hypotheses.
The null and alternative are two contradictory statements about a parameter and their union is the set of all possible parameter values. For example:

$$H_0 : \theta \leq 90 \qquad \textbf{vs.} \qquad H_1 : \theta > 90$$

where θ is a generic parameter. Usually, the null hypothesis H_0 is a current belief while the alternative hypothesis H_1 or H_A is the one that leads to new action, or the outcome you would like to prove wrong.

Example 1: H_0: The mean cable strength $\geq$ 3 tons.
H_1 : The mean cable strength $<$ 3 tons.

Example 2: H_0 : The sd in income $\leq$ \$5,000
H_1 : The sd in income $>$ \$5,000.

In example 1, accepting that mean cable strength $<$ 3 tons probably leads to a new action and for example 2, sd in income $>$ \$5,000 leads to a new action.

Hypotheses Tests

Step 2: Calculate the test statistic.

The test statisic is a one-number summary of all the information in the sample regarding the correctness of the alternative hypothesis Different kinds of hypothesis tests (e.g., about means, proportions, differences of means, differences of proportions, etc.) require different test statistics.
Soon we shall list many standard cases.

Step 3: Find the *P*-value (or significance probability).

Use a table to find the *P*-value. This is **"the probability of obtaining data that is as or more supportive of the alternative hypothesis than the data that were observed, when the null hypothesis is correct."**
This interpretation of the *P*-value is a bit subtle.

Hypotheses Tests

I: The Three Possible Pairs of Null and Alternative Hypotheses

1. $H_o: \theta = \theta_o$ versus $H_A: \theta \neq \theta_o$ (Simple null hypothesis, two-sided alternative).
2. $H_o: \theta \leq \theta_o$ versus $H_A: \theta > \theta_o$ (Composite null hypothesis, one-sided alternative).
3. $H_o: \theta \geq \theta_o$ versus $H_A: \theta < \theta_o$ (Composite null hypothesis, one-sided alternative).

Here θ represents a generic parameter. It could be a population mean, a population proportion, the difference of two population means, or many other things.

The θ_0 is the **null value**. Often it is a value specified in a contract, regulation, or clinical trial.

Hypotheses Tests

II Possible Test Statistics In this class, we are only going to consider simple test statistics of the form

$$ts = \frac{pe - \theta_0}{se} = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$$

Where

- pe or $\hat{\theta}$ is the point estimate for θ.
- θ_0 is the null value.
- se is the standard error of $\hat{\theta}$.

Clearly, when $\hat{\theta}$ is an average (or sum) and θ_0 is true, the law of large numbers and the central limit theorem kicks in, so that ts is like a "z-transformation" and it has a standard normal or student-t distribution and we can find the probability that it exceeds some critical value.

Hypotheses Tests

a. For a test on the population mean we take θ to be the population mean μ. If you know the population σ, or for $n > 31$ with

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum(X_i - \bar{X})^2}$$

as an estimate of the σ, then you get the significance probability from a z-table and the test statistic is:

$$ts = \frac{\bar{X} - \mu_o}{\hat{\sigma}/\sqrt{n}}.$$

b. For the previous case, if you have a sample of size $n \leq 31$ and you use $\hat{\sigma}$ to estimate the population σ, then the significance probability comes from a t_{n-1} table and again the test statistic is:

$$ts = \frac{\bar{X} - \mu_o}{\hat{\sigma}/\sqrt{n}}.$$

Hypotheses Tests

c. For a test about a proportion, $\theta = p$. The significance probability comes from a z-table and the test statistic is:

$$ts = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}.$$

d. For a test of the difference of two means, $\theta = \mu_1 - \mu_2$. Assuming that the sample sizes from each population satisfy $n_1 > 30$ and $n_2 > 30$, then the significance probability comes from a z-table and the test statistic is:

$$ts = \frac{(\bar{X}_1 - \bar{X}_2) - \theta_o}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

If $n_1 > 30$ and $n_2 > 30$ isn't satisfied, use a t-table with $n_1 + n_2 - 2$ degrees of freedom.

Hypotheses Tests

e. For a test of the difference of two proportions, take $\theta = p_1 - p_2$. Use a z-table for the significance probability and the test statistic:

$$ts = \frac{(\hat{p}_1 - \hat{p}_2) - \theta_o}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

f. For $n > 30$, with $\theta = \mu_1 - \mu_2$, and n paired differences $X_i - Y_i$, use the z-table for the significance probability. The test statistic is:

$$ts = \frac{(\bar{X} - \bar{Y}) - \theta_o}{\hat{\sigma}_d / \sqrt{n}}$$

Here $\hat{\sigma}_d$ is the estimated standard deviation of the n differences. This is called a **paired difference test** and can have greater power than the two-sample tests in II.d and II.e. We will talk about what power means later. If $n \leq 30$, use a t-table with $n - 1$ degrees of freedom.

Hypotheses Tests

III Significance Probability or P-value

The significance probability of the test statistic depends upon the hypothesis chosen in Part I. For that choice, let W be a random variable with a z or t_{n-1} distribution, as indicated in Part II. Then for each possible pair of null and alternative hypotheses in Part I,

1. The significance probability is $\mathbb{P}[W \leq -|ts|] + \mathbb{P}[W \geq |ts|]$.
2. The significance probability is $\mathbb{P}[W \geq ts]$.
3. The significance probability is $\mathbb{P}[W \leq ts]$.

The significance probability is **THE CHANCE OF OBSERVING DATA THAT SUPPORTS THE ALTERNATIVE HYPOTHESIS AS OR MORE STRONGLY THAN THE DATA YOU HAVE SEEN, WHEN THE NULL HYPOTHESIS IS CORRECT.**

Hypotheses Tests

With all the pieces, a decision can be reached on whether to reject the null hypothesis if there is too much evidence against it, or to fail to reject the null hypothesis if there isn't enough evidence to suggest it is false.

Note: We don't actually "accept the alternative hypothesis" since intuitively, our decision is based on a sample from the population. We could always collect more data and what if we find that we should in fact reject the alternative hypothesis after we accepted it?

Thus, formally, for each possible pair of null and alternative hypotheses in Part I, reject H_0 if the significance probability or p-value is less than a chosen error rate α (usually 0.05 or 0.01).

Remember that $\alpha = 1 - C$ was the error rate when we talked about confidence intervals?

Examples

Example 1: Suppose you have a new oil additive that may extend the life of an engine. You give it to 25 random motors and find that the average lifespan is 78 months, and the standard deviation in their lifespan is 12. You know that with unmodified oil, the mean lifespan is 72 months, and hope to show that your additive improves that.

The first step is to choose the null and alternative hypotheses. You put what you want to prove in the alternative, so this is case I.2 of the previous taxonomy:
H_0: The mean lifetime with the new additive $\leq$ 72 months.
H_A : The mean lifetime with the new additive $>$ 72 months.

The second step is to find the test statistic. We are in case II.b of the taxonomy, so:

$$ts = \frac{\bar{X} - \mu_0}{sd/\sqrt{n}} = \frac{78 - 72}{12/\sqrt{25}} = 2.5$$

Examples

The third step finds the significance probability. If you use hypotheses I.**2**, then you use rule III.**2**.

The significance probability, or P-value, is $\mathbb{P}[t_{24} > 2.5]$ and from the t-table in the book, 2.5 isn't actually on the table but we can provide bounds for it (note that we need to find 2.5 on the table and find the corresponding α value). Thus,

$$0.01 = \mathbb{P}[t_{24} > 2.492] > \mathbb{P}[t_{24} > 2.5] > \mathbb{P}[t_{24} > 2.797] = 0.005.$$

So if the null hypothesis is true and the additive does not help, then you have between a 1% chance and a 0.5% chance of observing the result in your experiment, which is so rare. This is pretty persuasive that the additive helps.

Also, if we let $\alpha = 0.05$, then the p-value is less than α and we can reject the null hypothesis that the new additive does not help and conclude that it does help.

Examples

Example 2: From decades of experience, the statistics department knows that 30% of students fall asleep in class. One of the professors wants to prove that his teaching is more lively. Suppose he collects data and out of 120 students, only 35 fall asleep. Is this evidence that his teaching is better?

First, we must find the null and alternative hypotheses. We put what he wants to show in the alternative.

$$\mathbf{H_0}: \ p \geq 0.3 \qquad \textbf{vs.} \qquad \mathbf{H}_A : p < 0.3$$

Next, we must find his test statistic, the one-number summary of all the information in the sample regarding the null hypothesis.

$$ts = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\frac{35}{120} - .3}{\sqrt{\frac{.3*.7}{120}}} = -.1992.$$

Examples

Finally, we need to find the significance probability, or P-value. Since we are in case I.**3** for the hypotheses, we use the rule III.**3** to determine the significance probability.

From the standard normal table, the P-value is:

$$\mathbf{P}[z < ts] = \mathbf{P}[z < -.1992] = .4211.$$

This result is not unlikely. Just by chance, 42% of the time he would get a result like this if his teaching were no better than anyone else's.

If $\alpha = 0.1$, we fail to reject the null hypothesis and conclude that his teaching were no better than anyone else's.

Working with the table is a crutch. The relationship between parts I and III are clear when you think about them: you are trying to decide whether your sample statistic is improbably different from the null, improbably larger than the null, or improbably smaller than the null, respectively.

Examples

Example 3a: You can purchase automobile tires from Firestone or Dunlop. You want to decide whether there is any difference in the way their tires wear.

You get 100 tires from Firestone and 100 tyres from Dunlop. You install them on your firm's vehicles, and after six months, measure their tread wear. You find that the sample average wear for Firestone is 4.75 mm, with a sample standard deviation of 1 mm, and the average lifespan of the Dunlop tyres is 5 mm, with a standard deviation of 2 mm.

What are your null and alternative hypotheses?

$$H_0 : \mu_F - \mu_D = 0 \qquad \textbf{vs.} \qquad H_A : \mu_F - \mu_D \neq 0$$

This is case I.1 among the hypotheses.

Examples

Your test statistic is from II.**f**:

$$ts = \frac{\bar{X}_F - \bar{X}_D - 0}{\sqrt{\frac{\hat{\sigma}_F^2}{n_F} + \frac{\hat{\sigma}_D^2}{n_D}}} = \frac{4.75 - 5}{\sqrt{\frac{1}{100} + \frac{4}{100}}} = -1.118.$$

To find the P-value, you go to a z-table. Since the hypotheses are case I.**1**, we use the signficance probability from case III.**1** to find

$$\mathbb{P}[Z \leq -1.118] + \mathbb{P}[Z \geq 1.118] = 0.267.$$

Since the significance probability is larger than the conventional values for α (0.05, or even 0.01) we fail to reject the null hypothesis. We do not have sufficient evidence to decide that tires wear differently from tyres.

Examples

Example 3b: You can purchase automobile tires from Firestone or Dunlop. You want to decide whether there is any difference in the way their tires wear.

You get 100 tires from Firestone and 100 tyres from Dunlop. **You install one of each on 100 cars, randomly assigning one to the left front wheel and one to the right front wheel. This controls for differences in use among cars in the fleet.**
You find that the average difference in wear between Firestone and Dunlop is 0.25 mm, with a sample standard deviation of 0.1 mm.

What are your null and alternative hypotheses?

$$H_0 : \mu_F - \mu_D = 0 \qquad \textbf{vs.} \qquad H_0 : \mu_F - \mu_D \neq 0$$

This is case I.1 among the hypotheses.

Examples

Your test statistic is

$$ts = \frac{\bar{X}_F - \bar{X}_D - 0}{\frac{\hat{\sigma}_d}{\sqrt{n}}} = \frac{-0.25}{\frac{0.1}{\sqrt{100}}} = -25.$$

To find the P-value, you go to a z-table. Since the hypotheses are case I.**1**, we use the signficance probability from case III.**1** to find

$$\mathbb{P}[Z \leq -25] + \mathbb{P}[Z \geq 25] \approx 0.$$

Here we reject the null hypothesis. There is strong evidence that there is a difference in average wear. The pairing of the tires on the cars increased the power of the test.

Recap

Today we learned about hypotheses testing and how to setup testing problems.

The duality between confidence intervals and hypotheses testing isn't emphasized in this lecture but you a few things should seem familiar to you from confidence intervals especially α.

In the next lecture, we will learn about what α really means and why we use it as a cut-off in hypotheses testing (and confidence intervals).