Hypothesis Tests

Hypothesis tests are the method classical statisticians use to make decisions.

Hypothesis testing is one of the two main techniques of classical statistical inference (the other is estimation and confidence intervals.) Problems for both confidence intervals and hypothesis tests will be on the final. You will see practice problems on inferential statistics for the final on my website***. Hypothesis tests have several steps and the logic behind them is sophisticated and indirect. Be sure to practice these problems a lot!***

Terminology

Hypothesis tests have several steps, each of which has some terminology involved with it. You will be expected to know the following terminology (good True/False question candidates), and how to calculate and use them when relevant:

- Null hypothesis
- Alternative hypothesis
- Type I error
- Type II error
- α (also known as the significance level or Probability of Type I error)
- β (also known as the probability of Type II error)
- Power of a test (Probability of rejecting the null hypothesis when it is actually false, also given as $1-\beta$)
- Test statistics (Z or t)
- Rejection region
- one-tailed test
- two-tailed test
- p-value (observed significance level)

As with confidence intervals, you will need to know when you should be using Z-tables to look up probabilities and when you should be using t-tables instead. You should also be able to draw the appropriate box model for the problem, and the appropriate Z or t curve for your calculations.

Example 1 – Fake, but hopefully easy to understand

Suppose you bump into an old friend in Las Vegas, someone you haven't seen in a long time. Turns out he's been living and gambling in Las Vegas for a few years now. He wants to play a game with you, in which he gets $10 from you for every head that turns up when you flip a coin, and you get $10 from him for every tail that turns up. He hands you a coin, and you agree to flip the coin 25x.

Suppose 20 heads turns up – 80% heads! You start wondering if the coin might be biased somehow. You have to make a decision whether to call him a cheater or let it go.

Actually, you could just accuse your friend of cheating, but you don't want to lose a friend just in case he happened to be lucky. On the other hand, if he's cheating, you don't particularly want him as a friend. We can lay this out formally as a decision table:

Decision/Reality	Friend not cheating	Friend cheating
Decide friend is cheating		
Decide friend is not cheating		

How can you make this decision? Intuitively, you feel he just has gotten too many heads. You remember somehow from that statistics class you thought so useless that you could actually calculate the probability of getting 20 heads or more.

But wait a minute. How do you do that? Wasn't there something the prof kept saying you had to assume about the coin? That it was _____? If it's _____, that means the probability of getting heads on one toss is _____. There was a box model she insisted on drawing. The box for this game should look like:

And she always was putting down something about μ's or σ's or p's for the box. What should it be for this box?

The null hypothesis is the value of the parameter (in our case, either μ or p) that describes the distribution of the box. Remember that the box represents the true population that you are sampling from. You want to say something about the population (box), not about the sample itself.

The alternative hypothesis also says something about the value of the box, usually that it is something other than the null hypothesis.

Hypotheses should be set up in words first, then the number assigned to the relevant parameter for the box.

Here's a big question: How do we know which is the null and which is the alternative hypothesis?

There are two general criteria used to set up the null and alternative hypotheses. The first is that the null hypothesis is the *status quo* – that is, the usual situation. The alternative then is that the status quo doesn't hold – the usual situation has changed. **The second criterion often used is that the alternative hypothesis is what you'd like to prove, and the null is the opposite of that.**

Why the second criterion ? It is not sportsman-like to assume what you're trying to show. You don't assume your friend is a thief and then try to show that. What you do is to give him the benefit of the doubt. In this case, it is the same as saying you expect the status quo: the coin isn't special – it's just any old coin.

You can also think of the second criterion in terms of the US legal system: a person is innocent (null hypothesis) until proven guilty (alternative hypothesis). The prosecution wants to show the person is guilty, but has to work from the legal assumption that the person is innocent.

What's the rationale behind the weird hypothesis test logic?

The idea is this: You can't assume what you want to show. That's cheating. Besides, it makes you look better if you give the opposition the benefit of the doubt, and then show that even after they got the benefit of the doubt, the evidence is more in your favor. Again, think of the US legal system. The idea is that we assume that the person is innocent until the evidence mounts up to the point where we just can't believe that anymore. Only our evidence is statistical and deals with numbers.

Choosing the significance level

We've got the null and alternative hypotheses in place, now we have to evaluate how strong the evidence has to be. This needs to be done before the test is conducted. Remember that it's possible (though intuitively rare) for your friend to get 20 heads in 25 tosses of a coin, even if the null hypothesis is true. How much do you value his friendship by calling him a cheat when he's really not?

This type of mistake, "rejecting the null hypothesis when it is really true" is called Type I error. The probability of Type I error is called the significance level and is traditionally denoted by α. The traditional levels used for α are: 10%, 5%, and 1%, with the 5% level being the most used. What this means is that you are willing to be wrong about 1 time in 20 using this rule, or that there's a 5% chance of calling your friend a cheat when he really isn't. The lower the significance level you choose, the less you want to make a Type I error. However, you are increasing your probability of Type II error – of saying

he's not a cheat when he really is a cheat. The point is, you don't know what the truth really is, and you are trying to make educated decisions.

In our present example, we will chose a significance level of _____, from class vote.

Setting up the rejection region

Now we have to get down to figuring out the rule in which we'll say your friend is a cheat or not. We have an estimate of the parameter _____, so we'll use the Central Limit Theorem to say that _____of _____ should be a _____ Curve.

We draw it. What should the center be? How do we measure the variation?_____ _____ What value should this have? _____

In which direction do we have evidence that the null hypothesis isn't true, and the alternative is? _____

Given our significance level α, what should our cutoff value be? _____ Color this in on the curve.

Can we state a rule that gives this? What is it?

Calculating the test statistic

We have the rejection region, and now we want to see if our data ends up in the rejection region or not.

We see that the result of this calculation falls inside/outside our rejection region. So we reject/do not reject the null hypothesis in favor of the alternative.

The way to interpret a statistic falling inside the non-rejection region is: The difference we see between our estimated value of the parameter and the hypothesized value of the parameter is copnsistent with chance variation.

If the statistic falls within the rejection region, we say that there is evidence that the difference is due to something other than chance variation. (However, we cannot state exactly what is causing that difference.)

Not quite done – the p-value

Earlier we chose the significance level based on our own personal acceptance of Type I error for this problem. But other people might have different acceptance levels of Type I error for this problem. In the old days, hypothesis tests were just reject/not reject. **Nowadays, we calculate the 'observed significance level,' typically called the 'p-value.' The p-value is a measure of the strength of the evidence against the null hypothesis. It allows people to do hypothesis tests using their own personal significance levels (**α's).

The way the p-value is calculated is to find the α that would correspond to having your rejection region be determined by your test statistic. (That's why it's called 'observed significance level.') If you look at papers using statistics to make decisions in medicine, business, etc., they will always include p-values as part of their results.

How do we calculate the p-value for this example?

Rule for using p-values for hypothesis tests: if the p-value is less than the significance level α, then you reject the null hypothesis.

Some rules of thumb about p-value interpretation:

.05 < p <= .10 – there is evidence against the null hypothesis
.01 < p <= .05 – there is strong evidence against the null hypothesis
p < = .01 - there is very strong evidence against the null hypothesis

Let's go through this again, with another example

Example 2 – Family income

Suppose that the average family income for students at a state university is $45,000 per year. A benefactor wants to set up a scholarship fund for students in financial need from the Department of Education, because she feels that their average family income is less than that of the university as a whole. However, she wants to test that out before she commits herself. She has the university administration pull 25 students' records at random from the Education Department list. They find that the average income of these students is $44,500, with an SD of $500. Can she then feel justified in establishing her scholarship for this department, or should she continue looking?

Step 1.

Set up the alternative and null hypotheses, in words and in numbers.

What is the box? Does this make sense?

Step 2. Draw the curve that is necessary. Should it be a Z or a t curve?

Step 3. What is the amount of Type I error that can be tolerated here?

Step 4. What is the rejection region? Should it be one-sided or two sided

Step 5. What is the test statistic using these data?

Step 6. Do we reject the null hypothesis? Or not? Why?

Step 7. What is the p-value? (What should the region of the p-value look like?)

Example 3. More variations on hypothesis tests

Suppose you are in charge of testing if the production line for a cereal manufacturer is filling boxes accurately. According to the box, there should be 20 oz of cereal. Historically, the standard deviation for the process has been 0.12 oz. You take a random sample of 36 from the day's production, and find that the average weight is 19.9 oz. Should your line be adjusted, or is this within chance variation?

Set up and do a hypothesis test, and pay special consideration about how the null and alternative hypotheses should look. Be sure to calculate the p-value.

Steps for hypothesis tests

1. What is the population you are trying to say something about? Draw a box model.
2. Decide what the parameter of interest is (that is,μ or p) for this population.
3. What are the null and alternative hypotheses about this parameter, in words and translated into math.
4. Decide whether a Z or a t is the appropriate test statistic, depending on the information you have available.
5. Choose an appropriate level of significance (α).
6. Draw the curve, and find the rejection region(s) and rejection rule, based on α and the alternative hypothesis.
7. Calculate the test statistic.
8. Compare the test statistic to the rejection rule/region.
9. Decide whether to reject the null hypothesis or not.
10. Calculate the p-value.