

## Hypothesis testing

So far, we've talked about inference from the point of estimation. We've tried to answer questions like "What is a good estimate for a 'typical' value?" or "How much variability is there in my best guess for a future value?" We've also noted, however, that sometimes questions are better phrased in the form of a claim, or a *hypothesis*. Questions like "Is the true mean very different from 100?" are statements about unknown parameters, and **investigating whether the data support a claim is called hypothesis testing**.

Hypothesis testing arises often in practice. Here are a few examples:

1. When a drug company wishes to introduce a new drug, they must first convince the Federal Drug Administration (FDA) that it is a worthwhile addition to the current set of drugs. That is, they must be able to demonstrate that the new drug is better than current drugs, with "better" possibly referring to greater effectiveness, fewer side effects, or quicker response by the patient. There are risks in introducing a new drug — is it worth it?
2. When an advertiser is considering launching a new ad campaign, they are first interested in knowing if the new campaign would be an improvement on the current campaign. Will it alienate current customers? Is it worth the cost?
3. Many advertisements compare one product to another, stating that their product is "better." The Federal Communications Commission (FCC) and Federal Trade Commission (FTC) require that these claims be backed up. How do we do that?

In statistical hypothesis testing, hypotheses of interest are converted to statements about unknown parameters (e.g., the mean survival rate using a new drug is higher than the mean survival rate using the old drug).

The hypothesis testing framework is characterized by the distinction between two kinds of hypotheses: the *null* hypothesis ( $H_0$ ) and the *alternative* hypothesis ( $H_a$ ). **The null hypothesis is the standard; it is the statement that we are going to believe unless it is proven otherwise. That is, the null gets the benefit of the doubt. The alternative hypothesis (sometimes called the *research* hypothesis) has the burden of proof; we are not going to believe it unless the data satisfy that burden.** Now consider the examples above:

1. FDA guidelines say that the FDA doesn't have to prove that a new drug is unsafe; rather, the manufacturer of a new drug must prove that the drug is safe and effective. This emphasis on safety has resulted in many drugs ultimately found to be dangerous not being made available in the United States; the anti-nausea drug thalidomide,

which caused many serious birth defects around the world in the early 1960's, is a prime example of this. Thus, FDA guidelines result in hypotheses

$$H_0 : \text{The new drug is not better than the current one}$$

and

$$H_a : \text{The new drug is better than the current one.}$$

2. What do you think?
3. Comparative advertisements generally take one of two forms. One is the *superiority claim*, an example of which is “New improved Sani-Flush cleans better than Vanish.” The makers of Sani-Flush made this claim, and by doing so, they put the burden of proof on their product to demonstrate that it is, in fact better. Thus, the relevant hypotheses are

$$H_0 : \text{Sani - Flush does not clean better than Vanish}$$

and

$$H_a : \text{Sani - Flush cleans better than Vanish.}$$

On the other hand, another type of claim is the *unsurpassed claim*, an example of which is “Nothing's better than Aleve for arthritis pain.” When the makers of Aleve made this claim, they put the burden of proof on other manufacturers to show that Aleve wasn't as effective as other products. Thus, the relevant hypotheses are

$$H_0 : \text{Aleve is as good as other pain relievers}$$

and

$$H_a : \text{Aleve is not as good as other pain relievers.}$$

Obviously, unsurpassed claims are a lot less likely to be challenged than superiority claims, and for that reason they are much more common.

Once we have developed this structure, there are two possible “states of nature”: the null is true, or it is false. There are also two possible decisions: reject the null as not being true, or don't reject the null (note that we don't really ever “accept” the null — since it gets the benefit of the doubt, the best we can say is that we haven't heard enough to reject it). This leads to four possibilities:

	Don't reject $H_0$	Reject $H_0$
$H_0$ is true	Correct	Type I error
$H_0$ is false	Type II error	Correct

In this table, not rejecting  $H_0$  when it is true and rejecting  $H_0$  when it is false are obviously correct decisions. **Rejecting  $H_0$  when it is in fact true is an error (a *Type I error*), and can be viewed as representing a “false alarm.” Not rejecting  $H_0$  when it is in fact false is also an error (a *Type II error*), and can be viewed as representing a “missed opportunity.”**

*Example:* Consider the U.S criminal justice system. The hypotheses refer to the guilt or innocence of the defendant. Which is the null? Which is the alternative? What would be a Type I error? What would be a Type II error? Which is more serious? Which type of error does the criminal justice system try to control, even if that makes the other type of error more likely? Can you imagine a legal situation where both errors might be viewed as equally important, and we might focus on minimizing  $P(\text{Type I error}) + P(\text{Type II error})$ , for example? How does this relate to the Scottish system, where there are three possible verdicts, guilty, not guilty, or not proven?

How do we implement these ideas in practice? The key is to recognize the connection with confidence intervals. If we were testing a null hypothesis that  $\mu = 1000$ , for example, and based on observed data constructed a 95% confidence interval for  $\mu$  that ended up being (973, 1034), presumably we would conclude that the data *were* consistent with  $H_0$ , since 1000 is in that interval. That is, we would not reject the null hypothesis. On the other hand, if a constructed 95% confidence interval was (1038, 1099), presumably we would be more likely to conclude that  $\mu$  is not equal to 1000, but rather was greater than 1000, and we would reject the null hypothesis. **That is, to construct a hypothesis test, just invert a confidence interval.**

We can formalize this. **A hypothesis test constructed to test the hypotheses**

$$H_0 : \mu = \mu_0$$

**versus**

$$H_a : \mu \neq \mu_0$$

**rejects the null hypothesis at the  $\alpha$  significance level if a constructed  $100 \times (1 - \alpha)\%$  confidence interval for the mean does not include  $\mu_0$ . A test is said to be**

at the  $\alpha$  significance level if it has probability of Type I error equal to  $\alpha$ , which is the case here.

This rejection condition is met if either

$$\mu_0 < \bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \quad \text{or} \quad \mu_0 > \bar{X} + z_{\alpha/2}\sigma/\sqrt{n},$$

which is equivalent to rejecting the null hypothesis if

$$z = \left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}.$$

**That is, reject the null hypothesis if the observed  $z$ -statistic is larger than the critical value  $z_{\alpha/2}$ .**

So, in the earlier example, let's say  $n = 100$  days worth of data resulted in  $\bar{X} = 1021.3$ , and say we knew that  $\sigma = 100$  (this is of course unrealistic, and we'll deal with it in a little while, in the just the way that (hopefully) you'd expect). Then

$$z = \left| \frac{1021.3 - 1000}{100/\sqrt{100}} \right| = 2.13;$$

comparing this to  $z_{.025} = 1.96$ , we would conclude that *the observed average sales are statistically significantly different from 1000 units at a .05 level* (equivalently, we would *reject the hypothesis of true average sales per day of 1000 units at a .05 level*). Of course, since the observed sample mean is greater than 1000, we would think that in fact the true mean is greater than 1000.

Now, let's generalize things a few ways:

- (1) The test described is a *two-tailed test*, in that rejection of the null hypothesis would come from observing a value of  $\bar{X}$  that is either low enough (the left tail) or high enough (the right tail). We can imagine a situation where the alternative is *one-tailed* — where the only question of interest is whether the true mean is larger (say) than  $\mu_0$ . The FDA drug testing example given earlier, for example, could be viewed as one-sided, since (if  $\mu$  represented average effectiveness in some way) we are only interested in identifying if the new drug is better than the standard (that is,  $\mu > \mu_0$ ). That is, we are testing the hypotheses

$$H_0 : \mu \leq \mu_0$$

versus

$$H_a : \mu > \mu_0.$$

The arguments used earlier still apply, except that now we would only reject the null on the basis of surprisingly large values of  $\bar{X}$ . This means that Type I error could only occur in one tail (the upper tail), so we need to set the critical value to assure that this rejection region corresponds to the entire  $\alpha$  Type I error level. **This leads to the rule to reject the null hypothesis if**

$$\bar{X} > \mu_0 + z_\alpha \sigma / \sqrt{n},$$

**or, equivalently,**

$$z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} > z_\alpha.$$

How can we tell if a problem is one-tailed or two-tailed? It comes from the context of the problem. **If the question of interest is of the form “Is the sample mean significantly different from 100?” the implication is that either small or large values are of interest; this is two-tailed. On the other hand, if the question of interest is of the form “Is the sample mean significantly larger than 100?” the implication is that only large values are of interest; this is one-tailed.**

Having said this, however, we should note that some people think that all tests should be conducted as two-tailed. The reason for this is that it is easier to reject the null hypothesis using a one-tailed test (since  $z_\alpha$  must be smaller than  $z_{\alpha/2}$ ), so to give the null the biggest benefit of the doubt, the more difficult two-tailed criterion is used. The FDA requires all tests to be two-tailed, for example, even if from a scientific view only one tail is actually of interest.

- (2) We could just as easily have been in a situation where the alternative was consistent with lower values of  $\mu$ , rather than higher ones. This is handled in exactly the same way as before, except that now attention is focused on the left tail, rather than the right tail. The test is simply to reject the null hypothesis if

$$\bar{X} < \mu_0 - z_\alpha \sigma / \sqrt{n},$$

or, equivalently,

$$z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} < -z_\alpha.$$

- (3) The rigid decision “Reject the null / don’t reject the null” seems a bit extreme. After all, say that  $\bar{X}$  above turned out to be 1019. This is not statistically significant at a

.05 level, but it is close; it would be nice to be able to note that it was a lot closer to being significant than if  $\bar{X}$  had been, say, 1002. This can be quantified using the *tail probability*, or *p-value*. **The tail probability is simply the significance level that would be appropriate if the observed  $z$ -statistic was precisely significant.** Consider first a one-tailed test, where we are only interested in the alternative that sales are higher than expected. If  $\bar{X} = 1014.1$ , the  $z$ -statistic is

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{1014.1 - 1000}{100/\sqrt{100}} = 1.41.$$

The tail probability is then the significance level that would correspond to a critical value of 1.41; that is,  $P(Z > 1.41)$ , which is  $p = .079$ . **This value can be viewed as a measure of the strength of evidence against the null hypothesis; the smaller it is, the less likely a value of  $\bar{X}$  that large would occur if, in fact, the sales were only at the expected level. A tail probability of .079 says that only about 8% of the time, in repeated sampling, would the sample mean be as large as (or larger than) 1014.1 if the true average productivity was only 1000 or less. Smaller values of  $p$  indicate more evidence against the null hypothesis.**

Tail probabilities also can be defined for two-tailed tests. Say  $\bar{X}$  turned out to be 1024.7. The  $z$ -statistic is then 2.47, so the tail probability is

$$P(|Z| > 2.47) = P(Z < -2.47) + P(Z > 2.47) = .0068 + .0068 = .0136$$

(we need to look at both tails, since the alternative hypothesis is concerned with both). Note that this is a pretty small  $p$ -value, indicating pretty strong evidence against the null hypothesis (it's almost significant at a .01 level).

- (4) Let's now recognize the somewhat unrealistic assumption that we don't know  $\mu$ , but do know  $\sigma$ . What if we don't know  $\sigma$ ? Of course, Gosset already answered this question for us. **As long as the underlying population is reasonably close to Gaussian distributed, the  $t$ -statistic**

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

**should be compared to a  $t$ -distribution on  $n - 1$  degrees of freedom.**

- (5) After all of this talk of statistical significance, it's important to recognize that that's only part of the story. Of equal (or greater) importance is the issue of **practical**

**importance** of the results. Let's say that, in fact,  $\bar{X} = 1002$ . Does this result imply any important result for the company? Does it suggest any meaningful change in the revenue of the company? Probably not. Chances are, management wouldn't think that this achievement was particularly noteworthy. But, if we gathered enough data, this value of  $\bar{X}$  would indicate a significant difference from 1000 (say we measured 10000 days of data, rather than 100, and got  $\bar{X} = 1002$ ; the  $z$ -statistic is then

$$z = \frac{1002 - 1000}{100/\sqrt{10000}} = 2,$$

which is significant at a .05 level). The point is that **statistical significance only tells us if it is likely that  $\mu$  is different from 1000, but not if that difference is of any practical importance**. The latter point is not a statistical question; it comes from the context of the problem. In general, if the sample size gets much larger than a few hundred, this becomes an important issue that should not be ignored.

A way of countering the tendency to focus on statistical significance at the expense of practical importance is to focus on the *magnitude* of an effect, not just its statistical significance. Confidence intervals provide an effective way to do this. Hypothesis tests and confidence intervals *together* can help to overcome the weaknesses of each, and give a fuller picture of what is really going on in the data. Remember — **a small  $p$ -value implies strong evidence of an effect; a confidence interval centered far away from  $\mu_0$  provides evidence of a strong effect**.

- (6) These one-sample tests apply immediately to the situation where *pairs* of observations are obtained for each observation, and we are interested in the difference in those paired values. For example, in a market research setting, respondents' opinions on a product might be compared before and after seeing a commercial.

**This paired-samples problem is equivalent to a one-sample problem, where the single variable of interest is the difference between the “before” and “after” values.** The typical hypotheses tested are

$$H_0 : \mu_{Before} = \mu_{After}$$

versus

$$H_a : \mu_{Before} \neq \mu_{After},$$

where  $\mu_{Before}$  ( $\mu_{After}$ ) is the mean response value before (after) the intervention (drug, commercial, etc.). This is equivalent to testing the hypotheses

$$H_0 : \mu_{Difference} = 0$$

versus

$$H_a : \mu_{Difference} \neq 0,$$

which leads to a standard  $t$ -test using the difference in values as the data.

- (7) The  $t$ -test depends on the data being a random sample from a normally distributed population, and if this assumption is reasonable it is the best test. For large samples the Central Limit Theorem takes over, and the  $t$ -test is still reasonably effective even if the population is not very Gaussian. For small samples from non-Gaussian populations other tests must be considered.

### Minitab commands

A  $z$ -test is obtained by clicking **Stat** → **Basic Statistics** → **1-Sample z**. Enter the variable of interest under **Variables:**, and the true population standard deviation in the box next to **Sigma:**. Click the radio button next to **Test mean:**, and change the entry in the box to the actual null mean value (if necessary). The form of the alternative hypothesis can be changed by changing the box next to **Alternative:**. To obtain a graphical display of the data, with null mean value and confidence interval for the mean, click on **Graphs**.

A  $t$ -test is obtained by clicking **Stat** → **Basic Statistics** → **1-Sample t**. Enter the variable of interest under **Variables:**. Click the radio button next to **Test mean:**, and change the entry in the box to the actual null mean value (if necessary). The form of the alternative hypothesis can be changed by changing the box next to **Alternative:**. To obtain a graphical display of the data, with null mean value and confidence interval for the mean, click on **Graphs**.

A paired  $t$ -test is obtained by clicking **Stat** → **Basic Statistics** → **Paired t**. Enter the variables of interest under **First sample:** and **Second sample:**. Clicking **Options** allows you to change the null mean value (if necessary); the form of the alternative hypothesis can be changed by changing the box next to **Alternative:**. To obtain a graphical display of the differences, with null mean value and confidence interval for the mean difference, click on **Graphs**.

“The critical ratio is  $Z$ -ness,  
But when samples are small, it is  $t$ -ness.  
Alpha means  $\alpha$ ,  
So does  $p$  in a way,  
And it’s hard to tell  $\alpha$ -ness from  $p$ -ness.”

— Anonymous

“It is the first duty of a hypothesis to be intelligible.”

— Thomas H. Huxley

“Calculating statistical significance is a tool, a step in the process of analysis . . . What is important, or ‘significant’ in the English sense, may be statistically significant or non-significant.”

— R. Allan Reese