

## 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

$H_0$  : The new drug is not better than the current one

and

$H_a$  : 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

$H_0$  : Sani – Flush does not clean better than Vanish

and

$H_a$  : 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

$H_0$  : Aleve is as good as other pain relievers

and

$H_a$  : 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. Consider the following situation. The target daily sales of a particular product is 1000 units per day. Three months (100 days) worth of data are gathered to determine if the actual sales are consistent with this target. Going in, we have no reason to think that the true average sales are different from the target (presumably this number was picked for a good reason), so the hypotheses being tested are

$$H_0 : \mu = 1000$$

versus

$$H_a : \mu \neq 1000,$$

where  $\mu$  is the true average daily sales. How would we decide if the data are consistent with  $H_0$ ? If we constructed a 95% confidence interval for  $\mu$ , and it turned out that the interval was (973, 1034), presumably we would conclude that the data *were* consistent with  $H_0$ ; after all, the confidence interval is giving a region within which we are 95% confident that the mean falls, and 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.

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. The confidence interval includes the true mean with probability  $1 - \alpha$  (in repeated sampling), so the true mean will fall outside the interval with probability  $\alpha$ ; under the null hypothesis the true mean is  $\mu_0$ , so under the null  $\mu_0$  is outside the interval  $\alpha$  of the time, and that is when we mistakenly reject the null and commit a Type I error. Note that the Type I error is split into two halves:  $\alpha/2$  occurring when  $\bar{X}$  is lower than would be typically be expected when  $\mu = \mu_0$ , and  $\alpha/2$  occurring when  $\bar{X}$  is higher than would be typically be expected when  $\mu = \mu_0$ .

This rejection condition is met if either

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

Rearranging terms gives equivalent inequalities:

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

Putting these two inequalities together gives a final form for the test: reject 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 our sales 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. We could have also gotten this result by noting that we should reject the null hypothesis if the sample mean satisfies either

$$\bar{X} > 1000 + (1.96)(100)/\sqrt{100} = 1019.6 \quad \text{or} \quad \bar{X} < 1000 - (1.96)(100)/\sqrt{100} = 980.4,$$

which of course it does.

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.

We seem to have lost our connection with confidence intervals here, but we actually haven't. A one-sided hypothesis test is equivalent to a one-sided confidence interval,

$$(\bar{X} - z_\alpha \sigma / \sqrt{n}, \infty),$$

but this is rarely used in practice.

- (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. For example, workers might get a bonus if their quality level is good enough, where quality is measured by the number of defects per 100 units produced. 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.

Here's one way to see how the adherence to an artificial standard of .05 (say) can be misleading. Say you take a random sample, do your experiment, and get an observed  $z = 1.645$ , which is precisely significant at a .05 level. Now, you take a new random sample, trying to *replicate* your result. Just by random chance,  $\bar{X}$  will be larger than it was the first time 50% of the time, and lower than it was the first time 50% of the time. Thus, with probability .5 the new  $z$ -statistic will be less than 1.645, and will not be significant at a .05 level! Using tail probabilities helps guard against this misleading result, since the tail probability for the second test, while greater than .05 half of the time, is still likely to be relatively small (less than .10, say).

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) From one point of view,  $p$ -values make significance testing (in the sense of reject / don't reject) obsolete, since knowing the  $p$ -value implies the significance test result, but also gives more information. For example, if  $p = .04$ , we know that the test would reject at a .05 level, but wouldn't reject at a .01 level. This is true, but we should recognize that a significance test is required if a decision or action needs to be taken. This requires the potentially arbitrary choice of significance level, but is unavoidable.
- (5) 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.

While **Minitab** gives tail probabilities for the  $t$ -distribution, if you don't have a computer available, it's possible to approximate  $t$  tail probabilities using the normal distribution (and hence a normal table). The approximation is that the tail probability for a  $t$ -statistic of  $x$  based on  $d$  degrees of freedom is approximated by the tail probability for a  $z$ -statistic of  $(4d + x^2 - 1)(x)/(4d + 2x^2)$ , as long as  $d \geq 3$ . So, for example, say that an observed  $t$ -statistic is 2.21, based on 5 degrees of freedom. The approximating  $z$ -statistic is  $(4 \times 5 + 2.21^2 - 1)(2.21)/(4 \times 5 + 2 \times 2.21^2) = 1.77$ . The normal table gives a one-tailed  $p$ -value of .0384 for  $z = 1.77$ , which can be compared to the true  $p$ -value for  $t = 2.21$  on five degrees of freedom, which is .0391. This (obviously very accurate) approximation is given in the 1999 paper "A corrected normal approximation for the Student's  $t$  distribution" by B. Li and B. De Moor, *Computational Statistics and Data Analysis*, **29**, 213–216.

- (6) So far, all that we've mentioned is Type I error. Naturally, we would like to use a test that is likely to identify when the null is actually not true; that is, one that has a small probability of Type II error. It turns out that, assuming the structure given here, no tests have smaller probability of Type II error than the tests given above. We define the term *power* to be  $1 - P(\text{Type II error})$ , so another way of saying this is that tests based on  $\bar{X}$  are the most powerful tests.
- (7) 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. This is not merely an idle point; I've seen quite a few papers in top finance journals that trumpet results

that clearly have little practical importance, but are based on data from thousands of stocks, and are therefore statistically significant.

Consider the following example. In the early 1990's, the Genentech corporation introduced a drug to treat heart disease that cost \$5000 per dose. There were two alternative drugs that cost between \$250 and \$500 per dose. A clinical trial (involving a few hundred subjects) found no evidence of difference in effectiveness between the more expensive drug and the cheaper ones. Genentech funded a very large multimillion dollar study, which reported in 1992 that the more expensive drug worked better than the cheaper ones in 1 in 5000 cases (with no difference otherwise). From a cost / benefit point of view, it's hard to argue in favor of a drug that is 10 to 20 times more expensive, and only more effective in 0.02% of all cases, but from a malpractice point of view, doctors and hospitals might not want to take the chance of not using it. In fact, the day that these results were released, the price of Genentech stock rose sharply.

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. Consider two possible outcomes related to a null hypothesis  $\mu = 100$ , versus a two-sided alternative  $\mu \neq 100$ , with  $\sigma = 10$ . In one case,  $\bar{X} = 100.2$ , based on a sample size  $n = 10000$ ; in the other,  $\bar{X} = 105$ , based on a sample of size  $n = 10$ . In the first case  $z = 2$ , with  $p = .0456$ . In the second case,  $z = 1.58$ , with  $p = .1142$ . The first test provides stronger evidence against the null, but a comparison of 95% confidence intervals ((100.004, 100.396) in the first case, (98.8, 111.2) in the second case) shows that the estimated effect size (that is, the estimated difference from the null) is much more important in the second case. A situation like this should lead to a desire for more data in the latter case, and little interest in investigating further in the former case, in direct contradiction to the  $p$ -values. Thus, 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*.

- (8) 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 clinical trial blood measurements might be taken

before and after a drug is administered, and we are interested in the change in certain blood levels after taking the drug. Another example would be in a market research setting, where respondents' opinions on a product before and after seeing a commercial are compared.

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.

- (9) The  $t$ -test depends on the data being a random sample from a normally distributed population, and if this assumption is reasonable they are the best (most powerful) tests. 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. But what if we have a small sample, and, for example, the data contain an outlier or two, are highly skewed (a long right or left tail), or are very fat-tailed (both long left and right tails)? In the first case we could consider omitting the outlier(s) from the data and rerunning the  $t$ -test, but that option is not appropriate in the latter two cases.

An alternative approach is to use tests that do not assume normality. These tests are called *nonparametric* tests, because they require far fewer assumptions than parametric tests like the  $t$ -test. We will not discuss these tests here, but some information about them can be found in the Appendix.

## Appendix — Nonparametric tests of location

The simplest nonparametric location test is the **sign test**. The (two-tailed) sign test tests the hypotheses

$$H_0 : m = m_0$$

versus

$$H_a : m \neq m_0,$$

where  $m$  is the population median. That is, our previous hypotheses based on means are now based on medians. The sign test counts the proportion of observations above  $m_0$ , the hypothesized median, and compares that to .5, the value we would expect under  $H_0$  (values exactly equal to  $m_0$  are discarded from the sample). If the observed proportion of observations above  $m_0$  is very far from .5, we reject  $H_0$ . The test is based on the test for binomial proportions that we will discuss in a little while.

The sign test is appropriate even if the data are highly skewed, but has relatively low power. If the population distribution is roughly symmetric (while still not necessarily normal), a better (more powerful) test is the **Wilcoxon signed rank test**. This also tests the hypotheses

$$H_0 : m = m_0$$

versus

$$H_a : m \neq m_0,$$

with the added (implicit) assumption is that the distribution is symmetric around its median  $m$ . First,  $m_0$  is subtracted from each observation. Then, the observations are ordered from smallest to largest, ignoring the sign (any observations exactly equal to  $m_0$  are discarded). The appropriate sign is then assigned back to the rank of each observation, and the smaller of the sum of the positive and negative ranks (call this  $T$ ) is compared to its expected value under  $H_0$ ,

$$z = \frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}.$$

Effectively this is a  $z$ -test on the signed ranks, since under the null hypothesis the sums of the positive and negative signed ranks should be roughly zero.

## 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**.

A sign test is obtained by clicking **Stat** → **Nonparametrics** → **1-Sample Sign**. Enter the variable of interest under **Variables:**. Click the radio button next to **Test median:**, and change the entry in the box to the actual null median value (if necessary). The form of the alternative hypothesis can be changed by changing the box next to **Alternative:**.

A signed rank test is obtained by clicking **Stat** → **Nonparametrics** → **1-Sample Wilcoxon**. Enter the variable of interest under **Variables:**. Click the radio button next to **Test median:**, and change the entry in the box to the actual null median value (if necessary). The form of the alternative hypothesis can be changed by changing the box next to **Alternative:**.

“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