Tests of Hypotheses
Why and How

Statistics is about more than calculating the probability of things or the proportions of things. It can be a powerful tool for making decisions, influencing people and proving or refuting claims. Good statistics can check new drugs for safety, sway voters, and re-allocate budgets. One of the simplest tools for using statistics to change the real world is the **test of hypotheses**. A test of hypotheses is used to help disprove a claim that someone else has made based on their own study's findings. Given that few people can tell the difference between bad statistics and valid statistics, this is a most useful tool.

Say that a cigarette company says that no more than 0.3% of its product's users get lung cancer, and therefore cigarettes aren't as dangerous as doctors say they are. They might present the data they used to draw this conclusion, but the data might be falsified, or their experiment might be biased. If you don't believe this claim, and you think the company is lying to the public, what can you do?

You could run your own statistical experiment. You could get test subjects who also use the company's cigarettes and see how many get lung cancer. But how do you use the results of your experiment to disprove the cigarette company's claim? How can you know, scientifically, that when you get a different result from their study — and you almost certainly will — that it's different *enough* not to be explained away by experimental variation?

A test of hypotheses goes like this:

- We set up a null hypothesis and alternative hypothesis based on the original claim.
- We do a statistical experiment similar to the one that produced the claim.
- We assume that the study results conform to the null hypothesis.
- We find the probability that our result could happen by chance, given the assumption.
- We decide about the validity of our assumption, thus possibly refuting the claim.

We'll look at an example to examine these steps.

Example 1: A medical study says that an allergy medication does not increase blood pressure in users, stating that the mean systolic pressure is 118.1. You survey 67 users of the medication. Your sample mean systolic blood pressure is 121.3 with an s of 10.4, suggesting prehypertension. Test the claim that the mean systolic blood pressure in users of the allergy medication is greater than 118.1 at the 0.05 level.

Solution: First we need to outline our test. We need a **null hypothesis**, a statement in statistical terms of what the "status quo" is. The null hypothesis always looks like this:

$$H_0: \mu = ?$$

It has an equal sign because the null hypothesis always represents the idea that nothing unusual is going on, that things are as they always were.

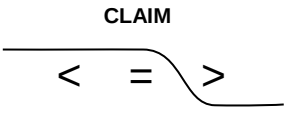
The **alternative hypothesis** (H_a or H_1) is an alternative to the status quo. It is a **one-tailed alternative** if H_a needs to be *strictly higher* or *strictly lower* than the stated value to be a support or refutation of the claim, and it is a **two-tailed alternative** if H_a only needs to be *different* from the stated value. The alternative hypothesis will look like this:

one-tailed alternative	$H_a: \mu < ?$
… or	$H_a: \mu > ?$
two-tailed alternative	$H_a: \mu \neq ?$

The one-tailed alternative will state which side refutes the null hypothesis, and the two-sided alternative simply says the mean is not equal to the stated value. In all cases, the number that replaces the "?" in H_a will be the same one that replaced it in H_0.

The claim in the study that we are testing must always be stated explicitly, and it will pick sides. You can use the following system to figure out how to properly frame the null and alternative hypotheses.

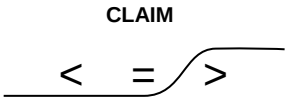
Start with the diagram at the right. The symbols at the bottom of the diagram represent those numbers less than, equal to, and greater than the number proposed by the question. Let's read the claim again carefully: "the claim that the … blood pressure … is greater than 118.1". We're going to draw a line separating those symbols into two sides: the top side is the claim and the bottom is the opposing side of the claim. Each of those three symbols will belong to one side or the other. Which of the symbols is the claim taking for its side? If the mean is a value less than 118.1, or equal to 118.1, then the claim is not met. The claim is true only if μ is greater than 118.1, so the symbols should be divided like this:



The null hypothesis is the side of the line that has the equal sign (whether that's the claim or not) and the alternative hypothesis is the other side. For the allergy medication example, our hypotheses are:

$$H_0: \mu = 118.1$$
$$H_a: \mu > 118.1$$

The claim that the cigarette company made was that "that no more than 0.3% of its product's users get lung cancer". If we want to test this claim, we need null and alternative hypotheses. The claim is true if the population proportion is either less than 0.003 or equal to 0.003, so the claim gets the less-than and equal signs. The greater-than sign goes to the other side. We split the signs as shown. The claim is the null hypothesis this time, since the claim takes the equal sign. The hypotheses would be H_0: $p = 0.003$ and H_a: $p > 0.003$. This alternative hypothesis also uses the greater-than sign, even though the claim said "no more than". It's all about the placement of the equal sign.



Before we do the experiment, we must also establish a **significance level**. This level is a benchmark that we will use to say whether the evidence is good enough to refute the claim. (We cannot establish it after the experiment, as that would bias our result.) We'll explain what the level means later in this worksheet.

Now we are ready to do the experiment. We get 67 of the allergy drug's users and do a survey similar to the original one. Their mean systolic blood pressure is 121.3. It's higher than the mean in the claim — if it weren't, we'd stop— but is it high enough?

We'll use *reductio ad absurdum*, a logical argument, to examine the null hypothesis. If we start with the assumption that the claim is valid, but it gives us an absurd result, then we have evidence that the assumption wasn't a good one.

For this example, we assume the null hypothesis, that $\mu = 118.1$, and we find the probability that we could get a result of 121.3 for $\bar{x}$ in a survey of $n = 67$. Since we have a survey of a large number of people, we can use the t distribution to determine probability.

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$
$$= \frac{121.3 - 118.1}{10.4 / \sqrt{67}}$$
$$= 2.5186\ldots$$

We can use technology to find out what probability is associated with this value. (If we don't have access to computerized help on a question, we're using Table A-3. See our worksheet *Working with t Statistics* for more help.) We learn that $P(T \leq 2.5186) = 0.9929$. Which side of the t-statistic do we want? This depends on whether we're doing a one-tailed test or a two-tailed test. In our example, we have a one-tailed alternative, so we want the side stated in the alternative hypothesis. It says that $\mu >$ 118.1, so we're interested in $P(T > 2.52)$. This is the probability that we would get a result that is *as far away from the mean* as ours was. Our probability is 0.0071. This is a good sign for us! It suggests that either we got a batch of freaks in our survey who all have high blood pressure by accident, or the assumption we started with, that $\mu = 118.1$, was flawed.

If we were doing a two-tailed test, then we would look at $P(T > 2.52 \text{ or } T < -2.52)$. This is the same as $2 \times P(T > 2.52)$. In this case, we always take the tail ends of the normal curve, away from the mean (since we want to know the probability of being far from the mean).

In the context of a test of hypotheses, these probabilities are called **P-values**. **Low P-values constitute evidence against H_0** since a high P-value can occur by random chance. A survey which gave a P-value of 21.3%, for example, would mean that a result similar to the survey would happen 1 time in 5 or more, which is consistent with the assumption that the null hypothesis is correct. We would have no reason to doubt it.

These P-values are clear; 0.71% is low enough to be evidence against the claim and 21.3% is not low enough. What does "low enough" mean? We decide what threshold makes a result significant. The **significance level** (α) is the number that we say is low enough to be significant before we do our experiment. **If the P-value is less than the significance level, then the result is considered significant**, and it's evidence that the claim isn't true. A lower α means a stricter definition of significance (since the P-value has to be even lower to qualify as significant). We compare the P-value to α: $0.0071 < 0.05$, so the result is significant.

So how do we finish? We need to make a statement expressing the conclusion we've drawn. Because our P is low, we "reject the null hypothesis". If our P were higher than α, then we would "fail to reject the null hypothesis". Our study did not give us any reason to doubt the status quo. *We can never say that we accept the null hypothesis*, at least not with this test! To do so would be to imply that we've proven that number of 118.1 was correct, and there's no way to offer that kind of proof with one study.

On the other hand, it's no good telling people that you've rejected the null hypothesis — after all, before this chapter *you* had no idea what that means, so neither will anyone else. We should report our findings using the language and context of the original problem. We were asked whether the allergy medication increases blood pressure dangerously, and we should answer *that* question.

We have a significant result; we have evidence of something. Go back to the diagram with the < = > symbols. It shows us that the claim is the alternative hypothesis (the side without the "="). The test came down on the claim's side, so we can say that "this study constitutes evidence to support the claim that the mean systolic blood pressure in users of the allergy medication is greater than 118.1." (In the case of the cigarette company's claim, the claim is the null hypothesis. If the test tells us to reject H_0, we still address the claim: "the test constitutes evidence to warrant rejection of the claim that…". If the test were not significant, either way we'd start instead with "the test does not constitute…".)

ANOTHER USE FOR TESTS OF HYPOTHESES

So far the examples you've seen in this worksheet start with the idea that the value for a statistic may have been reported incorrectly. In such cases, we are interested in knowing if the value in the claim stands up to scrutiny.

We also use tests of hypotheses to try to prove changes to the value of p or μ. If a car maker hopes that a new body design improves fuel economy, we might use the prior design's value as the basis for the null hypothesis, and test the new cars. If the test says that we should reject the null hypothesis (the status quo), it means the new design is effective. A result of "fail to reject" means the new design is no better than the old one.

EXERCISES

A. For each of these situations, draw the "< = >" diagram, and write the null hypothesis and the alternative hypothesis.

1) The advertised average lifespan of a Sta-Brite light bulb is 1000 hours. A consumer advocacy group is testing the claim that the mean lifespan is at least 1000 hours.

2) A generic brand of low-dose aspirin contains 50 mg of ASA per tablet. The factory's quality control department is testing a batch to test the claim that tablets contain 50 mg on average.

3) The city's water filtration system removes mercury from drinking water down to a standard of no more than 13 parts per million. A municipal employee makes sure the system is working and check the water to test the claim that mercury levels are not below the standard.

4) According to current legislation, a blood alcohol level of 0.09 impairs drivers to the point where they are a hazard on the road. An activist group is testing the claim that

the mean blood alcohol level for drivers who cause accidents where alcohol was a factor is lower than 0.09.

5) The manufacturer wishes to ensure that PVC pipe and fittings are produced with an inner diameter of 0.22 m. A quality control officer is testing the claim that the machines are not producing pipes with this diameter.

6) The most commonly prescribed cholesterol medication lowers LDL by 36 points. The pharmaceutical company is testing the claim that the new formulation of the medication improves on that.

7) A public health organization is concerned that cockroaches are becoming resistant to pesticides used to control them. In the past they've determined that their first-line pesticide kills 72% of roaches, and they want to test the claim that the poison kills a smaller percentage than that today.

8) A soft drink company is testing the claim that less than 12% of people detect an aftertaste when drinking diet cola made with aspartame.

9) New legislation has changed the way in which hospitals are allowed to collect and store data on their patients. They're testing a new intake form with the claim that the success rate for patients filling out the form with errors is at least 93% (the success rate for the old form).

B. For each situation, and using the information from question A, calculate a test statistic and a P-value (or a range of values for the t distribution with Table A-3).

1) Out of a SRS of 320 bulbs, the mean lifespan of a light bulb is 995 hours with a standard deviation of 72 hours.

2) When testing a batch of 48 aspirin tablets, the mean quantity of active ingredient ASA is 52 mg. The sample standard deviation is 4.2 mg.

3) A mean of 16 ppm of mercury was measured from 27 randomly chosen sources of tap water, with s = 8.2 ppm.

4) During one month, there were 35 accidents where the investigating officer reported that alcohol might have been involved. The drivers' BAC's (blood alcohol content) were normally distributed with a mean of 0.10, and a standard deviation of 0.021.

5) A simple random sample, n = 150, of pipes are measured. Mean inner diameter of the sample (which had a standard deviation of 0.08 m) was 0.208 m.

6) During a study of the replacement treatment, LDL levels went down an average of 38 points for a sample of 30 patients with high cholesterol, with a s.d. of 4.7 points.

7) Animal Control collects samples from around the city and determines that out of 412 roaches sprayed with pesticide, 278 of them were killed.

8) In taste tests, 23 out of 209 consumers said they detected an aftertaste from the diet cola with aspartame they were asked to try.

9) A suburban hospital trialed their new form on a sample of 364 incoming patients in the course of a month and, upon interviewing the patients, determined that 19 of the forms had errors.

C. Two of the questions from part B are quite different from the others. Which ones and what impact does that difference have on the test of hypotheses?

D. Which of the five remaining tests are significant at:
1) $\alpha = 0.05$? 2) $\alpha = 0.01$?

Is it possible for any test to get:
3) a yes to part 1 and a no to part 2? 4) a no to part 1 and a yes to part 2?

E. Write concluding statements for the nine tests in B with a 0.05 level of significance.

SOLUTIONS

A. (1) $<$ (CLAIM) $=$ $>$, H_0: $\mu = 1000$; H_a: $\mu < 1000$ (2) $<$ $=$ (CLAIM) $>$, H_0: $\mu = 50$, H_a: $\mu \neq 50$
(3) $<$ $=$ (CLAIM) $>$, H_0: $\mu = 13$, H_a: $\mu > 13$ (4) $\leq$ (CLAIM) $=$ $>$, H_0: $\mu = 0.09$, H_a: $\mu < 0.09$
(5) $\leq$ (CLAIM) $=$ $>$, H_0: $\mu = 0.22$, H_a: $\mu \neq 0.22$ (6) $<$ (CLAIM) $=$ $>$, H_0: $\mu = 36$, H_a: $\mu > 36$
(7) $\leq$ (CLAIM) $=$ $>$, H_0: $p = 0.72$, H_a: $p < 0.72$ (8) $\leq$ (CLAIM) $=$ $>$, H_0: $p = 0.12$, H_a: $p < 0.12$
(9) $<$ (CLAIM) $=$ $>$, H_0: $p = 0.93$, H_a: $\mu < 0.93$

B. (1) $t = -1.24$, $P(T < -1.24) = 0.1075$ [df = 300, $P > 0.10$]
(2) $t = 3.30$, $2 \times P(T > 3.30) = 0.0018$ [df = 45, $P < 0.01$]
(3) $t = 1.90$, $P(T > 1.90) = 0.0242$ [df = 26, $0.025 < P < 0.05$]
(4) $t = 2.82$, $P(T < 2.82) = 0.9960$ [df = 34, $P > 0.995$**]
(5) $t = -1.84$, $2 \times P(T < -1.84) = 0.0682$ [df = 100, $0.05 < P < 0.10$]
(6) $t = 2.33$, $P(T > 2.33) = 0.0135$ [df = 29, $0.01 < P < 0.025$]
(7) $z = -2.05$, $P(Z < -2.05) = 0.0202$ (8) $z = -0.44$, $P(Z < -0.44) = 0.3300$
(9) $z = 1.33$, $P(Z < 1.33) = 0.9082$

C. Questions 4 and 9 can be stopped early, as the sample statistic is on the wrong side of the mean to provide evidence against the null hypothesis. (In the case of Question 4, this is bad; the activist group has no leverage to get the law changed. In Question 9, this is good! The hospital would be hoping the new form wouldn't make things worse.)

D. (1) parts 2, 3, 6 and 7 (2) part 2 (3) Yes: parts 3, 6 and 7 did. (4) No: if a P-value is less than 0.01, it must necessarily be less than 0.05.

E. Other phrasings of the claims themselves are possible.
(1) This study does not constitute evidence to warrant rejection of the claim that the mean lifespan of a Sta-Bright light bulb is at least 1000 hours.
(2) …constitutes…warrant rejection of…the tablets contain 50 mg of aspirin.
(3) …constitutes…warrant rejection of…mercury levels are not below the standard of 13 ppm.
(4) …does not constitute…support…the BAC level of drivers who cause accidents is lower than 0.09.
(5) …does not constitute…support…the machines are failing to produce pipes with the correct diameter.
(6) …constitutes…support…the new medication lowers LDL by more than 36 points.
(7) …constitutes…support…the pesticide is effective on less than 72% of cockroaches.
(8) …does not constitute…support…less than 12% of people detect an aftertaste when drinking diet cola.
(9) …does not constitute…warrant rejection of…the new form is less error-prone than the old one.