Tests of Significance:  
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 significance**. A test of significance 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 only 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 significance 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 claim accurately represents the mean of our experiment.
- 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 claims that an allergy medication does not increase blood pressure in users, stating that the mean systolic pressure is 118.1 with a $\sigma$ of 10.4. You survey 67 users of the medication. Your sample mean systolic blood pressure is 121.3, suggesting prehypertension. Determine whether this result is significant 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 original claim is. The null hypothesis always looks like this:

$$H_0: \mu = ?$$

It has an equal sign because the original claim will always be that the mean is a particular value. There may be an implication that the mean could have a value to one side of the stated value, and this would be fine. Obviously, our cigarette company would be happy if the mean were less than 0.3%. This implication will help us state our view.
The alternative hypothesis (Hₐ or H₁) is an alternative to the original claim. It is a one-sided alternative if Hₐ needs to be strictly higher or strictly lower than the stated value to be a refutation of the claim, and it is a two-sided alternative if Hₐ only needs to be different from the stated value. The alternative hypothesis will look like this:

- **one-sided alternative**: Hₐ: µ < ?
- **… or** Hₐ: µ > ?
- **two-sided alternative**: Hₐ: µ ≠ ?

The one-sided 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 Hₐ will be the same one that replaced it in H₀.

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.

For the allergy medication example, we will use a one-sided alternative since a result lower than 118.1 doesn’t help our case:

H₀: µ = 118.1  
Hₐ: µ > 118.1

Now we are ready to do the experiment. We get 67 of the drug’s users and do a survey similar to the original one. We find that 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 µ = 118.1, and we find the probability that we could get a result of 121.3 for x in a survey of n = 67. Since we have a survey of a large number of people, we can use a z-score to determine probability.

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{121.3 - 118.1}{10.4/\sqrt{67}}$$

$$= 2.5185 \ldots \approx 2.52$$

Which side of the z-score do we want? This depends on whether we’re using a one-sided alternative or a two-sided alternative. In our example, we have a one-sided alternative, so we want the side stated in the alternative hypothesis. It says that µ > 118.1, so we’re interested in P(Z > 2.52). This is the probability that we would get a result that is as far away from the mean as ours was. We know that P(Z > 2.52) is going to be low even without looking it up. 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 µ = 118.1, was flawed.
If we had a two-sided limit, then we would look at \( P(Z > 2.52 \text{ or } Z < -2.52) \). This is the same as \( 2 \times P(Z > 2.52) \). 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).

We look up \( P(Z > 2.52) \) to get a probability of 0.0059, or 0.59\%. In the context of a test of significance, this probability is called a **P-value**, the probability that a survey would give a particular result through statistical variation, assuming \( H_0 \). **Low P-values constitute evidence against the original claim** 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 would not be considered a strange enough result to suggest that our assumption was wrong. A **test of significance can only be used to refute a claim, never to support it**.

These P-values are clear; 0.59\% 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** (\( \alpha \)) 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 \( \alpha \) means a stricter definition of significance (since the P-value has to be even lower to qualify as significant). To finish off our example, we compare the P-value to \( \alpha \): 0.0059 < 0.05, so the result is significant.

**ANOTHER USE FOR TESTS OF SIGNIFICANCE**

So far the examples you’ve seen in this worksheet start with the idea that the value for \( \mu \) 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 significance to try to prove changes to the value of \( \mu \). If a car maker hopes that a new body design improves fuel economy, we might use the old, established value as the basis for the null hypothesis, and test the new cars. If the test indicates that we should reject the null hypothesis, 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, state whether the alternative hypothesis should be one-sided or two-sided, 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 verifying the claim.

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.

3) The city’s water filtration system removes mercury from drinking water down to 13 parts per million. A municipal employee is testing the water.

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 lobbying for change.

5) The manufacturer wishes to ensure that PVC pipe and fittings are produced with an inner diameter of 0.22 m.

6) The most commonly prescribed cholesterol medication lowers LDL by 36 points. A new formulation of the medication is trying to improve on that.
B. For each situation, and using the information from question A, calculate a z-score and a P-value.

1) The manufacturer lists a σ of 72 hours. Out of a SRS of 320 bulbs, the mean lifespan of a light bulb is 995 hours.
2) When testing a batch of 48 aspirin tablets, the average quantity of active ingredient ASA is 52 mg. The established standard deviation is 4.2 mg.
3) A mean of 16 ppm of mercury was measured from 27 randomly chosen sources of tap water. Past tests suggest a standard deviation of 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. Reviewing historical accident reports gives a standard deviation of 0.021.
5) A simple random sample, n = 150, of pipes (which have had a standard deviation in diameters of 0.08 m over the entire manufacturing run) are measured. Average inner diameter of the sample was 0.208 m.
6) On the usual treatment, the s.d. of LDL decrease is 4.7 points for all users. 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.

C. One of the questions from part B is quite different from the others. Which one and what impact does that difference have on the test of significance?

D. Which of the five remaining tests are significant at:

1) the 5% level?  2) α = 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?

SOLUTIONS

A. (1) one-sided, H₀: μ = 1000; Hₐ: μ < 1000  (2) two-sided, H₀: μ = 50, Hₐ: μ ≠ 50
(3) one-sided, H₀: μ = 13, Hₐ: μ > 13  (4) one-sided, H₀: μ = 0.09, Hₐ: μ < 0.09
(5) two-sided, H₀: μ = 0.22, Hₐ: μ ≠ 0.22  (6) one-sided, H₀: μ = 36, Hₐ: μ > 36
B. (1) z = −1.24, P = 0.1075  (2) z = 3.30, P = 0.0010  (3) z = 1.90, P = 0.0287
(4) z = 2.82, P = 0.9976  (5) z = −1.83, P = 0.0672  (6) z = 2.33, P = 0.0099
C. Question 4 can be thrown out because the result supports the null hypothesis; x̄ is on the opposite side of the mean from the side we’re examining.
D. (1) parts 2, 3 and 6  (2) part 2 and 6  (3) Yes: part 3 did.  (4) No: if a P-value is less than 0.01, it must necessarily be less than 0.05.

© 2013 Vancouver Community College Learning Centre.
Student review only. May not be reproduced for classes.