UNIVERSITY OF WATERLOO

Faculty of Mathematics
Waterloo, Ontario N2L 3G1

Centre for Education in
Mathematics and Computing

Grade 7/8 Math Circles

March 3, 2021

Significance Testing

Introduction

Significance tests are a very common way to draw conclusions from data in statistics. The process of significance testing defines and tests claims about a parameter, based upon evidence from collected data. In other words, it's the mathematical basis for deciding whether a set of evidence means that something is probably true or probably false. If you've ever heard the phrase "statistically significant" used, this is where it comes from!

Like with confidence intervals, this lesson we will be focusing on the reasoning behind significance testing and how to apply it rather than actually doing the associated calculations. Each section of this lesson focuses on a component of significance testing.

Defining Hypotheses

The first step in carrying out a significance test is defining the claims that we intend to compare.

In a court trial, a person is assumed innocent until proven guilty. Evidence is presented against this assumption of innocence until it is determined that either there is sufficiently compelling evidence to convict them as guilty, or there is insufficiently compelling evidence and the assumption of innocence holds.

Significance tests work within the same structure, where we assume a *null hypothesis* and evaluate evidence against it. The null hypothesis is abbreviated as H_0. The claim that we gather evidence in favour of is called the *alternative hypothesis*, abbreiviated as H_a.

- **null hypothesis (H_0)**—the claim about the population that we weigh evidence *against* during significance testing; it is what we assume to be true by default
- **alternative hypothesis (H_a)**—the claim about the population that we weigh evidence *for* during significance testing

Example 1: A study reports that in a survey at several high schools, 58% of students admitted to plagiarism. In hopes that your classmates are more honest students, you conduct a survey at school to see if there is convincing evidence that the rate of plagiarism is lower at your school than at the schools in the original study.
In this case, the null hypothesis is a statement of "no difference." We assume that the proportion of students who have plargiarised at your school are the same as in the study, 58%. On the other hand, the alternative hypothesis for this scenario is that the plargiarism rate at your school is lower than 58%. This gives us:

- H_0: the proportion of students who have plagiarised at your school $= 0.58$.
- H_a: the proportion of students who have plagiarised at your school < 0.58.

When stating hypotheses, it's extremely important to define what they are *before* examining the data. Like when carrying out the scientific method, you should never conduct your experiment before stating the hypothesis—that would be cheating!

Question 1: The city is making temporary changes to their bike lanes to see if they cause a significant increase in the number of bikers in the area each day. Currently, the mean number of bikers each day is 9500. State the null and alternative hypotheses.

Solution: Since our hypothesis of interest is that there may be a significant increase from the current number of bikers in a day, we have

- H_0: mean number of bikers each day $= 9500$
- H_a: mean number of bikers each day > 9500

P-Values

Significance tests evaluate how strong sample data is as evidence against the null hypothesis and in favour of the alternative hypothesis. In particular, a test answers the question, "When the null hypothesis is true, what is the probability that we would observe the given data?"

When H_0 is true, there is a range of possibilities that the data we observe could look like. Some of these possibilities are less likely than others. Significance testing lets us calculate the probability of how unlikely those possibilities are when H_0 is true.

This probability has a specific name: the *p-value*. The p-value produced by a significance test measures the strength of the evidence from sample data against H_0.

- **p-value**—the probability that, when H_0 is true, sample data would be as extreme or more extreme than the data actually observed

When the p-value is low, that tells us that it would be unlikely for us to obtain our observed data when H_0 is true. Thus, it can serve as evidence that H_0 is not true, and that instead we should reject it in favour of H_a.

Using more concrete ideas, we can imagine a scenario where we have the following hypotheses:

- H_0: the population proportion is A
- H_a: the population proportion is not A

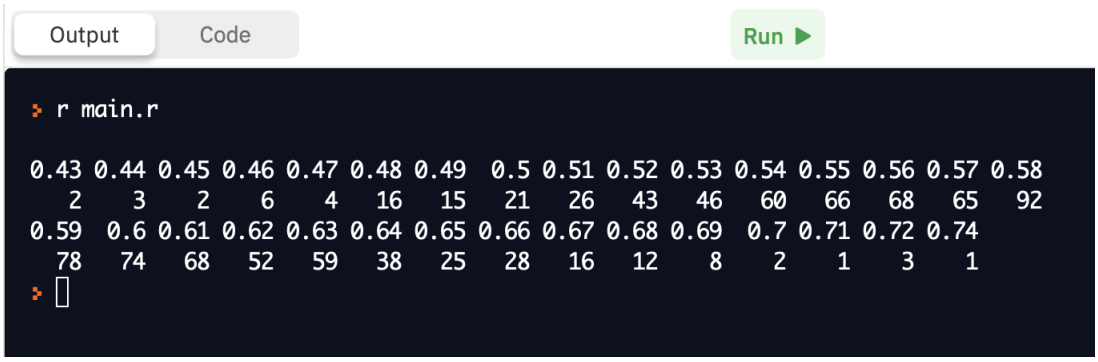
To test this alternative hypothesis, we conduct a survey that then gives us a sample proportion of B. The results of a significance test give a very small p-value. This indicates that *if the population proportion was actually A*, then it would be very unlikely for us to have observed a sample proportion of B. This gives us convincing evidence that the population proportion *is not actually A*.

Example 2: Returning to the earlier scenario about plagiarism, you survey an appropriately-selected random sample of 100 students at your school. You find that out of those 100 students, 47 of them answer yes, giving you a sample proportion of 47% for students at your school who have plagiarised.
The program linked here: `https://repl.it/@cemc/plagiarism-p-values#main.r` simulates the results of surveying 1000 different random samples of 100 students when the population proportion is 0.58. After clicking "Run," it will display a table that shows how many times each indicated sample proportion was observed. *(If the link does not work, copy-and-paste the url into your browser.)*
Recall our hypotheses:

- H_0: the proportion of students who have plagiarised at your school $= 0.58$.
- H_a: the proportion of students who have plagiarised at your school < 0.58.

Refer to the output of the program. Out of 1000, how many of the sample proportions were 0.47 or less?
The probabilty $\frac{\text{number of sample proportions that were 0.47 or less}}{1000}$ is an approximation of the p-value: the probability that we could observe a sample proportion of 0.47 or less when the population proportion is 0.58, as in H_0.

To look at a specific case, here is one randomly-generated output of the program:



Notice that out of 1000 simulated samples, only $2 + 3 + 2 + 6 + 4 = 17$ out of 1000 times did we see a resulting sample proportion of 47% or less. That means that, if the population proportion was actually 58%, the chances of having 47% or less of your classmates admit to plagiarism would be somewhere around $\frac{17}{1000}$. This number, $0.017 = 1.7\%$, is the (approximate) p-value. If the null hypothesis is true, there was only a 1.7% chance that you would get a sample proportion that small by chance alone.

That leaves us with two possibilities:

1. H_0 is correct: the population proportion is actually 0.58, and by chance, you just ended up with a random sample that produced very unlikely results.
2. H_a is correct: the population proportion is less than 0.58.

It's possible that either Possibility 1 or 2 could be correct. However, the probability of Possibility 1 (the population proportion is 0.58) being true is very low—having a sample proportion this small would only happen in 1.7% of samples in the population when H_0 is true. Thus, this low p-value serves as evidence that Possibility 1 (and thus H_0) is probably false, and thus, H_a is probably true.

Question 2: A study aims to determine whether a newly-launched health guide has increased the population mean of servings of fruits and vegetables in student lunches each week. Before beginning their survey, hypotheses are defined; H_0: mean number of servings $= 20$, H_a: mean number of servings > 20.
After completing the survey, we have have a sample mean of 23 servings.
Using this program: `https://repl.it/@cemc/healthy-p-values#main.r`, determine an approximate p-value for the scenario. What does this p-value represent?

Solution: Solutions may vary, since the program uses random generation. This video goes over the steps to solve this question using a specific answer: `https://youtu.be/F1LPn-zXiAM`.

What we have done in this section is develop an understanding of what p-values mean. In that process, we have also approximated p-values by simulating many random samples of a population. During formal significance tests, we would produce exact p-values using various formulas and information that we have about the sample and population, but we will not be investigating those calculations in this lesson.

Statistical Significance

After obtaining a p-value for the observed outcome, there are then two options for a conclusion to draw from your significance test:

1. **reject the null hypothesis** because the observed outcome is too unlikely while assuming that H_0 is true; or

2. **fail to reject the null hypothesis** because we do not have convincing enough evidence to reject H_0 in favour of H_a.

How small the p-value has to be to reject H_0 is chosen depending on context, but is defined prior to carrying out the study like the hypotheses. This threshold value is called the *significance level*, often denoted α (alpha). When the p-value is lower than α, then the results are *statistically significant at level* α.

- **statistically significant at level** α—when a result is less than α likely to occur when H_0 is true, giving evidence against H_0 and for H_a

Example 3: With a significance level of $\alpha = 0.05$, we can conclude that data that produces a p-value of 0.44 do *not* provide statistically significant evidence against H_0. We thus fail to reject H_0.
We can justify this conclusion with our understanding of p-values: a p-value of 0.44 means that when H_0 is true, you'll observe data at least as extreme as the data from your sample 44% of the time—that's not rare at all! We do not reject the null hypothesis, because the data we observed is perfectly normal when H_0 is true.

Question 3: At the end of Example 2, we estimated a p-value of 0.017. Is this result statistically significant at $\alpha = 0.05$? What conclusion would we draw?

Solution: The results are statistically significant at $\alpha = 0.05$, because $0.017 < 0.05$. Since the results are significant, we choose to reject H_0 in favour of H_a. Thus, we conclude that the population proportion of students at your school who have plagiarised is less than 58%.
We can justify this conclusion with our understanding that having a p-value of 0.017 means that if H_0 *was actually true*, we would only see sample data as extreme as the sample proportion 1.7% of the time. Less than 1 out of 50 is pretty rare, indicating to us that H_0 is probably *not* true. The results *do* support H_a though!

Question 4: A clinical trial has H_0: proportion of patients who get sick = 0.4, H_a: proportion of patients who get sick < 0.4. After conducting the trial, the results yield a p-value of 0.025. Trials are required to have results that are statistically significant at $\alpha = 0.01$ for studies like this. Is this result statistically significant at α? What conclusion would we draw?

Solution: The results are not statistically significant at $\alpha = 0.01$, because $0.025 > 0.01$. Since the results are not significant, we fail to reject the null hypothesis. This would mean concluding that the proportion of patients who got sick did not decrease a statistically significant amount during the trial.

Significance levels of $\alpha = 0.05$, $\alpha = 0.01$, and $\alpha = 0.10$ are the most common, but any probability can be used. There's no universal rule for what α should equal—in each case, it depends on how strong you want evidence against H_0 to be before rejecting H_0.

At $\alpha = 0.05$, there can be up to a 5% chance of observing results at least as extreme as your sample data when H_0 is true. That means, 1 out of 20 times, you will incorrectly reject the null hypothesis. At $\alpha = 0.01$, there can be up to only a 1% chance of the same error.

However, by requiring stronger evidence to reject H_0, you also increase the chances of failing to reject the null hypothesis when it is actually false.

By drawing a conclusion from a p-value, you accept the inherent risk that the conclusion may be wrong. Despite this, significance tests remain a common and useful tool for evaluating evidence from a sample against different hypotheses for the parameter value.