Hypothesis Testing

Part 1 - p-values, null distributions, and errors

Ryan Miller

Grinnell College
Statistics

Introduction

- The first portion of the semester we focused on *describing relationships* in our data
  - Next, we introduced *confidence intervals* as a statistical tool to assess the uncertainty in *generalizing* sample data to a broader population
- Confidence intervals allow us to identify relationships with statistical confidence, but they aren't the most natural tool for this purpose
  - For example, suppose we find the 99% confidence interval for the correlation between hours worked and annual income is (-0.05, 0.23), what can we conclude?
  - What could we conclude if the 90% confidence interval in the same scenario is (0.03, 0.15)?

Hypothesis Tests

Statistical tests involve two main components:

1) Proposing a **null hypothesis**, H_0, and an **alternative hypothesis**, H_a
   - The null hypothesis is a *statement about the population of interest* that is falsifiable and that researchers aim to disprove
   - The alternative hypothesis represents the conclusion the researchers would like to establish

Hypothesis Tests

Statistical tests involve two main components:

1) Proposing a **null hypothesis**, H_0, and an **alternative hypothesis**, H_a
   - The null hypothesis is a *statement about the population of interest* that is falsifiable and that researchers aim to disprove
   - The alternative hypothesis represents the conclusion the researchers would like to establish
2) Deciding whether the sample data provide *sufficient evidence* to falsify the null hypothesis
   - A **null distribution** is a probability model for the outcomes that might have occurred *had the null hypothesis been true*
   - Evidence against H_0 comes from comparing the outcome observed in the *real data* against the null distribution

Example

An experiment published in the scientific journal *Nature* explored whether infants have preference towards friendly behavior. $n = 16$ infants repeatedly watched demonstrations of two scenarios:

- A "helper" toy assisting the main character
- A "hinderer" toy blocking the main character

After watching these demonstrations, 14 of 16 infants chose the "helper" toy. The researchers were careful to randomize the color and shape of each character to prevent biases. Do the results of this study suggest that the infants can understand friendly behavior?

Example (cont.)

Based upon the two main components of statistical tests, answer the following:

1. Using statistical symbols, what do the researchers in this study want to disprove? (ie: what is their null hypothesis?)
2. Using statistical symbols, what would they like to establish? (ie: what is their alternative hypothesis?)
3. Consider using coin tosses to *simulate* the null distribution. How could this be done? What would the null distribution look like?

Example (cont.)

- The website StatKey allows us to simulate outcomes for random processes that produce a single proportion
  - We can simulate the null distribution for the Nature study by providing the null hypothesis $H_0 : p = 0.5$ and $n = 16$
  - This simulation gives us a distribution of outcomes that we would expect to see *had the null hypothesis been true*.
- Next, we need to assess how compatible the observed outcome, $\hat{p} = 14/16 = 0.875$, is with these simulated outcomes.

Example (cont.)

- Using our StatKey simulation, sample proportions of 0.875 or larger happens in less than 1% of simulations, which suggests observing 14 or more many choices of the "helper" toy would be *very unlikely* if the null hypothesis were true.
  - Because the probability of observing at least 14/16 "helper" choices is so small, these data provide *strong evidence* against the null hypothesis
- The **p-value** is defined as the probability of observing an outcome at least as unusual as the one observed in the sample data under the assumption that the null hypothesis is true
  - In our Nature example, we'd estimate the p-value to be less than 0.01, because there was less than a 1% chance of observing a sample proportion further away from the null value of $H_0 : p = 0.5$ than our observed proportion $\hat{p} = 0.875$

What is "Unusual"

- The p-value is based upon outcomes that are at least as extreme as the one observed in the sample data
  - In the Nature study, it's possible to interpret this as observing 14, 15, or 16 "helper" choices
    - This is a "one-sided" alternative (ie: $H_a : p > 0.5$)
  - It's also possible to consider "at least as extreme" to include 0, 1, or 2 helper choices (in addition to 14, 15, and 16)
    - This is a "two-sided" alternative (ie: $H_0 : p \neq 0.5$)
- Two-sided alternatives are overwhelmingly more common in practice, and we will exclusively use them in our course

Statistical Tests and p-values

- The premise of statistical testing is to propose a "straw man" theory for the results observed in a study. This typically entails a null hypothesis of "no association":
  - For a binary categorical variable, $H_0 : p = 0.5$ implies each possible outcome is equally likely
  - For comparisons of groups, $H_0 : p_1 - p_2 = 0$ implies both groups have the same conditional proportion (risk) and $H_0 : \mu_1 - \mu_2$ implies both groups have the same mean outcome
  - For two quantitative variables, $H_0 : \rho = 0$ implies the variables are independent in the population (zero correlation)

Statistical Tests and *p*-values

- The *p*-value quantifies the strength of evidence that the sample data provide *against* the null hypothesis
  - A *p*-value of 0.5 means that if the null hypothesis were true, data like the sample we observed (or more unusual) would happen 50% of the time, so null hypothesis and the sample data are *compatible*
  - A *p*-value of 0.001 means that if the null hypothesis were true only 1 of 1000 samples would resemble the observed sample, so the null hypothesis and the sample are *incompatible* as it's unlikely that we observed such a rare outcome

Decision Thresholds

Many scientific fields have adopted $\alpha = 0.05$ as a threshold for "statistical significance"

- Data yielding a p-value smaller than $\alpha = 0.05$ are seen as *sufficient evidence* for rejecting H_0
- Data yielding a p-value larger than $\alpha = 0.05$ provide *insufficient evidence* and result in a "failure to reject H_0"

This black and white approach has its flaws, but it's still very widely used. It's also important to note that the p-value is merely a measure of compatibility between the data and the null hypothesis, it doesn't definitively tell you whether the hypothesis is true or false.

Practice #1

A study conducted by Johns Hopkins University found that 31 out of 39 babies born 15 weeks early go on to survive. Do these data provide compelling evidence that a majority of babies born 15 weeks early survive?

1) Propose a null hypothesis and an alternative hypothesis using both words and statistical symbols.
2) Use StatKey to create a null distribution and find the p-value measuring the evidence this study's observed outcome provides against the null hypothesis.
3) Use a threshold of $\alpha = 0.05$ to make a decision regarding the null and alternative hypotheses.

Practice #1 (solution)

1) $H_0 : p = 0.5$, the null hypothesis is that 50% of babies born 15 weeks early survive vs. $H_a : p > 0.5$, the alternative that more than 50% survive
2) The 1-sided p-value should be nearly zero
3) Because the p-value is so small, we reject H_0 in favor of H_a and conclude that these data provide *strong evidence* that more than 50% of babies born 15 weeks prematurely will survive

Practice #2

Wikipedia claims that 70% of babies born 15 weeks early will survive. Do the data in the Johns Hopkins University study (where 31 of 39 babies survived) provide compelling evidence against Wikipedia's claim?

1) Propose a null hypothesis and an alternative hypothesis
2) Use StatKey to create a null distribution and find the p-value measuring the evidence this study's observed outcome provides against the null hypothesis
3) Use a threshold of $\alpha = 0.05$ to make a decision regarding the null and alternative hypotheses

Practice #2 (solution)

1) This time, $H_0 : p = 0.7$, and $H_a : p \neq 0.7$ (since either a larger than expected or smaller than expected result would disprove Wikipedia's claim)
2) This time, the 2-sided p-value is approximately 0.13
3) Because this p-value is larger than 0.05, there's *insufficient evidence* to reject H_0. It's unclear whether Wikipedia's claim is true, but these data are relatively compatible with it.

Misconception #1

As a silly example, suppose Prof. Miller and Steph Curry compete in a 3-point shooting contest. Further, suppose that Prof. Miller makes 3 of 5 and Steph Curry makes 5 of 5.

- We might use these data to evaluate the null hypothesis $H_0 : p_1 - p_2 = 0$, which reflects Prof. Miller and Steph Curry making the same proportion of 3-pt shots (not in this sample, but in the long run)
- If we perform a test of this hypothesis using the sample data, the p-value is 0.17
  - So are Prof. Miller and Steph Curry equally good shooters?

Misconception #1

- A high p-value indicates a *lack of evidence* against the null hypothesis
  - This is not the same as evidence in support of the null hypothesis
- In the Steph Curry example, the lack of evidence is due to the small sample size. If each participant took more shots we would quickly find sufficient evidence to reject the null hypothesis.

Misconception #2

Suppose a large-scale randomized experiment assigns 10,000 people to drink one glass of wine each day and another 10,000 people to completely abstain from consuming alcohol.

- Researchers found a lower risk of heart attack in the abstinence group, and the p-value was less than 0.001
  - Based upon this result, would you encourage complete abstinence from wine?

Misconception #2

- A low p-value provides no indication of *effect size*
  - It's possible that the risk of heart attack was only slightly higher (say 0.01%) in the wine-drinking group
    - This reduction in risk has little "practical significance" as its not a large enough effect to be worried about
- You should be careful to *avoid* interpreting small p-values as highly important results or large effects
  - Instead, small p-values simply indicate incompatibility between the sample data and the null hypothesis, and when the sample size is very large it's possible for a small amount of incompatibility to produce a very small p-value

Providing a Proper Conclusion

When interpreting the results of a hypothesis test in this class you must include the following information to receive full credit:

1. Scientific context about the study and the involved data
2. An indication of significance or strength of evidence
3. The direction of the relationship (if one was found)

You *should not* explicitly mention the null hypothesis unless you're specifically asked to do so.

Practice

A study on vaccine efficacy randomly assigned participants to receive the influenza vaccine in either the morning or the afternoon. Researchers measured the change in antibody levels for a specific flu strain over the next month. The morning group experienced an average increase of 105 ng/mL and the afternoon group saw an average increase of 120 ng/mL. The researchers performed a statistical test to evaluate the difference in mean antibody increases between these groups and the associated p-value was 0.16.

On the next slide you'll see several one-sentence conclusions summarizing this statistical test. You are to categorize each conclusion as "good", "okay", or "bad".

Practice (cont.)

Possible conclusions:

1. Whether a vaccine is received in the morning or afternoon most likely doesn't impact vaccine uptake success.
2. The study provides insufficient statistical evidence of a link between vaccine timing and successful uptake.
3. The study cannot reject that the success of vaccination is unrelated to the timing of the vaccine.
4. The p-value is large enough to suggest the researcher's null hypothesis is true.
5. The researchers can conclude with a very high level of confidence that the vaccine makes no difference.

Conclusion

- Hypothesis tests are a statistical tool used to measure the evidence that sample data provides against a null hypothesis
  - The *p*-value is the probability of obtaining an outcome at least as extreme as the one observed in the sample data under the assumption that the null hypothesis is true
  - The *p*-value *does not* tell you how likely it is that the null hypothesis is true
- Hypothesis tests and confidence intervals are *complementary tools*
  - Confidence intervals suggest plausible values for the parameter of interest, something that is not given by a hypothesis test