Notes 14 - Hypothesis Tests for Differences

STS 2300 Introduction to Data Analytics

Updated: 2025-05-01

Table of Contents

[Reading for Notes 14 1](#_Toc196977735)

[Learning Goals for Notes 14 1](#_Toc196977736)

[Hypothesis tests for 2](#_Toc196977737)

[Step 1 - Setting up the hypotheses 3](#_Toc196977738)

[Step 2 - Gathering and summarizing data 3](#_Toc196977739)

[Step 3 - Simulating a null distribution and finding a p-value 5](#_Toc196977740)

[Step 4 - Conclusions for hypothesis tests about 6](#_Toc196977741)

[Hypothesis tests for 6](#_Toc196977742)

[Step 1 - Setting up the hypotheses 7](#_Toc196977743)

[Step 2 - Gathering and summarizing data 7](#_Toc196977744)

[Step 3 - Simulating a null distribution and finding a p-value 8](#_Toc196977745)

[Step 4 - Conclusions for hypothesis tests about 9](#_Toc196977746)

[Theory-based hypothesis tests for or 9](#_Toc196977747)

[Using prop.test() for hypotheses about 9](#_Toc196977748)

[Using t.test() for hypotheses about 10](#_Toc196977749)

[Revisiting the Learning Goals for Notes 14 10](#_Toc196977750)

# Reading for Notes 14

Read [Chapter 9 (Sections 9.3 - 9.5) of the Modern Dive textbook](https://moderndive.com/9-hypothesis-testing.html) to supplement Notes 14.

# Learning Goals for Notes 14

* Be able to set up the hypotheses for hypothesis tests about differences (e.g., or ).
* Be able to use the infer package to conduct simulation-based hypothesis tests for differences.
* Be able to use R to calculate p-values for theory-based hypothesis tests for and .
* Be able to write conclusions for hypothesis tests for differences

I’ll be using the following packages in this set of notes, so I’ll load them before I get started.

library(infer)  
library(ggplot2)  
library(dplyr)

# Hypothesis tests for

In 2012, Pew Research Center asked Americans how they felt about the individual mandate portion of the Affordable Care Act. However, they asked the question in two different ways to see how question wording might (or might not) impact approval of the act. The question was phrased:

“As you may know, by 2014 nearly all Americans will be required to have health insurance. **Statement 1** while **Statement 2**. Do you approve or disapprove of this policy?”

where **Statement 1** and **Statement 2** were the two following statements presented in either order

* “people who do not buy insurance will pay a penalty”
* “people who cannot afford it will receive financial help from the government”

The main variable of interest is whether the respondent approves of the individual mandate or not. However, we will compare the proportion of people who approve between the two possible question wordings. The researchers will use as their significance level for their hypothesis test.

**Notice that for a test of it means we have *two categorical variables***

(Note: The original example is [here](https://www.pewresearch.org/politics/2012/03/26/public-remains-split-on-health-care-bill-opposed-to-mandate/) and was covered in the textbook [Introductory Statistics with Randomization and Simulation](https://www.openintro.org/book/isrs/)). A text file of the data is [here](https://www.openintro.org/data/tab-delimited/healthcare_law_survey.txt).

## Step 1 - Setting up the hypotheses

In hypothesis tests comparing two proportions, our null hypothesis will still have an equal sign (like it did for a single proportion). Again, the alternative hypothesis will look exactly the same as the null hypothesis except the equal sign will be swapped out for a <, >, or ≠ sign depending on our question of interest.

**Practice:** For the example above, Pew Research is interested in whether approval of the Affordable Care Act differs based on the ordering of the statements in the question. What are our hypotheses? I have defined the parameters and for us, but typically this would be part of defining your hypotheses.

p1 – p2 = 0 (or p1 = p2)

vs.

p1 – p2 ≠ 0 (or p1 ≠ p2)

where

= the population proportion of people who approve of the individual mandate when “cannot afford” wording goes second

= the population proportion of people who approve of the individual mandate when “penalty” wording goes second

(Note: We could put either group first/second as long as we are consistent throughout the whole hypothesis test.)

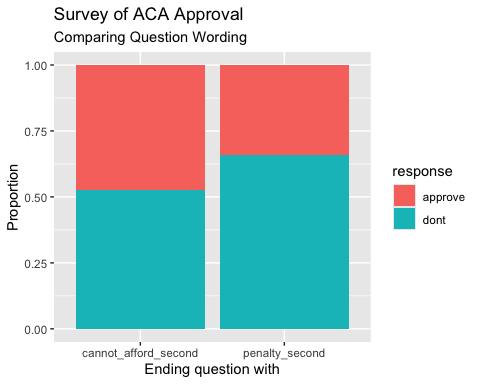
## Step 2 - Gathering and summarizing data

You can access this data by running the code below. The data contains two variables order (which says which part was asked second) and response (which says whether the person approved). (Note: I’ve used mutate() to change the response variable so that it only contains two groups instead of three.)

library(readr)  
ACA\_survey <- read\_delim("https://www.openintro.org/data/tab-delimited/healthcare\_law\_survey.txt", "\t",   
 escape\_double = FALSE,   
 trim\_ws = TRUE) |>   
 mutate(response = ifelse(response == "approve", "approve", "dont"))

We may want to look at a graph of our data. We saw earlier that stacked bar graphs work well for comparing two proportions.

**Practice:** See if you can recreate the graph below that visually summarizes our data in relation to our question of interest.



This data is best summarized numerically with the difference in sample proportions. We can use the specify() and calculate() functions from the infer package to calculate it.

ACA\_obsphatdiff <- ACA\_survey |>   
 specify(formula = response ~ order, success = "approve") |>   
 calculate(stat = "diff in props", order = c("cannot\_afford\_second", "penalty\_second"))  
ACA\_obsphatdiff

## Response: response (factor)  
## Explanatory: order (factor)  
## # A tibble: 1 × 1  
## stat  
## <dbl>  
## 1 0.133

Our observed statistic is . In other words, people asked the question with “cannot afford” second reported approving of the individual mandate 13.3 percentage points more often in our sample.

However, we want to know what is true of the *population*. To do this, we will see if results like this are likely (or unlikely) to happen in a world where there is no difference in **population** proportions between these groups.

## Step 3 - Simulating a null distribution and finding a p-value

We can again use the infer package functions to generate a null distribution. We will use null = "independence" in the hypothesize() function to say that in our null hypothesis the response variable is not impacted by which group someone is in. To create our null distribution, we use type = "permute" in the generate() function to shuffle up our response variable and randomly assign it to a group. The rest of the code will look like when we generated bootstrap distributions for differences in proportions.

null\_dist <- data |>   
 specify(formula = response ~ explanatory, success = "level") |>   
 hypothesize(null = "independence") |>   
 generate(reps = 1000, type = "permute") |>   
 calculate(stat = "diff in props", order = c("first", "second"))

The [promotions example](https://moderndive.com/9-hypothesis-testing.html#ht-activity) in Section 9.1 of the textbook gives another illustration of how “permute” is used to let us explore a world where our two population proportions are the same.

For our specific example, our code will look like:

set.seed(100731)  
ACA\_null <- ACA\_survey |>   
 specify(formula = response ~ order, success = "approve") |>   
 hypothesize(null = "independence") |>   
 generate(reps = 1000, type = "permute") |>   
 calculate(stat = "diff in props", order = c("cannot\_afford\_second", "penalty\_second"))

**Practice:** Write code to graph this distribution with a p-value shaded on it. Then calculate the p-value. (Note: You will get a warning along with your p-value. What is it telling you and why?)

**Answer:** We got a warning because our p-value was 0. The warning was to remind us that it doesn’t mean it’s impossible to get results like our (difference in p-hats of 0.133) in a world where the null is true, but it is rare enough that it didn’t happen in the 1000 samples we looked at from that world.

## Step 4 - Conclusions for hypothesis tests about

We started this problem by saying that the Pew Research Center would use a significance level of . With a p-value of essentially 0, we saw that the data they collected was not at all compatible with our null hypothesis.

**Practice:** Use this information to write an appropriate conclusion in context of this example.

**Conclusion:** Our p-value that’s close to 0 is well below the 0.05 significance level, so we reject the null hypothesis. We have very strong evidence to conclude that the population proportion of people who approve of the individual mandate is different depending on whether the part about not affording it or the part about a penalty is included second.

Recall, our conclusions should do the following…

* compare the p-value to the significance level
* stated what this means for the null hypothesis
* follow that with a sentence about the alternative hypothesis *in context* of the problem

In general, we can say we have sufficient evidence for anytime we have a p-value < . However, if our p-value is very small, we might communicate that this suggests even stronger evidence than just sufficient.

# Hypothesis tests for

There is a long history of people creating studies to test whether ESP (extrasensory perception) exists. Two researchers wanted to see if whether someone was a believer in ESP (or a skeptic of it) was related to their ESP ability. Specifically, they devised an experiment where participants would try to guess what was on 50 Zener cards. There are 5 possible symbols on each card, so someone guessing at random should get around 10 (20% of the 50 cards) correct. The researchers planned to test whether the population mean score would be higher for believers than for skeptics. Suppose the researchers have decided to use as their significance level. (You can find more information [here](http://users.stat.ufl.edu/~winner/data/para_skeptic.txt) if you’re interested.)

## Step 1 - Setting up the hypotheses

If nothing is going on, believers and skeptics will have the same population mean number of correct guesses. The researchers are interested in trying to show that believers have a *higher* population mean. Let’s say we consider believers to be group 1 and skeptics to be group 2. In that case, our two hypotheses would be (**fill in the end of** ):

vs.

> 0

where

= the population mean number of correct guesses for believers

and

= the population mean number of correct guesses for skeptics

## Step 2 - Gathering and summarizing data

The researchers gathered a sample of people and separated them into whether they were believers or skeptics of ESP. They then had them guess what was on 50 of the Zener cards and recorded the results. Since our question of interest is about the difference in population means, we can calculate the difference in sample means to see what happened in our data. (Note: The code below will read the data from our course code repository).

esp <- read.csv("https://raw.githubusercontent.com/nbussberg/STS2300-Spring2025/refs/heads/main/Data/esp.csv")

**Practice:** Calculate the difference in sample means along with an appropriate graph to visualize the data. What do you notice?

**Answer:** The two distributiosn look VERY similar. We get the same Q1, median, and Q3 for believers and skeptics. The difference in sample means is 0.218, which is pretty small for something on a scale from 0 to 50.

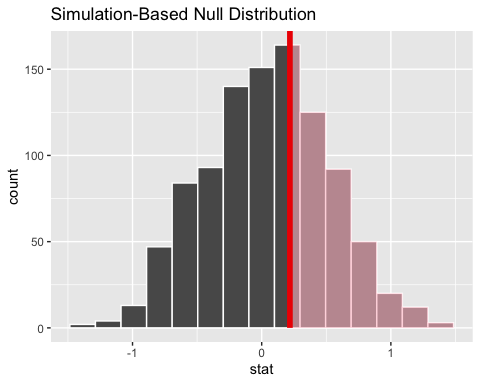
## Step 3 - Simulating a null distribution and finding a p-value

Below is generic code to utilize the infer package for a problem where we’re comparing two population means for our hypotheses.

null\_dist <- data |>   
 specify(formula = response ~ explanatory) |>   
 hypothesize(null = "independence") |>   
 generate(reps = 1000, type = "permute") |>   
 calculate(stat = "diff in means", order = c("first", "second"))

This code will look almost identical to our code for a difference in proportions. Again, our null hypothesis is that our response and explanatory variables are independent of one another (i.e., unrelated), which we can simulate by permuting (i.e., shuffling) one variable and rematching values to one another.

**Practice:** Generate a null distribution for our ESP example. Include the shaded p-value on the graph. Then calculate a p-value and write a sentence explaining what the p-value means in context.



**p-value =** 0.341

**Sentence:** In a world where there is no difference in the population mean number of card matches for believers and skeptics, we would see a sample mean difference of 0.218 or larger around 34% of the time. In other words, our data is very compatibvle with what the null hypothesis says.

## Step 4 - Conclusions for hypothesis tests about

From the p-value above, we can see that results like this would not be that unusual if these two groups had the same population mean. Since that’s our status quo, we haven’t really seen enough evidence to convince us we should change our minds. Our formal conclusion might look something like this.

**Conclusion**: Because my p-value is quite a bit bigger than , I will fail to reject the null hypothesis. I have insufficient evidence to conclude the population mean number of cards correct is higher for believers in ESP than it is for skeptics.

It looks like this data doesn’t really help to convince us that people who believe in ESP will have more of the ability than those who don’t.

**Question:** Suppose that 100 people conducted the exact same experiment. In other words, they each took a sample and had the believers and skeptics guess what was on the 50 cards. Do you think any of these people would end up with p-values under ? If so, how many? What does this tell us about the role of our significance level in a hypothesis test?

**Answer:** In a world where the null hypothesis is true, we will get p-values below 0.10 around 10% of the time, so I would expect around 10 of the 100 people doing this experiment to reject the null and conclude believing in ESP helps match the cards. Thus, the significance level controls how often we will reject the null when the null is actually true. This might make it seem like we should use a small a value as possible, but this would make it harder for us to reject the null when the null is in fact not true.

# Theory-based hypothesis tests for or

Again, we can use prop.test() (for tests of ) and t.test() (for tests of ). These functions will use the same arguments that we used in the previous set of notes.

## Using prop.test() for hypotheses about

Below is a table of values in our two variables from our Affordable Care Act data.

table(ACA\_survey$order, ACA\_survey$response)

##   
## approve dont  
## cannot\_afford\_second 365 406  
## penalty\_second 249 483

**Practice:** Use the results above along with prop.test() to carry out a test of vs. . Remember that your x and n arguments will each have two values (using c() to enter them as a vector). You can use the alternative argument to specify the symbol in .

**p-value =** 0.000000199 (written as 1.99e-07 in the R output)

## Using t.test() for hypotheses about

**Practice:** Write code to get a theory-based p-value for the esp example using t.test(). Use the alternative argument to specify the symbol in . (Hint: Think about which order R will use for the groups.)

**p-value =** 0.3176

# Revisiting the Learning Goals for Notes 14

* Be able to set up the hypotheses for hypothesis tests about differences (e.g., or ).
* Be able to use the infer package to conduct simulation-based hypothesis tests for differences.
* Be able to use R to calculate p-values for theory-based hypothesis tests for and .
* Be able to write conclusions for hypothesis tests for differences