Notes 15 - Other Hypothesis Tests (Chi-Squared & ANOVA)

STS 2300 Introduction to Data Analytics

Updated: 2025-05-03

Table of Contents

[Reading for Notes 15 1](#_Toc197164390)

[Learning Goals for Notes 15 1](#_Toc197164391)

[Chi-Squared Hypothesis Tests 2](#_Toc197164392)

[Step 1 - Setting up the hypotheses 3](#_Toc197164393)

[Step 2 - Gathering and summarizing data 3](#_Toc197164394)

[Step 3 - Simulating a null distribution and finding a p-value 4](#_Toc197164395)

[Step 4 - Conclusions for Chi-Squared hypothesis tests 5](#_Toc197164396)

[ANOVA tests comparing many population means 5](#_Toc197164397)

[Step 1 - Setting up the hypotheses 6](#_Toc197164398)

[Step 2 - Gathering and summarizing data 6](#_Toc197164399)

[Step 3 - Simulating a null distribution and finding a p-value 8](#_Toc197164400)

[Step 4 - Conclusions for ANOVA hypothesis tests 9](#_Toc197164401)

[Theory-based ANOVA and Chi-Squared hypothesis tests 9](#_Toc197164402)

[Revisiting the Learning Goals for Notes 15 10](#_Toc197164403)

# Reading for Notes 15

You can reference the infer package documentation for the [Chi-Squared test](https://infer.netlify.app/articles/chi_squared) and [ANOVA test](https://infer.netlify.app/articles/anova) to supplement Notes 15.

# Learning Goals for Notes 15

* Be able to set up the hypotheses Chi-Squared and ANOVA tests
* Be able to use the infer package to conduct simulation-based Chi-Squared and ANOVA hypothesis tests
* Be able to use R to calculate p-values for theory-based Chi-Squared and ANOVA hypothesis tests
* Be able to write conclusions for Chi-Squared and ANOVA hypothesis tests

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)

Both Chi-Squared and ANOVA tests are extensions of tests that we have already learned about.

In a Chi-Squared test, we have two categorical variables (like we do in a difference in proportions test) and are looking at whether they are related. Unlike a difference in proportions test, each categorical variable may have more than two categories.

In an ANOVA test, we have one quantitative variable and one categorical variable (like we do in a difference in means test) and are comparing the means of the groups. Unlike a difference in means test, we may have more than two means that we are comparing.

# Chi-Squared Hypothesis Tests

Chi-squared tests are used when we want to see if there is a relationship between two categorical variables. A test for a difference in sample proportions would be the simplest version of this, but with a Chi-Squared test we may have more than two categories with one or both of our variables.

The Elon Poll contacted a random sample of recent college graduates (34 or younger) to ask if they thought college was “worth it” for them (broken into four categories). They also recorded whether the person was a first generation student or not. Suppose we’re interested in whether or not there’s a relationship between these two variables.

[Full report here](https://eloncdn.blob.core.windows.net/eu3/sites/819/2019/07/2019_7_31-ElonPoll_Report.pdf)

## Step 1 - Setting up the hypotheses

Our hypothesis for Chi-Squared tests can be difficult to write in symbols, so we will write them out in sentences. Our null hypothesis will say there is *no relationship* between our two variables (i.e., the variables are independent). The alternative hypothesis will say there is *some sort of relationship* between the variables (i.e., the variables are not independent).

For the Elon Poll example that might look like:

For recent college grads, there is **no relationship** between first-generation status and opinion on whether college was “worth it”

For recent college grads, there is **some sort of relationship** between first-generation status and opinion on whether college was “worth it”

## Step 2 - Gathering and summarizing data

Below is code to read the data from the Elon Poll into R and to display it in a table. The levels argument I included in data.frame() is used to maintain the ordering of the categories in R (instead of putting them alphabetically by default).

college <- data.frame(worthit = factor(c(rep("def\_yes", 263), rep("prob\_yes", 197),  
 rep("prob\_no", 90), rep("def\_no", 47),  
 rep("def\_yes", 527), rep("prob\_yes", 331),  
 rep("prob\_no", 88), rep("def\_no", 29)),  
 levels = c("def\_yes", "prob\_yes",  
 "prob\_no", "def\_no")),  
 firstgen = c(rep("yes", 263 + 197 + 90 + 47),  
 rep("no", 527 + 331 + 88 + 29)))  
  
table(college$worthit, college$firstgen)

##   
## no yes  
## def\_yes 527 263  
## prob\_yes 331 197  
## prob\_no 88 90  
## def\_no 29 47

**Practice:** Create an appropriate graph to visualize this data. (Hint: What have we done for difference in proportions examples that also involved two categorical variables?)

To summarize data in a Chi-Squared test, we use a statistic. The statistic compares what we observed to what we would expect if the null hypothesis were true. This is done for each of the category combinations for our two variables. Then we add the results for each cell together.

The formula is

We can use the infer package or the chisq.test() functions to calculate this value. For either method, it doesn’t matter what order we use for the variables. Below is generic code for both approaches:

# Using infer  
data |>   
 specify(var1 ~ var2) |>   
 calculate(stat = "Chisq")  
  
# Using chisq.test()  
chisq.test(data$var1, data$var2)

**Practice:** Use one of the two methods to calculate a statistic.

**Answer:** We get 37.8 regardless of which method we use. However, we don’t have context yet for what this number means.

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

Below is generic code to utilize the infer package to create a null distribution for a Chi-squared test. The only difference between this code and code for a difference in proportions is:

* We do not need to include a success argument in specify() because we aren’t focusing on a particular category.
* We use stat = "Chisq" instead of stat = "diff in props" to calculate a chi-squared statistic.

null\_dist <- data |>   
 specify(formula = var1 ~ var2) |>   
 hypothesize(null = "independence") |>   
 generate(reps = 1000, type = "permute") |>   
 calculate(stat = "Chisq")

**Question:** Which way will we shade our p-value on our null distribution and why? (Hint: Consider the formula for the statistic and what it would look like when vs.  are true.)

**Answer:** We will shade to the right. We always shade in the direction that favors the alternative hypothesis. Since our statistic looks at the difference between what we observed and what we expect when H0 is true, we will get bigger values of the statistics when the alternative is true.

**Practice:** Generate a null distribution for our Elon Poll example. Include the shaded p-value on the graph. Then calculate a p-value. Write a sentence explaining what it means in context.

**p-value =** 0

**Sentence:** In a world where there is no relationship between first generation status and how someone feels about whether college was “worth it,” we would almost never see results as extreme as the ones in our sample. (Note: This is among recent college graduates.)

## Step 4 - Conclusions for Chi-Squared hypothesis tests

**Practice:** Write a conclusion for our test. Consider the same things we have considered when writing our other hypothesis test conclusions this semester.

**Conclusion:** Since our p-value is approximately 0, we reject the null hypothesis. There is very strong evidence of some sort of relationship between first generation status and opinions on whether college is “worth it” among recent college graduates.

Note: we haven’t said WHAT the relationship is, but we could use something like our graph from earlier to show people what it looked like in our sample.

# ANOVA tests comparing many population means

As we said at the start, ANOVA tests are used to compare many different population means. Suppose that a cereal box company is planning a new cereal box design. Obviously, their goal is to choose a design that will make more people buy their cereal (or at least not deter people). Their marketing department has come up with four possible designs (called A, B, C, and D), and they’ve devised an experiment to test how well the designs sell. They’ve chosen twenty stores that are relatively similar in size, location, etc. Each of the four designs will be sent to five stores for a couple weeks after which the researchers will record the number of cases of cereal sold at each store.

## Step 1 - Setting up the hypotheses

In an ANOVA test, our null hypothesis says that all of our groups have the same population mean (nothing going on) while our alternative hypothesis says that at least one of the groups has a different population mean than another. Notice that we are **not** saying all of the population means are different. This means it can be difficult to write our alternative hypothesis in symbols. For this class, we will just focus on writing the two hypotheses in words.

**Practice:** Set up a null and alternative hypothesis for the cereal box example. Make sure you include context for this specific example.

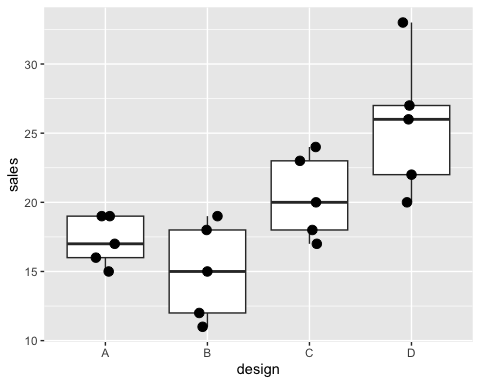
The population mean sales are the same for all four package designs.

At least one package design has a different population mean sales than another.

## Step 2 - Gathering and summarizing data

Below is code to read in the cereal data and to visualize it using boxplots. I’ve added the points to the plot using geom\_jitter(). The height and width arguments are used to control how much random noise to add to the data. By setting height to 0, the points maintain their correct sales value on the graph (y axis). By setting width to 0.1, a small amount of side-to-side variation is added so that we can see points that would otherwise overlap.

cereal <- data.frame(design = c(rep("A", 5), rep("B", 5),   
 rep("C", 5), rep("D", 5)),  
 sales = c(19, 17, 16, 19, 15, 12, 18, 15, 19, 11,  
 23, 20, 18, 17, 24, 27, 33, 22, 26, 20))  
  
ggplot(cereal, aes(x = design, y = sales)) +  
 geom\_boxplot() +  
 geom\_jitter(height = 0, width = 0.1, size = 3)



To summarize data in an ANOVA test, we use an statistic. The statistic compares average variability between groups (called Mean Squares Between or MSB) and average variability within group (called Mean Squares Within or MSW).

The formula is:

If the null hypothesis is true, we will see similar variability (due to random chance) both within and between groups. If the alternative hypothesis is true, we would expect larger F statistics because the groups are differing for reasons other than random variability.

(Note: To learn more about how to calculate MSB and MSW as well as the many extensions of ANOVA tests, you can take a class like *STS 3250: Design and Analysis of Experiments*, which is offered each Fall)

We can use the infer package or the aov() and summary() functions to calculate the statistic. Below is generic code for both

The response is the quantitative variable. The explanatory is the categorical variable.

data |>   
 specify(response ~ explanatory) |>   
 calculate(stat = "F")  
  
aov(response ~ explanatory, data = data) |>   
 summary()

**Practice:** Calculate an F statistic using each of the methods. How does the output differ?

**Answer:** The first approach gives us just our F statistic. The second approach gives something called an ANOVA table that includes multiple pieces. The F statistic and the p-value are the two most notable ones we would use.

cereal\_F <- cereal |>   
 specify(formula = sales ~ design) |>   
 calculate(stat = "F")

We should end up with an F statistic of 8.4. This tells us that the average variability between stores was over 8 times as large as the average variability within stores. Like other hypothesis tests, we will use a p-value to decide whether or not random chance is a reasonable explanation for data like this.

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

Our code for the null distribution will look very similar to our code for a difference in means problem. The main differences are that we use stat = "F" and we do not need to specify an order for our groups.

null\_dist <- data |>   
 specify(formula = response ~ explanatory) |>   
 hypothesize(null = "independence") |>   
 generate(reps = 1000, type = "permute") |>   
 calculate(stat = "F")

**Question:** Which direction should we shade for the p-value and why?

**Answer:** We will shade to the right because our F statistic will get bigger when there are differences between group means (i.e., when the alternative hypothesis is true).

**Practice:** Create a null distribution for the cereal example (remember to set a seed). Shade your p-value on the graph and then calculate the value.

**Review:** What does our p-value mean in context of this example?

**Answer:** In a world where all four package designs have the same population mean sales, we would see an F statistic at least as big as 8.42 only 0.01% of the time.

## Step 4 - Conclusions for ANOVA hypothesis tests

Like all hypothesis tests, we will use our p-value to reach a conclusion. Our conclusion will:

* compare the p-value to the significance level
* state what this means for the null hypothesis (reject or don’t reject)
* include a sentence about the level of evidence for the alternative hypothesis *in context* of the problem

**Practice:** Write a conclusion for our hypothesis test in context of this example. It may help to refer back to our hypotheses.

**Conclusion:** Since the p-value of 0.0001 was well below our 0.05 significance level, we will reject the null hypothesis. There is very strong evidence that at least one of the four package designs has a different population mean sales than another design.

Note: The next step would be to use simultaneous confidence intervals to explore *which* population means differ from one another. There is no good way to do this with the infer package or without installing new packages in R. I’m happy to answer questions on the topic, but you should take STS 3250 if you’d like to learn more.

# Theory-based ANOVA and Chi-Squared hypothesis tests

As we saw above, the aov() and summary() functions can be combined to conduct a theory-based ANOVA hypothesis test. This test has the following assumptions/conditions:

1. Each group is a random sample from the population (or the sample is randomly assigned to groups)
2. Each group comes from a population with a normal distribution (or has a large enough sample size)
3. Samples come from populations with equal variances

Condition 1 is also needed for simulation-based tests. Condition 2 would be assessed by looking at graphs to see if our *sample* seems close to bell-shaped. There are separate tests that can be done to assess Condition 3.

The chisq.test() function can be used to conduct a theory-based Chi-Squared hypothesis test. It has the following assumptions / conditions:

1. The data is a random sample from the population
2. The sample size is “big enough”

Condition 1 is needed for simulation-based tests also. Condition 2 is assessed in different ways but often is considered to be met if most (e.g., at least 80%) of the cells have expected counts of at least 5.

# Revisiting the Learning Goals for Notes 15

* Be able to set up the hypotheses Chi-Squared and ANOVA tests
* Be able to use the infer package to conduct simulation-based Chi-Squared and ANOVA hypothesis tests
* Be able to use R to calculate p-values for theory-based Chi-Squared and ANOVA hypothesis tests
* Be able to write conclusions for Chi-Squared and ANOVA hypothesis tests