Lab 8. Chi Squares

# Introduction

Today you will be running both Goodness-of-Fit Chi Squared tests and Chi Squared Tests for Independence. You’re going to use these to evaluate a few fake data questions, and then do a walk through with real data to determine whether Tucson or Phoenix pilots are more efficient at using their giant planes to murder different species of birds - as well as their overall overall bird murder rates. Or, conversely, you’ll be testing which city has birds that are best at flying into planes - it could be their fault I suppose.

The real data is not in a contingency table, so you will be learning some tricks to make it into the right format. Then as your final portion of the lab, you’ll be picking two (or more) new airports to look at and cross-compare.

**Note:** I got a bit carried away with synonyms and slang for diarrhea. So if you don’t understand a term… it’s probably a slang term for diarrhea. This will make more sense shortly.

# Learning Outcomes

By the end of today’s class you should be able to do the following in R:

* Use the chisq.test() function to calculate goodness of fit
* Use the chisq.test() function to determine if data independence
* Interpret Chi Square p values and make recommendations
* Use dplyr and reshape2 to create contingency tables
* Use ggplot2 to make a filled barplot
* Use mutate() to make new columns
* Use filter() to, well, filter
* Use select() to pick different columns or drop them
* Use dcast() to make your data into wider format

# Part 1: Revisiting Chi Square Test Types

Before we begin looking at how to run these, let’s make sure you can tell the difference between a goodness of fit test and a test of independence. If you are unsure on the answers to any of the following questions, you might want to revisit the lecture before you continue.

We will be practicing with fake data from two different examples.

##Part 1.1 Example 1 (Cars) White cars can be hard to see. You are trying to see if white cars get into more accidents then black cars or cars of other colors. You have data from 2019, the total number of cars that are white, black, or all other colors and the number of cars that got into accidents for each color category.

### QUESTION 1: What type of chi-squared question is Example 1?

1. Goodness of fit
2. Independence

### QUESTION 2: How would you calculate your degrees of freedom for Example 1?

1. n - 1, leaving you with 2 degrees of freedom
2. (R-1) \* (C-1), leaving you with 4 degrees of freedom

## Part 1.2 Example 2 Poosplosions

You are working with a group of people who have an intestinal disease that causes frequent and volcanic diarrhea. You are trying to see if a particular medication causes fewer cases of violent diarrhea. You have people on the medication and people off the medication, and their relative numbers of folks with and without diarrhea.

### QUESTION 3: What type of chi-squared question is Example 2?

1. Goodness of fit
2. Independence

### QUESTION 4: How would you calculate your degrees of freedom for Example 2?

1. n - 1, leaving you with 3 degrees of freedom
2. (R-1) \* (C-1), leaving you with 1 degrees of freedom

# Part 2: Fake Examples

To run today’s code you will need the ggplot2, reshape2 and dplyr packages installed. Once they are installed, use the library() function to remind R to load them.

library(dplyr)  
library(ggplot2)  
library(reshape2)

## Part 2.1 Make The Data

To begin with, we will run both types of chi-squared tests using these fake data examples. The below code will create a data frame for each example – and I’m fairly sure that by the names you can tell which is which. The reason we are starting with these is because this data is already in **contingency table** format - when we work with real data in Part 3, we have to do a lot more work to get your data into the right format.

cars <- data.frame(Color = c("White", "Black", "Everything Else"), Accidents = c(252, 109, 450), TotalCars = c(2023, 4005, 12124))  
  
poopoo <- data.frame(DevilsCoffeeStatus = c("Butt Barfing", "Enjoying Life"), Control = c(35, 78), Medication = c(53, 13))

## Part 2.2 Cars Example

In this question we are looking to see if the proportion of cars of each color has an equal number of accidents. We have the total number of cars for each color, but we are more interested in asking whether or not white cars have a lot of accidents versus whether or not people prefer purchasing white cars. So to start off with we need to calculate what our expected percentages of accidents would be given the number of cars in each category.

To calculate percentages, we first need to find the total number of cars (**all**) using the sum() function.

all <- sum(cars$TotalCars)

Now we need to find the overall percentage of cars that fall into each category. We can do this using a new function from the dplyr package: mutate(). This essentially is a way of quickly doing column math without having to write a lot of code. The mutate function takes two arguments: the name of the new column, and whatever you wanted to have done.

We want to have the overall proportion of each type of car (**Proportion.Of.Cars**).

cars <- cars %>%  
 mutate(Proportion.Of.Cars = TotalCars/all)

The chisq.test() function takes two arguments: **x**, the actual number of accidents, and **p**, the probabilities of accidents expected for each type. Though the hand calculation of Chi Square requires expected observations rather than proportions, R likes to do that for you. It instead wants you to specify the expected probabilities.

chisq.test(x= cars$Accidents, p = cars$Proportion.Of.Cars)

This is statistically significant. What does that actually mean? We need to look at effect size to find out. To get this, we need the number of accidents, and to do a little column math.

total.accidents <- sum(cars$Accidents)  
  
cars <- cars %>%   
 mutate(ExpectedAccidents = total.accidents\*Proportion.Of.Cars) %>% #how many accidents we expected by category  
 mutate(EffectSize = Accidents/ExpectedAccidents)

### QUESTION 5: Look at the chi square results and the effect size column. Are white cars more likely to be involved in accidents? Use your p value and your effect size to explain your answer.

## Part 2.3 Pants Gravy Problem

In this case, we are interested in looking both at the people who took the medication, and those that didn’t, to see if there’s interaction between the medication and the likelihood of having a Downstairs Exxon Valdez. That means we’re not actually calculating expected values, we’re looking for a relationship.

To do that, we have to make this data into a format R likes, which is a matrix(). They look like data frames, but in a matrix all the data has to be the same type (no categories, etc.). We can use the select() function to only take the columns we want, and then the as.matrix() function to turn it into a matrix.

muddymatrix <- poopoo %>%  
 select(Control, Medication) %>%  
 as.matrix()

And now place that matrix in the chisq.test() function. No need to add anything else!

chisq.test(muddymatrix)

Again, we will want to look at effect size to interpret these results. In this case, our expected values are calculated a little differently - across columns, rather than compared to the whole. You can still use mutate() for this.

poopoo <- poopoo %>%  
 mutate(EffectSize = Medication/Control)

### QUESTION 6: If you wanted to lower diarrhea rates in this group of people, should you put them on this medication? Use your p value and effect size to defend your answer.

# Part 3: Real (Complicated) Data

We will be using some real data on collisions between wildlife and aircraft between 1990 and 2015, collected from <https://www.kaggle.com/faa/wildlife-strikes>.

Import this dataset (angrybirds.csv) and call it **birds**. 

If you are having problem loading it or downloading it, use the smolangrybirds.csv file instead, which has fewer states.

When I was initially looking at this data I thought it was only birds. It isn’t. There are coyotes and bats and other non-avian wildlife in here too (no shark tornadoes though). I just want to clarify before you do this part that I do have a degree in biology and understand that coyotes and bats are not birds, but I also have a dumb sense of humor and think it is funnier to pretend that they are.

## Part 3.1 Subsetting Data

There is a lot of data in here, but what we really want is just Tucson and Phoenix airport data and we need it in a contingency table. So we have to do a lot of work to get it there. We are going to use the filter() function from dplyr to subset our data. We are giving the data set a new name (**birbs1**) instead of overwriting **birds** because we are going to come back to this data set later.

birbs1 <- birds %>%  
 filter(Airport %in% c("TUCSON INTL","PHOENIX SKY HARBOR INTL ARPT"))

We now need to have a total count by species for each airport. We can use the group\_by() and tally() functions from dplyr to get this information.

birbs1 <- birbs1 %>%  
 group\_by(Species.Name, Airport) %>%  
 tally()

But right now this data is long and hard to read, and won’t work well for a chi square test. Instead, we need to make it wide format. We can do this using dcast() from reshape2.

dcast() works as a formula, where you put the things you want to have stay in rows on the left-hand side and the thing you want to have be new columns on the right hand side. You do have to tell R what sort of data you want to have fill the new columns - sometimes it will figure that out automatically but it can be unpredictable. You use the **value.var** argument and put the column name in quotation marks (**"n"**). Finally, use the **fill** argument to tell R that any species&airport combination that doesn’t have a data point should be a 0.

widebirbs <- birbs1 %>%  
 dcast(Species.Name~Airport, value.var = "n", fill = 0)

### QUESTION 7: If you remove the fill = 0 portion of this function, what happens to missing values?

1. They are sad and lost forever
2. They become zeros
3. They become NA values

### QUESTION 8: If you remove the quotations around “n” from this function, what happens?

1. Your code runs totally perfectly and RStudio gives you a high five
2. R thinks that n is a function and can’t find it
3. R thinks n is an object, not a column
4. R recommends that n search out a life partner via Match.com

## Part 3.2 For Chi Square of Independence and for ggplot

We want to remove any species that doesn’t have an observation at both airports, as zeros do not work well in a chi-square test. Additionally, there are a lot of species here and we really only want to look at more common ones. So we will use mutate() to create a new column adding together both airports, and then use filter() to only take observations that have more than five between both airports. We will name this new data set **somewidebirbs** as we will need the full data down below.

somewidebirbs <- widebirbs %>%  
 mutate(Total =`TUCSON INTL`+`PHOENIX SKY HARBOR INTL ARPT` ) %>% #add species together  
 filter(`TUCSON INTL` != 0 & `PHOENIX SKY HARBOR INTL ARPT`!= 0) %>% #Remove 0-observation species  
 filter(Total >5) #filter to only be reasonably common animals

Ahh but we want to look at it in ggplot2, and that uses long format! Use melt() to put your data in long format (**somelongbirbs**). We also do not want the **Total** column, so we will use select() with a minus sign in front of the column to get rid of it.

somelongbirbs <- somewidebirbs %>%   
 select(-Total) %>%   
 melt()

Now we will look at our data before going any further. This code creates a “filled” barplot, which looks at relative frequency rather than counts per category. You can use a regular barplot as well if that makes more sense to you.

ggplot(somelongbirbs, aes(y = Species.Name, x = value, fill =variable ))+  
 geom\_bar(stat="identity", position = "fill")

This dataset is a good example of how seemingly the same data can be very inconsistent between different data collectors. For example, according to this graphic Tucson seems to be hitting a lot of Doves, yet no Pigeons.

### QUESTION 9: Why is the difference between Dove and Pigeon strikes between the two airports most likely not real:

1. Tucson hits more birds just generally
2. Tucson seems to have a harder time identifying birds (they have more unknown bird hits)
3. A Rock Pigeon and a Dove are the same bird with different names
4. Phoenix is better at reporting their numbers
5. Tucson has more symbolic “release the doves” events
6. A Dove is just a Pigeon that got bleached in the sun from being further south in Tucson

We now need to make a matrix of our data. We can then plug this straight into a chi-squared formula to look at our results.

matrixbirbs <- somewidebirbs %>%  
 select(`TUCSON INTL`,`PHOENIX SKY HARBOR INTL ARPT` ) %>%  
 as.matrix()  
  
chisq.test(matrixbirbs)

### QUESTION 10: Are species and airport related in this dataset? Use your p value to support your answer.

## Part 3.3 Chi Square for Goodness of Fit

What about the overall proportion of collision rates? Is it equal between the two airports?

While we have been using the filtered data set (**somewidebirbs**) for our chi-squared before we want to make sure we are using the full data set (**widebirbs**) here as we are combining all of the observations and so it doesn’t matter if a species only got hit once.

We need a vector of the total number of birds hit for each airport (**deadbirbs**) and a vector of probability (**birbs.p**).

P.birbs <- sum(widebirbs$`PHOENIX SKY HARBOR INTL ARPT`)  
T.birbs <- sum(widebirbs$`TUCSON INTL`)  
deadbirbs <- c(P.birbs, T.birbs)  
birbs.p <- c(.5, .5)

Now we can use these values in a chi-squared test for goodness of fit. Essentially, we are determining if he would expect to see this number of bird related accidents for both airports if the probability of accidents at both airports was equal.

chisq.test(x = deadbirbs, p = birbs.p)

### QUESTION 11: According to your results, which of the below is true?

1. We fail to reject the null hypothesis that the two airports have the same number of bird related accidents
2. We fail to reject the null hypothesis that the two airports have the same probability of bird related accidents
3. We accepted the alternate hypothesis that the two airports have different numbers of bird related accidents
4. We conclude that the birds in Phoenix have better eyesight than in Tucson

Now, Wikipedia says that Tucson had ~124,000 aircraft operations in 2020 while Phoenix Sky Harbor had ~310,000 aircraft operations in the same year. This information changes your overall probability for each category. Below is the code to create probability for each airport given the number of flights they operate in a year.

It is very important here to be careful to make sure that you have put them in the right order to match what the order of airports in your frequency vector (**widebirbs**)!

P.p <-310000/(124000 + 310000) #phoenix probability  
T.p <- 124000/(124000 + 310000) #tucson probability  
planes.p <- c(P.p, T.p)  
chisq.test(deadbirbs, p=planes.p)

### QUESTION 12: Which of the following is a good measure of effect size for showing which airport has the higher liklihood of hitting an animal?

A <- (P.birbs/(P.birbs + T.birbs))/P.p  
B <- P.birbs/(P.birbs + T.birbs)  
C <- T.p/P.p  
D <- P.birbs - T.birbs

As a hint, Phoenix is 1.08 times as likely to hit birds as expected by the number of their flights.

# Part 4: Your Turn!

Pick two or more airports from the dataset. They cannot be Tucson or Phoenix. Use Wikipedia to find how many flights your airports ran last year.

### QUESTION 13: Which airports are you looking at, and what does Wikipedia say is their number of aircraft operations last year?

Then analyze your data using the steps outlined in Part 3 to answer the following questions: 1) Is there a relationship between species accident count and airport? (Chi Square of Independence) 2) Is it fair to assume that they have equal probability of wildlife accidents? (Chi Square of Goodness of Fit, 50:50 chances) 3) Given their number of flights per year, can you tell if either airport is more likely to hit wildlife than the other? (Chi Square of Goodness of Fit, Using Overall Number of Flights for Proportions)

### QUESTION 14: Copy-paste your R script for JUST Part 4.

It should include all the steps you need to do this part, and nothing else. You don’t have to supply a graphic, but it is helpful!

### QUESTION 15: Is there a relationship between species accident count and airport? Use your p value to defend your answer.

### QUESTION 16: Is it fair to assume that they have equal probability of wildlife accidents? Use your p value and effect size measurement to defend your answer.

### QUESTION 17: Given their number of flights per year, can you tell if either airport is more likely to hit wildlife than the other? Use your p value and effect size measurement to defend your answer.