**2-way Tables, Row/Column Percents, and Tests**

In this lab, we will begin by reviewing some data summaries for categorical variables that we did at the end of the Summaries of Data lab. We’ll create 2-way tables and generate row and column percentages to get some sense of the way 2 categorical variables are related. Then, we’ll get some practice implementing a test for the significance of such relationships.

**#1** Consider the following data (counts) for the relationship between number of ear piercings and whether a student has a tattoo for 1014 male Penn State students.

|  |  |  |  |
| --- | --- | --- | --- |
| **Ear Piercings** | **No tattoo** | **Tattoo** | **Total** |
| **0** | 699 | 68 | 767 |
| **1** | 87 | 27 | 114 |
| **2 or more** | 87 | 46 | 133 |
| **Total** | 873 | 141 | 1014 |

**a.** Use the counts given above to calculate **column percentages** for this table. Fill in the following table with those column percentages.

|  |  |  |
| --- | --- | --- |
| **Ear Piercings** | **No tattoo** | **Tattoo** |
| **0** |  |  |
| **1** |  |  |
| **2 or more** |  |  |
| **Total** |  |  |

You can achieve this using R:

ex1=matrix(c(699,68,87,27,87,46),nrow=3,byrow=T)

prop.table(ex1,2) ##give proportions over the second dimension (columns) 🡪 using a

##1 would give row%

##If you prefer to be more explicit, try the following:

ex1[,1]/sum(ex1[,1])

ex1[,2]/sum(ex1[,2])

##You can even **bind** these **c**olumns of info into a matrix!

cbind(ex1[,1]/sum(ex1[,1]), ex1[,2]/sum(ex1[,2]))

**b.** Explain why the answer to part a of this activity suggests that there may be a relationship between number of ear piercings and whether a student has a tattoo.

**c.** This is just suggestive of a relationship, but the only way to know if this is truly meaningful is to perform the hypothesis test. To do this in R, simply run:

chisq.test(ex1,correct=FALSE)

Write the null and alternative hypotheses that are being tested by this test

On the basis of the test, what are we able to conclude?

**d.** The result in part **c.** was “statistically significant.” Explain what that means.

**#2** At the course website, access the folder for today’s lab, where again I have placed a dataset called chi2\_ex2.R. On a Mac, you can double-click on this .R file directly. If R is not open, it will launch it and produce a line of code, and if it is already open, look in the console and see that it produced a line of code for you with the load() command. On a Windows machine, you must use the load command. You can use **load(file.choose())** and then navigate to the location of the file in the resulting window. Alternatively, first download the data (say to your Desktop), and then change the working directory of R to the location where you downloaded the data (File<Change dir…<). Then simply use

load("chi2\_ex2.R")

The data now exists in R, and to see what it is called, use the ls() command to get a list of all objects currently in R. The data are from an introductory statistics class at UC Davis.

**a.** The variable ***Cheat*** gives responses to the question “Have you ever cheated on a significant other?” Use **with(ex2,table(Sex,Cheat))** to fill in the following two-way table with counts for the relationship between ***Sex*** and ***Cheat***. Be sure to disregard the missing values in the first column and their effects on the totals.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Cheated on SO?** | |  |
| **Gender** | No | Yes | Total |
| Female |  |  |  |
| Male |  |  |  |
| Total |  |  |  |

Those missing values are going to be problematic for us in having the software do our calculations. If we want to omit them and make this cleaner, we can do the following:

ex2.1=subset(ex2,ex2$Cheat %in% c("Y","N"))

t=with(ex2.1,table(Sex,Cheat))

Now let’s fill in this two-way table with row percentages via prop.table(t,1).

|  |  |  |
| --- | --- | --- |
|  | **Cheated on SO?** | |
| **Gender** | No | Yes |
| Female |  |  |
| Male |  |  |

**b.** Write a sentence that interprets the row percentage given in the “Female” row and “No” column in part a.

**c.** Explain whether you think that the table of percents found in part **a**. is evidence of a relationship between the two variables in this activity or not.

**d.** Now let’s peform the chi-square test. Use: chisq.test(t,correct=FALSE)

What is the *p*-value for the test?

Is the observed result “statistically significant?” Explain why or why not.

**e.** Write a sentence than gives a conclusion about the two variables in this activity (gender and whether student has cheated on significant other.).

**f.** Refer back to the chi-square test done in part **d.** Write the null and alternative hypotheses being tested by that chi-square test. Make your answer specific to the variables in this example.

**#3** Continue to use the ex2 dataset. The variable ***Hand*** records handedness.

**a.** Use steps like those in #2a to fill in the following tables with counts *and* (row) percentages by gender.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Handedness** | | |
| **Gender** | Left | Right | Total |
| Female |  |  |  |
| Male |  |  |  |
| Total |  |  |  |

|  |  |  |
| --- | --- | --- |
|  | **Handedness** | |
| **Gender** | Left | Right |
| Female |  |  |
| Male |  |  |

**b.** On the basis of percentages that you found for part **a**., describe the relationship between the two variables in this activity.

**c.** Do a chi-square test for statistical significance of the observed relationship (see #2d).

(i) Give a *p*-value for the test.

(ii) Explain whether the observed relationship is statistically significant.

**#4** Continue to use the ex2 data. The variable ***Seat*** gives student responses to the question “Where do you normally sit in the classroom? **B**ack of the room, **M**iddle of the room, or **F**ront of the room.” Analyze the relationship between ***Gender*** (female, male) and ***Seat***. Write a paragraph that summarizes your analysis. In your summary compare relevant conditional percentages to describe the relationship and report the result of a chi-square test of statistical significance. State a clear conclusion for the test of significance. NOTE: You might want to get rid of some missing Seat values using the following code:

t=with(ex2,table(Sex,Seat))

t=t[,-1]

**#5** At the course website, access the folder for today’s lab, where again I have placed a dataset called chi2\_ex5.R, which, on a Mac, you can double-click directly to load it into R (with the load command produced in the console). On a Windows machine, use the load command again (if the data are saved in the same location), or again use **load(file.choose())**. It gives data for a few questions asked in the Spring 2006 PSU Stat 200 survey. The variable ***FavMusic*** gives responses to the question, “What is your favorite type of music (Rock, Rap/HipHop, Pop, Country, Other)?” In this activity, we’ll look at how this variable relates to gender.

**a.** For each sex, determine the sample percentages in each category of favorite music. (Percentages should add to 100% within each sex.) Give those percentages as the answer to this part.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Country | Other | Pop | Rap/Hip Hop | Rock |
| Female |  |  |  |  |  |
| Male |  |  |  |  |  |

**b.** Using your answer to part **a**. as the basis, write a brief description of how males and females differ with regard to their favorite type of music.

**c.** Do a chi-square test of the relationship between gender and favorite type of music.

What is the *p*-value for the test?

State a conclusion for the chi-square test.

**d.** The dataset also includes measurements (in centimeters) on left forearm lengths and left foot lengths of the students. Explain why we should not use a two-way table and a chi-square test to analyze the relationship between those two variables. What would we do instead?