Journal of Statistics Education

Does Eye Color Depend on Gender? It Might Depend on Who or How You Ask.

Amy G. Froelich
W. Robert Stephenson
Iowa State University

Journal of Statistics Education Volume 21, Number 2 (2013),
www.amstat.org/publications/jse/v21n2/froelich_ds.pdf

Key Words: Student data; Contingency table; Independence; Probability

Abstract

As a part of an opening course survey, data on eye color and gender were collected from students enrolled in an introductory statistics course at a large university over a recent four year period. Biologically, eye color and gender are independent traits. However, in the data collected from our students, there is a statistically significant dependence between the two variables. In this article, we present two ideas for using this data set in the classroom, and explore the potential reasons for the dependence between the two variables in the population of our students.

1. Introduction

To increase student interest and to obtain a source of real data to be used in the classroom (Gnanadesikan, Scheaffer, Watkins & Witmer 1997; Neumann, Neumann & Hood 2010), an opening course survey was administered to all students enrolled in an introductory statistics course at a large university in the Midwest over a recent four year period. The anonymous survey was available on the course website and contained questions on student demographic variables, such as gender, height, eye color, whether or not a student exercises and for how many hours per week, etc. After seven semesters, the full data set contains 2,068 records on 14 different categorical and quantitative variables. This data set has been used in introductory and intermediate level statistics courses as a source of examples for data analysis and inference and as an example of data cleaning. See Holcomb & Spalsbury (2005) for an example of a project using summary statistics and graphical analysis for data cleaning.

The focus of this article is the relationship between the variables reported eye color and gender. Biologically, several traits are known to be sex-linked, including color blindness. While the genetic basis of eye color is not yet fully known, none of the known genes for eye color are located on either the X or Y chromosomes (Duffy, et al. 2007). Thus, biologically, eye color and gender are considered to be independent traits.

We were surprised to find, in the data collected from our students, strong evidence of dependence between gender and reported eye color. If treated as a representative sample from a larger population, this data set can be used to illustrate concepts such as conditional distributions, populations, samples and sampling variability, and tests of independence. Alternatively, considering the data as the population of interest, this example can be used to illustrate probability rules for different events based on selecting a student at random from the population. In both cases, the importance of reliable data collection procedures can be discussed by comparing our results to results from studies on the genetic basis of eye color (Duffy, et al. 2007).

In Section 2 of this paper, the data on reported eye color and gender are described. A detailed analysis of the data is presented in Section 3 along with notes for incorporating this example in the classroom. A discussion of the results appears in Section 4 with conclusions in Section 5.

2. Description of Data

Data on eye color and gender were collected on a web-based survey using drop down boxes with pre-determined categories for each variable. The categories for eye color were determined by investigating eye color classification charts. Eye colors can range from very light blue (gray) to very dark brown (black) and are determined by the relative amount, structure and quality of the melanin in the iris of the eye (Sturm and Frudakis 2004). Typically, in studies of the genetic basis of eye color, eye color is classified by researchers into one of three categories – blue, green/hazel, or brown (Duffy, et al. 2007). However, we were concerned a significant number of students would leave the question unanswered if only three eye color choices were given. We were also interested in how students would differentiate between green and hazel eye colors. After some deliberation, we selected the eye color categories of Blue, Brown, Green, Hazel, and Other for the survey. Very few students (only 42 out of 2,068) ultimately chose the eye color Other, so these responses are excluded from the analyses below.

The data set accompanying this article contains a total of 2,068 records; one for each student who responded to the course survey. Among the 14 variables collected from each student are the student's Gender and Eye Color. The data set *eyecolorgenderdata.csv* and its description *eyecolorgender.txt* can be downloaded from the *JSE* website.

3. Analysis of Data Set

In this section, we present two ideas for using this data set in the classroom. The first treats the data as a representative sample from a larger population to illustrate statistical methods related to the analysis of one and two categorical variables. The second uses the data as the population of

interest to illustrate concepts in probability. All analyses including graphs were produced in JMP®, Version 10.

> *Helpful Hint: We recommend you use this data set to illustrate statistical analysis of a sample or to illustrate probabilities of events based on random selection from a population, but not both. Using the data set both as a sample in one part of the course and as a population in another should be avoided as this blurs the important distinction between samples and populations.*

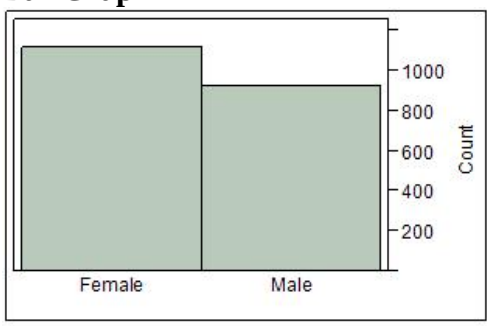
3.1 Using Data as a Representative Sample from Larger Population

This example is presented to students during the first week of the introductory statistics course as a part of a chapter on descriptive statistics for one categorical variable and the relationship between two categorical variables. In this chapter, we cover bar and pie graphs and frequency and relative frequency tables for describing the distribution of one categorical variable. For the relationship between two categorical variables, we cover contingency tables, marginal and conditional distributions, mosaic plots or segmented bar graphs, and introduce the idea of independent and dependent relationships. During the last week of the semester, if time allows, we return to this example in discussing Pearson's Chi-square test of independence for two categorical variables. We have also used this example in other classes as part of a unit on descriptive and inferential analyses of the relationship between two categorical variables.

After presenting the structure of the larger data set described in the Introduction above, we show students how to obtain the distribution of a single categorical variable from this data set, using Gender and Eye Color as two examples. The bar graphs and frequency tables for both variables are given in Figures 1 and 2 below.

Figure 1. Bar Graph and Frequency Table of Distribution of Gender

Bar Graph

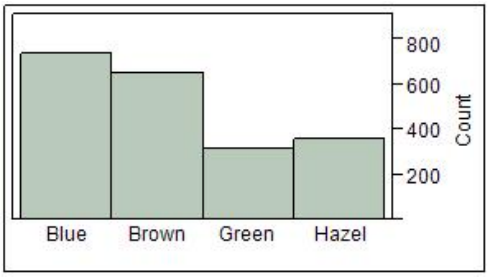


Frequency Table

Gender	Number	Proportion
Female	1107	0.5464
Male	919	0.4536
Total	2026	1.0000

Figure 2. Bar Graph and Frequency Table of Distribution of Eye Color

Bar Graph



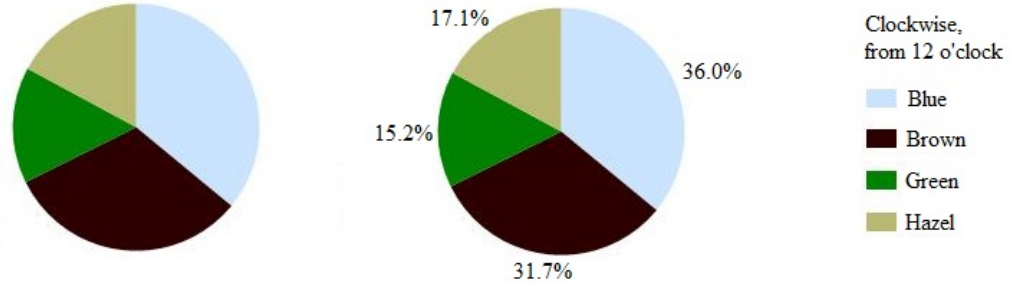
Frequency Table

Eye Color	Number	Proportion
Blue	729	0.35982
Brown	642	0.31688
Green	308	0.15202
Hazel	347	0.17127
Total	2026	1.00000

We then discuss interesting features of the two distributions, such as the approximate 55% – 45% split of females to males, and the roughly equal percentages of Blue (35.98%) and Brown (31.69%) eyes and Green (15.20%) and Hazel (17.13%) eyes. Finally, we present the pie chart for the distribution of eye colors, and make comparisons between the information contained in the bar graph, pie chart, and frequency table.

Helpful Hint: The distribution of eye color is a good example to show how information can be lost in a pie graph if percentages are not included. The differences in percentages between the eye colors Blue and Brown and Green and Hazel are difficult to determine by comparison in a pie graph, while these differences are easily apparent in the bar graph (see Figure 3 below).

Figure 3. Pie Graph of Eye Color with and without percentages



After analyzing the distribution of each categorical variable, we then look at the distribution of eye color by gender. The cross-classification of all 2,026 students on the two variables gender and eye color is given below in Table 1.

Table 1. Contingency Table of Number of Students Classified by Gender and Eye Color

	Eye Color				
Gender	**Blue**	**Brown**	**Green**	**Hazel**	**Total**
Female	370	352	198	187	1107
Male	359	290	110	160	919
Total	729	642	308	347	2026

Using the data in Table 1, students calculate the conditional distribution of eye color given the two genders and the conditional distribution of gender given the four eye colors. These distributions, along with the marginal or unconditional distributions of the two variables are given in Tables 2 and 3 respectively.

Between the two genders, the conditional distribution of eye color is similar for Brown (≈32%) and Hazel (≈17%) with differences within one percentage point. However, the conditional distribution of eye color is different for the Blue and Green eye colors, with differences of around six percentage points. A larger percentage of Males report Blue eyes (39.06% to 33.42%) and a larger percentage of Females report Green eyes (17.89% to 11.97%).

Table 2. Conditional Distribution of Eye Color Given Gender

	Eye Color			
Gender	**Blue**	**Brown**	**Green**	**Hazel**
Given Female	33.42%	31.80%	17.89%	16.89%
Given Male	39.06%	31.56%	11.97%	17.41%
Unconditional	35.98%	31.69%	15.20%	17.13%

Table 3. Conditional Distribution of Gender Given Eye Color

	Eye Color				
Gender	**Given Blue**	**Given Brown**	**Given Green**	**Given Hazel**	**Unconditional**
Female	50.75%	54.83%	64.29%	53.89%	54.64%
Male	49.25%	45.17%	35.71%	46.11%	45.36%

Similar information can be obtained by looking at the conditional distribution of gender given eye color. Between the four eye colors, Brown and Hazel have a split of females to males similar to the overall percentages in the sample (54.64% to 45.36%). The split between the genders for Blue is tilted more towards Males than the overall percentage (49.25% to 45.36% overall), while the split between the genders for Green is tilted more towards Females (64.29% to 54.64% overall).

Helpful Hint: In discussing these results, it is important to note the differences in the split between genders for the colors relative to the overall split between genders. Students

generally expect to see the conditional distribution of gender given eye color split approximately 50-50 for each eye color. However, this would only be expected if the split of the genders in the sample is close to 50% - 50% females to males. In this case, we expect the split to be similar to the marginal or unconditional distribution of gender in the sample, roughly 55% females to 45% males.

The two conditional distributions can also be compared visually with mosaic plots as shown in Figures 4 and 5. Using the data from Tables 2 and 3 and the mosaic plots from Figures 4 and 5, we then ask students to discuss the relationship between the two variables. Using other examples, students see the similarity between the conditional distributions for unrelated variables (ex. gender and whether or not a student exercises; also from the opening course survey) and the differences for related variables (ex. class of ticket and rescued status on the RMS Titanic; DeVeaux, Velleman, and Bock 2009, Chapter 3). Students are then asked to judge whether or not the conditional distributions are different enough for these two variables to be related. Because many students believe, based on previous experience and knowledge, that gender and eye color are independent, this judgment is difficult for most students.

Figure 4. Mosaic Plot of the Conditional Distribution of Eye Color Given Gender

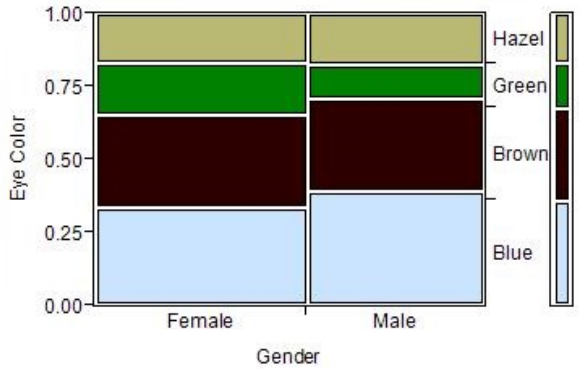
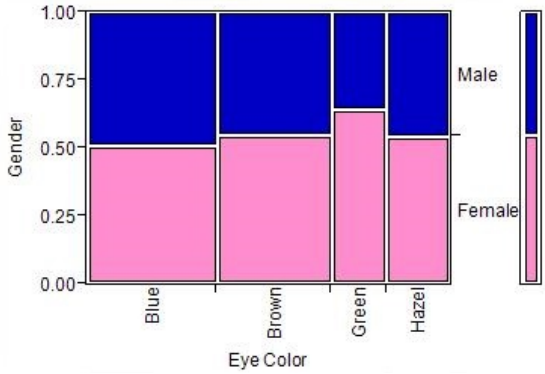


Figure 5. Mosaic Plot of the Conditional Distribution of Gender Given Eye Color



Alternative Application: While both conditional distributions and mosaic plots are included above, our classroom presentation uses only the conditional distribution and mosaic plot for eye color given gender.

At the end of the semester, we return to this example when discussing Pearson's Chi-square test for the independence of two categorical variables. Now students have a method for determining, using a test of hypothesis, if reported eye color and gender are independent or dependent variables in the population of interest. For this example, we discuss the definition of the population (students enrolled in this introductory statistics course) and how to view the sample that produced the data (representative sample from the population). We then present the hypothesis test and calculate the typical Pearson's Chi-square test statistic. In our sample, the value of this test statistic is 16.292 with an associated p-value of 0.001. The conclusion of our

hypothesis test: reported eye color and gender are dependent variables in the population of students enrolled in this introductory statistics course.

> *Alternative Application: While we split the presentation between descriptive and inferential topics, the data set could be presented only once when discussing the test of independence of categorical variables. However, the descriptive analysis should still be presented before the hypothesis test.*

3.2 Using Data as the Population

An alternative use of this data set is to consider the 2,026 students as the population of interest. For this population we have two pieces of information; gender, with events male and female; and eye color, with events blue, brown, green and hazel. Selecting a student at random from the population insures that each individual student is equally likely to be chosen. We can then explore the usual probability rules for events by asking questions in the form: "If a student is selected at random from the population, what is the probability the student will …?"

> *Helpful Hint: When reporting probabilities we have students round their final answers to two decimal places.*

Probabilities for simple events, e.g. the student will be female, the student will have green eyes, are very easy to calculate. Probabilities for more complicated events, e.g. the student will be male or have green eyes, can be found in two ways. One way is to add up the numbers of students that meet one or both of the conditions and divide by the total number of students in the population.

$$Pr(Male\ or\ Green\ Eyes) = \frac{359 + 290 + 110 + 160 + 198}{2026} = \frac{1117}{2026} = 0.55$$

The other way is to use the probability rule for the union of the events Male and Green Eyes.

$$\begin{aligned} Pr(\text{Male } or \text{ Green Eyes}) &= Pr(\text{Male}) + Pr(\text{Green Eyes}) - Pr(\text{Male } and \text{ Green Eyes}) \\ &= \frac{919}{2026} + \frac{308}{2026} - \frac{110}{2026} = 0.45 + 0.15 - 0.05 = 0.55 \end{aligned}$$

The calculation of conditional probabilities and the definition of independent events can be easily illustrated using this population. Asking if two events, e.g. Female and Green Eyes or Male and Brown Eyes, are independent requires calculating the conditional and unconditional probabilities of events. For example, are the two events Green Eyes and Female independent?

$$\Pr(\text{Green Eyes}) = \frac{308}{2026} = 0.15$$

$$\Pr(\text{Green Eyes}|\text{Female}) = \frac{198}{1107} = 0.18$$

Because the unconditional and conditional probabilities of the events are not equal, the two events are not independent. What about the two events Male and Brown Eyes?

$$Pr(\text{Male}) = \frac{919}{2026} = 0.45$$

$$\Pr(\text{Male}|\text{Brown Eyes}) = \frac{290}{642} = 0.45$$

Because the unconditional and conditional probabilities of the events are equal, the two events are independent.

In a similar way, students can verify that gender and the eye colors Brown and Hazel are independent, while gender and the eye colors Blue and Green are not. Because Blue and Green eyes are not independent of gender, we would conclude that the variables eye color and gender are not independent overall.

4. Classroom Discussion of Results

Whether you treat this data set as the population of interest or as a representative sample from a larger population, both analyses reach the same conclusion: gender and reported eye color are dependent variables in our population of students. Clearly, this conclusion is at odds with existing knowledge on the biological basis of eye colors. What are some potential explanations for this surprising and conflicting result?

When posing this question to students, many will express misconceptions about the source of our result. In both analyses, some students will blame our result on having more females than males in our data. However, all calculations, e.g. conditional distribution, Pearson's Chi-square test statistic, and conditional probabilities, are based on the percentages or proportions **within** each gender. With the same conditional distributions or probabilities, instructors can easily show a more even split between the two genders would not change our overall result. In the first analysis using the data as a representative sample, some students believe our result is just a case of "bad luck" – obtaining an unusual sample. While this is of course possible, it is not probable. Reminding students of the definition of the p-value and the results for our test (p-value = 0.001) usually convinces students "bad luck" is not a probable explanation for our result.

In our study of this data set and eye colors in general, we have identified three possible explanations for the results in our data. One, our survey gave students the option to choose between four different eye colors, although eye colors are defined in studies of the genetic basis for eye color (Duffy, et al. 2007) as belonging to only three categories: Blue, Green/Hazel or Brown. While we cannot determine the exact effect of this decision, we can estimate it by combining the Green and Hazel categories into one and reanalyzing the data. We find the Pearson's Chi-square test statistic for these data is 8.988 with p-value of 0.0112 and the conditional probability of eye color given gender is equal to the unconditional probability of eye color only for the color Brown. With this change, we still find that reported eye color and gender are dependent variables in this population of students (although the evidence is not as strong).

Two, in studies of the genetic basis of eye color (Duffy, et al. 2007), researchers use expert opinion to obtain the eye colors of the participants. Our survey, however, asked students to self-report their eye colors. There was no outside corroboration or expert opinion used in the data collection procedure.

This outside corroboration of eye color could play a role in negating the effect of color perception differences among people. For example, roughly 8-10% of males and less than 1% of females (Neitz & Neitz 2000) are affected by the sex-linked trait color blindness. Collectively called red-green color vision defects, people with color blindness are missing one class of photopigments, either medium wavelength (green) or long wavelength (red). The short wavelength photopigments (blue) are not affected in this class of color blindness (Neitz & Neitz 2000). People with red-green color vision defects see the color blue with no difficulty. However people with this defect have difficulties with the perception and differences among the colors green and red and all colors in between them in the color spectrum (yellow and orange). This defect can be slight, with the inability to distinguish between similar shades in this spectrum (olive green and brown, for example) or can be severe, with the inability to distinguish between green and red (Neitz & Neitz 2000).

Our survey did not include a question about color blindness. In our data, based on the overall prevalence rates of color blindness of 8-10% of males, we can expect anywhere from around 73 to 92 of the 919 male students responding to the survey have a red-green color vision defect with the standard deviation of this estimate ranging from 8.2 to 9.1 males. People with color blindness would be more likely to have trouble distinguishing green and hazel eye colors from each other and from blue. Since this trait is much more common in males than females, without outside corroboration, males could be more likely to over-report blue eyes. In our data, 28 more males reported having blue eyes than would be expected under the null hypothesis of independence. This number is well within the number of males expected to have a red-green color vision defect in our sample. Instructors wishing to collect their own data on eye color and gender could include a question on color blindness on their survey. A cross classification table of students by eye color and color blindness can then be analyzed to determine if color blind students are more likely to report having blue eyes than students with normal color vision.

Finally, the distributions of true eye colors between the two genders in our population of introductory statistics students could actually be different. While this would seem like the correct conclusion given our analyses in Section 3, the data collection procedures for our survey prevent us from making this final conclusion. We can only reach this conclusion by collecting more data using different data collection procedures (different eye color categories and corroboration of reported eye colors). Even if after further investigation this conclusion is reached, the result could be due to unknown differences between the racial and ethnic backgrounds of female and male students in our population, as the distribution of eye colors varies across people of different race and ethnic backgrounds (Duffy, et al. 2007).

> *Helpful Hint: In our classroom discussions, we usually have to suggest the three possible explanations for our result to get the conversation started. However, once started, students generally are quick to take over the discussions. For example, to illustrate*

potential difficulties with data collection, students have mentioned the phenomenon where people have two different eye colors. Other students have talked about gender based differences in color descriptions (ex. females may be more likely to describe a color as blue-green where males would simply describe it as blue or green.)

5. Conclusions

As a part of an opening course survey, data on reported eye color and gender were collected from students enrolled in our introductory statistics course over a four-year period. Surprisingly, the two variables are dependent in our data, even though eye color and gender are thought to be independent traits. The data and analyses in this article can be easily incorporated into an introductory course, or instructors could collect their own data through a web-based survey program, such as Survey Monkey, or through a course management system, such as Moodle or Blackboard.

"What is your eye color?" seems like a simple question. However, through this exercise, students learn that very few survey questions are straight forward and the data collection procedures used in a survey or study must match the intended use of the data. In this case, the determination of possible eye color categories, the influence of color perception differences between genders and the data collection method used all created potential problems in our survey with this "simple" question. While these problems do not cause serious consequences given the purpose of our survey, students learn data collection for research purposes must be much more precise.

Acknowledgements

The authors would like to thank our students for supplying the data on which this paper is based.

References

DeVeaux, R.D., Velleman, P.F., & Bock, D.E. (2009), *Intro Stats*, Boston: Pearson, Addison Wesley.

Duffy, D.L., Montgomery, G.W., Chen W., Zhao, Z.Z., Le, L., James, M.R., Hayward, N.K., Martin, N.G., Sturm, R.A. (2007), "A three-single-nucleotide polymorphism haplotype in intron 1of *OCA2* explains most human eye-color variation," *American Journal of Human Genetics*, Vol. 80, pp. 241 - 252.

Gnanadesikan, M., Scheaffer, R.L., Watkins, A.E., Witmer, J.A. (1997), "An activity-based statistics course," *Journal of Statistics Education* v.5, n.2, www.amstat.org/publications/jse/v5n2/gnanadesikan.html.

Holcomb, J. & Spalsbury, A. (2005), "Teaching students to use summary statistics and graphics to clean and analyze data," *Journal of Statistics Education* v.13, n.2, www.amstat.org/publications/jse/v13n3/datasets.holcomb.html.

JMP®, Version 10. SAS Institute Inc., Cary, NC, 1989-2012.

Neitz, M. & Neitz, J. (2000), "Molecular Genetics of Color Vision," *Archives of Ophthalmology*, Vol. 118, No. 5, pp. 691 – 700.

Neumann, D.L., Neumann, M.M. & Hood. M. (2010), "The development and evaluation of a survey that makes use of student data to teach statistics," *Journal of Statistics Education* v.18, n.1, www.amstat.org/publications/jse/v18n1/neumann.pdf.

Sturm, R.A. & Frudakis, T.N. (2004), "Eye colour: portals into pigmentation genes and ancestry," *Trends in Genetics*, Vol. 20, No. 8, pp. 327 – 332.

Amy G. Froelich
Department of Statistics
Iowa State University
3109 Snedecor Hall
Ames, IA 50011-1210
amyf@iastate.edu

W. Robert Stephenson
Department of Statistics
Iowa State University
3111 Snedecor Hall
Ames, IA 50011-1210
wrstephe@iastate.edu
