

8 Analysis of Categorical Data from the Ashe Center Student Wellness Survey

Before starting this lab, you should be familiar with: the difference between categorical and quantitative variables, and how to describe distributions of categorical variables using frequency counts.

About the Data

The dataset that we will be using in this lab comes from the Ashe Center Student Wellness Survey.

Load this dataset into *Stata*

```
. use http://www.stat.ucla.edu/labs/datasets/ashedata.dta
```

There should be five variables in the dataset:

drink:

The survey asks “Have you ever drunk alcohol, other than a sip or two?” Those who answered yes are represented by a “1”, while those who answered no, are represented by a “0”.

howmany:

The survey asks “During the academic year, about how many times would you say you have gotten drunk after drinking?” The response is recorded as a numerical value.

BMI:

Body-mass index. Participants in the survey gave self-reported heights and weights. From this the BMI can be calculated by weight in Kg / height squared.

gender:

A “0” represents female, while a “1” represents a male.

bodysat:

This is the self-reported weight satisfaction of the participant. A “0” rep-

resents unsatisfied with body weight, while a “1” represents satisfied with body weight.

Variables “drink”, “gender”, and “bodysat” are categorical variables. “BMI” and “howmany” are numerical variables.

Constructing and Interpreting a Two-Way Table

In order for us to evaluate the categorical variables in the dataset we will use two-way tables. Type:

```
. tabulate gender drink
```

The chart that you see shows the frequencies for gender compared with whether they have ever drank alcohol (other than a sip or two). The values in the first row represent the number of females in the survey and the values in the second row represent the number of males in the survey. For example, there are 106 females in the survey (gender = “0”) that have not used alcohol (drink = “0”), while 332 females in the survey have used alcohol (drink = “1”). There are 153 students in the dataset who have never drank and 438 of the students in the dataset are women.

Question 1: Interpret the values in the second row.

Question 2: How many students in the dataset have drank alcohol?

Now we will consider the row relative frequencies. Type:

```
. tabulate gender drink, row
```

In addition to the counts from before, we now have row percentages. Consider the first column of relative frequencies. The percentages can be interpreted as follows: Out of females in the survey, 24.20% of them have never drank alcohol. This percentage is calculated using $(106/438)$. 23.86% (calculated

47/197) of males in the survey have never drank alcohol. 24.09% of all the students in the dataset have not used alcohol.

Question 3: Interpret the second column of relative frequencies.

Question 4: According to the SHS survey are college-age men or college-age women more likely to drink alcohol, or are they the same? What numbers did you use to reach your conclusion?

Now lets consider the column percentages. Type:

```
. tabulate gender drink, column
```

Again, we have the counts from before, but now we have another set of percentages, the relative frequencies for the columns. These percentages can be interpreted as followed: 69.28% (106/153) of the students that have never drank alcohol are women. Of the students that have been exposed to drinking, 68.88% (332/482) were women and 68.98% (438/635) of the students in the dataset are women.

Question 5: Interpret the second row of relative frequencies.

Stata can perform one other set of percentages, cell frequencies. Type:

```
. tabulate gender drink, cell
```

Again, you will see the counts, but now we have cell frequencies. All of the frequencies are out of the entire dataset. For example there are 332 females who have drank alcohol before, therefore the cell percentage will be 52.28% (332/635). (Remember there are 635 total respondents in the survey.) All cell relative frequencies are calculated by dividing the cell frequency by the total number of students in the dataset.

Question 6: Interpret three of the cell frequencies.

Another questions posed in the survey, regarded how often the respondents had gotten drunk in the past academic year. We can continue to investigate whether men or women are comparable in their drinking habits, however we cannot use a two way table for this because “howmany” is not a categorical variable, it is numerical. We will use side-by-side box plots.

First we must sort using the variable gender.

```
. sort gender  
. graph howmany, box by(gender)
```

Question 7: Describe the box plots.

Question 8: According to the box-plots, do college age men drink more than college age women?

Here is a list of *Stata* commands we used in this Analysis of Categorical Data lab. Use the space next to each command to make notes on what that command does.

tabulate

tabulate, row

tabulate, column

tabulate, cell

sort

Assignment

Also included in this the survey and dataset are questions regarding body image. The variable “bodysat” represents body satisfaction. A “0” means the person is unsatisfied with their body weight and a “1” represents satisfaction with their body weight. The variable “BMI” is body-mass index, a numerical variable. Using these variables and the variables from before, answer the following questions.

What percent of women are satisfied with their weight? What percent of men are unsatisfied with their weight? Who is more satisfied with their weight, men or women? Do those students who are unsatisfied with their weight have higher BMI's?