MODULE 7

UNIVARIATE EDA - CATEGORICAL

Contents

Interpreting summaries of a single categorical variable is more intuitive and less defined than that for quantitative data. Specifically, one DOES NOT describe shape, center, dispersion, and outliers for categorical data. In this module, methods to construct tables and graphs for categorical data are described and the interpretation of the results demonstrated.

◇ Do not describe shape, center, dispersion, and outliers for a categorical variable.

These concepts are illustrated with three data sets. First, data recorded about MTH107 students in the Winter 2010 semester will be used. Specifically, whether or not a student was required to take the courses and the student's year-in-school will be summarized. Whether or not a student was required to take the course for a subset of individuals is shown in Table 7.1.

Table 7.1. Whether (Y) or not (N) MTH107 was required for eight individuals in MTH107 in Winter 2010.

Individual	1	2	3	4	5	6	7	8
Required	Y	N	N	Y	Y	Y	N	Y

Second, the General Sociological Survey (GSS) is a very large survey that has been administered 25 times since 1972. The purpose of the GSS is to gather data on contemporary American society in order to monitor and explain trends in attitudes, behaviors, and attributes. One question that was asked in a recent GSS was "How often do you make a special effort to sort glass or cans or plastic or papers and so on for recycling?" These data are found in the *recycle* variable in GSSEnviroQues.csv.

7.1 Summaries

7.1.1 Frequency and Percentage Tables

A simple method to summarize categorical data is to count the number of individuals in each level of the categorical variable. These counts are called frequencies and the resulting table (Table 7.2) is called a frequency table. From this table, it is seen that there were five students that were required and three that were not required to take MTH107.

Table 7.2. Frequency table for whether MTH107 was required (Y) or not (N) for eight individuals in MTH107 in Winter 2010.

Required	Freq
Y	5
N	3

The remainder of this module will use results from the entire class rather than the subset used above. For example, frequency tables of individuals by sex and year-in-school for the entire class are in Table 7.3.

Table 7.3. Frequency tables for whether (Y) or not (N) MTH107 was required (Left) and year-in-school (Right) for all individuals in MTH107 in Winter 2010.

Required	Freq
Y	38
N	30

Year	Freq
Fr	19
So	12
Jr	29
Sr	9

Frequency tables are often modified to show the percentage of individuals in each level. **Percentage tables** are constructed from frequency tables by dividing the number of individuals in each level by the total number of individuals examined (n) and then multiplying by 100. For example, the percentage tables for both whether or not MTH107 was required and year-in-school (Table 7.4) for students in MTH107 is constructed from Table 7.3 by dividing the value in each cell by 68, the total number of students in the class, and then multiplying by 100. From this it is seen that 55.9% of students were required to take the course and 13.2% were seniors (Table 7.4).

Table 7.4. Percentage tables for whether (Y) or not (N) MTH107 was required (Left) and year-in-school (Right) for all individuals in MTH107 in Winter 2000.

Required	Perc
Y	55.9
N	44.1

Year	Perc
Fr	27.9
So	17.6
Jr	42.6
Sr	13.2

7.1.2 Bar Plots

Bar plots, or bar charts, are used to display the frequency or percentage of individuals in each level of a categorical variable. Bar plots look similar to histograms in that they have the frequency of individuals on

the y-axis. However, category labels rather than quantitative values are plotted on the x-axis. In addition, to highlight the categorical nature of the data, bars on a bar plot do not touch. A bar plot for whether or not individuals were required to take MTH107 is in Figure 7.1-Left. This bar plot does not add much to the frequency table because there were only two categories. However, bar plots make it easier to compare the number of individuals in each of several categories as in Figure 7.1-Right.

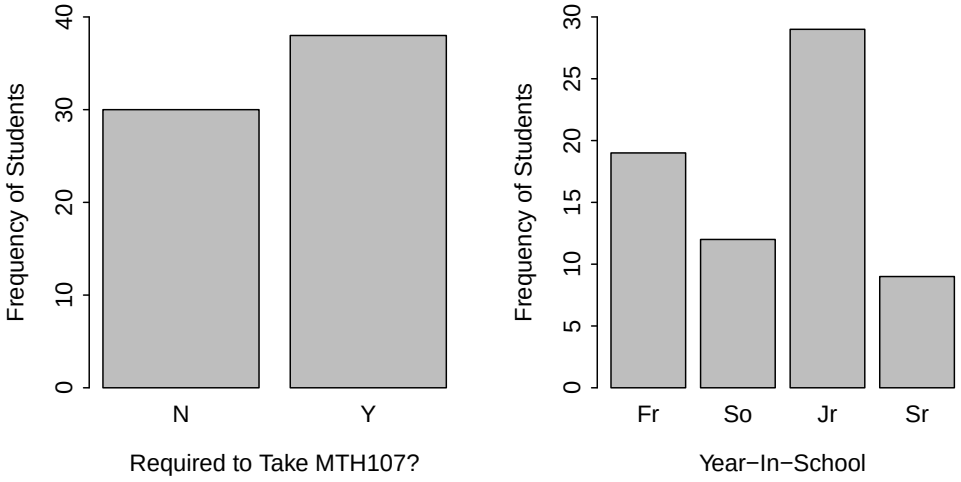

Figure 7.1. Bar charts of the frequency of individuals in MTH107 during Winter 2010 by whether or not they were required to take MTH107 (**Left**) and year-in-school (**Right**).

◇ **Bar charts are used to display the frequency of individuals in the categories of a categorical variable. Histograms are used to display the frequency of individuals in classes created from quantitative variables.**

7.1.3 Using in R

The General Sociological Survey (GSS) data are loaded, the structure of the data.frame is examined, and the levels of the *recycle* variable are shown below. These results show the five levels in the *recycle* factor variable, ordered alphabetically as is the default in R. However, the levels should be "Always", "Often", "Sometimes", "Never", and "Not Avail" to follow the natural order of this ordinal variable.

```
> GSS <- read.csv("data/GSSEnviroQues.csv")
> str(GSS)
'data.frame': 3539 obs. of  2 variables:
 $ recycle: Factor w/ 5 levels "Always","Never",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ tempgen: Factor w/ 5 levels "Extremely","Not",..: 1 1 1 1 1 1 1 1 1 1 ...
> levels(GSS$recycle)
[1] "Always"    "Never"     "Not Avail" "Often"     "Sometimes"
```

The order of a factor variable is controlled by including the ordered level names within a vector given to `levels=` in `factor()`. The names of the levels in this vector must be exactly as they appear in the original variable and they must be contained within quotes. The levels of *recycle* were reordered below. The

advantage of correcting this order is that when the summary table is made, the order will follow the natural order of the variable rather than the alphabetical order.

```
> lvls <- c("Always","Often","Sometimes","Never","Not Avail")
> GSS$recycle <- factor(GSS$recycle,levels=lvls)
> levels(GSS$recycle)
[1] "Always"    "Often"     "Sometimes" "Never"     "Not Avail"
```

◇ **When changing the order of the levels with the `levels=` argument, the level names must be contained within quotes and they must be spelled exactly as they were spelled in the original variable.**

A frequency table of a single categorical variable is computed with `xtabs()`, where the first argument is a one-sided formula of the form `~var` and the corresponding data.frame is in `data=`. The result from `xtabs()` should be assigned to an object for further use. For example, the frequency table is produced, stored in `tabRecycle`, and displayed below. Thus, 1289 respondents answered "Always" to the recycling question.

```
> ( tabRecycle <- xtabs(~recycle,data=GSS) )
recycle
   Always     Often Sometimes     Never Not Avail
     1289       850       823       448       129
```

A percentage table is computed by including the saved frequency table as the first argument to `percTable()`.[1] The number of digits of output is controlled with `digits=`. Thus, 36.4% of respondents answered "Always" to the recycling question.

```
> percTable(tabRecycle,digits=1)
recycle
   Always     Often Sometimes     Never Not Avail
     36.4      24.0      23.3      12.7       3.6
```

A bar plot is produced by giving the saved `xtabs()` object as the first argument to `barplot()`. The x- and y-axes may be explicitly labeled with `xlab=` and `ylab=`, respectively. For example, the bar plot for the recycling data (Figure 7.2) is produced below.

```
> barplot(tabRecycle,ylab="Frequency",xlab="Recycle Response")
```

7.2 Example Interpretations

For categorical data, an appropriate EDA consists of identifying the major characteristics among the categories. Shape, center, dispersion, and outliers are NOT described for categorical data because the data is not numerical and, if nominal, no order exists. In general, the major characteristics of the table or graph are described from an intuitive basis. For example, there were more males than females in the Winter 2010 MTH107 class and mostly juniors and Freshmen. Other examples are below.

[1] Thus, `xtabs()` must be completed and saved to an object before `percTable()`.

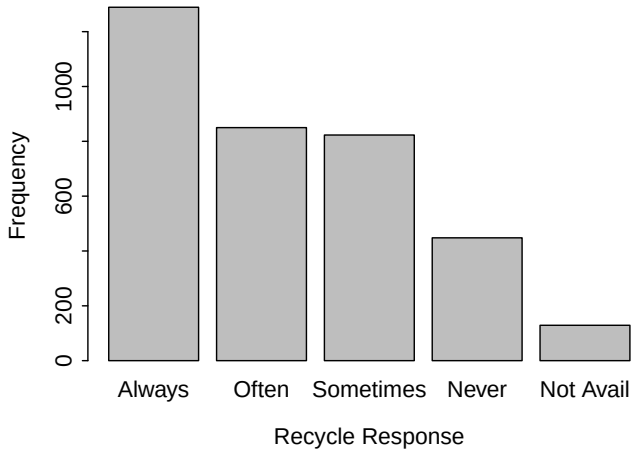

Figure 7.2. Bar chart of the frequency of responses to the recycling question on the GSS.

7.2.1 Mixture Seed Count

A bag of seeds was purchased for seeding a recently constructed wetland. The purchaser wanted to determine if the percentage of seeds in four broad categories – "grasses", "sedges", "wildflowers", and "legumes" – was similar to what the seed manufacturer advertised. The purchaser examined a 0.25-lb sample of seeds from the bag and recorded the results in WetlandSeeds.csv. Use these data to describe the distribution of seed counts into the four broad categories.

The majority of seeds were either sedge or grass with sedge being more than twice as abundant as grass (Table 7.5; Figure 7.3). Very few legumes or wildflowers were found in the sample.

Figure 7.3. Barplot of the percentage of wetland seeds by type.

Table 7.5. Percentage distribution of wetland seeds by type.

grass	legume	sedge	wildflower
27.9	1.6	64.5	6.0

R Appendix:

```
ws <- read.csv("data/WetlandSeeds.csv")
str(ws)
wtbl <- xtabs(~type,data=ws)
percTable(wtbl,digits=1)
barplot(wptbl[-5],ylab="Percentage of Total Seeds",xlab="Seed Type")
```