
MODULE 6

UNIVARIATE EDA - CATEGORICAL

Objectives:

1. Construct frequency and percentage tables with categorical data.
2. Construct bar-charts with categorical data, and
3. Use tables and graphs to describe the categorical data.

Contents

6.1	Summary Tables	72
6.2	Bar Plots	74
6.3	Example Interpretations	76

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. These concepts are illustrated with data recorded about MTH107 students in the Winter 2010 semester. Whether or not a student was required to take the course for a subset of individuals is shown in Table [6.1](#).

Table 6.1. 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

6.1 Summary Tables

A simple method to summarize categorical data is to count the number of individuals in each category (or level) of the categorical variable. These counts are called frequencies and the resulting table (Table 6.2) 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 6.2. 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

◊ Frequency tables show the number of individuals in each category of a categorical variable.

The remainder of this module will use the results from the entire class rather than the subset used above. For example, the frequency tables of individuals by sex and year-in-school for the entire class is in Table 6.3.

Table 6.3. 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	Year	Freq
Y	38	Fr	19
N	30	So	12
		Jr	29
		Sr	9

Frequency tables are often modified to show the percentage of individuals in each category. These modified tables are called **percentage tables**. Percentage tables are constructed from frequency tables by dividing the number of individuals in each category 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 6.4) for students in MTH107 is constructed from Table 6.3 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 6.4. 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	Year	Perc
Y	55.9	Fr	27.9
N	44.1	So	17.6
		Jr	42.6
		Sr	13.2

◊ Percentage tables show the percentage of all individuals in each category of a categorical variable.

6.1.1 Tables in R

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](#).

```
> 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"
```

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. 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.

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

♦ The order of the levels of a factor are controlled with the `levels=` argument in the `factor()` function.

♦ 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 in R 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   Sum
   36.4    24.0    23.3    12.7    3.6    100.0
```

6.2 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 6.1-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 6.1-Right.

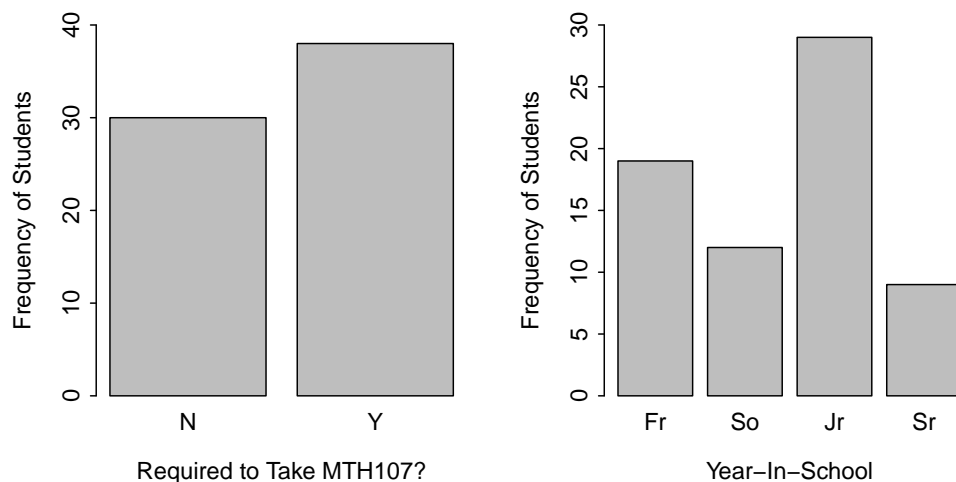


Figure 6.1. Bar charts of the frequency of individuals in MTH107 during Winter 2010 by whether or not they were required to take MTH107 (**Left**) and year-in-school (**Right**).

◇ 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.

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

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

6.2.1 Bar Plots in R

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 6.2) is produced below.

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

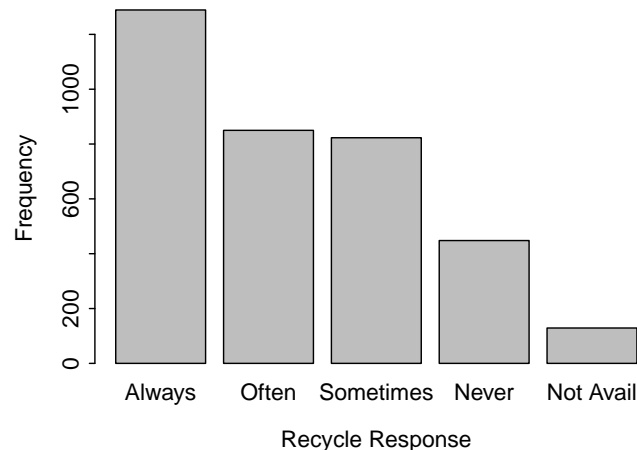



Figure 6.2. Bar chart of the frequency of responses to the recycling question on the GSS.

Review Exercises

6.1  Use the [Arsenic.csv](#) data in Exercise 5.40 to answer the questions below. [Answer](#)


- Construct a univariate EDA for the assessment of drinking water usage.
- Construct a univariate EDA for the assessment of cooking water usage.

6.2 The Environmental Protection Agency (EPA) commissioned the Gallup Organization to conduct a nationwide telephone survey of 1000 households during August and September of 2002 regarding consumer knowledge and satisfaction with drinking water quality. Of the 1000 respondents surveyed, 751 knew that their drinking water came from a public or commercial water supplier. Of these 751 respondents, the following percentages knew precisely where that water was derived:

Ground-water	Lake/Reservoir	River	Multiple Sources	Don't Know	Refused Answer
15.9%	29.2%	9.6%	15.7%	29.4%	0.2%

Use these data to answer the questions below. [Answer](#)


- Construct a frequency table of these data (note percentages above were rounded).
- Write a brief conclusion derived from these data.

- 6.3**  A neighborhood in Honolulu conducted a survey to determine if residents participated in the curbside recycling program. One question on their survey was, “How much has curbside recycling reduced your regular refuse? 0%, 25%, 50%, 75%, 100%, or ‘too early to tell’?” The individual responses for the returned surveys are shown below with letters corresponding to the category choices offered (e.g., A=0%, B=25%, and so on).

C, C, B, B, B, B, C, E, B, B, C, B, C, B, C, C, C, E, B, B, B,
 C, B, B, C, C, C, B, C, B, C, B, B, C, B, C, B, B, B, C, E, B,
 E, B, B, C, C, B, B, B, E, B, C, C, C, B, B, C, B, B, B, B, B

Use these data to answer the questions below. [Answer](#)

- Construct a frequency table of these data.
- Construct a percentage table of these data.
- Write a brief conclusion derived from these data.

- 6.4**  Students in a senior level environmental studies class at Rice University conducted a voluntary response survey regarding water usage by their peers. They received returned surveys from a total 130 students. One question on their survey was, “On average, for how many minutes do you let the water run each time you take a shower? 0-5, 6-10, 11-15, or over 15 minutes?” The individual responses for this survey are shown below with letters corresponding to the category choices offered (e.g., A=“0-5”, B=“6-10”, and so on).

[Answer](#)

D, C, B, B, C, C, B, B, C, C, C, B, D, B, C, C, B, C, D, D,
 B, C, C, A, B, C, C, A, C, C, D, A, C, C, B, B, B, B, B, C,
 D, B, D, B, C, B, C, C, D, C, B, B, D, C, B, C, B, B, C, C,
 B, C, B, C, B, B, C, D, B, C, D, C, B, C, D, C, C, B, C, B,
 D, B, B, D, B, C, B, B, C, B, C, D, D, C, D, B, B, C, B, C,
 A, A, B, C, B, C, D, D, C, B, D, C, C, C, C, A, C, D, B, C,
 B, B, D, C, B, B, A, B, C, B

Use these data to answer the questions below.

- Construct a frequency table of these data.
- Construct a percentage table of these data.
- Write a brief conclusion derived from these data.

6.3 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.

6.3.1 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 6.5; Figure 6.3). Very few legumes or wildflowers were found in the sample.

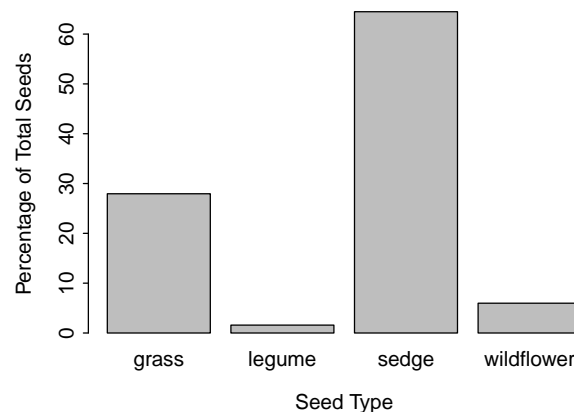


Figure 6.3. Barplot of the percentage of wetland seeds by type.

Table 6.5. Percentage distribution of wetland seeds by type.

grass	legume	sedge	wildflower	Sum
27.9	1.6	64.5	6.0	100.0

R Appendix:

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

Review Exercises

- 6.5** The data in [Zoo1.csv](#) contains a list of animals found in several different zoos. In addition, each animal was classified into broad “type” categories (“mammal”, “bird”, and “amph/rep”). Perform a univariate EDA on the `type` variable. [Answer](#)