
MODULE 10

BIVARIATE EDA - CATEGORICAL

Objectives:

1. Describe bivariate data.
2. Distinguish between response and explanatory variables.
3. Construct two-way contingency tables from raw data.
4. Identify marginal distributions.
5. Construct row-, column-, and table-percentage tables from two-way tables.
6. Interpret two-way contingency tables.

Contents

| | |
|------------------------|----|
| 10.1 Frequency Tables | 86 |
| 10.2 Percentage Tables | 87 |
| 10.3 Which Table? | 89 |
| 10.4 Tables in R | 91 |

TWO-WAY FREQUENCY TABLES summarize two categorical variables recorded on the same individual by displaying the categories of the first variable as rows and the categories of the second variable as columns. Each cell in this table contains a count of the number of individuals that were in the corresponding categories of each variable. Frequency tables are often converted to percentage tables for ease of summarization and comparison among populations. This module explores the construction and interpretation of these types of tables.

The following data from the General Sociological Survey (GSS) will be considered throughout this module. Two questions asked to 3955 respondents were:

- What is your highest degree earned? [choices – “less than high school diploma”, “high school diploma”, “junior college”, “bachelor’s”, or “graduate”; labeled as *degree*]
- How willing would you be to accept cuts in your standard of living in order to protect the environment? [choices – “very willing”, “fairly willing”, “neither willing nor unwilling”, “not very willing”, or “not at all willing”; labeled as *grnsol*]

The data in [GSSWill2Pay.csv](#) are loaded into R and examined below.

```
> gss <- read.csv("data/GSSWill2Pay.csv")
> str(gss)
'data.frame': 3955 obs. of 2 variables:
 $ degree: Factor w/ 5 levels "BS","grad","HS",...: 5 5 5 5 5 5 5 5 5 5 ...
 $ grnsol: Factor w/ 5 levels "neither","un",...: 4 4 4 4 4 4 4 4 4 4 ...
> headtail(gss)
      degree grnsol
1      ltHS  vwill
2      ltHS  vwill
3      ltHS  vwill
3953   grad    vun
3954   grad    vun
3955   grad    vun
```

The *degree* and *grnsol* variables are both *ordinal* categorical variables. By default the levels of factor variables are ordered alphabetically in R (as seen below with `levels()`).

```
> levels(gss$degree)
[1] "BS"    "grad"  "HS"    "JC"    "ltHS"
> levels(gss$grnsol)
[1] "neither" "un"     "vun"    "vwill"  "will"
```

The order of levels for these variables can be specified with `levels=` in `factor()`. The variable to be reordered is the first argument to `factor()` and the object to the left of the assignment operator. The code below specifies the correct orders for the *degree* and *grnsol* variables in GSS.

```
> gss$degree <- factor(gss$degree,levels=c("ltHS","HS","JC","BS","grad"))
> gss$grnsol <- factor(gss$grnsol,levels=c("vwill","will","neither","un","vun"))
> levels(gss$degree)
[1] "ltHS" "HS"   "JC"   "BS"   "grad"
> levels(gss$grnsol)
[1] "vwill" "will"  "neither" "un"    "vun"
```

If the variables had been nominal or if the natural order of levels is alphabetical, then `factor()` would not be needed.

◇ Levels for a factor variable are ordered alphabetically by default in R. You may need to use `factor()` with `levels=` to control the order of levels if the factor variable is ordinal.

10.1 Frequency Tables

A common method of summarizing bivariate categorical data is to count individuals that have each combination of levels of the two categorical variables. For example, how many respondents had less than a HS degree and were very willing, how many had a high school degree and were willing, and so on. The count of the number of individuals of each combination is called a frequency. A two-way frequency table offers an efficient way to display these frequencies (Table 10.1). For example, 40 of the respondents had less than a high school degree and were very willing to take a cut in their standard of living to protect the environment. Similarly, 542 respondents had a high school degree and were willing to cut their standard of living.

Table 10.1. Frequency table of respondent's highest completed degree and willingness to cut their standard of living to protect the environment.

| | vwill | will | neither | un | vun | Sum |
|------|-------|------|---------|------|-----|------|
| ltHS | 40 | 145 | 132 | 151 | 178 | 646 |
| HS | 87 | 542 | 512 | 557 | 392 | 2090 |
| JC | 15 | 61 | 64 | 54 | 44 | 238 |
| BS | 42 | 199 | 179 | 187 | 75 | 682 |
| grad | 24 | 104 | 83 | 64 | 24 | 299 |
| Sum | 208 | 1051 | 970 | 1013 | 713 | 3955 |

A two-way frequency table may be augmented with a column of row totals and a row of column totals (as in Table 10.1). This row and column is called the marginal row and the marginal column, respectively. Each marginal total represents the distribution of one of the categorical variables while ignoring the other categorical variable. The total column represents the distribution of the row variable; in this case, the highest degree completed, in this case, the number of respondents according to their willingness to cut their standard of living to protect the environment. Thus, for example there were 238 respondents whose highest completed degree was junior college and there were 713 respondents who were very unwilling to cut their standard of living to protect the environment.

Review Exercises

- 10.1** Marine biologists studied the foraging ecology of Northern Elephant Seals off the California coast (?). Part of their analysis required that they record, for each observed seal, the month that it was observed and the sex of the seal. Their results from 47 seals are listed below. Construct a two-way frequency table, with marginal totals, of these data (use months as columns). [Answer](#)

| | | | | | | | | | | | | | | | | |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| indiv | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Mon | Jun | Jun | Jun | Jun | Jun | Jun | Jun | Jun | Jul | Jul | Jul | Jul | Jul | Jul | Jul | Aug |
| Sex | M | M | M | M | M | M | M | M | M | M | M | M | M | M | M | M |

| | | | | | | | | | | | | | | | | |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| indiv | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |
| Mon | Aug | Aug | Aug | Aug | Aug | Aug | Aug | Aug | Aug | Aug | Aug | Jun | Jun | Jun | Jun | Jun |
| Sex | M | M | M | M | M | M | M | M | M | M | M | F | F | F | F | F |

| | | | | | | | | | | | | | | | |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| indiv | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 |
| Mon | Jul | Jul | Aug | Aug | Aug | Aug | Aug | Aug | Aug | Aug | Aug | Aug | Aug | Aug | Aug |
| Sex | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F |

10.2 Percentage Tables

Two-way frequency tables are often converted to percentage tables to allow for ease of comparison between levels of the variables and also between populations. For example, it is difficult to determine from Table 10.1 if respondents with a high school degree are more likely to be very willing to cut their standard of living than respondents with a graduate degree, because there are approximately seven times as many respondents with a high school degree in the sample. This comparison is easily made, however, if the frequencies are converted to percentages. Three types of percentage tables are constructed from a frequency table.

10.2.1 Row-Percentage Table

A **row-percentage table** is computed by dividing each cell of the frequency table by the total in the same row of the frequency table and multiplying by 100 (Table 10.2). For example, the value in the “vwill” column and “ltHS” row of the row-percentage table is computed by dividing the value in the “vwill” column and “ltHS” row of the frequency table (i.e., 40; Table 10.1) by the “Sum” of the “ltHS” row of the frequency table (i.e., 646) and multiplying by 100.

Table 10.2. Row-percentage table of respondent’s highest completed degree and willingness to cut their standard of living to protect the environment.

| | vwill | will | neither | un | vun | Sum |
|------|-------|------|---------|------|------|-------|
| ltHS | 6.2 | 22.4 | 20.4 | 23.4 | 27.6 | 100.0 |
| HS | 4.2 | 25.9 | 24.5 | 26.7 | 18.8 | 100.1 |
| JC | 6.3 | 25.6 | 26.9 | 22.7 | 18.5 | 100.0 |
| BS | 6.2 | 29.2 | 26.2 | 27.4 | 11.0 | 100.0 |
| grad | 8.0 | 34.8 | 27.8 | 21.4 | 8.0 | 100.0 |

The value in each cell of a row-percentage table is the percentage OF ALL individuals in that row that also have the characteristic of that column. For example, 6.2% of the respondents with less than a high school degree are very willing to cut their standard of living to protect the environment. This needs to be read very closely and literally. OF THE RESPONDENTS WITH LESS THAN A HIGH SCHOOL DEGREE, not of all respondents, 6.2% were very willing to cut their standard of living.

◇ Each value in a row-percentage table is computed by dividing the value in the same cell of the frequency table by the sum of the same row of the frequency table and multiplying by 100.

◇ The value in each cell of a row-percentage table is the percentage OF ALL individuals with the characteristic of that row that also have the characteristic of that column.

10.2.2 Column-Percentage Table

A **column-percentage table** is computed by dividing each cell of the frequency table by the total in the same column of the frequency table and multiplying by 100 (Table 10.3). For example, the value in the “vwill” column and “ltHS” row on the column-percentage table is computed by dividing the value in the “vwill” column and “ltHS” row of the frequency table (i.e., 40; Table 10.1) by the “Sum” of the “vwill” column of the frequency table (i.e., 208) and multiplying by 100.

The value in each cell of a column-percentage table is the percentage OF ALL individuals in that column that also have the characteristic of that row. For example, 19.2% of respondents who were very willing to cut their standard of living had less than a high school degree. Again, this is a very literal statement. OF

Table 10.3. Column-percentage table of respondent's highest completed degree and willingness to cut their standard of living to protect the environment.

| | vwill | will | neither | un | vun |
|------|-------|-------|---------|-------|-------|
| ltHS | 19.2 | 13.8 | 13.6 | 14.9 | 25.0 |
| HS | 41.8 | 51.6 | 52.8 | 55.0 | 55.0 |
| JC | 7.2 | 5.8 | 6.6 | 5.3 | 6.2 |
| BS | 20.2 | 18.9 | 18.5 | 18.5 | 10.5 |
| grad | 11.5 | 9.9 | 8.6 | 6.3 | 3.4 |
| Sum | 99.9 | 100.0 | 100.1 | 100.0 | 100.1 |

THE RESPONDENTS WHO WERE VERY WILLING TO CUT THEIR STANDARD OF LIVING, not of all respondents, 19.2% had less than a high school degree.

◊ Each value in a column-percentage table is computed by dividing the value in the same cell of the frequency table by the sum of the same column of the frequency table and multiplying by 100.

◊ The value in each cell of a column-percentage table is the percentage OF ALL individuals with the characteristic of that column that also have the characteristic of that row.

10.2.3 Table-Percentage Table

Each value in a **table-percentage table** is computed by dividing each cell of the frequency table by the total number of ALL individuals in the frequency table and multiplying by 100. For example, the value in the “vwill” column and “ltHS” row of the table-percentage table (Table 10.4) is computed by dividing the value in the “vwill” column and “ltHS” row of the frequency table (i.e., 40; Table 10.1) by the “Sum” of the entire frequency table (i.e., 3955) and multiplying by 100.

Table 10.4. Table-percentage table of respondent's highest completed degree and willingness to cut their standard of living to protect the environment.

| | vwill | will | neither | un | vun | Sum |
|------|-------|------|---------|------|------|------|
| ltHS | 1.0 | 3.7 | 3.3 | 3.8 | 4.5 | 16.3 |
| HS | 2.2 | 13.7 | 12.9 | 14.1 | 9.9 | 52.8 |
| JC | 0.4 | 1.5 | 1.6 | 1.4 | 1.1 | 6.0 |
| BS | 1.1 | 5.0 | 4.5 | 4.7 | 1.9 | 17.2 |
| grad | 0.6 | 2.6 | 2.1 | 1.6 | 0.6 | 7.5 |
| Sum | 5.3 | 26.5 | 24.4 | 25.6 | 18.0 | 99.8 |

The value in each cell of a table-percentage table is the percentage OF ALL individuals that have the characteristic of that column AND that row. For example, 1.0% of ALL respondents had less than a high school degree AND were very willing to cut their standard of living to protect the environment. Compare this interpretation to the interpretations from the row and column-percentage tables above. This interpretation DOES refer to all respondents.

◊ Each value in a table-percentage table is computed by dividing the value in the same cell of the frequency table by the total number of ALL individuals in the frequency table and multiplying by 100.

◇ The value in each cell of a table-percentage table is the percentage OF ALL individuals that have the characteristic of that column AND that row.

Review Exercises

- 10.2** Construct a row-, column-, and table-percentage table from the frequency table for the seal data in Review Exercise 10.1. [Answer](#)

10.3 Which Table?

Determining which table to use comes from applying one simple rule and practicing with several tables. The rule stems from determining if the question restricts the frame of reference to a particular level or category of one of the variables. If the question does restrict to a particular level, then either the row or column-percentage table that similarly restricts the frame of reference must be used. If a restriction to a particular level does not appear to be made, then the table-percentage table is used.

For example, consider the question – “What percentage of respondents with a bachelor’s degree were very unwilling to cut their standard of living to protect the environment?” This question refers to only respondents with bachelor’s degrees (i.e., “... of respondents with a bachelor’s degree ...”). Thus, the answer is restricted to the “BS” row of the frequency table. The ROW-percentage table restricts the original table to the row levels and is, thus, used to answer this question. Therefore, 11.0% of respondents with bachelor’s degrees were very unwilling to cut their standard of living to protect the environment (Table 10.2).

Now consider the question – “What percentage of all respondents had a high school degree and were very willing to cut their standard of living?” This question does not restrict the frame of reference because it refers to “... of all respondents ...”. Therefore, from the table-percentage table (Table 10.4), 2.2% of respondents had a high school degree and were very willing to cut their standard of living.

Also consider this question – “What percentage of respondents who were neither willing nor unwilling to cut their standard of living had graduate degrees?” This question refers only to respondents who were neither willing nor unwilling to cut their standard of living and, thus, restricts the question to the “neither” column of the frequency table. Thus, the answer will come from the COLUMN-percentage table. Therefore, 8.6% of respondents who were neither willing nor unwilling to cut their standard of living had graduate degrees (Table 10.3).

Finally, consider this question – “What percentage of all respondents were very willing to cut their standard of living to help the environment?” This question has no restrictions so the table-percentage table should be used. In addition, this question is only concerned with one of the two variables in the frequency table; thus, the answer will come from a marginal distribution. Therefore, 208 out of all 3955 respondents, or 5.3%, were very willing to cut their standard of living to help the environment.

◇ To determine which percentage table to use determine what type of restriction, if any, has been placed on the frame of reference for the question.

◊ If a question does not refer to one of the two variables, then the answer will generally come from the marginal distribution of the other variable.

It should be noted that if one of the two categorical variables is determined to be a response variable, then this variable is usually used to define the columns and the row-percentage table becomes the main table of interest. In this example, the “willingness to cut” would be considered the response variable and it was, appropriately, placed as the column variable in the frequency table. Thus, the questions answered from the row-percentage table (i.e., “Of respondents with a certain degree ...”) make “more sense” than the questions answered from the column-percentage table (i.e., “Of respondents with a certain willingness ...”).

◊ The response variable is typically used to define the columns of the two-way table.

Review Exercises

10.3 Use the frequency and percentage tables for the seal data constructed in Review Exercises 10.1 and 10.2 to answer the questions below. [Answer](#)

- (a) What percentage of elephant seals were male?
- (b) What percentage of male elephant seals were observed in July?
- (c) What percentage of elephant seals were observed in August?
- (d) What percentage of elephant seals were females observed in July?

10.4 ? conducted a survey of general and family practitioners, pediatricians, and obstetrician-gynecologists in the cities of Phoenix and Tucson, Arizona. In one part of the study, each physician was classified according to religion and whether they supported genetic counseling for parents or not. A summary of their responses for Jewish, Protestant, and Catholic physicians is shown in the table below. Use these results to answer the questions below. [Answer](#)

- (a) What percentage of Jewish physicians support genetic counseling?
- (b) What percentage of Catholic physicians don't support genetic counseling?
- (c) What percentage of all physicians surveyed were Protestant?
- (d) What percentage of those physicians not supporting genetic counseling were Catholic?
- (e) What percentage of all physicians supported genetic counseling?

| | Jewish | Protestant | Catholic |
|---------------|--------|------------|----------|
| Support | 21 | 36 | 10 |
| Don't Support | 26 | 142 | 52 |

10.5 The two-way table below depicts the results of an observational study concerned with the timing (i.e., month) of death for young herring gulls (after fledging) in three locations. Each cell in the table is the number of dead herring gulls in each month-location combination. Use the table to answer the questions below. [Answer](#)

- (a) What percentage of the gulls that died in New Jersey died in July?

- (b) What percentage of all gulls died in July?
 (c) What percentage of all gulls died in September and in The Netherlands?

| Month | Location | | | Total |
|-------|------------|-------------|---------|-------|
| | New Jersey | Netherlands | England | |
| Jul | 4 | 4 | 10 | 18 |
| Aug | 7 | 28 | 60 | 95 |
| Sep | 19 | 130 | 89 | 238 |
| Oct | 9 | 150 | 39 | 198 |
| Nov | 2 | 61 | 31 | 94 |
| Dec | 1 | 32 | 12 | 45 |
| Total | 42 | 405 | 241 | 688 |

- 10.6** In an attempt to study rainfall patterns in West Africa caused by El Nino weather events, ? constructed a two-way table that relates the number of days rainfall that occurred each month to the amount of rain in inches that fell on those days (categorized as less than 1 inch and more than 1 inch). Use the modified version of their table below to answer the questions further below. [Answer](#)

| | Jun | Jul | Aug |
|-------|-----|-----|-----|
| <1 in | 7 | 11 | 20 |
| >1 in | 5 | 9 | 10 |

- (a) How many days did it rain in July?
 (b) In the months of June and August, how many days did it rain more than 1 inch?
 (c) What percentage of rainy days in August had less than 1 inch of precipitation?
 (d) If there are 31 days in July, on what percentage of those days did it rain?
 (e) What percentage of rainy days did more than 1 inch of rain fall?
 (f) What percentage of rainy days were in June?

10.4 Tables in R

Two-way frequency tables are constructed in R with `xtabs()`. The first argument is a formula of the form `~rowvar+colvar`, with the corresponding data.frame in `data=`. The result of `xtabs()` should be assigned to an object for further use.

```
> ( tbl11 <- xtabs(~degree+grnsol,data=gss) )
      grnsol
degree vwill will neither un vun
1tHS    40  145    132 151 178
HS      87  542    512 557 392
JC      15   61     64  54  44
BS      42  199    179 187  75
grad    24  104     83  64  24
```

Totals may be added to the margins of the saved table with `addMargins()`.


```
> addMargins(tbl1)
      grnsol
degree vwill will neither  un  vun  Sum
1tHS   40  145   132  151  178  646
HS     87  542   512  557  392 2090
JC     15   61    64   54   44  238
BS     42  199   179  187   75  682
grad   24  104    83   64   24  299
Sum    208 1051   970 1013  713 3955
```

Percentage tables are constructed by submitting the saved `xtabs()` object to `percTable()`. The number of decimals to display is controlled with `digits=`. A row-percentage table is constructed by including `margin=1`.

```
> percTable(tbl1,margin=1,digits=1)
      grnsol
degree vwill will neither  un  vun  Sum
1tHS   6.2  22.4   20.4  23.4  27.6 100.0
HS     4.2  25.9   24.5  26.7  18.8 100.1
JC     6.3  25.6   26.9  22.7  18.5 100.0
BS     6.2  29.2   26.2  27.4  11.0 100.0
grad   8.0  34.8   27.8  21.4   8.0 100.0
```

A column-percentage table is constructed by including `margin=2`.



```
> percTable(tbl1,margin=2,digits=1)
      grnsol
degree vwill will neither  un  vun
1tHS  19.2  13.8   13.6  14.9  25.0
HS    41.8  51.6   52.8  55.0  55.0
JC     7.2   5.8    6.6   5.3   6.2
BS    20.2  18.9   18.5  18.5  10.5
grad  11.5   9.9    8.6   6.3   3.4
Sum   99.9 100.0  100.1 100.0 100.1
```

Finally, a table-percentage table is constructed by omitting `margin=`.

```
> percTable(tbl1,digits=1)
      grnsol
degree vwill will neither  un  vun  Sum
1tHS   1.0  3.7    3.3  3.8  4.5 16.3
HS     2.2 13.7   12.9 14.1  9.9 52.8
JC     0.4  1.5    1.6  1.4  1.1  6.0
BS     1.1  5.0    4.5  4.7  1.9 17.2
grad   0.6  2.6    2.1  1.6  0.6  7.5
Sum    5.3 26.5   24.4 25.6 18.0 99.8
```

◇ The table submitted as the first argument to `percTable()` must be a frequency table WITHOUT margin totals.

Review Exercises

- 10.7**  Using the data provided in Review Exercise 10.1. Construct a two-way frequency table, including the marginal totals, of these data with month as the column variable. [Answer](#)
- 10.8** Construct a row-, column-, and table-percentage table from the frequency table for the seal data in Review Exercise 10.7. [Answer](#)
- 10.9** Use the [Arsenic.csv](#) data introduced in Review Exercise ?? to construct a bivariate EDA for the drinking and usage variables. [Answer](#)
- 10.10** In the General Social Survey (GSS), two questions were asked – “How often do you make a special effort to sort glass or cans or plastic or papers and so on for recycling?” and “In general, do you think that a rise in the world’s temperature caused by the greenhouse effect, is extremely likely, very likely, somewhat likely, not very likely, or not at all likely?”. Both of these variables are recorded in the [GSSEnviroQues.csv](#) file. Use these data to answer the questions below. [Answer](#)
- (a) What percentage of all respondents recycle often and feel that it is very likely that the greenhouse effect has caused the rise in world’s temperature?
 - (b) What percentage of those respondents that recycle often feel that it is very likely that the greenhouse effect has caused the rise in world’s temperature?
 - (c) What percentage of those respondents that think it is very likely that the greenhouse effect has caused the rise in world’s temperature also recycle often?
 - (d) What percentage of all respondents recycle often?
 - (e) What percentage of all respondents think it is very likely that the greenhouse effect has caused the rise in world’s temperature?
- 10.11**  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”). The researchers that collected these data wanted to examine if the distribution of broad animal types differed among zoos. Use these data to answer the questions below. [Answer](#)
- (a) What is the “response” variable in this analysis?
 - (b) What percentage of all animals were birds?
 - (c) What percentage of animals in the Minnesota zoo were birds?
 - (d) What percentage of animals in the Chicago zoo were amphibians/reptiles?
 - (e) What percentage of animals were in the Chicago zoo?
 - (f) What percentage of birds were in the Minnesota zoo?
-