MODULE 10

BIVARIATE EDA - CATEGORICAL

Contents

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

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. Data from the following two questions on the GSS are used throughout this module.

- What is your highest degree earned? [choices – "less than high school diploma", "high school diploma", "junior college", "bachelors", 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*]

These data, stored 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 can be specified using `factor()`. The variable to be reordered is the first argument to `factor()`, as well as the object to the left of the assignment operator. The desired order of the levels is listed in a vector that is given to `levels=`. It is important that the levels in this vector are "spelled" exactly as they appeared originally. Correct orders for *degree* and *grnsol* in the `gss` data.frame are created below.

```
> 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 natural order of levels is alphabetical or the variable is nominal, then `factor()` is not needed.

◇ **Levels for a factor variable are ordered alphabetically by default in R. If the factor variable is ordinal, then `factor()` with `levels=` may be needed to specify the correct order of levels.**

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

The margins of a two-way frequency table may be augmented with row and column totals (as in Table 10.1). Each marginal total represents the distribution of one of the, while ignoring the other, categorical variable. The total column represents the distribution of the row variable; in this case, the highest degree completed. The total row represents the distribution of the column variable; in this case, 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.

If one variables can be considered as the response, then this variable should form the columns of the frequency table. For example, "willingness to cut" could be considered the response variable and it was, appropriately, placed as the column variable in Table 10.1.

Frequency Tables in R

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

```
> ( tbl1 <- xtabs(~degree+grnsol,data=gss) )
      grnsol
degree vwill will neither  un vun
  ltHS    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 a saved table with `addMargins()`. For example, `addMargins()` was used to construct Table 10.1 from `tbl1`.

```
> addMargins(tbl1)
```

10.2 Percentage Tables

Two-way frequency tables may be converted to percentage tables 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. However, if the frequencies are converted to percentages, then this comparison is easily made. Three types of percentage tables may be 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 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 statement must be read carefully. 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.

If the response variable formed the columns, then the row-percentage table allows one to compare percentages in levels of the response (i.e., columns) across groups (i.e., rows). For example, one can see that there is a general decrease in the percentage of respondents that were "very unwilling" to cut their standard of living to protect the environment as the level of education increased (Table 10.2).

Row-Percentage Table in R

Percentage tables are constructed in R 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`. For example, the code below produced Table 10.2.

```
> percTable(tbl1,margin=1,digits=1)
```

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.

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 value in each cell of a column-percentage table is the percentage OF ALL individuals in that column that 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 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.

Column-Percentage Table in R

A column-percentage table is constructed by submitting the saved `xtabs()` object to `percTable()` with `margin=2`. For example, the code below produced Table 10.3.

```
> percTable(tbl1,margin=2,digits=1)
```

10.2.3 Total-Percentage Table

Each value in a **total-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.

Total-Percentage Table in R

A table-percentage table is constructed by submitting the saved `xtabs()` object to `percTable()` and omitting `margin=`. For example, the code below produced Table 10.4.

```
> percTable(tbl1,digits=1)
```

10.3 Which Table to Use?

Determining which table to use comes from applying one simple rule and practicing with several tables. The rule comes 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 is not made, then the total-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 would be 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 total-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 total-percentage table would be used. In addition, this question is only concerned with one of the two variables; 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.**