TOOL

Summarizing and Visualizing Categorical Variables

Table of Contents

When you are interested in answering questions that involve only categorical variables, you will want to use a specific set of summary statistics and plots to gain an understanding of your data. When your question uses only categorical data, you should:

1. Summarize the data by calculating frequencies and proportions of each group.
2. Visualize these summary statistics using different types of barplots.

Use this tool to help you summarize and visualize questions that involve only categorical variables. This tool is divided into three sections based on the type of question you want to answer: univariate, bivariate, or multivariate.

In the examples below, we use the **`titanic`** data set you examined in the course.

Using R With This Tool

The portions of this tool with a gray background are code text that you can use to do the examples included in this tool or modify to work with your own data. To use these examples, type the lines of code that don't begin with a pound sign (#) into R to carry out the command. Commented text begins with one pound sign (#) and explains the lines of code. The code output begins with two pound signs (##).

Data Set Information

The **`titanic`** data set contains demographic information about passengers on the RMS Titanic, which sank in the Atlantic Ocean in 1912. The **`titanic`** data set has data on each passenger in the rows and on passenger characteristics in the columns. To use the **`titanic`** data set with this tool, download the data set, set your working directory to the location of the data set, and run the following code:

```
titanic <- read.table("titanic.txt", header = TRUE) # read in the data

# Change the Survived variable to a factor with the factor() function by
# telling R to replace 0 with No and 1 with Yes, then replacing
# titanic$Survived with SurvivedFactor:

SurvivedFactor <- factor(titanic$Survived, levels = c("0", "1"),
                         labels = c("No", "Yes"))
titanic$Survived <- SurvivedFactor

head(titanic) # display the first 6 rows of data
```

	Name	PClass	Age	Sex	Survived
1	Allen, Miss Elisabeth Walton	1st	29.00	female	Yes
2	Allison, Miss Helen Loraine	1st	2.00	female	No
3	Allison, Mr Hudson Joshua Creighton	1st	30.00	male	No
4	Allison, Mrs Hudson JC (Bessie Waldo Daniels)	1st	25.00	female	No
5	Allison, Master Hudson Trevor	1st	0.92	male	Yes
6	Anderson, Mr Harry	1st	47.00	male	Yes

Univariate Questions

In the following sections, you will summarize and visualize data to help you answer the univariate question "What proportion of passengers survived the sinking of the Titanic?" Use these techniques when you want to understand a single categorical variable.

Summarizing Categorical Data With Frequency Tables

Summarize the distribution of a categorical variable in your data with a *frequency table* that displays the counts or proportions of each category.

R Functions to Use:
`table()`
`prop.table()`

The following code summarizes the **`Survived`** variable into a frequency table and a frequency table of proportions.

```
# Create a frequency table:

tbl.titanic <- table(titanic$Survived)
tbl.titanic

##  No Yes
## 863 450
# Create a frequency table of proportions:

prop.table(tbl.titanic)

##    No         Yes
## 0.6572734 0.3427266
```

Based on these tables, you can see that 450 passengers (34.27%) in the data set survived the sinking of the Titanic.

Visualizing Categorical Data With Barplots

Once you've summarized a categorical variable, you can use barplots to visualize its distribution.

R Functions to use:

`table()`

`prop.table()`

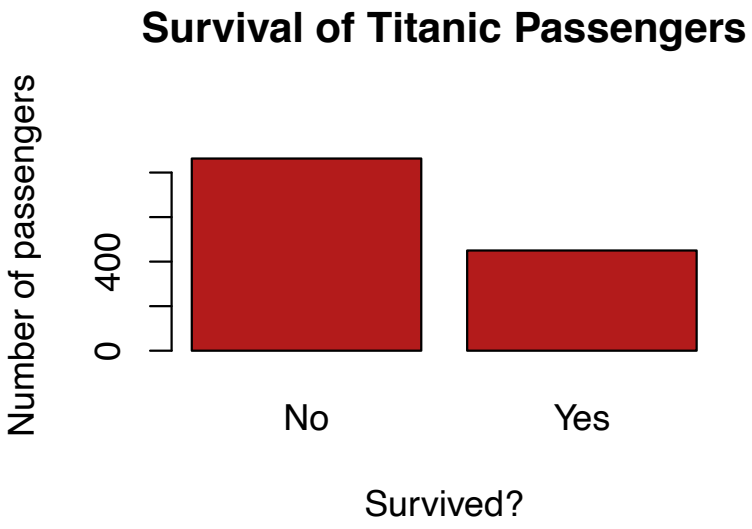
`barplot()`

For example, the following barplot shows passenger counts:

```
# Create a barplot that shows the number of passengers that survived:

counts = table(titanic$Survived)
crimson = "#b31b1b" # Create the color crimson

barplot(counts,            # tell R to use the counts table
        main= "Survival of Titanic Passengers",
        col = crimson,
        names = c("No", "Yes"),    # Label the bars on the x-axis
        ylab= "Number of passengers", xlab = "Survived?")
```



Based on this plot, it is easy to see that fewer passengers survived the sinking of the Titanic than did not survive.

Bivariate Questions

In the following sections, you will summarize and visualize data to help you answer the bivariate question "What was the association between a passenger's sex and their survival when the Titanic sank?" Use these techniques when you want to know whether two categorical variables are associated.

Summarizing Categorical Data With Contingency Tables

You can summarize the distribution of two categorical variables in the same table with a *contingency table* of counts that shows the number or proportion of data points in each possible outcome for each group.

R functions to do this:
`table()`
`prop.table()`

For example, if you want to see whether sex and survival of passengers are associated, you can use the `table()` function to count or find the proportion of:

- How many male passengers survived.
- How many male passengers did not survive.
- How many female passengers survived.
- How many female passengers did not survive.

Setting your table up this way makes it easy to look for a difference between the survival rates of male and female passengers.

```
# Create a contingency table
table(titanic$Sex, titanic$Survived)

##           No Yes
##    female 154 308
##    male   709 142
# Calculate proportions based on rows by setting margin = 1:
prop.table(table(titanic$Sex, titanic$Survived), margin = 1)

##                No       Yes
##    female 0.3333333 0.6666667
##    male   0.8331375 0.1668625
# Calculate proportions based on rows by setting margin = 2
prop.table(table(titanic$Sex, titanic$Survived), margin=2)

##                No       Yes
##    female 0.1784473 0.6844444
##    male   0.8215527 0.3155556
```

From the table of counts, you can see that of the total number of passengers, 308 female passengers survived, whereas only 142 male passengers survived. From the first table of proportions, you can see that 66.67% of female passengers survived, compared to only 16.69% of the males. Similarly, in the second table of proportions, you can see that 68.4% of all survivors were female.

These results suggests a large disparity, and therefore an association, in survival rate across sex.

Visualizing Categorical Data With Stacked Barplots

You can visualize the association between two categorical variables by creating a stacked barplot.

R Functions to use:

`table()`
`prop.table()`
`barplot()`
`t()`

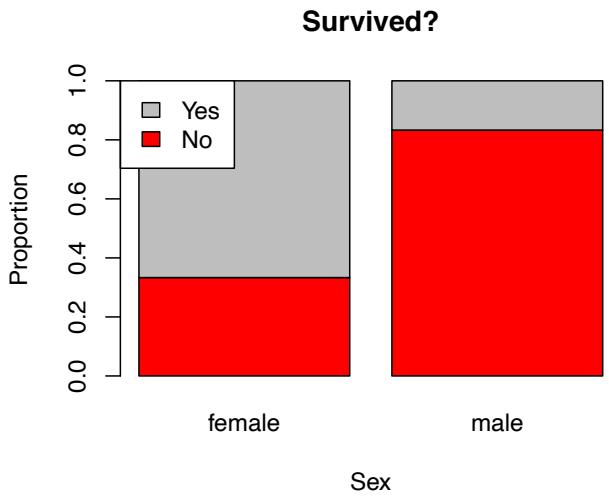
For example, you can look at the proportion of each of the four categories of sex and survival in the **titanic** data set.

```
# Create a contingency table of proportions, then use the transpose function
# to set the table up so Sex is in the rows:
pt = prop.table(table(titanic$Sex,titanic$Survived),1) # contingency table
transposedTable <- t(pt) # switch rows to columns

##          female      male
##   No  0.3333333 0.8331375
##   Yes 0.6666667 0.1668625

crimson = "#b31b1b" # Crimson
darkGray = "#606366" # darkGray
# Create a barplot:

barplot(transposedTable,
        main = "Survived?",
        col = c("red", "gray"),
        ylab = "Proportion",
                  xlab = "Sex",
        legend.text = c("No", "Yes"),
        args.legend = list(x = "topleft")) # Create a legend in the
                                           # top left of the plot
```



By visualizing all four categories at once, you can see that a higher proportion of females survived.

Multivariate Questions

You can visualize the impact one variable, *Z*, has on the relationship between two other variables, *X* and *Y*, by examining your data by group and determining whether *Z* influences the relationship between *X* and *Y* consistently. If *Z* does not influence the relationship between *X* and *Y*, you would not expect to see consistent differences across group *Z*.

In the following sections, you will use summarization and visualization techniques to help you answer the multivariate question "Does passenger class influence the relationship between passenger survival and passenger sex?" Use these techniques when you want to know how the outcome of two categorical variables was impacted by another categorical variable.

Summarizing Categorical Data With the aggregate() Function

You can summarize multivariate questions by finding a summary statistic of all the relevant categories.

R Functions to use:

`ifelse()`

`aggregate()`

For example, to summarize `Sex`, `Survived`, and `PClass`, you would first transform survival status to 0s and 1s, then calculate the proportions that fall into each category.

```
# Create a new variable in the titanic data set. The ifelse() function
# converts Yes values in the titanic$Survived column to 1 and No to 0:

titanic$SurvBin = ifelse(titanic$Survived == "Yes", 1, 0)

prop = aggregate(SurvBin ~ PClass + Sex,    # The formula for your aggregation
    FUN = mean,                             # The operation you want to perform
    data = titanic)                         # The data set your formula uses


prop # display the table:
##   PClass    Sex   SurvBin
## 1    1st female 0.9370629
## 2    2nd female 0.8785047
## 3    3rd female 0.3773585
## 4    1st   male 0.3296089
## 5    2nd   male 0.1445087
## 6    3rd   male 0.1162325
```

You can now see that more females traveling first class survived than any other category, but it is difficult to look at each of these categories and make a solid conclusion.

Visualizing Categorical Data With Side-by-Side Barplots

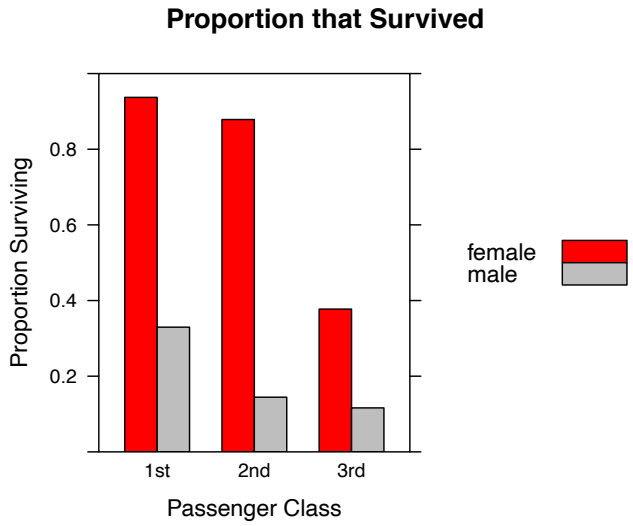
You can visualize multivariate questions with a **side-by-side barplot**, which lets you see many categories at once.

R functions to use:
`library(lattice)`
`barchart()`

For example, to summarize **`Sex`**, **`Survived`**, and **`PClass`**, you would create a side-by-side barplot that shows the survivors in each class by sex based on the aggregated table you made in the previous section.

```
library(lattice) # Load this library to use the barchart() function
crimson = "#b31b1b" # Crimson
darkGray = "#606366" # darkGray

barchart(SurvBin ~ PClass, # The formula your bar chart uses
    groups = Sex,          # The groups in your bar chart
    dat = prop,            # Your data set, which you created
                           # in the summary section.
    ylim = c(0, 1),
    main = "Proportion that Survived",
    xlab = "Passenger Class",
    ylab = "Proportion Surviving",
    par.settings = list(superpose.polygon =
        list(col = c("red", "gray"))),     # Creates the colors and legend
    auto.key = list(space = "right"))      # Put the legend
                                           # on the right side of the plot
```



By examining this barplot, you can see that the disparity between male and female survival rates is very high in the first and second classes, but not as high in the third class. This suggests that the categorical variable **`PClass`** may influence the association between **`Sex`** and **`Survived`**. Notice that presenting the data as a graph makes it easier to interpret than presenting it in a table.