Single Variable Visualization

At this time, we will focus in on a single variable, i.e., for each particular trait we have n observations on this trait, namely x_{1j}, x_{2j}, …, x_{nj}. With what tools can we analyze this set of values?

Data Frame:

Variables $\longrightarrow$

$$\begin{array}{c|c} \textbf{Units} \downarrow & \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} \end{array}$$

j **th variable** $\longrightarrow$ column $x_{1j}, \ldots, x_{ij}, \ldots, x_{nj}$

Our first set of tools include pictorial and frequency methods from **Descriptive Statistics**.

The `Cars93` Data Frame

The `Cars93` data frame (in the `MASS` package) contains information on 93 cars sold in the U.S. in the year 1993. It has 93 rows and 27 columns. Additional information is available using `?Cars93` or `help(Cars93)`

The command `head` displays the first few lines (the default is 6) of an object in R.

```
> head(Cars93)
  Manufacturer   Model    Type Min.Price Price Max.Price MPG.city MPG.highway            AirBags DriveTrain
1        Acura Integra   Small      12.9  15.9      18.8       25          31               None      Front
2        Acura  Legend Midsize      29.2  33.9      38.7       18          25 Driver & Passenger      Front
3         Audi      90 Compact      25.9  29.1      32.3       20          26        Driver only      Front
4         Audi     100 Midsize      30.8  37.7      44.6       19          26 Driver & Passenger      Front
5          BMW    535i Midsize      23.7  30.0      36.2       22          30        Driver only       Rear
6        Buick Century Midsize      14.2  15.7      17.3       22          31        Driver only      Front
  Cylinders EngineSize Horsepower  RPM Rev.per.mile Man.trans.avail Fuel.tank.capacity Passengers Length
1         4        1.8        140 6300         2890             Yes               13.2          5    177
2         6        3.2        200 5500         2335             Yes               18.0          5    195
3         6        2.8        172 5500         2280             Yes               16.9          5    180
4         6        2.8        172 5500         2535             Yes               21.1          6    193
5         4        3.5        208 5700         2545             Yes               21.1          4    186
6         4        2.2        110 5200         2565              No               16.4          6    189
  Wheelbase Width Turn.circle Rear.seat.room Luggage.room Weight  Origin          Make
1       102    68          37           26.5           11   2705 non-USA Acura Integra
2       115    71          38           30.0           15   3560 non-USA  Acura Legend
3       102    67          37           28.0           14   3375 non-USA       Audi 90
4       106    70          37           31.0           17   3405 non-USA      Audi 100
5       109    69          39           27.0           13   3640 non-USA      BMW 535i
6       105    69          41           28.0           16   2880     USA Buick Century
```

Distribution of a Variable

The ***distribution*** of a variable provides the possible values that a variable can take on and how often (frequently) these possible values occur. The distribution of a variable shows the **pattern** of variation of the variable.

The distribution of a variable can be summarized graphically, numerically, or with a model.

Displaying Distributions

Categorical Variables

Categorical variables are usually not measured on a numerical scale. Typically, the frequency or percentage of observations in each category is displayed.

Definition: A ***frequency*** of a category is the number of times it occurs in the data set.

Definition: A ***frequency distribution*** is a table that presents the frequency for each category.

Example: The data frame `Cars93` contains data from 93 cars on sale in the USA in 1993. We can use the `table` function to find the frequency distribution for the standard airbag option.

```
> table(Cars93$AirBags)

Driver & Passenger        Driver only                None
                16                 43                  34
```

Displaying Distributions

Categorical Variables

Definition: The ***relative frequency*** of a category is the frequency of the category divided by the sum of all the frequencies.

Definition: A ***relative frequency distribution*** is a table that presents the relative frequency of each category. Often the frequency is presented as well.

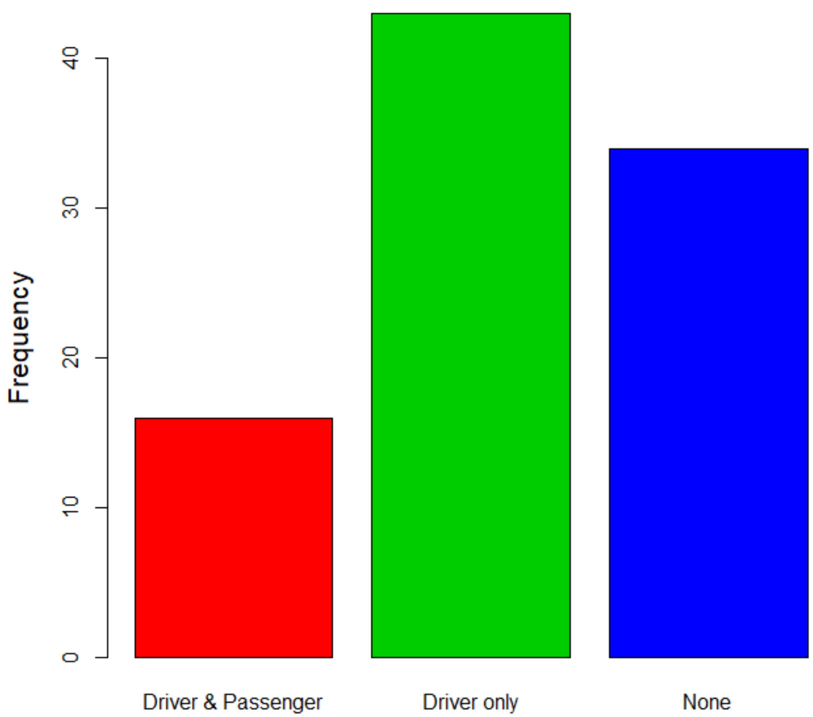
Example: We can use the `table` function to display the relative frequency for the standard airbag option in the `Cars93` dataset.

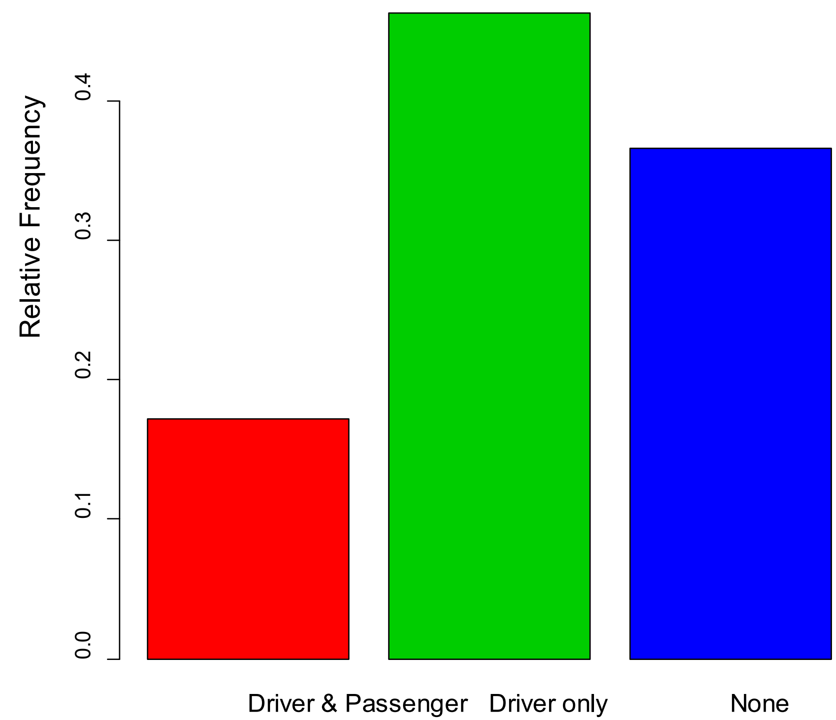
```
> table(Cars93$AirBags)/nrow(Cars93)

Driver & Passenger        Driver only               None
         0.1720430          0.4623656          0.3655914

> round(table(Cars93$AirBags)/nrow(Cars93),3)

Driver & Passenger        Driver only               None
             0.172              0.462              0.366
```

Bar Graphs

- A ***bar graph*** is a graphical representation of a frequency distribution.
- One bar is displayed for each category, and the height of each bar is the frequency (count) or relative frequency (proportion) in each category.
- The width of the bars has <u>no meaning</u>.

```
> barplot(table(Cars93$AirBags),ylab="Frequency",cex.lab=1.3,col=c(2,3,4))
> barplot(table(Cars93$AirBags)/nrow(Cars93),ylab="Relative Frequency",
+ cex.lab=1.3,cex.names=1.2,col=c(2,3,4))
```





A Note on Colors

You will often want to add color to a graphic (lines, plotting characters, fill, …) and R has a large variety of color possibilities. As is often the case in R, there are multiple ways to specify colors.

The available built-in color names can be accessed with the `colors` function. Here are the first 20 (of 657)

```
> colors()[1:20]
 [1] "white"          "aliceblue"      "antiquewhite"   "antiquewhite1"  "antiquewhite2"
 [6] "antiquewhite3"  "antiquewhite4"  "aquamarine"     "aquamarine1"    "aquamarine2"
[11] "aquamarine3"    "aquamarine4"    "azure"          "azure1"         "azure2"
[16] "azure3"         "azure4"         "beige"          "bisque"         "bisque1"
```

The **color palette** tells R which color name is referred to by a specific integer. It can be viewed using the `palette` function.

```
> palette()
[1] "black"   "red"     "green3"  "blue"    "cyan"    "magenta" "yellow"  "gray"
```

This shows that in the current palette (the default) 1 indicates black, 2 gives red, 3 gives green3, 4 gives blue, etc.

A Note on Colors

You can also set the palette with the `palette` command.

```
> palette(c("red2","orchid1","yellow4","tomato2"))
> palette()
[1] "red2"    "orchid1" "yellow4"  "tomato2"
```

The color palette now assigns 1 to red2, 2 to orchid1, 3 to yellow4, etc. We can restore the default at any time using

```
> palette("default")
> palette()
[1] "black"   "red"     "green3"  "blue"    "cyan"    "magenta" "yellow"  "gray"
```

When setting the `col` parameter, either use the color names (with quotes around them) or first set the palette and then use the mapped integers.

Additional colors in R can be created using primitives `rgb`, `hsv`, and `hcl` or the derived `rainbow`, and `heat.color`.

Bar Graphs

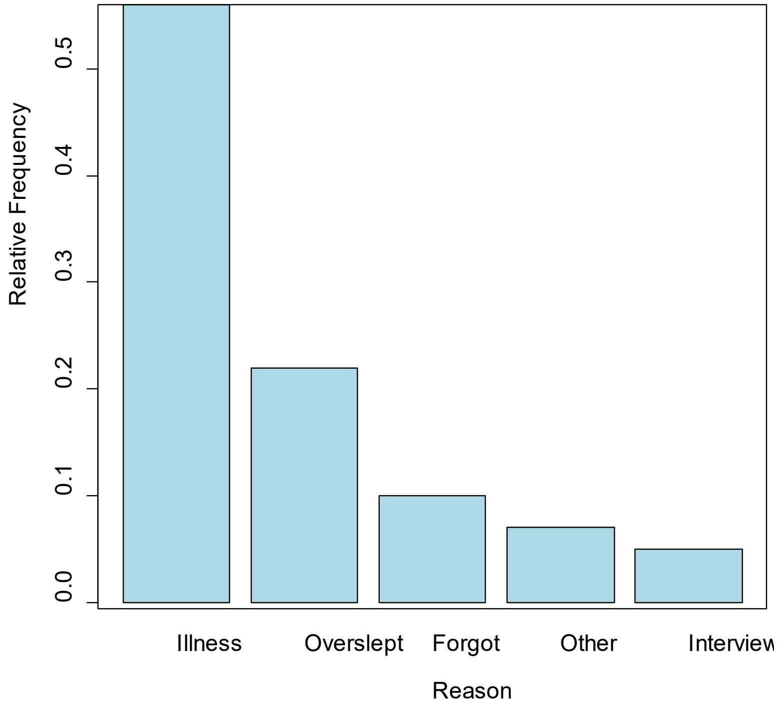
A bar graph is called a ***Pareto Chart*** if it is constructed with the categories presented in order of frequency or relative frequency with the largest value to the left.

– This can be used to highlight important issues

Example: The following is a Pareto chart for the reasons given for missing my Single Variable Calculus class last semester.

```
> barplot(sort(table(Reason)/length(Reason),decreasing=T),ylab="Relative Frequency",
+ xlab="Reason",col="lightblue",cex.names=1.2,cex.axis=1.2,cex.lab=1.2)
> box()
```

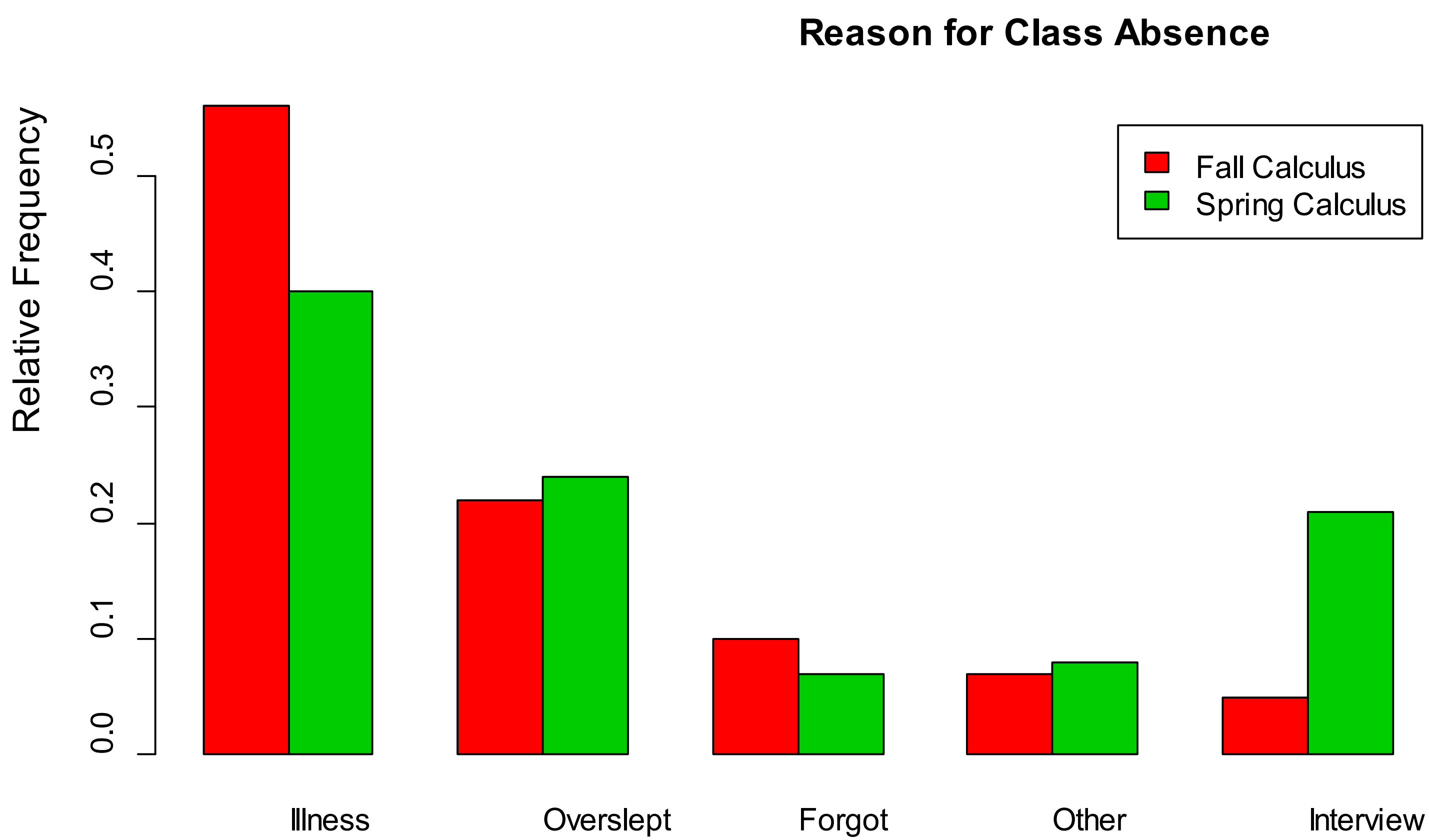
The function `box` draws a box around the current plot



Bar Graphs

To compare two bar graphs with the same categories, we can construct a ***side-by-side bar graph***.

<u>Example</u>: The following side-by-side bar graph shows the reasons given for missing my Single Variable Calculus class during the spring and fall semesters last academic year.



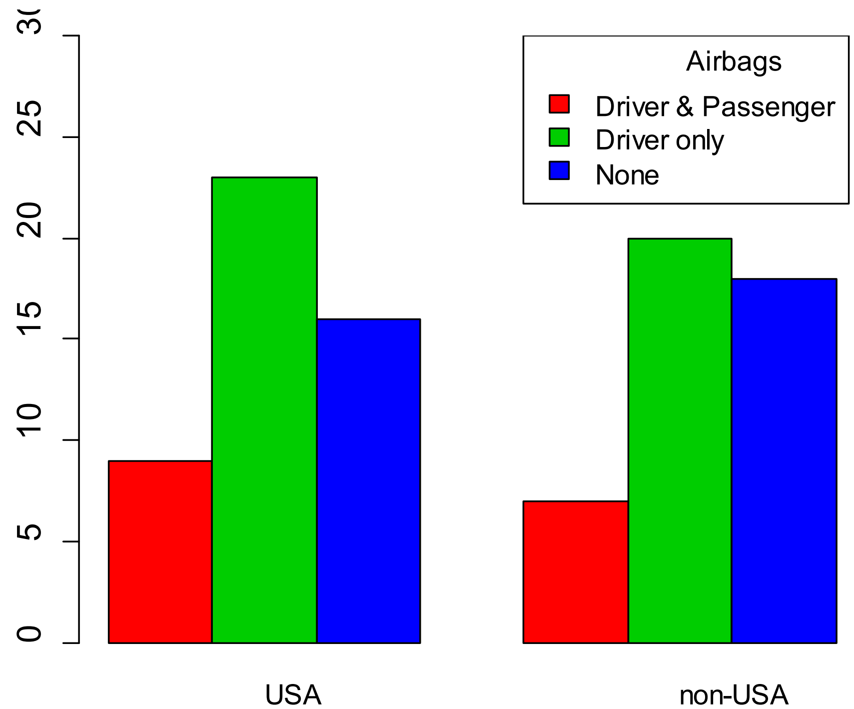
Example

We can obtain the two-way frequency table of Airbags by Origin using:

```
> table(Cars93$AirBags,Cars93$Origin)

                    USA non-USA
  Driver & Passenger   9       7
  Driver only         23      20
  None                16      18
```

```
> barplot(table(Cars93$AirBags,Cars93$Origin),col=c(2,3,4),beside=T,
+ ylim=c(0,30), cex.axis=1.2)
> legend(x=5,y=30,title="Airbags",legend=sort(unique(Cars93$AirBags)),
+ fill=c(2,3,4))
```

The script above provides the figure to the right depicting the `AirBags` variable as a function of the `Origin` variable.



Pie Charts

- A ***pie chart*** is a graphical method for displaying the distribution of a qualitative variable.
- The circle or pie represents the whole (all the units). The pie is divided into slices, one for each category of the qualitative variable.

```
> pie(table(Cars93$AirBags),main="Standard Air Bags",col=c(2,3,4),cex=1.3)
```

Note that the eye is good at judging linear measures and bad at judging relative areas. A bar chart or is often a preferable way of displaying this type of data.