**Summarizing Data – Graphical Methods**

**Topics**

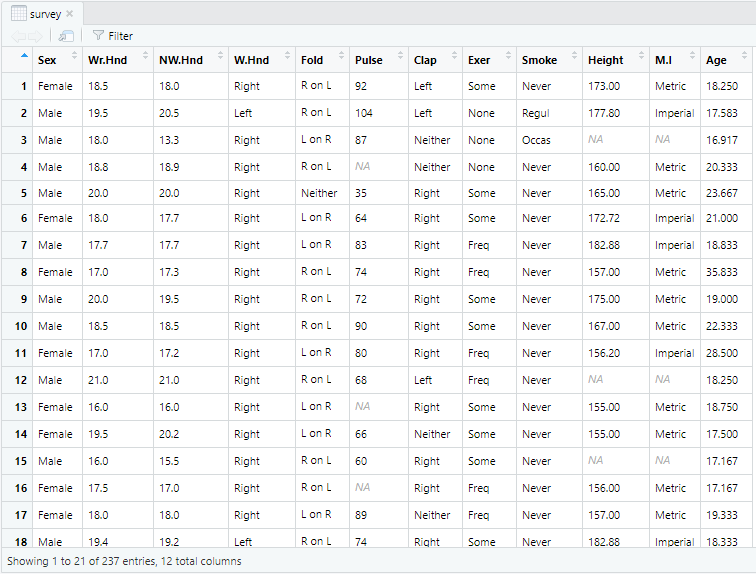
**Introduction – Descriptive versus Inferential Statistics**

* **What is/are statistics?**

**Statistics is the branch of mathematics concerned with *making sense of data*: collecting, summarizing, organizing, presenting, and analyzing numerical information.**

**Descriptive statistics covers methods for organizing and summarizing possibly rather large collections of data.**

**Example – student survey in R (built-in dataset).**

****

**Inferential statistics (or statistical inference) which provides methods for (1) estimating values of population parameters (statistical estimation) or (2) evaluating the degree of support for a statement or claim about some characteristic of a population (hypothesis testing) using data obtained from a random sample of that population.**

**(1)**

**(2)**

**Descriptive Statistics**

* **Graphical Methods**
* **Numerical Methods (Numerical measurements of position and variation)**

**Graphical ways of summarizing data**

**Pie Chart**

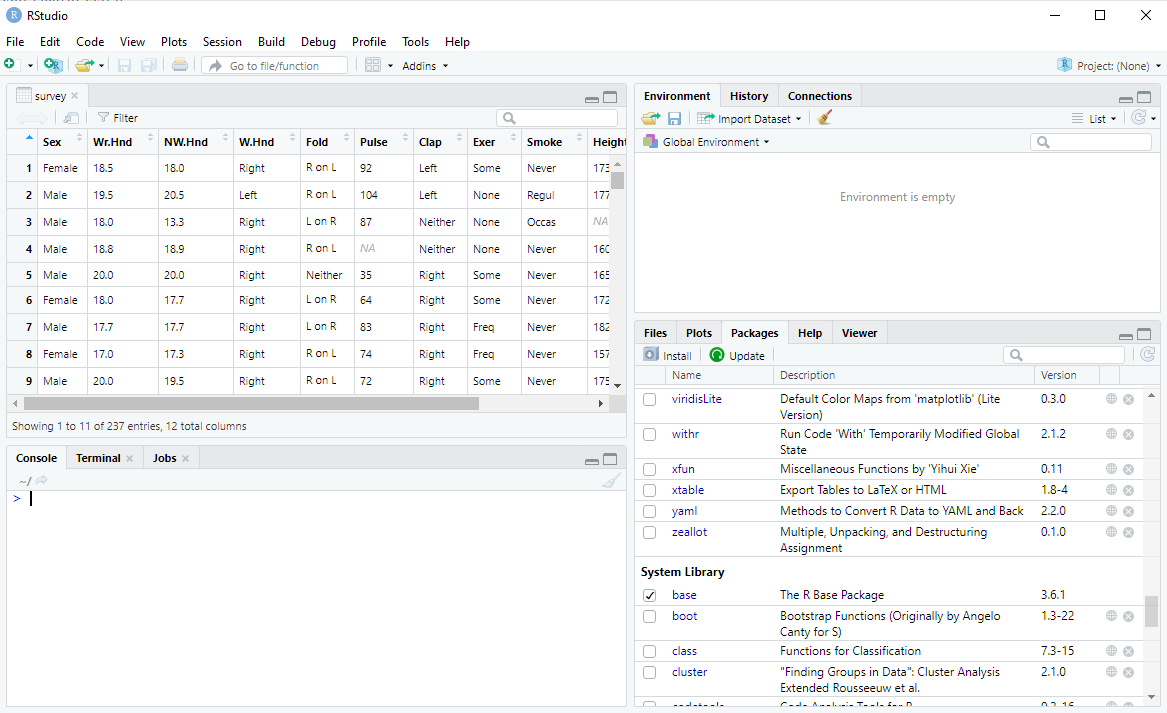
**Stem-and-Leaf Plots**

**Frequency Histograms**

**Ogive**

**Scatter Diagrams**

**We will be using the software R to create graphs and charts. R is a programming language and free, open-source software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.**

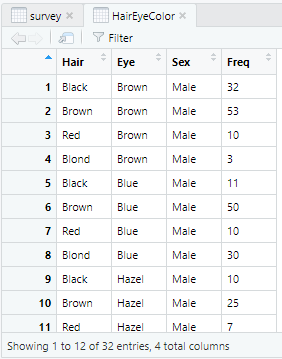
****

### Qualitative data:

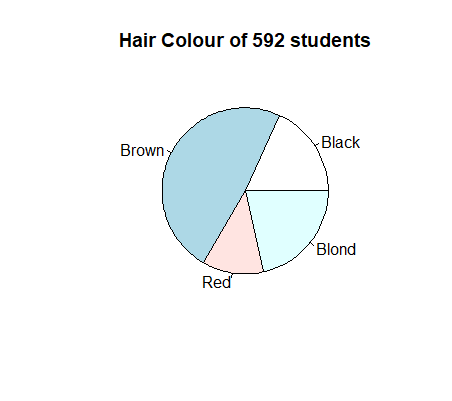
### Pie chart

**Pie charts are generally a poor way to represent quantitative data. However, they provide a simple way to visualize qualitative data.**

**The built-in dataset *HairEyeColor* in R gives the hair and eye colour of 592 students.**

****

**If we are interested in visualizing the proportion of students who have each different hair colour, we can easily represent that information in a pie chart.**

****

### Quantitative data

### We will be focusing on quantitative data in this course.

### When we have a raw data set, it is often difficult to get information from it, especially if the data set is large. Consider, for example, the built-in dataset *faithful*, which gives the duration of eruptions as well as the time between eruptions (both in minutes) from the geyser Old Faithful.

|  |  |
| --- | --- |
|  |  |

**It is hard to obtain any meaningful information from this dataset in its current form.**

**How can we organize the data?**

### The data comes from the built-in dataset *faithful* in R. We can view the data in R with the command *View(faithful)*.

**It is clear that we are mostly interested in roughly how long the eruptions were – “between 1.8 and 2 minutes” is sufficient information. This will allow us to group the data.**

**A *stem-leaf plot* allows us to group the data without losing any information. To create one, we divide the data into stems – based on the leftmost significant figures in the data – and the leaves. Here is the stem-leaf plot produced by the command *stem(faithful$eruptions)*:**

The decimal point is 1 digit(s) to the left of the |

16 | 070355555588

18 | 000022233333335577777777888822335777888

20 | 00002223378800035778

22 | 0002335578023578

24 | 00228

26 | 23

28 | 080

30 | 7

32 | 2337

34 | 250077

36 | 0000823577

38 | 2333335582225577

40 | 0000003357788888002233555577778

42 | 03335555778800233333555577778

44 | 02222335557780000000023333357778888

46 | 0000233357700000023578

48 | 00000022335800333

50 | 0370

**R helpfully tells us how to interpret the plot. The first row, for instance, gives the following times for eruptions (in minutes): 1.60, 1.67, 1.70, 1.73, 1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 1.78, 1.78.**

**What can we say about the eruptions?**

**We can customize our stem plot by adjusting parameters.**

**Frequency distribution**

**For large data sets, a stem and leaf plot is ridiculous to construct. Even the Old Faithful stem and leaf plot is a bit unwieldy and gives us more information than we need. Instead we can form intervals (classes) and tally the number of values within each interval. We can do this by sorting the data into *classes* with a frequency distribution.**

**Tips: The classes should all be the same width.**

**The classes should capture all the data.**

**The classes should not overlap.**

**If we use the same classes of 0.2 minutes apiece that we used for the stem plot, R can give us the following frequency table:**

[1.6,1.8) 12

[1.8,2) 39

[2,2.2) 20

[2.2,2.4) 18

[2.4,2.6) 3

[2.6,2.8) 3

[2.8,3) 2

[3,3.2) 1

[3.2,3.4) 4

[3.4,3.6) 6

[3.6,3.8) 10

[3.8,4) 16

[4,4.2) 31

[4.2,4.4) 29

[4.4,4.6) 35

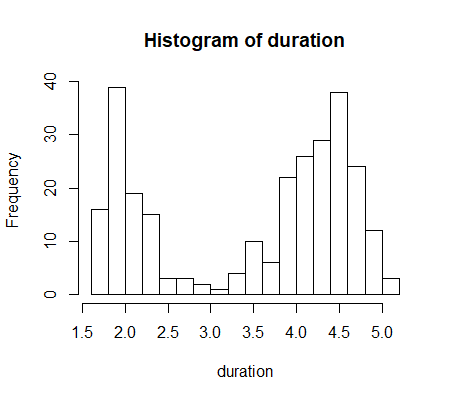
[4.6,4.8) 28

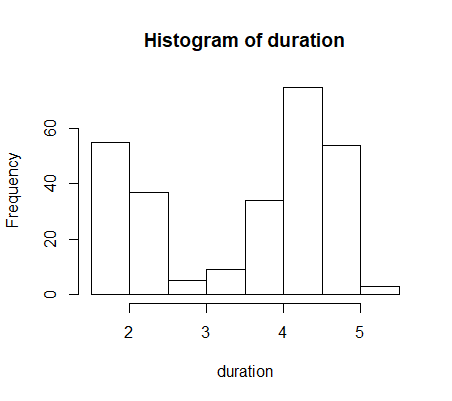
[4.8,5) 11

[5,5.2) 4

**For example, there were 12 eruptions out of 272 that were between 1.6min and 1.8min in length. Note the brackets: [1.6,1.8) denotes a class that includes 1.6 minute eruptions but excludes 1.8 minute ones. This is so that there is no overlap between classes.**

**It is still hard to make sense of this data in the above form, so we can represent it graphically with a histogram.**

****

**This histogram still looks a bit cluttered. We can change the number of classes as well. We want a number of classes so that we can obtain meaningful information about the data at a glance. Between 5 and 20 classes is a standard guideline, with the classes chosen so that the numbers are “nice”. (eg, a class of [1.6,1.8) is nicer than a class of [1.56,1.87).)**

**Cumulative Frequency Table**

**Often we are interested in knowing how much of our data is *less* than a certain value. For instance, we might be interested in knowing how many of Old Faithful’s eruptions were less than 4 minutes. We can create a *cumulative frequency* table with that information.**

**Using the intervals of 0.2 minutes again, R gives us the following cumulative frequency table:**

[1.6,1.8) 12

[1.8,2) 51

[2,2.2) 71

[2.2,2.4) 89

[2.4,2.6) 92

[2.6,2.8) 95

[2.8,3) 97

[3,3.2) 98

[3.2,3.4) 102

[3.4,3.6) 108

[3.6,3.8) 118

[3.8,4) 134

[4,4.2) 165

[4.2,4.4) 194

[4.4,4.6) 229

[4.6,4.8) 257

[4.8,5) 268

[5,5.2) 272

**This table tells us that \_\_\_\_\_\_ of the eruptions were less than 4 minutes in duration. Note that the cumulative frequencies increase.**

**By dividing each of these cumulative frequencies by the total number of eruptions (which is equal to the last of these entries, 272), we can obtain a *cumulative relative frequency table*:**

[1.6,1.8) 0.04

[1.8,2) 0.19

[2,2.2) 0.26

[2.2,2.4) 0.33

[2.4,2.6) 0.34

[2.6,2.8) 0.35

[2.8,3) 0.36

[3,3.2) 0.36

[3.2,3.4) 0.38

[3.4,3.6) 0.40

[3.6,3.8) 0.43

[3.8,4) 0.49

[4,4.2) 0.61

[4.2,4.4) 0.71

[4.4,4.6) 0.84

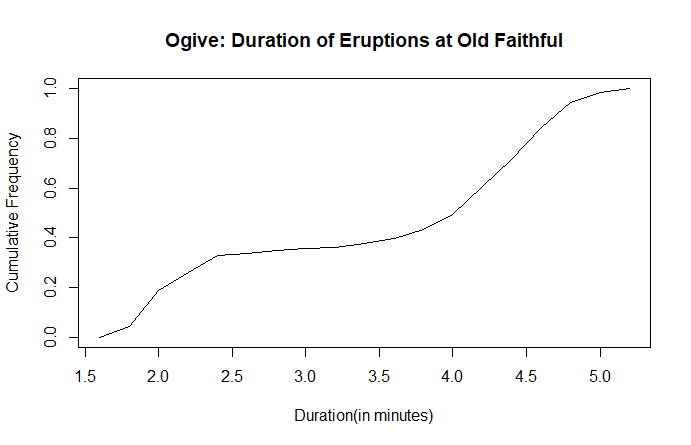
[4.6,4.8) 0.94

[4.8,5) 0.99

[5,5.2) 1.00

**Here we can see that \_\_\_\_\_\_ of eruptions were less than 4 minutes in duration. (Note that these numbers are rounded to two decimal places. By default R displays 8 digits, which is excessive.)**

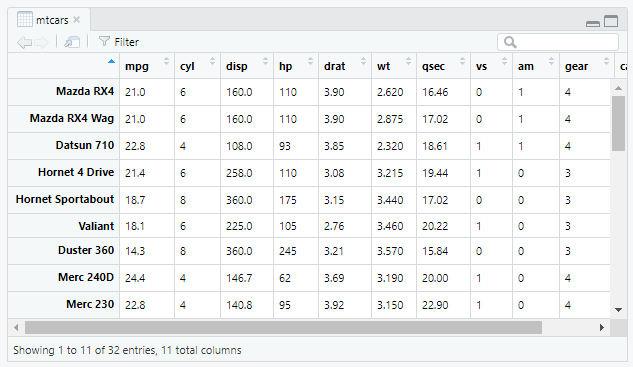
**We can use the table above to form an ogive (“oh-jive”). An ogive is a line graph that depicts cumulative frequencies or cumulative percent frequencies.**

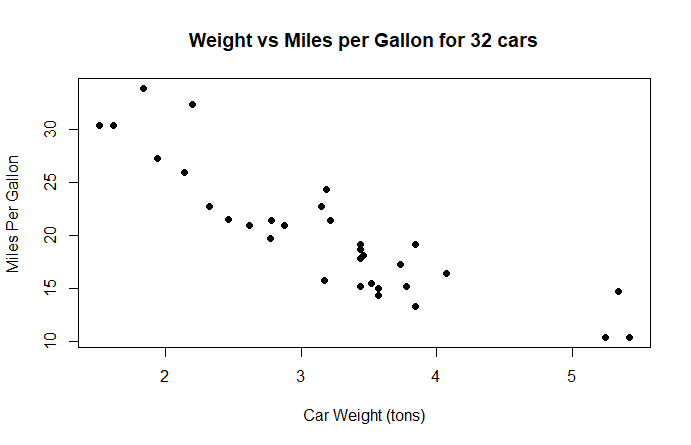
****

**Approximately what percentage of eruptions were less than 3 minutes?**

**Approximately 30% of eruptions were less than how long?**

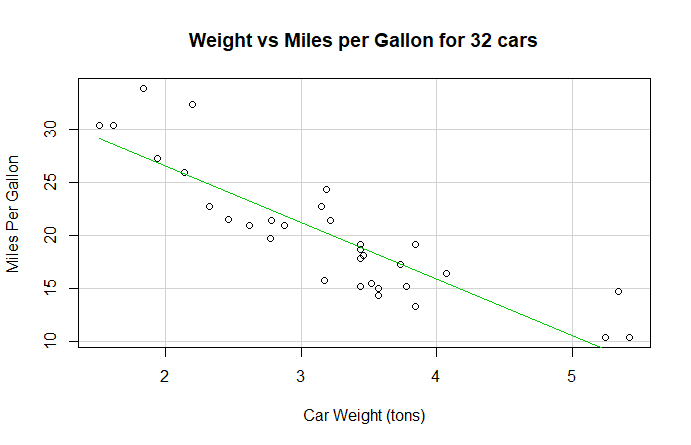
**Scatter Diagrams**

**Often we are interested in knowing how two variables in a dataset are related. The file *01 – cars.csv,* based on R’s dataframe *mtcars*,contains information about 32 makes of cars.**

****

**We can see how the cars’ weights and their efficiencies (in miles per gallon) by creating a *scatter diagram*, or *scatterplot*.**

**What trend can we see from the scatterplot?**

**We can get R to create a *line of best fit*.**

**Here, R tells us that the equation of the line of best fit is**

**y=-5x+37**

**where x is the car’s weight in tons, and y is the number of miles per gallon. Since the data doesn’t fall exactly on the line, there is some error associated with this equation. Later, we will see when it is appropriate to construct a line of best fit, and when other methods are necessary.**