

In this second session, we will learn:

1. How to access Prof Poje's R data files and functions easily over the web.
2. How to find the **RANGE**, **MAX**, and **MIN** of a data set in R .
3. How to customize the **Histograms** produced by R .
4. How to create **Stem and Leaf** diagrams in R.
5. Get ready for the first, full-blown R **HOMEWORK/LAB** assignment.

1 Accessing Prof. Poje's R stuff

From time to time, I will put data and collections of R commands on my computer so that you may access them. This should save you time typing in lots of numbers and/or complicated R commands. R makes it easy for you to get at them with its built-in web-access.

For example lets open up R and try getting the test exam data set we discussed in class. This is done by typing:

```
> scores = read.table(file=url("http://www.math.csi.cuny.edu/~poje/R_stuff/TestData"))
```

To see what we got, look at `scores`,

```
> names(scores)
> attach(scores)
> plot(Exam1,Exam2)      (Did students who did well on Exam1 do well on Exam2?)
> hist(Exam1)
```

You should get a *dataframe* called `scores` with the results from two tests (`Exam1` and `Exam2`). From now on, you should be able to access any data in Prof Poje's `R_stuff` directory by using the above `read.data` command.

Sometimes, I will post new R functions in the `R_stuff` directory. The command for accessing new functions is very similar to reading data. To access a silly R function, try

```
> source(file=url("http://www.math.csi.cuny.edu/~poje/R_stuff/Silly.r"))
```

You should now have a brand new R function called `Silly`. To see what it does, try

```
> Silly()
```

.... pretty SILLY, hah? (Note: Don't forget that R insists that ALL functions have those pesky parentheses!)

2 R 's RANGE,MAX,MIN commands

We know that the first thing we need to do to make a frequency distribution (histogram) is to compute the **range** of the data. This all built-in to R . For example, to find the highest score on Exam1, try

```
> max(Exam1)
```

Pretty straightforward. To find the lowest grade on Exam2

```
> min(Exam2)
```

To find the complete **range** of scores on Exam1, try

```
> range(Exam1)
```

What happens when if you type:

```
> range(scores)
```

Explain the answer.

3 More about plotting data in R

Now that we know something more about histograms, let's look a bit deeper into R 's **hist** command. For example, try the following

```
> hist(Exam1,main="My Name is Lucille!")
```

What did the **main=** argument do to the picture? How would you label a histogram with the information: "Produced by *your name here.*" ? This works for most of R 's plotting routines, you can create a title for the graph by using the **main=** argument.

Now, back to statistics. Does the histogram R produced look exactly like the one we produced in class? How did R decide how to pick the intervals? What if we don't like the one's it picked? How can we draw a histogram with more, or less, intervals?

All this can be taken care of by setting the **breaks=** argument of the **hist** function. For example, suppose we want more, smaller intervals. Try

```
> hist(Exam1,breaks=10)
```

or

```
> hist(Exam1,breaks=20)
```

What changes in the picture? Which picture is 'right'?)

We can go one step further. Suppose we want to set exactly where the intervals are. For example, we may want to divide the Exam1 data into 12 intervals, starting at 45 and counting by 5 until 100. Try

```
> hist(Exam1,breaks=c(45,50,55,60,65,70,75,80,85,90,95,100))
```

Actually, we could have done this with much less typing. Try

```
> mybreaks = seq(45,100,5)      (This assigns numbers from 45 to 90 by 5)
```

Check out `mybreaks`, and use it in the `hist` plot.

```
> mybreaks
> hist(Exam1,breaks=mybreaks)
```

Now add your name and some fancy colors:

```
> hist(Exam1,breaks=mybreaks,main="Pugsley Addams", col="red")
```

4 Frequency Distributions versus Relative Frequency Distributions

We talked in class about computing the **Relative Frequency** of a class of samples. In other words, we compute what fraction of the total number of samples in each interval. This changes the scale of the scale of the histogram (the y axis) but not the shape. Also this allows us to approximate the **probability** that a random sample would take values in a given interval.

Of course, R can do this too. Try making some data and looking at the histogram

```
> junk = c(2,2,3,3,1)
> hist(junk)
```

Yuck! Not a pretty histogram, you shouldn't like the intervals R has chosen. Change them to intervals *centered* on the data.

```
> hist(junk,breaks=(0.5,1.5,2.5,3.5))
```

Much better. Now let's compute the **relative frequency**. This is easily done by adding an argument `probability=T` to the `hist` function.

```
> hist(junk,breaks=(0.5,1.5,2.5,3.5),probability=T)
```

Check out the y -axis now. It says that the relative frequency of a 1 in the data is $1/5 = 0.2$, for a 2, this is $2/5 = 0.4$, for a 3 this is again, $2/5 = 0.4$.

5 Stems and Leaves in R

Now that we know what a **stem and leaf diagram** is, let's again let R do the dirty work. Try:

```
> stem(Exam1)
```

Easy! Let's see what information we can read straight off the stem and leaf tells us. What was the highest grade on `Exam1`? What was the lowest? How does the stem and leaf compare to the **histogram**?

Assignment!

Do the following and hand in before leaving the lab today:

1. Use R to load the file `TestData` from Prof. Poje's directory.
2. Make a histogram of the test scores on Exam1 and Exam2. For each histogram, add your name as a title and PRINT the output.
3. For each set of grades, make a *strange* histogram. Set five intervals corresponding to letter grades:
 - F: 0-59
 - D: 60-69
 - C: 70-79
 - B: 80-89
 - A: 90-100

NOTE: R will try to make a relative frequency distribution, dont let it! Set the argument `probability=F` in the `hist()` function.

Print out the histograms with your name as a title. On each, write down how many students recieved A's,B's,C's etc on each test.

4. Use the R function `mean()` to find the average student grade on each exam. Write this on the graph as well. How many students recieved a grade *near* (say with 5 points on either side of the mean) the average on each test?
5. Explain, in words, the overall results of each test. In your opinion, was one test more *reasonable* than the other? Why?
6. Explain why the average of a set of numbers may not be a good way to describe the members of the set.