R TUTORIAL, #1:
DATA, FREQUENCY TABLES, and HISTOGRAMS

The (>) symbol indicates something that you will type in.
A bullet (•) indicates what the R program should output (and other comments).

ENTERING DATA

> Type (in the R Console window): scores <- c(42, 8, 59)
> Press RETURN or ENTER on your computer.
    • This creates a "data vector" called "scores." The 'c' command stands for "combine values."

> Type: scores
    • You will see the three data values. The '[1]' indicates that 42 is the 1st data value.

> Type: scores[3]
    • You will see the third score in the data vector.

> Type: sum(scores)
    • You will see the sum of the scores.

HELP SCREENS

> Type: ?c
    • You will see a very technical help file on the command 'c'. Go ahead and close this window.

SEQUENCES

> Type: seq(0, 100)
    • You will see the sequence of integers from 0 through 100.

> Type: seq(0, 100, by=5)
    • Here, the values jump by 5s.

PRESIDENTS' AGES DATA SET

The following is a list of the ages of the U.S. presidents when they became president. (The ages are in years, rounded down.)

57 61 57 57 58 57 61 54 68 51 49 64 50 48 65 52 56 46 54 49 51 47 55 55
54 42 51 56 55 51 54 51 60 62 43 55 56 61 52 69 64 46 54

(President Obama was 47, but he is not included in the table presented in the class notes. The notes were written before 2009.)

ENTERING DATA USING SCAN( )

> Type: ages = scan( )
- You should see '1:' in your console window.

> Copy and paste the 43 numbers in the PRESIDENTS' AGES DATA SET above.
- Notice that, unlike for the 'c' command, commas (,) are NOT used with 'scan'.
- Copying can be done by using CTRL-C.
- Pasting can be done by using CTRL-V.

> Press RETURN or ENTER on your computer.
* You should see '44:' in your console window. This means that, if you were to enter another value, it would be the 44^{th} data value.

> Press RETURN or ENTER again, since we are not entering in any more values.
- You should see 'Read 43 items' in your console window.

> Type: ages[16]
- You should see '[1] 52'. This indicates that Abraham Lincoln, our 16^{th} president, was 52 years old when he became president.

FREQUENCY TABLES

> Type: table(ages)
- You should see a sorted list of the data values, listed with their frequencies.

> Type: boundaries = seq(34.5, 69.5, by=5)
- The sequence of numbers we will use to separate our classes will be the numbers from 34.5 through 69.5, jumping by 5s. These numbers are called "class boundaries."

> Type: boundaries
- You will see the list of class boundaries.

> Type: table(cut(ages, boundaries))
- You will see a frequency table for the ages.

> Type: table(cut(ages, c(boundaries, Inf)))
- This includes the last class of "70+" years. The "Inf" indicates that the last class goes off to infinity.

RELATIVE FREQUENCY TABLES

> Type: length(ages)
- This tells you that there are 43 ages in our data set.

> Type: table(cut(ages, boundaries)) / 43
- You will see a relative frequency table for the ages.

BARPLOTS

> Type: barplot(ages)

- You will see (in a separate window) bars corresponding to the ages in the order we entered them in, but this is NOT a correct histogram.

HISTOGRAMS

> Type: hist(ages)

- You will see a histogram of the ages.

> Type: hist(ages, breaks=boundaries)

- You will see a histogram similar to what we have in our notes.

RELATIVE FREQUENCY HISTOGRAMS and LABELS

> Type: hist(ages, prob=TRUE)

- You will see a relative frequency histogram of the ages.
- The histogram you see is different from the one we have in our notes, because relative frequencies correspond to areas here, not heights. Different sources do the histograms differently.
- Here, "prob" means "probability that a randomly selected data value lies in the class." Probabilities are often related to relative frequencies.
- 'TRUE' can be abbreviated as 'T'.

> Type: hist(ages, breaks=boundaries, prob=T, main="Relative frequency histogram", ylab="Relative frequencies")

- You will see a relative frequency histogram of the ages similar to what we have in our notes, except that relative frequencies correspond to areas, not heights.
- 'main' means "main title" here.
- 'ylab' means "label for *y*-axis."
- 'xlab' means "label for *x*-axis." We didn't use that here.