Stats for Clinical Trials, Math 150
Jo Hardin
Lab # 1, Due: Tuesday, Sep 7, 2010
Homework # 1, Due: Thursday, Sep 9, 2010

Lab # 1: Activities (1) - (8) in Chp 2, Stat2Labs (available on Sakai)
HW # 1: Book 2.1, 2.4, 3.2; Chp 2 Stat2Labs K2.34

True for all assignments all semester: graphs must be described in a sentence or two and must have labels with units. Numerical summaries must have units. Confidence intervals and hypothesis tests must have one to two sentence interpretations in the words of the problem.

Important note: I highly recommend you writing every single line of code into Notepad or Wordpad (you don't want to use Word, because it will correct your grammar and syntax). Really, I promise that you need to get into the habit of typing EVERYTHING into a separate file and copy & pasting it into R.

Some analysis I did on the Hodgkins data

The data set `Hodgkins` contains plasma bradykininogen levels (in micorgrams of bradykininogen per milliliter of plasma) in three types of subjects (normal, in paitents with active Hodgkin's disease, and in patients with inactive Hodgkin's disease). The globulin bradykininogen is the precursor substance for bradykinin, which is thought to be a chemical mediator of inflammation. (Taken from *Stat2Labs*, Kuiper.)

- After going through the R tutorial, to work on the data, the first thing I did was copy the data into the folder called "Statistics". Then, in R, I typed the following commands:

```
> hodg.dat <- read.table("Hodgkins.csv",header=T, sep=",")
> attach(hodg.dat)    # allows me to use the column names as variables
> names(hodg.dat)     # prints out the column names that are my variables
> dim(hodg.dat)       # tells me there are 65 rows and 2 columns
> hodg.dat            # prints out the actual data for me
```

- I wonder what the distribution of subjects looks like?

```
> table(subject)
```

- How high are the highest bradykininogen levels?

```
> sort(brady)
```

  Are they higher for different subjects?

```
> sort(brady[subject=="active"])  # active Hodgkins
> sort(brady[subject=="normal"])  # healthy
```

  It's hard to tell with just the numbers, let's look at some summary statistics:

```
> summary(brady)  # note, there are also min, max, median, and mean functions
> summary(brady[subject=="active"])
> summary(brady[subject=="normal"])
> sd(brady)
> sd(brady[subject=="active"])
> sd(brady[subject=="normal"])
```

  Fortunately, the variability numbers look similar for bradykininogen levels across the normal and active groups. Why did I start the previous sentence with the word fortunately?

- What if I want to see graphs for baseline temperature (broken down by treatment):

```
> hist(brady)
> hist(brady[subject=="active"])
> hist(brady[subject=="normal"])
> boxplot(brady)
> boxplot(brady~subject)
```

  But it seems like it would be nicer to have all the graphs on one page... Type the following before typing the plot commands above:

```
> par(mfrow=c(2,3))  # makes a matrix of plots which is 2 x 3
```

- Residuals for the t-test model would simply be the values minus their sample means.

```
> tapply(brady, subject, mean)
  active inactive   normal
  4.305625 6.856667 6.095000
> table(subject)
  subject
    active inactive   normal
        16       27       22
> rep(tapply(brady,subject,mean),table(subject))    # repeat
 [1] 4.305625 4.305625 4.305625 4.305625 4.305625 4.305625 4.305625 4.305625 4.305625
[10] 4.305625 4.305625 4.305625 4.305625 4.305625 4.305625 4.305625 6.856667 6.856667
[19] 6.856667 6.856667 6.856667 6.856667 6.856667 6.856667 6.856667 6.856667 6.856667
[28] 6.856667 6.856667 6.856667 6.856667 6.856667 6.856667 6.856667 6.856667 6.856667
[37] 6.856667 6.856667 6.856667 6.856667 6.856667 6.856667 6.856667 6.095000 6.095000
[46] 6.095000 6.095000 6.095000 6.095000 6.095000 6.095000 6.095000 6.095000 6.095000
[55] 6.095000 6.095000 6.095000 6.095000 6.095000 6.095000 6.095000 6.095000 6.095000
[64] 6.095000 6.095000
> hodg.resid <- hodg.dat - rep(tapply(brady,subject,mean),table(subject))
```

- Let's test whether the bradykininogen levels are different for people with active Hodgkin's compared with healthy people:

```
> ?t.test     # to find out info about the t-test
> t.test(brady[subject=="normal"],brady[subject=="active"],paired=F,alt="two.sided")
```

  Why paired = F? Why alt="two.sided"? How do you interpret the results? What about the confidence interval?

- A scatterplot is created with two numerical variables.

```
> plot(myexplan, myresp)
```

- To create a linear model, we use the `lm` command:

```
> myreg <- lm(myresp ~ myexplan)
> resid(myreg)     # to get the residual values
> fitted(myreg)    # to get the predicted values
> abline(myreg)    # to add the regression line to your plot
```