36-463/663: Multilevel & Hierarchical Models
Fall 2016
HW02 – Due Tue, 13 Sept 2016

Announcements

- Homework is due Tues 13 Sept on Blackboard. Submit a single file only. pdf preferred, doc or docx acceptable.
- You should have read or be finishing Gelman & Hill (G&H), Chapters 1 and 2. Please get started on reading G&H Chapters 3–4.
- This assignment involves both
  - Material from Ch 4 of *Using R for Data Analysis and Graphics: Introduction, Code and Commentary*, by J. H. Maindonald, Centre for Mathematics and Its Applications, Australian National University (`usingR.pdf`), and the accompanying data file (`usingR.RData`). You can get these at

    `http://www.stat.cmu.edu/~brian/463/hw01`

    just like last week.

    You will also need the package "`ggplot2`"[1], which you can install from the "Packages" menu in the R console window. After you install the lattice package, use the command `library(ggplot2)` to make ggplot functions like `ggplot()` or `qplot()` available for use. You can get a list of functions in the lattice library with `library(help=ggplot2)`, and help on specific functions like `qplot()` with `help(qplot)`.

  - Examples and data sets from G&H, Chapters 1 and 2. You can get these from

    `http://www.stat.columbia.edu/~gelman/arm/examples/`

Exercises

1. [5 parts] Read all of Chapter 4 of Maindonald. Since the examples are in `lattice` rather than `ggplot2` you do not have to try the things in the chapter, unless you are curious. You will need some trial-and-error to make sense of things. Then please do and turn in the following exercises from Chapter 4, using basic R graphics and `ggplot2` as needed[2] (instead of `lattice`):

   (a) Chapter 4, #1. Hint: To enter the data try something like this *(see also additional hint on next page)*:

```
yieldvec <- c(621, 793, 593, 545, 753, 655, 895, 767, 714, 598, 693)
n0 <- length(yieldvec)
yieldvec <- c(yieldvec, 947, 945, 1086, 1202, 973, 981, 930, 745, 903,
          899, 961)
n1 <- length(yieldvec)
typevec <- c(rep("Smoking",n0),rep("Nonsmoking",n1-n0))
milk.vol <- data.frame(yield=yieldvec, type=typevec)
```

[1] Recall, we are using `ggplot2` instead of the `lattice` package discussed in the Maindonald pdf.

[2] Please consult `02a-intro-ggplot-graphs.r` and `02b-intro-ggplot-london.r` from last week's lecture for examples of the kinds of plotting functions you'll need.

You will need commands like

- `ggplot(...) + geom_boxplot(...) + xlab(" ") + coord_flip()`
- `ggplot(...) + geom_dotplot(...)`

(b) Chapter 4, #2. Despite the use of the plural ("histograms", "normal probability plots", "density plots"), make one plot for each of the three parts of #2. Your plotting commands will be something like this (using `ggplot2`):

- `ggplot(...) + geom_histogram(...)`
- `ggplot(...) + stat_qq(...)`
- `ggplot(...) + geom_density(...)`

(c) Re-do part (1b) but now using the built-in plotting functions:

- `hist()`
- `qqnorm()`
- `density()`

Comment on any differences between (1b) and (1c).

(d) Chapter 4, #12. Hint: First, try something like

```
ggplot(possum,aes(x=hdlngth)) + geom_histogram() + facet_grid(sex ~ Pop)
```

This makes a set of histograms of `hdlngth`, taking into account `sex` and `Pop`. You want to do something similar for each part of this problem.

(e) Chapter 4, #13. Use `ggplot2` commands, of course.

2. [4 parts] Please read and try everything in G&H Chapter 2. Note that some of the code is written out for you in the file `book_codes.zip` at `http://www.stat.columbia.edu/~gelman/arm/examples/`. Then please do and turn in:

(a) G&H Chapter 2, #2. Note that part (b) asks you to conduct a hypothesis test, similar to the discussion in sections 2.3 and 2.4, but now comparing to a chi-squared distribution (`qchisq()`) under the null hypothesis.

(b) G&H Chapter 2, #3. Hint: Here are two possible correct solutions for this problem:

```
sims <- matrix(runif(20*1000),nrow=1000)
sums <- apply(sims,1,sum)
hist(sums,probability=TRUE)
m <- mean(sums)
s <- sd(sums)
x <- seq(min(sums),max(sums),length=100)
lines(x,dnorm(x,m,s))
```

```
sums <- NULL
for (i in 1:1000) {
  data <- runif(20)
  newsum <- sum(data)
  sums <- c(sums,newsum)
  }
hist(sums,probability=TRUE)
m <- mean(sums)
s <- sd(sums)
x <- seq(min(sums),max(sums),length=100)
lines(x,dnorm(x,m,s))
```

Whether you copy mine or create your own, write in detail about what each line of the code you use does, and how it helps move toward a solution to the problem. Use the `help()` *function liberally, experiment with the functions to see what they do, and use any other tools you need, to give a good explanation of each line.*

(c) G&H Chapter 2, #4. *Hints:* The first part of this problem is basically a more elaborate version of problem 2b: 1000 times, you need to do something like this:

```
mens_heights <- rnorm(n=100,mean=69.1,sd=2.9)
womens_heights <- rnorm(n=100,mean=63.7,sd=2.7)
x <- mean(mens_heights)
y <- mean(womens_heights)
height_diff <- x - y
```

Then make a histogram of your 1000 values of `height_diff`. Is the exact mean of `x - y` roughly in the middle of the histogram? Is it roughly equal to the mean of the 1000 values of `height_diff`? Is the exact standard deviation (SD) roughly equal to the sample SD of the 1000 values of `height_diff`? Is the histogram approximately six SD's wide *[why is this an appropriate question?]*?

(d) G&H Chapter 2, #5. Use the following information to complete this problem: The heights of men are approximately normally distributed with mean 69.1 inches and SD 2.9 inches. The heights of women are approximately normally distributed with mean 63.7 inches and SD 2.7 inches.

Hints:

- You should first find $\mathrm{Cov}\,(x, y)$ from the fact that $\mathrm{Cor}\,(x, y) = 0.3$, using the well-known formula

$$\mathrm{Cor}\,(x,y) = \frac{\mathrm{Cov}\,(x,y)}{\mathrm{SD}\,(x)\mathrm{SD}\,(y)}$$

- Then you will need to apply the well-known formulas

$$\begin{aligned} E[ax+by+c] &= aE[x] + bE[y] + c \\ \mathrm{Var}\,(ax+by+c) &= a^2\mathrm{Var}\,(x) + 2ab\mathrm{Cov}\,(x,y) + b^2\mathrm{Var}\,(y) \end{aligned}$$

No R is required for this problem. It is a straight calculation, like in 36-226 or a similar math stat course, using a little algebra and a calculator.