2010 R Workshop 2: Introduction to R and Statistics

Worked Example: Marginal, conditional, and multivariate distributions

In this example, we'll be looking at how multi-variate, marginal and conditional distributions relate to one another, and examine some of the methods for data processing and plotting available R.

We're going to try and debunk a common slander: that drinking coffee stunts your growth.

1. First, let's collect the data we're going to work on. Go to the Google Docs survey at https://spreadsheets.google.com/viewform?formkey=dFlXaW82TFduTUJEeFV3cWk0cEhRT0E6MQ and enter:

- You're height, in the form of <inches> in, or <cm> cm (Make sure to label which type of measurement it is... we'll have the code fix inconsistencies afterwords).
- your sex
- Do you drink coffee every day?

2. Once everyone in the class has answered the survey, go to the Google Docs spreadsheet summarizing all the data:

https://spreadsheets.google.com/ccc?key=0AoaOyrnls2xEdFlXaW82TFduTUJEeFV3cWk0cEhRT0E&hl=en&authkey=CIbkmuoK#gid=0

In Google Docs, click 'File' - 'Download as' - 'CSV (current sheet)'. Save the csv file to your current working directory, with the name "coffee height.csv".

3. Enter the following code in R:

```
coffee_height_data <- read.csv("coffee height.csv",header = T)
#type coffee_height_data to view whole data set
```

4. Now we'll create a new variable, in the same data frame, to correct for the different height measurements:

```
#first, we'll use an if statement to see if we've entered the height types
#correctly. This is one major way to use R: to detect mis-entered data
#The single vertical line "" means "OR", "!=" means "NOT", so this statement says
#"If any of the data in the variable "type" are not equal to either "cm" or
#"in", raise an error, otherwise run the rest of the code.
#If there is an error, try typing fix( coffee_height_data), and look for where
#the problem in the data is.
if(any(!(coffee_height_data$type == "cm" | coffee_height_data$type == "in"))){
      print("One or more of the type entries is wrong.")
}else{
      coffee_height_data$fixed_height <- ifelse(coffee_height_data$type == "cm",
coffee_height_data$height,  coffee_height_data$height*2.54)
}
```

5. Let's look at the marginal distribution of height, without worrying about our other variables for now. We'll look at two kinds of plot: histograms, which work by binning data points into discrete boxes so that the total area in the boxes equals 1, and kernel density estimators, which fit a smooth curve to the data which is constrained to integrate to 1 :

```
hist(coffee_height_data$fixed_height, freq=F)
points(density(coffee_height_data$fixed_height),type="l")
```

6. Let's focus now on the conditional distribution we're worried about: how does the distribution of height change between coffee and non-coffee drinkers? We'll be using a specialized plotting package from R, designed just for this sort of complex data exploration: GGplot2 (we'll be discussing how to use this package far more extensively in the course on data visualization)

```
require(ggplot2)
#This package uses a special graphical language for plotting; rather than setting
#up a grid, plotting histograms, and adding lines for the density after each plot
#(which we can also do in R, if necessary) we instead tell GGPlot2 each element we
#want to use in a separate command in an
#equation, and the package figures out how to make it look good. Try and figure out
#what each term in this equation is doing.
coffee_grid_plot <- ggplot(coffee_height_data, aes(fixed_height)) +
      geom_histogram(binwidth= 10,aes(y=..density..)) +
      geom_density() +
      facet_grid(.~coffee)+
      scale_x_continuous("height (cm)")

coffee_grid_plot
```

7. Now we can plot the multi-variate distribution of these three variables.

```
coffee_sex_grid_plot <- ggplot(coffee_height_data, aes(fixed_height)) +
      geom_histogram(binwidth= 10,aes(y=..density..)) +
      geom_density() +
      facet_grid(sex~coffee)+
      scale_x_continuous("height (cm)")

coffee_sex_grid_plot
```