

INFX 573 Lab: Conditional Probability

Pierre Augustamar

October 27th, 2016

Collaborators: Derrick Phoebe

Don't forget to list the full names of your collaborators!

Instructions:

Before beginning this assignment, please ensure you have access to R and/or RStudio.

1. Download the `week5b_lab.Rmd` file from Canvas. Open `week5b_lab.Rmd` in RStudio (or your favorite editor) and supply your solutions to the assignment by editing `week5b_lab.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name.
3. Be sure to include code chunks, figures and written explanations as necessary. Any collaborators must be listed on the top of your assignment. Any figures should be clearly labeled and appropriately referenced within the text.
4. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click Knit, rename the R Markdown file to `YourLastName_YourFirstName_lab5b.Rmd`, and knit it into a PDF. Submit the compiled PDF on Canvas.

In this lab, you will need access to the following R packages:

```
# Load some helpful libraries
library(tidyverse)
```

Problem: If a baseball team scores X runs, what is the probability it will win the game?

This is the question we will explore in this lab (ddapted from Decision Science News, 2014). We will use a dataset of baseball game statistics from 2010-2013.

Baseball is a played between two teams who take turns batting and fielding. A run is scored when a player advances around the bases and returns to home plate. More information about the dataset can be found at <http://www.retrosheet.org/>.

Note: More information about the dataset can be found at <http://www.retrosheet.org/>

```
colNames <- read.csv("cnames.txt", header=TRUE) # load cnames.txt data in colnames
baseballData <- NULL # set baseballdata to null
for (year in seq(2010,2013,by=1)){ #generating a loop from using a sequence that starts in 2010 and end
  mypath <- paste('GL',year,'.TXT',sep='') # concatenate using gl, year and has an extension txt
  # cat(mypath,'\n')
  baseballData <- rbind(baseballData,read.csv(mypath, # combine dataframes into basellbadat
  col.names=colNames$Name)) # retrieve columnname
  baseballData <- tbl_df(baseballData) #create a dataframe babesall data
}
# baseballData
```

Data files can be found on Canvas in the lab folder. Download the files and load them into one data.frame in R as shown below. Comment this code to demonstrate you understand how it works.

Select the following relevant columns and create a new data frame to store the data you will use for your analysis.

- Date
- Home
- Visitor
- HomeLeague
- VisitorLeague
- HomeScore
- VisitorScore

```
# Load some helpful libraries
baseballAnalysis <- baseballData %>% select(Date,
  Home, HomeLeague, VisitorLeague, HomeScore,
  VisitorScore)
```

```
# Load some helpful libraries get teams from
# NL
nlTeams <- baseballData %>% filter(VisitorLeague ==
  "NL" & HomeLeague == "NL")
```

Considering only games between two teams in the National League, compute the conditional probability of the team winning given X runs scored, for $X = 0, \dots, 10$. Do this separately for Home and Visitor teams.

- Design a visualization that shows your results.
- Discuss what you find.

Extra Credit: Repeat the above problem, but now consider the probability of winning given the number of hits.