606 Project Proposal

Ahsanul Choudhury

October 16, 2016

```
# DO NOT REMOVE
# THIS IS FOR SETTING SOME PLOTTING PARAMETERS SO THAT YOUR PLOTS DON'T TAKE UP TOO MUCH SPACE
# IF YOU WOULD LIKE TO CHANGE THESE, SEE HELP FILES ON THE par() FUNCTION
# OR ASK FOR HELP
library(knitr)
## set global chunk options
opts_chunk$set(fig.path='figure/manual-', cache.path='cache/manual-', fig.align='center', fig.show='hol
## tune details of base graphics (http://yihui.name/knitr/hooks)
knit_hooks$set(par=function(before, options, envir){
if (before && options$fig.show!='none') par(mar=c(4,4,.2,.1),cex.lab=.95,cex.axis=.9,mgp=c(2,.7,0),tcl=
})
```

```
# load data
data <- read.csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/bad-drivers/bad-drivers
head(data)
```

```
##        State
## 1    Alabama
## 2     Alaska
## 3    Arizona
## 4   Arkansas
## 5 California
## 6   Colorado
##   Number.of.drivers.involved.in.fatal.collisions.per.billion.miles
## 1                                                             18.8
## 2                                                             18.1
## 3                                                             18.6
## 4                                                             22.4
## 5                                                             12.0
## 6                                                             13.6
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Speeding
## 1                                                                  39
## 2                                                                  41
## 3                                                                  35
## 4                                                                  18
## 5                                                                  35
## 6                                                                  37
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Alcohol.Impaired
## 1                                                                          30
## 2                                                                          25
## 3                                                                          28
## 4                                                                          26
## 5                                                                          28
## 6                                                                          28
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Not.Distracted
## 1                                                                         96
## 2                                                                         90
```

```
## 3                                                                    84
## 4                                                                    94
## 5                                                                    91
## 6                                                                    79
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Had.Not.Been.Involved.In.Any.Previous.Accid
## 1
## 2
## 3
## 4
## 5
## 6
##   Car.Insurance.Premiums....
## 1                     784.55
## 2                    1053.48
## 3                     899.47
## 4                     827.34
## 5                     878.41
## 6                     835.50
##   Losses.incurred.by.insurance.companies.for.collisions.per.insured.driver....
## 1                                                                       145.08
## 2                                                                       133.93
## 3                                                                       110.35
## 4                                                                       142.39
## 5                                                                       165.63
## 6                                                                       139.91
```

```
dim(data)
```

```
## [1] 51  8
```

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for. ##I want to manipulate the data to find out, ##-The probability of a randomly chosen driver involved in a fatal accident is from a particular state" ##-The probability of a randomly chosen driver involved in a fatal accident had been involved in previous accidents by states ##-Likely proportion of fatal accidents by each state from a randomly chosen sample of 100 fatal accidents with a 95% confidence interval, plot 30 confidence intervals for the highest and lowest ranking states based on the sample size of 100

Cases

What are the cases, and how many are there? ##There are 51 cases, 1 for each of 50 states plus 1 for District of Columbia with 8 variables

Data collection

Describe the method of data collection. ##The data was collected from following sources: -Number of drivers involved in fatal collisions per billion miles(National Highway Traffic Safety Administration, 2012) -Percentage Of Drivers Involved In Fatal Collisions Who Were Speeding (National Highway Traffic Safety Administration, 2009) -Percentage Of Drivers Involved In Fatal Collisions Who Were Alcohol-Impaired (National Highway Traffic Safety Administration, 2012) -Percentage Of Drivers Involved In Fatal Collisions Who Were

Not Distracted (National Highway Traffic Safety Administration, 2012) -Percentage Of Drivers Involved In Fatal Collisions Who Had Not Been Involved In Any Previous Accidents(National Highway Traffic Safety Administration, 2012) -Car Insurance Premiums ()($National Association of Insurance Commissioners$, 2011) – $Losses incurred by insurance companies for collisions per insured driver$()(National Association of Insurance Commissioners, 2010)

Type of study

What type of study is this (observational/experiment)? ##This study is observational

Data Source

If you collected the data, state self-collected. If not, provide a citation/link. ##URL:https://github.com/fivethirtyeight/data/blob/master/bad-drivers/bad-drivers.csv

Response

What is the response variable, and what type is it (numerical/categorical)? #Numarical

Explanatory

What is the explanatory variable, and what type is it (numerical/categorival)?

The response variable is "Number of drivers involved in fatal collisions per billion miles" and it is a numerical variable.

Relevant summary statistics

Provide summary statistics relevant to your research question. For example, if you're comparing means across groups provide means, SDs, sample sizes of each group. This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
##Retrive names of columns
names(data)
```

```
## [1] "State"
## [2] "Number.of.drivers.involved.in.fatal.collisions.per.billion.miles"
## [3] "Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Speeding"
## [4] "Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Alcohol.Impaired"
## [5] "Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Not.Distracted"
## [6] "Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Had.Not.Been.Involved.In.Any.Previous.Ac
## [7] "Car.Insurance.Premiums...."
## [8] "Losses.incurred.by.insurance.companies.for.collisions.per.insured.driver...."
```

```
##Select only required columns
sumdata <- data.frame(data$State, data$Number.of.drivers.involved.in.fatal.collisions.per.billion.miles
##Rename required columns
names(sumdata) <- c("state", "fatal_accident", "percent_no_previous")
head(sumdata)
```

```
##        state fatal_accident percent_no_previous
## 1    Alabama           18.8                  80
## 2     Alaska           18.1                  94
## 3    Arizona           18.6                  96
## 4   Arkansas           22.4                  95
## 5 California           12.0                  89
## 6   Colorado           13.6                  95
```

```
##Add new column
sumdata1 <-mutate(sumdata, previous_accident = round(fatal_accident*(percent_no_previous/100),1))
head(sumdata1)
```

```
##        state fatal_accident percent_no_previous previous_accident
## 1    Alabama           18.8                  80              15.0
## 2     Alaska           18.1                  94              17.0
## 3    Arizona           18.6                  96              17.9
## 4   Arkansas           22.4                  95              21.3
## 5 California           12.0                  89              10.7
## 6   Colorado           13.6                  95              12.9
```

```
##Delete unnecessary column
sumdata2 <- select(sumdata1, state, fatal_accident, previous_accident)
head(sumdata2)
```

```
##        state fatal_accident previous_accident
## 1    Alabama           18.8              15.0
## 2     Alaska           18.1              17.0
## 3    Arizona           18.6              17.9
## 4   Arkansas           22.4              21.3
## 5 California           12.0              10.7
## 6   Colorado           13.6              12.9
```

```
##Standard deviation of Number of drivers involved in fatal collisions per billion miles
sd(sumdata2$fatal_accident)
```

```
## [1] 4.122002
```

```
##Standard deviation of Number of drivers who had previous accidents involved in fatal collisions per bi
sd(sumdata2$previous_accident)
```

```
## [1] 3.768068
```

```
hist(sumdata2$fatal_accident)
```