BIOST 2041 Intro to Statistical Methods

R Learning Material 1– Getting Started

TA Crystal Zang

8/23/2020

Contents

I. Basics

Here we are creating a variable and storing values in the variable. We say score A is a vector that contains 3 elements, and it's a numeric variable. The first element of score A is 85, the second element is 89, and the third element is 92. We say the data type of scoreA and scoreB are numeric. Next, we create a student variable, where the student names, Alex, Jordan, and Ryan, are stored. The data type of student vector is character.

```
#numeric
scoreA <- c()
scoreA <- c(85, 89, 92) #assign values to a variable

scoreA #print the variable
```

```
## [1] 85 89 92
```

```
class(scoreA) #check the data type of the variable
```

```
## [1] "numeric"
```

```r
scoreB <- c(75, 80, 95)
scoreB
```

```
## [1] 75 80 95
```

```r
class(scoreB)
```

```
## [1] "numeric"
```

```r
scoreA[1] < scoreB[1]
```

```
## [1] FALSE
```

```r
1 > 2
```

```
## [1] FALSE
```

```r
scoreB > 80
```

```
## [1] FALSE FALSE  TRUE
```

```r
pass <- scoreB > 80
class(pass)
```

```
## [1] "logical"
```

```r
#character
student <- c("Alex", "Jordan", "Ryan")
class(student)
```

```
## [1] "character"
```

II. Numeric Operations

```r
1 + 1

2 * 9

9/2

9^2

log(9)
```

Now we know the basic numeric operations, we can perform these operations on the vectors we previously created.

```r
scoreA + scoreB

#calculate the sum of the elements in scoreA
sum(scoreA)

#find the number of elements in the vector scoreA
length(scoreA)

min(scoreA) #minimum

max(scoreA) #maximum

#calculate the mean of scoreA
mean(scoreA)
sum(scoreA)/length(scoreA)

#calculate the median of scoreA
median(scoreA)

#calculate standard deviation of scoreA
sd(scoreA)

#inter-quatile range of scoreA
IQR(scoreA)

#summary statistics of scoreA
summary(scoreA)
```

Question 1.

What's the difference in the mean score A and mean score B

III. R package Installation

fivethirtyeight R package hosts many datasets. You can find detailed information on these datasets here: https://fivethirtyeight-r.netlify.app/articles/fivethirtyeight.html.

We need to install an R package using the commend **install.packages()**. You only install package once. But everytime you open you R Markdown file, you have to read in the package using the commend **library()**. To find documentation of any R package use **?** with the name of the R package. For example **?fivethirtyeight**.

Packages you need to install: **fivethirtyeight**(dataset), **ggplot2**(visualization tool).

```r
#install from a published R package
#install.packages("fivethirtyeight")

#install from the repository to get the most recent updates
#install.packages('fivethirtyeightdata', repos ='https://fivethirtyeightdata.github.io/drat/',type = 's

library(fivethirtyeight)
data(bad_drivers)
```

```r
#gives you the top rows of the dataset
head(bad_drivers)
```

```
##        state num_drivers perc_speeding perc_alcohol perc_not_distracted
## 1    Alabama        18.8            39           30                  96
## 2     Alaska        18.1            41           25                  90
## 3    Arizona        18.6            35           28                  84
## 4   Arkansas        22.4            18           26                  94
## 5 California        12.0            35           28                  91
## 6   Colorado        13.6            37           28                  79
##   perc_no_previous insurance_premiums losses
## 1               80             784.55 145.08
## 2               94            1053.48 133.93
## 3               96             899.47 110.35
## 4               95             827.34 142.39
## 5               89             878.41 165.63
## 6               95             835.50 139.91
```

```r
#data documentation including variable explainations
?bad_drivers

#View drug use dataset in a seperate tab
View(bad_drivers)
```

IV. Dataframe

We will use **bad_drive** dataframe as an example. The raw data behind the story "Dear Mona, Which State Has The Worst Drivers?" https://fivethirtyeight.com/features/which-state-has-the-worst-drivers/

First, we want to know the dimension of the dataset. There are 51 U.S. states and 8 variables in the dataset. This dataset is in a "tidied" format, meaning that each row is an observation, preferable unique, and each column is variable. In this dataset, have have each row indicating a state.

```r
#find the number of row and column
dim(bad_drivers)
```

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

We are interested in the percentage of drivers involved in fatal collisions who were speeding using the variable **perc_speeding** in the dataset. To extract a variable from a dataframe, we use **$**. An extracted variable is in a vector, similar to the scoreA variable that we showed before.

```r
bad_drivers$perc_speeding #print the variable
```

```
##  [1] 39 41 35 18 35 37 46 38 34 21 19 54 36 36 25 17 27 19 35 38 34 23 24 23 15
## [26] 43 39 13 37 35 16 19 32 39 23 28 32 33 50 34 38 31 21 40 43 30 19 42 34 36
## [51] 42
```

```r
length(bad_drivers$perc_speeding) #find the length of the variable
```

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

```r
bad_drivers[1, ] #extract one row
```

```
##     state num_drivers perc_speeding perc_alcohol perc_not_distracted
## 1 Alabama        18.8            39           30                  96
##   perc_no_previous insurance_premiums losses
## 1               80             784.55 145.08
```

Question 2.

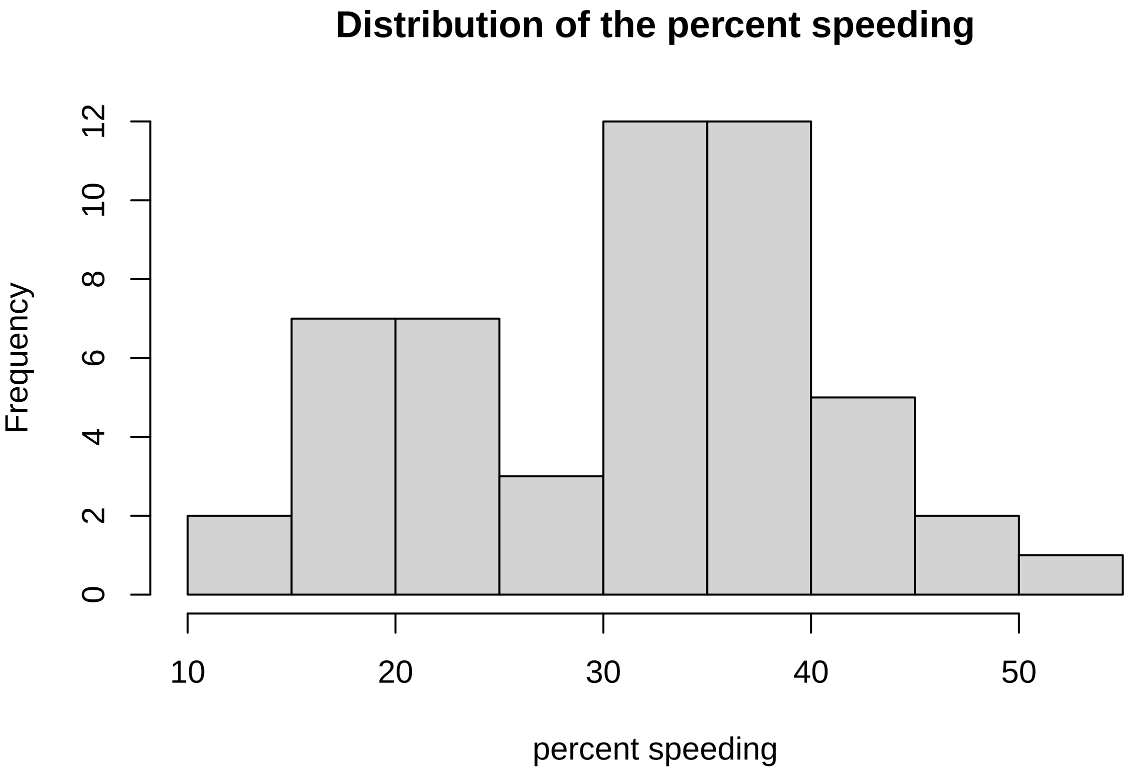
Report the summary statistics (mean, standard deviation, IQR) of the percent speeding.

Visualization

Here we are visualizing the distribution of the percentage of drivers involved in fatal collisions who were speeding among all 51 states.

1. Visualization using Base R

```r
#using base R to plot the histogram
hist(bad_drivers$perc_speeding,
     main = "Distribution of the percent speeding",  #title
     xlab = "percent speeding")  #title of x-axis
```

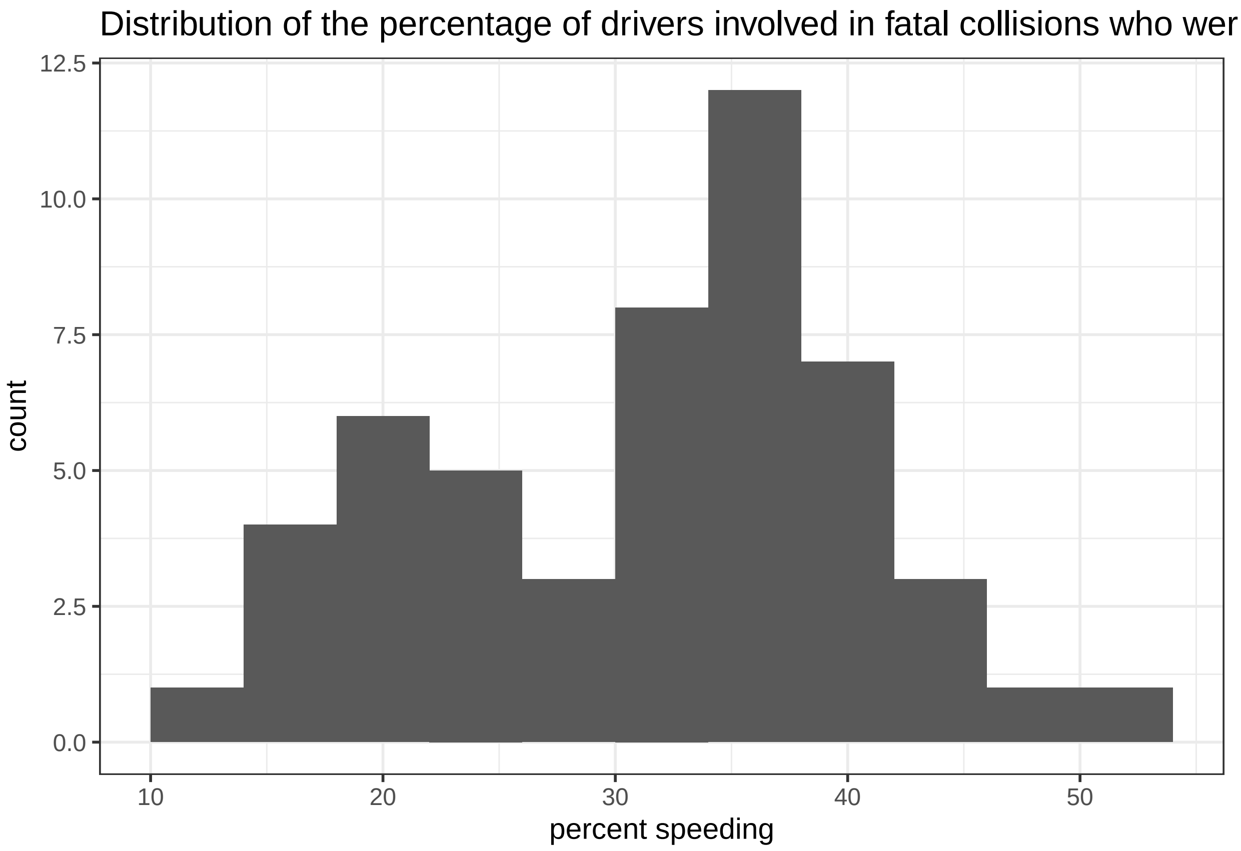


2. Visualization using ggplot2

```
#useing ggplot2 package to plot the histogram
#install.packages("ggplot2")

library(ggplot2)

ggplot(bad_drivers, aes(x = perc_speeding))+
  geom_histogram(binwidth = 4) +  #can manually change binwidth
  labs(title = "Distribution of the percentage of drivers involved in fatal collisions who were speedin
       x = "percent speeding") +
  theme(title =element_text(size=9)) +#size of the title
  theme_bw() #make the plot looks pretty
```



Question 3.

Explore a different variable of your interest in the “bad_drivers” dataset. Calculate the summary statistics and plot the distribution of the variable