**Statistical Methods for Data Science (Fall 2018)**

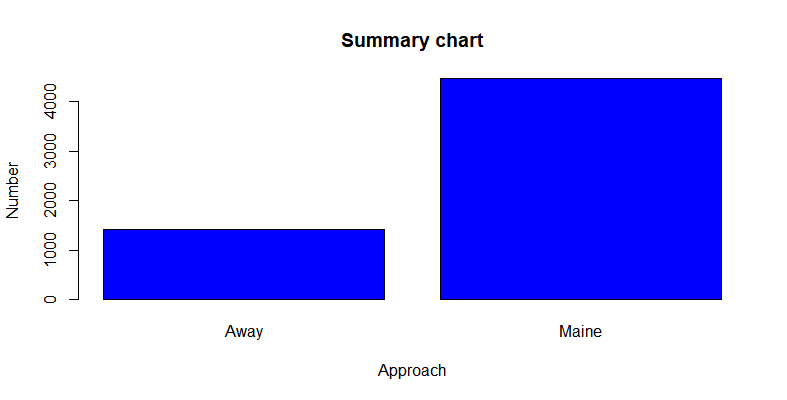
**Mini Project 2 (Solution)**

# Haoran Lou

# hxl180012

1.

(a).



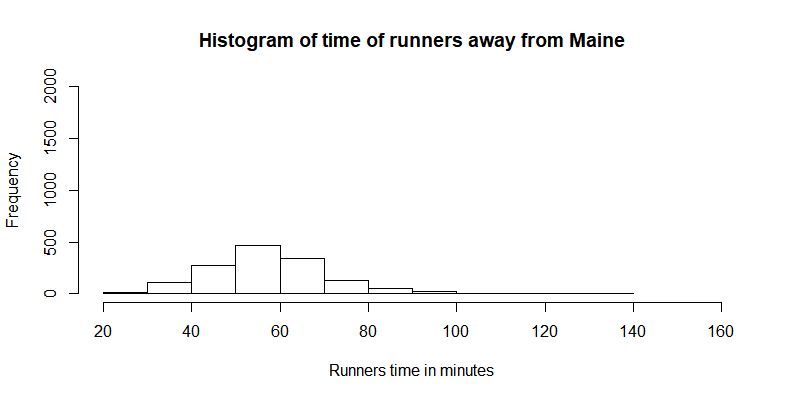
The created bar graph bar graph of the variable Maine is above. We can see that the number of runners from Maine is about 3 times of runners from Away.

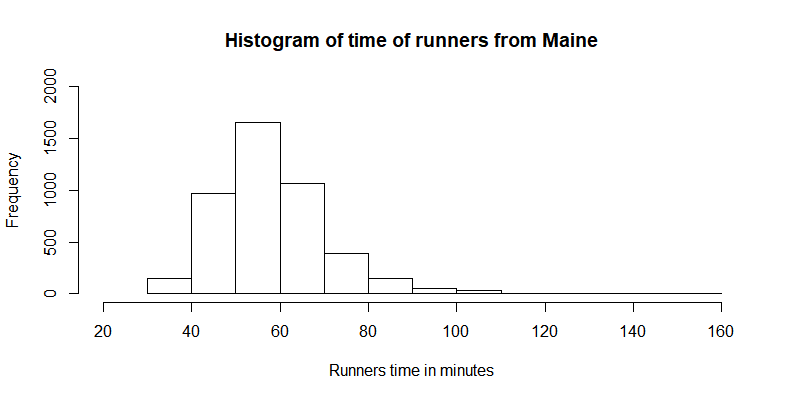
Summary:



So, the accurate numbers are about 1:3.

(b)





|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | mean | standard deviation | median | range | IQR |
| Away | 57.82181 | 13.83538 | 56.92 | 27.782-133.710 | 15.674 |
| Maine | 58.19514 | 12.18511 | 57.0335 | 30.567-152.167 | 14.24775 |

The mean of two grams are almost the same means that the runners of two different approach spend almost same average time, which is to say, the approach has less effect on time runner spend.

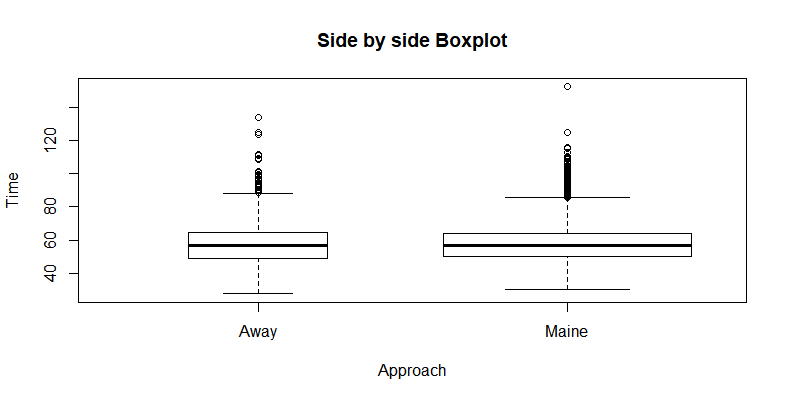
Away groups’ standard deviation is bigger than another group means that the gap of time usage of runners among Away group is little bigger that group “Maine”.

Medians of the two groups are about same.

The range of “Maine” group is bigger, means that the gap between maximum and minimum of Maine is larger.

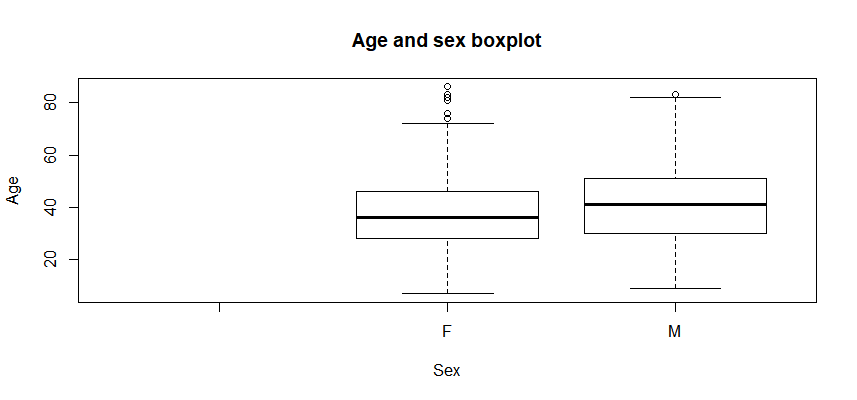
IOR of Maine is smaller than IQR of away means that the variability of time usage of “Maine” in the middle 50% is bigger that another group.

(c)



The conclusions are consisting with the conclusions we made from the histogram.

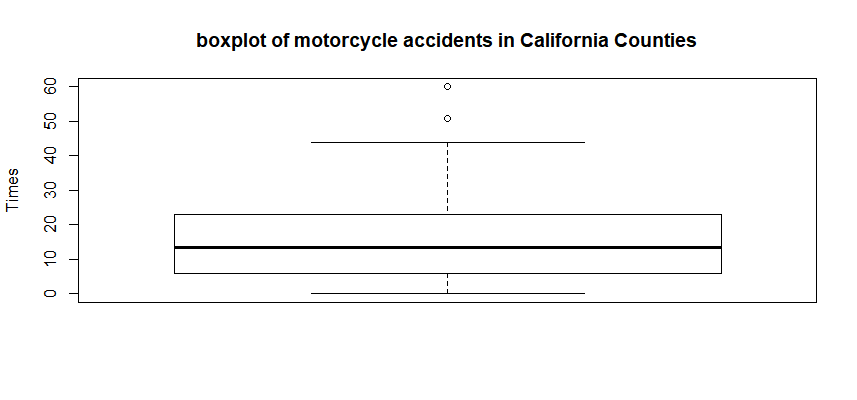
(d)



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | mean | standard deviation | median | range | IQR |
| Male Age | 40.4468 | 13.99289 | 41 | 9-83 | 21 |
| Female Age | 37.23653 | 12.26925 | 36 | 7-86 | 18 |

We can see that the average age of male runners is older than female runners from the mean of the dataset. The median of male is also bigger than female. The standard deviation of male age is bigger than female age, the gap and variability is bigger in male age. The range of female age is bigger means that there are females of bigger gap of ages participate in the competition. IQR of male is smaller than male means that the age of female is more concentrate at the Q3-Q1 area than male.

2.



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | mean | standard deviation | median | range | IQR |
| counts | 17.02083 | 13.81256 | 13.5 | 0-60 | 17 |

We know the data distribution is right skewed from the boxplot, and the numbers of motorcycle fatalities of counties are most near 13.5+-8.5 times. GREENVILLE and HORRY may be considered as outliers. There many reasons may contribute to this:

1. It is too easy to qualify for a driver’s license in these two places.
2. The number of Highway Patrol officers may too little at a rate comparable to growing traffic volume.
3. Maybe there are some problems with the road.
4. Failure to yield the right of way, driving left of the center line of some driver of less consciousness.
5. ##########################
6. # R code for Miniproject2#
7. ##########################
9. #1.
10. #a. Create a bar graph of the variable Maine.
11. #import data from roadrace.csv
12. data=read.csv("C:/Users/lou/Desktop/R/R exercises/roadrace.csv",na.strings ="\*" )
14. #get data of numbers of Away and Maine group and transfer it to length.
15. datax=c(length(which(data**$Maine**=="Away")),length(which(data**$Maine**=="Maine")))
16. labels=c("Away","Maine")
18. #make a barplot
19. barplot(datax,names.arg=labels,xlab="Approach",ylab="Number",col="blue",main="Summary chart",border="black")

22. #b.Create two histograms the runners' times (given in minutes), one for the Maine group and
23. # the minitues histgram for the Away group.
24. Away\_Time=c(data**$Time**..minutes.[which(data**$Maine**=="Away")])
25. hist(Away\_Time,
26. xlab = "Runners time in minutes",
27. main = "Histogram of time of runners away from Maine",
28. xlim = c(20,160),
29. ylim = c(0,2000))
31. # compute statistic values of Away\_Time
32. mean(Away\_Time)
33. sd(Away\_Time)
34. median(Away\_Time)
35. range(Away\_Time)
36. IQR(Away\_Time)


40. #the minitues histgram for the maine group.
41. Maine\_Time=c(data**$Time**..minutes.[which(data**$Maine**=="Maine")])
42. hist(Maine\_Time,
43. xlab = "Runners time in minutes",
44. main = "Histogram of time of runners from Maine",
45. xlim = c(20,160),
46. ylim = c(0,2000))
48. # compute statistic values of Maine\_Time
49. mean(Maine\_Time)
50. sd(Maine\_Time)
51. median(Maine\_Time)
52. range(Maine\_Time)
53. IQR(Maine\_Time)
55. #c.Create a side-by-side boxplot the runners' times (given in minutes), one for the Maine group and
56. #the second for the Away group.
57. boxplot(Time..minutes. ~ Maine,
58. data=data,
59. varwidth=TRUE,
60. xlab="Approach",
61. ylab="Time",
62. main="Side by side Boxplot")
64. #d. Create a box plot of Ages of female and male
65. boxplot(Age ~ Sex,
66. data=data,
67. varwidth=TRUE,
68. xlab="Sex",
69. ylab="Age",
70. main="Age and sex boxplot")
71. #Summary statisic of Sex and Age
72. #Male age
73. Male\_Age=c(data**$Age**[which(data**$Sex**=="M")])
74. mean(Male\_Age)
75. sd(Male\_Age)
76. median(Male\_Age)
77. range(Male\_Age)
78. IQR(Male\_Age)
80. #Female age
81. Female\_Age=c(data**$Age**[which(data**$Sex**=="F")])
82. mean(Female\_Age)
83. sd(Female\_Age)
84. median(Female\_Age)
85. range(Female\_Age)
86. IQR(Female\_Age)

89. #2.
90. #Create a boxplot of Motorcycle accident of counties in California.
92. #import data from motorcycle.csv
93. data1 = read.csv("C:/Users/lou/Desktop/R/R exercises/motorcycle.csv")
94. AccidentCount = data1**$Fatal**.Motorcycle.Accidents
96. #create a barplot of accident counts.
97. boxplot(AccidentCount,
98. varwidth=TRUE,
99. main="boxplot of motorcycle accidents in California Counties",
100. ylab="Times")
102. #summary statistics
104. mean(AccidentCount)
105. sd(AccidentCount)
106. median(AccidentCount)
107. range(AccidentCount)
108. IQR(AccidentCount)