anova_gender.R

poojasingh

Wed Nov 25 10:04:12 2015

```
#Read Marathon data dile
#fname = file.choose()
fname = "/Users/poojasingh/Documents/HStatE139/git/statse139-project/Previous Boston Marathon study/BAA
dfm = read.csv(fname, header=T,sep=" ")
names(dfm)
```

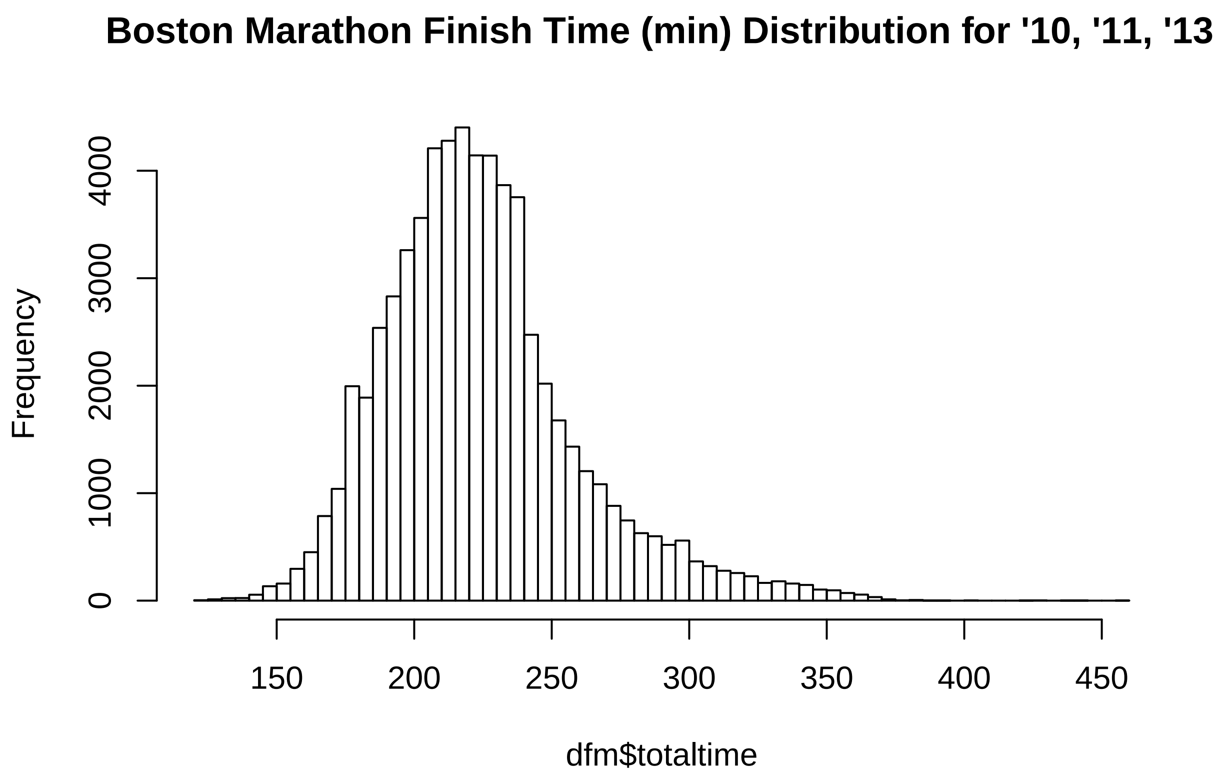
```
##  [1] "BibNum"     "Year"       "Age"        "Gender1F2M" "StartHr"
##  [6] "StartMin"   "K0.5"       "K5.10"      "K10.15"     "K15.20"
## [11] "K20.25"     "K25.30"     "K30.35"     "K35.40"     "K40.Fin"
## [16] "HalfMar"    "X"          "Age2014"
```

```
n= length(dfm)
attach(dfm)
#Add all times
times = as.matrix(dfm[,7:15], ncol=9)
dfm$totaltime = rowSums(times)

#Distinct Years
unique(Year)
```

```
## [1] 2013 2011 2010
```

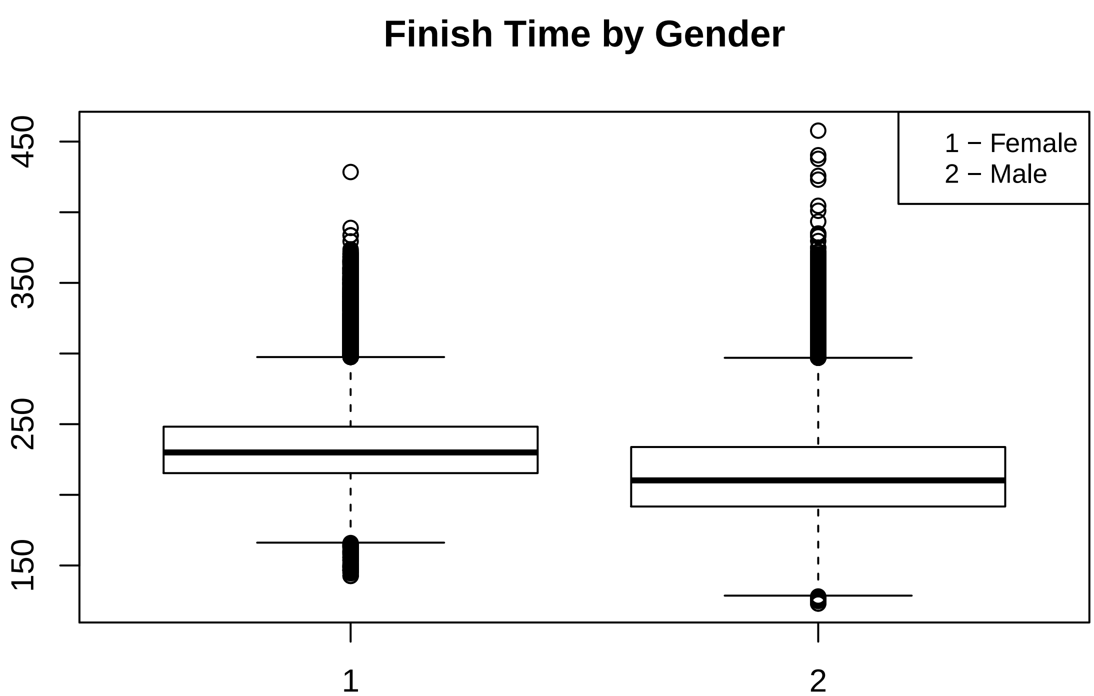
```
#Plot the data first
hist(dfm$totaltime,breaks=50, main="Boston Marathon Finish Time (min) Distribution for '10, '11, '13")
```



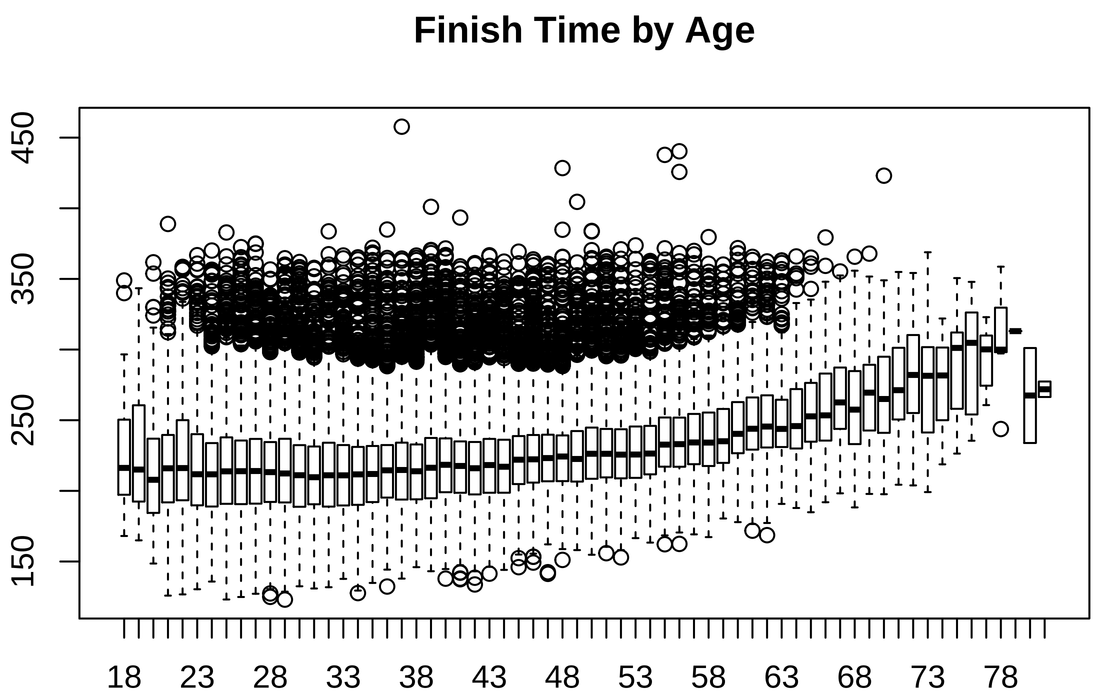

```r
length(dfm$totaltime) #sanity check
```

```
## [1] 69923
```

```r
#ANOVA - boxplot by Gender
boxplot(dfm$totaltime~dfm$Gender1F2M, data=dfm, main="Finish Time by Gender")
legend("topright",c("1 - Female","2 - Male"), cex = 0.8)
```



```r
#ANOVA - boxplot by Age
boxplot(dfm$totaltime~dfm$Age, data=dfm, main="Finish Time by Age")
```




```r
#F-test to copmare variances by Gender
var.test(dfm$totaltime~dfm$Gender1F2M, data=dfm)
```

```
##
##  F test to compare two variances
##
## data:  dfm$totaltime by dfm$Gender1F2M
## F = 0.7713, num df = 26567, denom df = 37598, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7544116 0.7886875
## sample estimates:
## ratio of variances
##          0.7713426
```

```r
#T-test by Gender
t.test(dfm$totaltime[Gender1F2M==1], dfm$totaltime[Gender1F2M==2])
```

```
##
##  Welch Two Sample t-test
##
## data:  dfm$totaltime[Gender1F2M == 1] and dfm$totaltime[Gender1F2M == 2]
## t = 71.1751, df = 61229.14, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  18.50427 19.55226
## sample estimates:
```

```
## mean of x mean of y
##  235.4841  216.4559
```

```r
#Two-way ANOVA by Gender and Age
model.gender.age = aov(dfm$totaltime ~ dfm$Gender1F2M*dfm$Age)
summary(model.gender.age)
```

```
##                            Df   Sum Sq Mean Sq F value Pr(>F)
## dfm$Gender1F2M              1  5636658 5636658  5370.7 <2e-16 ***
## dfm$Age                     1  7001273 7001273  6670.9 <2e-16 ***
## dfm$Gender1F2M:dfm$Age      1   293562  293562   279.7 <2e-16 ***
## Residuals               64163 67340227    1050
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5756 observations deleted due to missingness
```

```r
#Distinct Age
age = as.vector(unique(Age))
sort(age)
```

```
##  [1] 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
## [24] 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63
## [47] 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83
```

```r
#Split data into agegroup
dfm.age.18.25 = subset(dfm, (dfm$Age>=18 & dfm$Age>=25))
dfm.age.26.35 = subset(dfm, (dfm$Age>=26 & dfm$Age>=35))
dfm.age.36.45 = subset(dfm, (dfm$Age>=36 & dfm$Age>=45))
dfm.age.46.55 = subset(dfm, (dfm$Age>=46 & dfm$Age>=55))
dfm.age.56.65 = subset(dfm, (dfm$Age>=56 & dfm$Age>=65))
dfm.age.66.75 = subset(dfm, (dfm$Age>=66 & dfm$Age>=75))
dfm.age.76.83 = subset(dfm, (dfm$Age>=76 & dfm$Age>=83)) #TotalTime is NA

#(Todo: Paiwise) T-test by Agegroup
t.test(dfm.age.18.25$totaltime, dfm.age.26.35$totaltime)
```

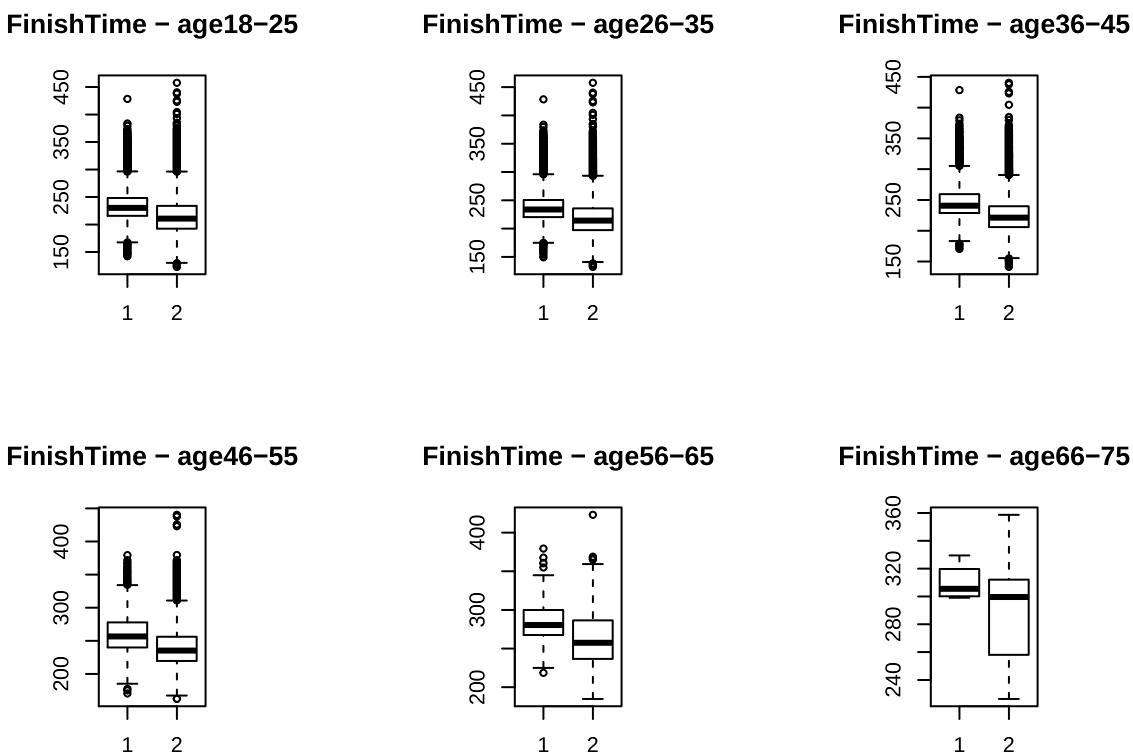
```
##
##  Welch Two Sample t-test
##
## data:  dfm.age.18.25$totaltime and dfm.age.26.35$totaltime
## t = -11.8697, df = 101496.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.919392 -2.091899
## sample estimates:
## mean of x mean of y
##  224.5623  227.0679
```

```r
t.test(dfm.age.18.25$totaltime[dfm.age.18.25$Gender1F2M==1], dfm.age.26.35$totaltime[dfm.age.18.25$Gend
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  dfm.age.18.25$totaltime[dfm.age.18.25$Gender1F2M == 1] and dfm.age.26.35$totaltime[dfm.age.18
## t = 16.8833, df = 36226.72, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4.989533 6.300187
## sample estimates:
## mean of x mean of y
##  235.6643  230.0195
```

```
#Boxplots FinishTime by Gender per ager group
par(mfrow = c(2, 3), mar=c(5.1, 4.1, 4.1, 8.1))
#par(xpd=FALSE) # this is usually the default
boxplot(dfm.age.18.25$totaltime~dfm.age.18.25$Gender1F2M, data=dfm, main="FinishTime - age18-25")
boxplot(dfm.age.26.35$totaltime~dfm.age.26.35$Gender1F2M, data=dfm, main="FinishTime - age26-35")
boxplot(dfm.age.36.45$totaltime~dfm.age.36.45$Gender1F2M, data=dfm, main="FinishTime - age36-45")
boxplot(dfm.age.46.55$totaltime~dfm.age.46.55$Gender1F2M, data=dfm, main="FinishTime - age46-55")
boxplot(dfm.age.56.65$totaltime~dfm.age.56.65$Gender1F2M, data=dfm, main="FinishTime - age56-65")
boxplot(dfm.age.66.75$totaltime~dfm.age.66.75$Gender1F2M, data=dfm, main="FinishTime - age66-75")
```




```
#boxplot(dfm.age.76.83$totaltime~dfm.age.76.83$Gender1F2M, data=dfm, main="FinishTime - age76-83")  #To
#legend("topright",c("1 - Female","2 - Male"), cex = 0.8)

#ANOVA by Age
model.age = aov(dfm$totaltime ~ dfm$Age)
summary(model.age)
```

```
##                  Df    Sum Sq Mean Sq F value Pr(>F)
```

```
## dfm$Age          1  4156398 4156398    3504 <2e-16 ***
## Residuals   64165 76115322    1186
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5756 observations deleted due to missingness
```

```
#ANOVA by AgeGroup
model.age.18.25 = aov(dfm.age.18.25$totaltime ~ dfm.age.18.25$Age)
summary(model.age.18.25)
```

```
##                        Df   Sum Sq Mean Sq F value Pr(>F)
## dfm.age.18.25$Age       1  4479943 4479943    3885 <2e-16 ***
## Residuals           61121 70483674    1153
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5454 observations deleted due to missingness
```

```
#Two ANOVA by Gender and AgeGroup
model.gender.age.18.25 = aov(dfm.age.18.25$totaltime ~ dfm.age.18.25$Gender1F2M * dfm.age.18.25$Age)
summary(model.gender.age.18.25)
```

```
##                                               Df   Sum Sq Mean Sq F value
## dfm.age.18.25$Gender1F2M                       1  5161563 5161563  5059.5
## dfm.age.18.25$Age                              1  7263938 7263938  7120.3
## dfm.age.18.25$Gender1F2M:dfm.age.18.25$Age     1   186009  186009   182.3
## Residuals                                  61119 62352107    1020
##                                            Pr(>F)
## dfm.age.18.25$Gender1F2M                   <2e-16 ***
## dfm.age.18.25$Age                          <2e-16 ***
## dfm.age.18.25$Gender1F2M:dfm.age.18.25$Age <2e-16 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5454 observations deleted due to missingness
```

```
#Plot by year
# par(mfrow = c(1, 3))
# hist(dfm$totaltime, subset=(Year==2010), breaks=50, main="Boston Marathon Finish Time (min) Distribut
# hist(dfm$totaltime, subset=(Year==2011), breaks=50, main="Boston Marathon Finish Time (min) Distribut
# hist(dfm$totaltime, subset=(Year==2013), breaks=50, main="Boston Marathon Finish Time (min) Distribut

#Split data by Year
# dfm.2010 = subset(dfm, Year==2010)
# dfm.2011 = subset(dfm, Year==2011)
# dfm.2013 = subset(dfm, Year==2013)
#
# n.2010 = length(dfm.2010$Gender1F2M)
# n.2011 = length(dfm.2011$Gender1F2M)
# n.2013 = length(dfm.2013$Gender1F2M)
# n.2010 + n.2011 + n.2013 #equals  69923 sanity check

# par(mfrow = c(1, 3))
```

```
# #For 2010
# boxplot(dfm.2010$totaltime~dfm.2010$Gender1F2M, data=dfm.2010, main="Finish time for 2010")
# legend("topright",c("1 - Female","2 - Male"), cex = 0.8)
#
# #For 2011
# boxplot(dfm.2011$totaltime~dfm.2011$Gender1F2M, data=dfm.2011, main="Finish time for 2011")
# legend("topright",c("1 - Female","2 - Male"), cex = 0.8)
#
# #For 2013
# boxplot(dfm.2013$totaltime~dfm.2013$Gender1F2M, data=dfm.2013, main="Finish time for 2013")
# legend("topright",c("1 - Female","2 - Male"), cex = 0.8)

# dfms = dfm[!is.na(dfm$totaltime), ]
# dfms = dfms[order(dfms$totaltime),]
# model = lm(totaltime~Age+Gender1F2M+K0.5, data=dfms)
# summary(model)
# model.resid = resid(model)
#
# plot(dfms$totaltime,model.resid,ylim=c(-100,100))

# #Summary statistics for FinishTime by Gender
# summarystats.finishtime.gender = cbind(by(dfm$totaltime/60, Gender1F2M, mean, na.rm=TRUE),
#                        by(dfm$totaltime/60, Gender1F2M, sd, na.rm=TRUE),
#                        by(dfm$totaltime/60, Gender1F2M, var, na.rm=TRUE)) #Divided by 60 to report in h
# colnames(summarystats.finishtime.gender)=c("mean", "sd", "var")
# rownames(summarystats.finishtime.gender)=c("female", "male")
# summarystats.finishtime.gender
#
# #Split data by Gender
# dfm.female = subset(dfm, dfm$Gender1F2M==1)
# dfm.male = subset(dfm, dfm$Gender1F2M==2)
# length(dfm.female$Gender1F2M) + length(dfm.male$Gender1F2M) #69923 sanity test
```