Analysis of FEV Data Set

Michael Williams

May 28, 2015

Overview

The accompanying R code gives a demonstration of data manipulation, graphics, and statistical procedures. The outputs contain data displays (histograms, boxplots, scatterplots) and statistical analysis (categorical data analysis, tTest for comparison of means, linear regression).

The data correspond to the following variables and descriptions.

- **id** for *Identification Number*
- **age** for *Age* (in years)
- **fev** for *Forced Expiratory Volume* (in liters)
- **height** for *Height* (in inches)
- **sex** for *Sex* ('male' or 'female')
- **smoke** for *Smoking Status* ('non-current smoker' or 'current smoker')

The data was obtained from B. Rosner's book *Introduction to Biostatistics*.

Exploratory Analysis of FEV

In this section, we explore the data by producing summary statistics and insightful graphics. For the graphics, the ggplot2 package will be used.

```
library(ggplot2)
```

Load the FEV data set via

```
load(url('http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/FEV.sav'))
```

The first six rows of FEV.

```
head(FEV)
```

```
##     id age   fev height    sex              smoke
## 1  301   9 1.708   57.0 female non-current smoker
## 2  451   8 1.724   67.5 female non-current smoker
## 3  501   7 1.720   54.5 female non-current smoker
## 4  642   9 1.558   53.0   male non-current smoker
## 5  901   9 1.895   57.0   male non-current smoker
## 6 1701   8 2.336   61.0 female non-current smoker
```

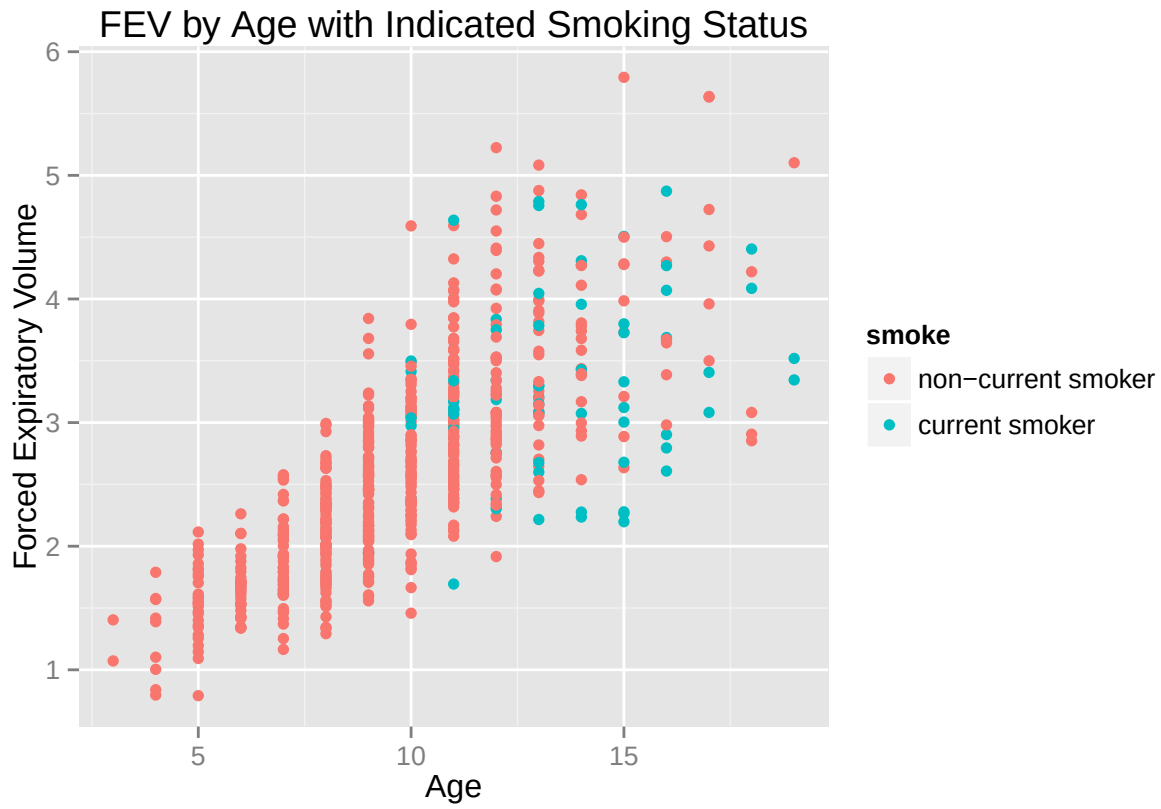
Summary statistics for all variables.

```
summary(FEV)
```

```
##        id              age              fev             height
##  Min.   :  201   Min.   : 3.000   Min.   :0.791   Min.   :46.00
##  1st Qu.:15811   1st Qu.: 8.000   1st Qu.:1.981   1st Qu.:57.00
##  Median :36071   Median :10.000   Median :2.547   Median :61.50
##  Mean   :37170   Mean   : 9.931   Mean   :2.637   Mean   :61.14
##  3rd Qu.:53638   3rd Qu.:12.000   3rd Qu.:3.119   3rd Qu.:65.50
##  Max.   :90001   Max.   :19.000   Max.   :5.793   Max.   :74.00
##      sex                      smoke
##  female:318   non-current smoker:589
##  male  :336   current smoker    : 65
##
##
##
##
```

Scatter plot of fev by age with smoke indicated.

```
plot1 <- ggplot(FEV, aes(x=age, y=fev, colour=smoke)) + geom_point()
plot1 <- plot1 + labs(x="Age", y="Forced Expiratory Volume",
                      title="FEV by Age with Indicated Smoking Status")
plot1
```



It will be useful to find out the age of the youngest male smokers.

```
summary(
        subset(
                FEV,
                subset = smoke=="current smoker" & sex=="male",
                select = age
               )
        )
```

```
##       age
##  Min.   : 9.00
##  1st Qu.:12.00
##  Median :14.00
##  Mean   :13.92
##  3rd Qu.:16.00
##  Max.   :18.00
```

We do the same for the female smokers.

```
summary(
        subset(
                FEV,
                subset = smoke=="current smoker" & sex=="female",
                select = age
               )
        )
```

```
##       age
##  Min.   :10.00
##  1st Qu.:11.50
##  Median :13.00
##  Mean   :13.26
##  3rd Qu.:15.00
##  Max.   :19.00
```

Create a contingency table for smoke by sex.

```
cats <- subset(FEV, select = c(sex,smoke))
tcats <- table(cats)
tcats
```

```
##         smoke
## sex      non-current smoker current smoker
##   female                279              39
##   male                  310              26
```

The code `prop.table(tcats, margin=1)` gives proporitons by sex.

```
prop.table(tcats, margin=1)
```

```
##         smoke
## sex      non-current smoker current smoker
##   female         0.87735849     0.12264151
##   male           0.92261905     0.07738095
```

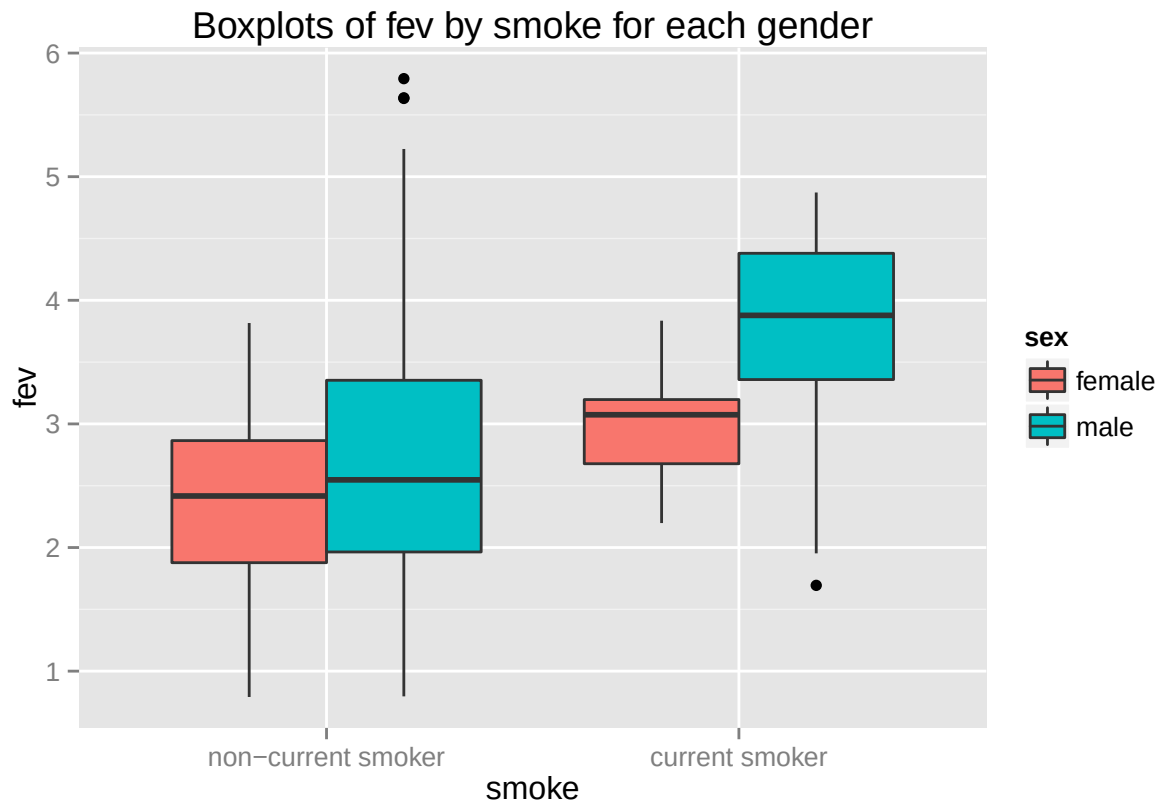
Use a chi-square test for proportions without Yates correction.

```
chisq.test(tcats, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tcats
## X-squared = 3.739, df = 1, p-value = 0.05316
```

Boxplot of fev by smoke and sex.

```
plot2 <- ggplot(FEV, aes(smoke, fev)) + geom_boxplot(aes(fill = sex))
plot2 <- plot2 + labs(title="Boxplots of fev by smoke for each gender")
plot2
```



Finally, we give a linear regression model of the response variable fev and predictor variables age, height, sex, and smoke.

```r
fit <- lm(fev ~ age + height + sex + smoke, data=FEV)
summary(fit)
```

```
##
## Call:
## lm(formula = fev ~ age + height + sex + smoke, data = FEV)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37656 -0.25033  0.00894  0.25588  1.92047
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -4.456974   0.222839 -20.001  < 2e-16 ***
## age                  0.065509   0.009489   6.904 1.21e-11 ***
## height               0.104199   0.004758  21.901  < 2e-16 ***
## sexmale              0.157103   0.033207   4.731 2.74e-06 ***
## smokecurrent smoker -0.087246   0.059254  -1.472    0.141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4122 on 649 degrees of freedom
## Multiple R-squared:  0.7754, Adjusted R-squared:  0.774
## F-statistic:   560 on 4 and 649 DF,  p-value: < 2.2e-16
```

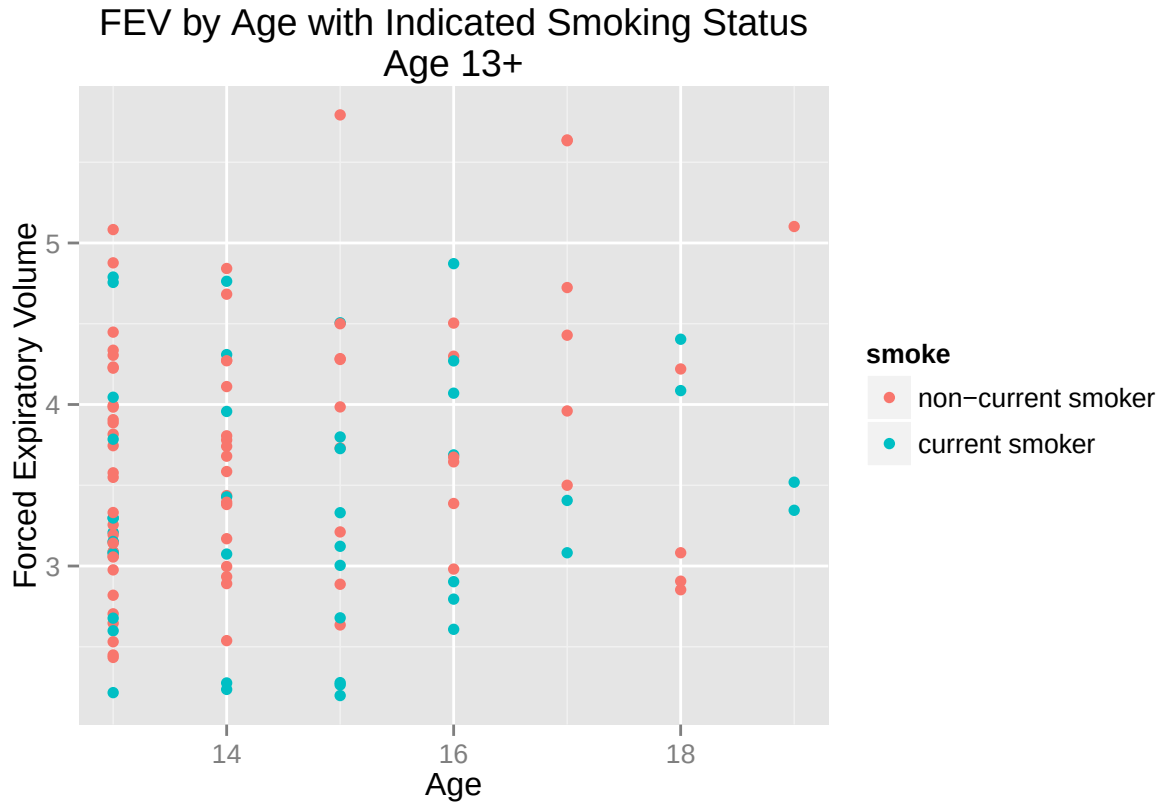
FEV Restricted to Teenagers

In this section, we restrict the FEV data set to teenagers, that is, those subjects age 13 to 19. This restricted data set is called FEV2.

```r
FEV2 <- subset(FEV, subset = age>= 13)
```

Scatterplot of fev by age with smoke coding.

```r
plot3 <- ggplot(FEV2, aes(x=age, y=fev, colour=smoke)) + geom_point()
plot3 <- plot3 + labs(x="Age", y="Forced Expiratory Volume",
                      title="FEV by Age with Indicated Smoking Status\nAge 13+")
plot3
```



A table for smoke status by sex.

```
cats2 <- subset(FEV2, select = c(sex,smoke))
tcats2 <- table(cats2)
tcats2
```

```
##         smoke
## sex      non-current smoker current smoker
##   female                 30             25
##   male                   44             18
```

The code `prop.table(tcats2, 1)` gives proporitons by sex.

```
prop.table(tcats2, 1)
```

```
##         smoke
## sex      non-current smoker current smoker
##   female          0.5454545      0.4545455
##   male            0.7096774      0.2903226
```

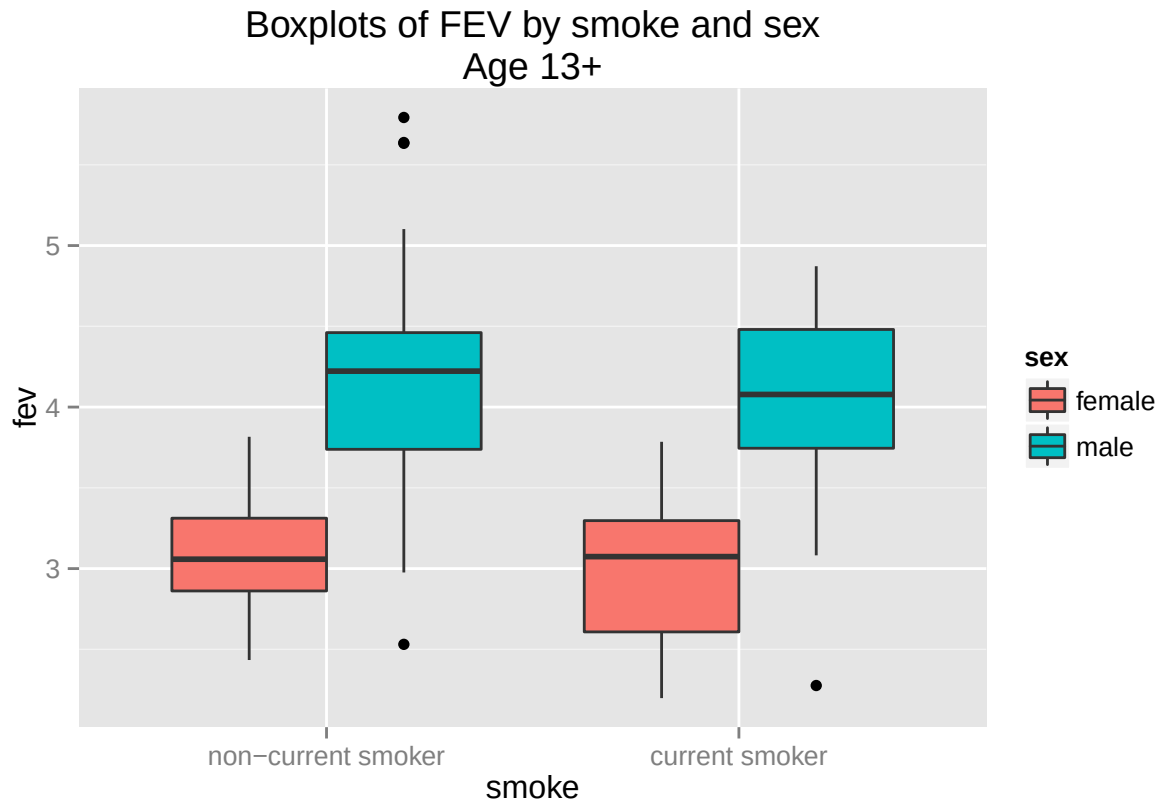
Use a chi-square test for proportions without Yates correction.

```
chisq.test(tcats2, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tcats2
## X-squared = 3.3815, df = 1, p-value = 0.06593
```

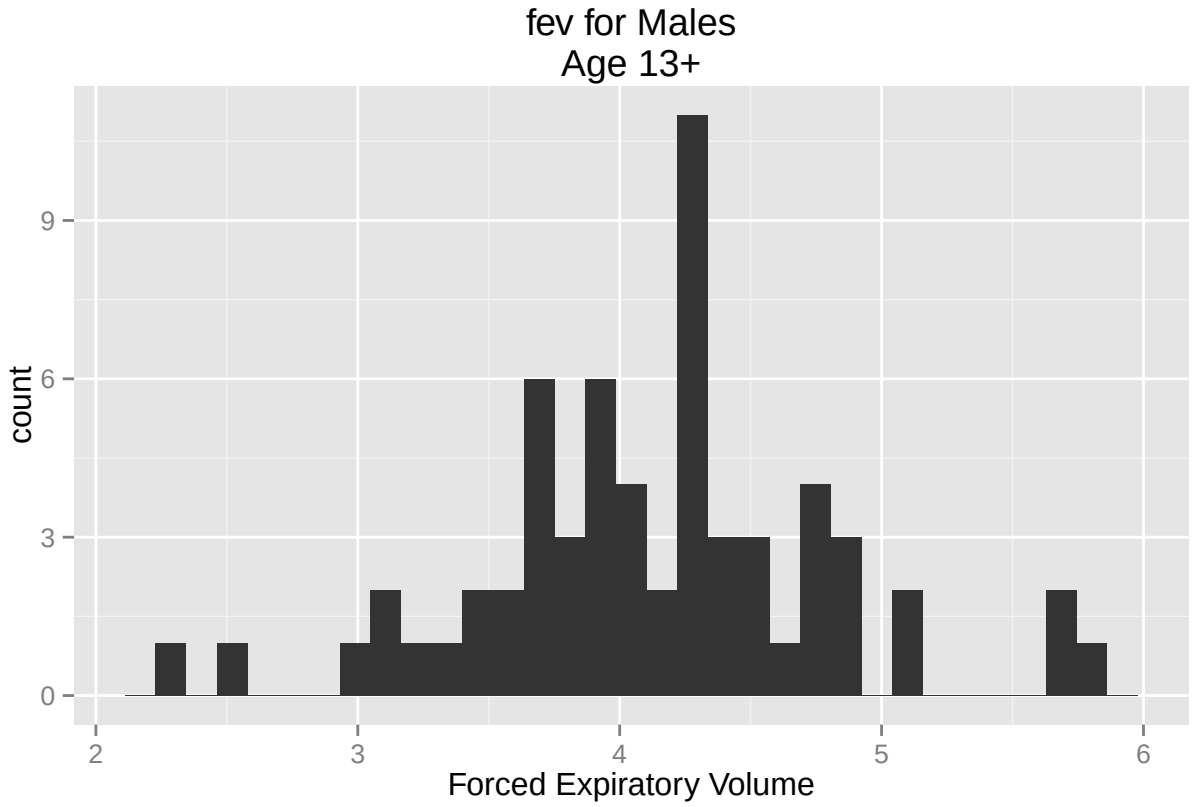
Boxplot of FEV by smoke and sex.

```
plot4 <- ggplot(FEV2, aes(smoke, fev)) + geom_boxplot(aes(fill = sex))
plot4 <- plot4 + labs(title="Boxplots of FEV by smoke and sex\nAge 13+")
plot4
```



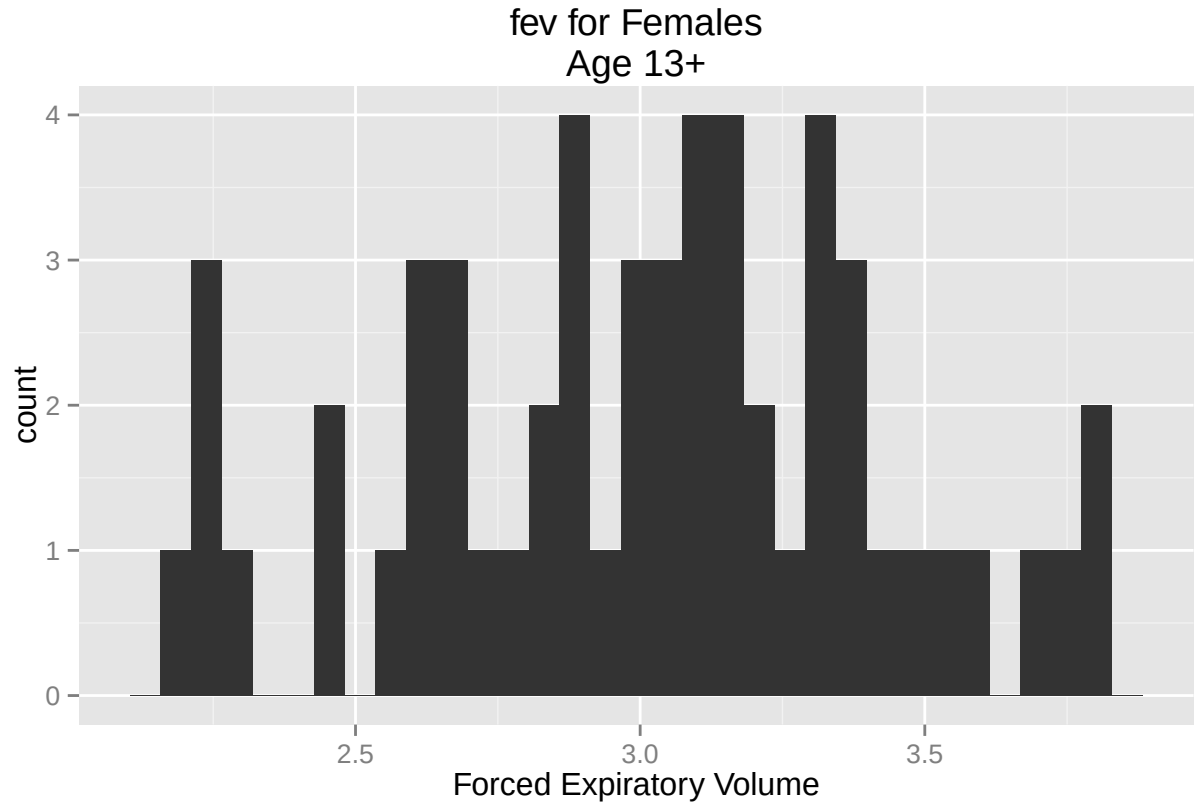
Histogram of fev for males.

```
plot5 <- ggplot(subset(FEV2, subset = sex == "male"),
                aes(fev)) + geom_histogram()
plot5 <- plot5 + labs(x="Forced Expiratory Volume",
                      title="fev for Males\nAge 13+")
plot5
```



Histogram of fev for females.

```
plot6 <- ggplot(subset(FEV2, subset = sex == "female"),
                aes(fev)) + geom_histogram()
plot6 <- plot6 + labs(x="Forced Expiratory Volume",
                      title="fev for Females\nAge 13+")
plot6
```



t-test for fev by sex.

```
t.test(fev ~ sex, data = FEV2)
```

```
##
##  Welch Two Sample t-test
##
## data:  fev by sex
## t = -10.924, df = 102.56, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.3204148 -0.9146286
## sample estimates:
## mean in group female   mean in group male
##             3.007236             4.124758
```

Split-up the data set FEV2 by sex

```
FEV2m <- subset(FEV2, subset = sex == "male")
FEV2f <- subset(FEV2, subset = sex == "female")
```

t-test for fev by smoke.

```
t.test(fev ~ smoke, data = FEV2m)
```

```
##
##  Welch Two Sample t-test
##
## data:  fev by smoke
## t = 0.60154, df = 32.643, p-value = 0.5516
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2688169  0.4943724
## sample estimates:
## mean in group non-current smoker     mean in group current smoker
##                         4.157500                         4.044722
```

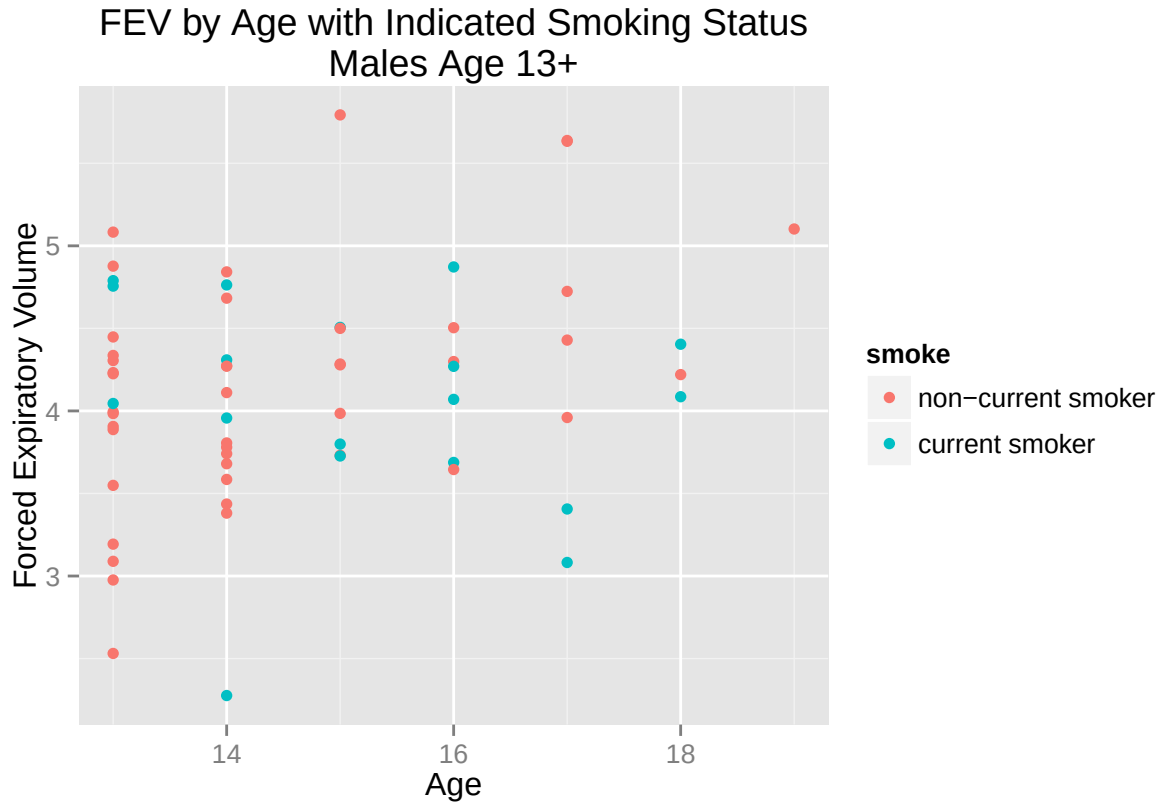
t-test for fev by smoke.

```
t.test(fev ~ smoke, data = FEV2f)
```

```
##
##  Welch Two Sample t-test
##
## data:  fev by smoke
## t = 1.2998, df = 46.317, p-value = 0.2001
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.08054486  0.37430486
## sample estimates:
## mean in group non-current smoker     mean in group current smoker
##                          3.07400                          2.92712
```

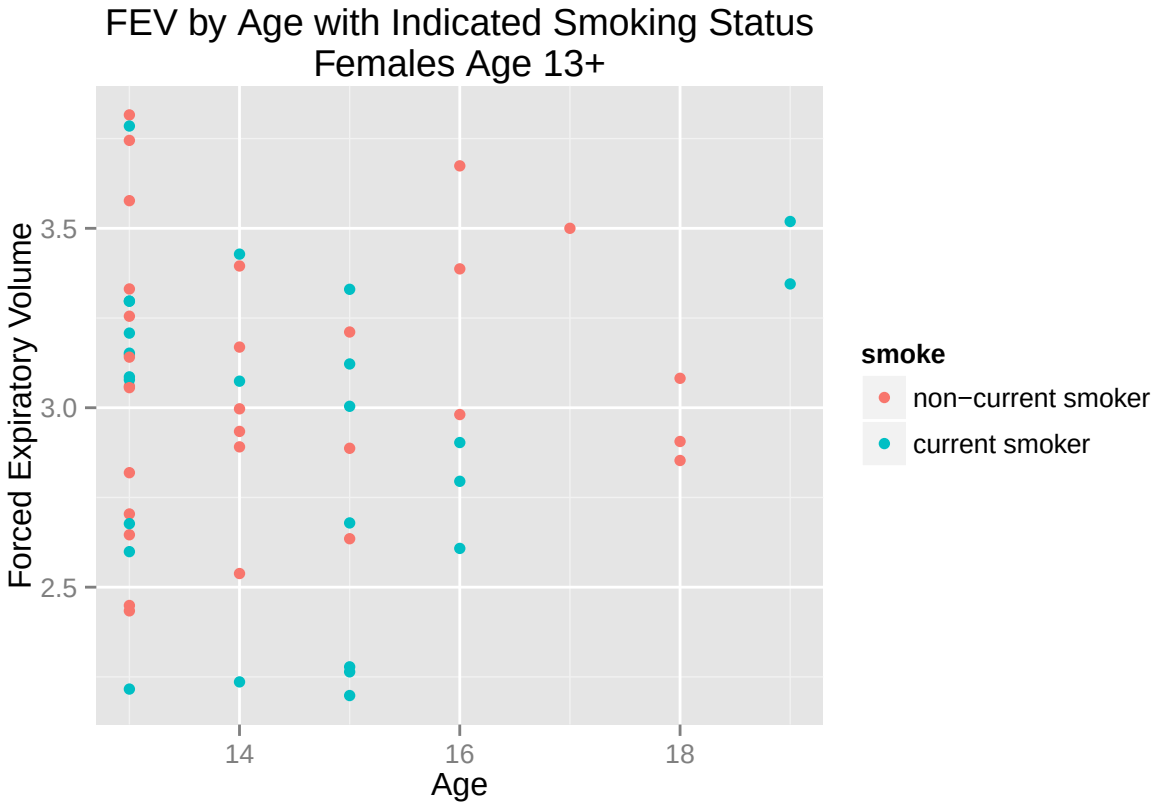
scatterplot of fev by age with smoke coding for males.

```
plot7 <- ggplot(FEV2m, aes(x=age, y=fev, colour=smoke)) + geom_point()
plot7 <- plot7 + labs(x="Age", y="Forced Expiratory Volume",
                      title="FEV by Age with Indicated Smoking Status\nMales Age 13+")
plot7
```



scatterplot of fev by age with smoke coding for females.

```
plot8 <- ggplot(FEV2f, aes(x=age, y=fev, colour=smoke)) + geom_point()
plot8 <- plot8 + labs(x="Age", y="Forced Expiratory Volume",
                      title="FEV by Age with Indicated Smoking Status\nFemales Age 13+")
plot8
```


FEV by Age with Indicated Smoking Status
Females Age 13+
Forced Expiratory Volume
3.5
3.0
2.5
14
16
18
Age
smoke
non−current smoker
current smoker