DASI_Exploratory Data Analysis_Project Phase I

Nishant Upadhyay

Sunday, September 21, 2014

```
library(datasets)
data(InsectSprays)
```

```
summary(InsectSprays)
```

```
##      count      spray
##  Min.   : 0.0   A:12
##  1st Qu.: 3.0   B:12
##  Median : 7.0   C:12
##  Mean   : 9.5   D:12
##  3rd Qu.:14.2   E:12
##  Max.   :26.0   F:12
```

```
head(InsectSprays)
```

```
##   count spray
## 1    10     A
## 2     7     A
## 3    20     A
## 4    14     A
## 5    14     A
## 6    12     A
```

To know the class of each variable:=

```
sapply(InsectSprays,class)
```

```
##     count     spray
## "numeric"  "factor"
```

Since we have one Numeric variable and one categorical variable we can see the summary by each factor/group levels:-

```
by(InsectSprays$count,InsectSprays$spray,summary)
```

```
## InsectSprays$spray: A
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     7.0    11.5    14.0    14.5    17.8    23.0
## ------------------------------------------------------------
## InsectSprays$spray: B
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     7.0    12.5    16.5    15.3    17.5    21.0
## ------------------------------------------------------------
## InsectSprays$spray: C
```

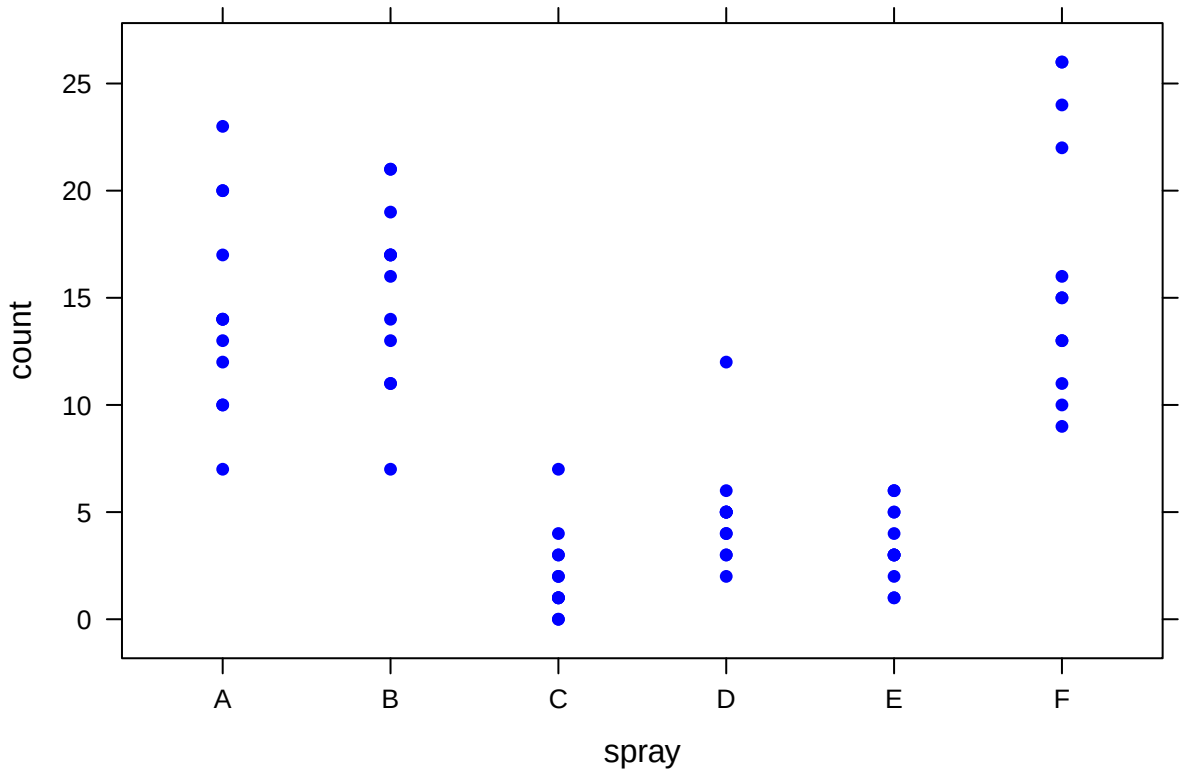
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    1.50    2.08    3.00    7.00
## ------------------------------------------------------------
## InsectSprays$spray: D
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    3.75    5.00    4.92    5.00   12.00
## ------------------------------------------------------------
## InsectSprays$spray: E
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.75    3.00    3.50    5.00    6.00
## ------------------------------------------------------------
## InsectSprays$spray: F
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0    12.5    15.0    16.7    22.5    26.0
```

We first plot the response variable'count' vs explanatory variable'spray'

```
library(lattice)
```

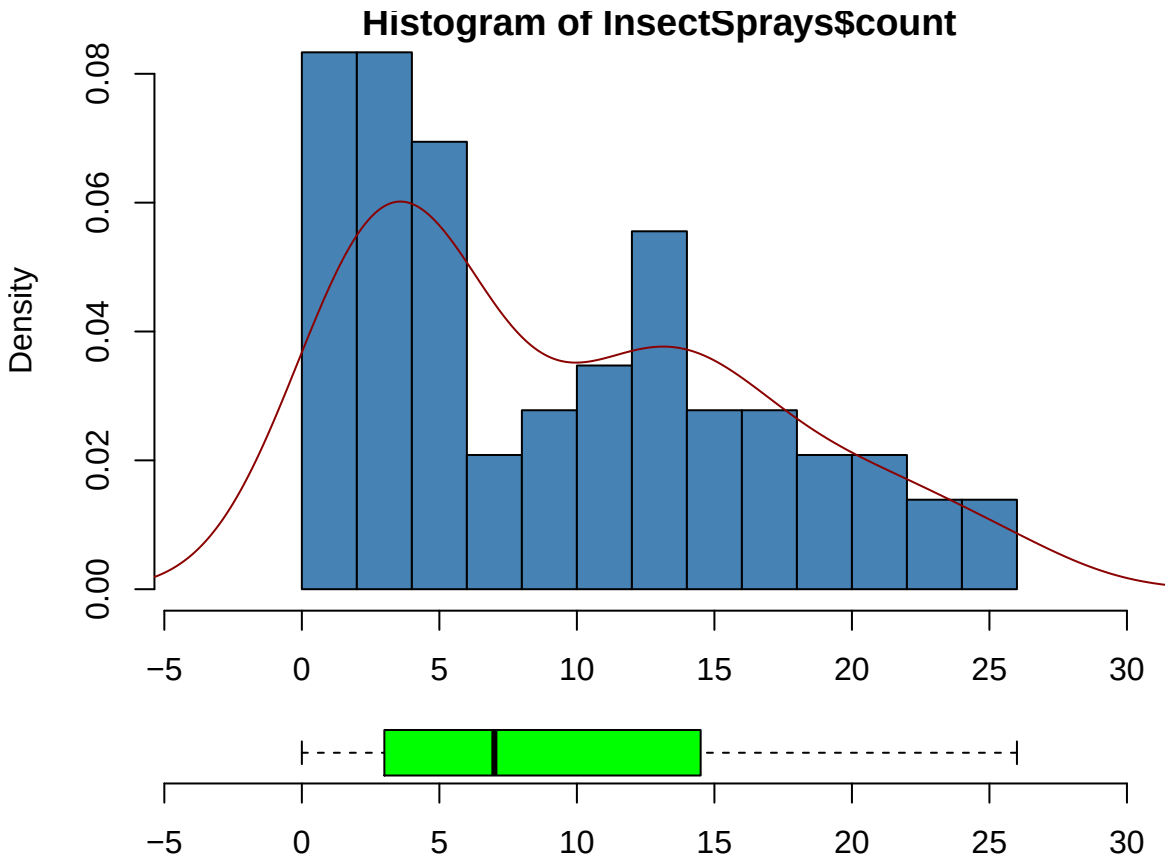
```
## Warning: package 'lattice' was built under R version 3.1.1
```

```
xyplot(count~spray,data=InsectSprays,pch=16,col="blue")
```



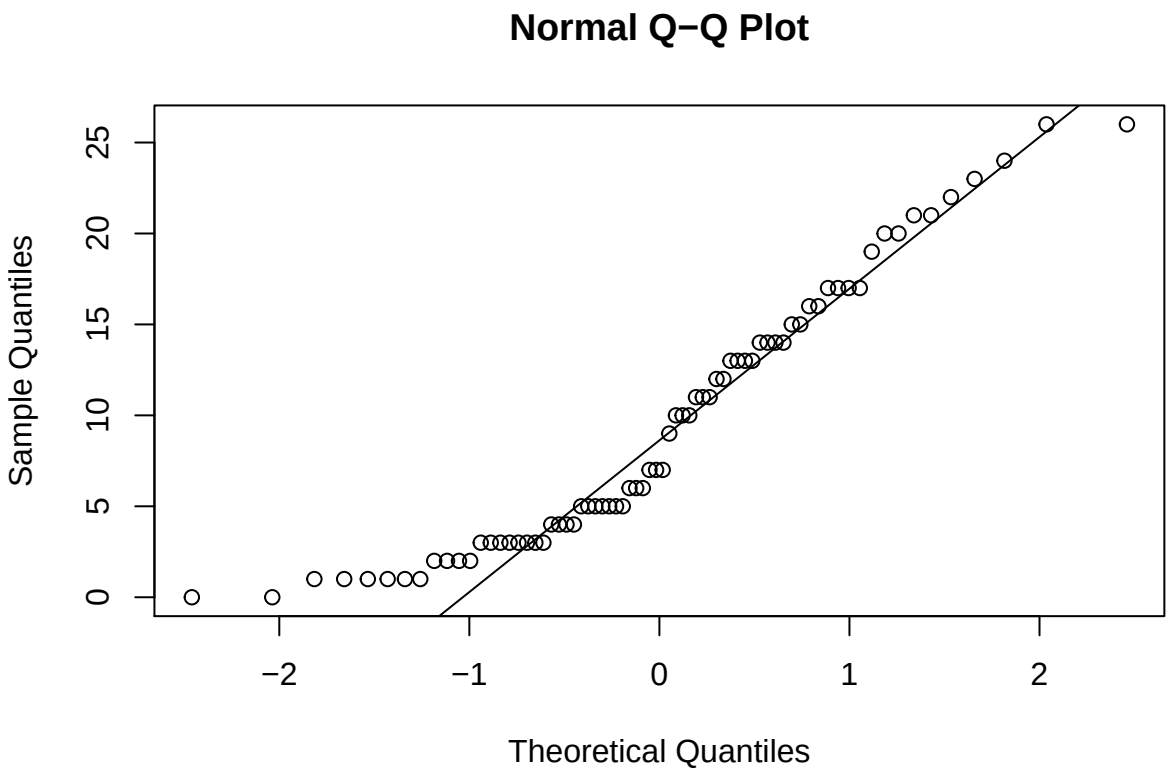
Let's now see the distribution of numeric variable"count":-

```r
# layout boxplot is at the bottom
nf <- layout(mat = matrix(c(1,2),2,1, byrow=TRUE),  height = c(2,0.5))
par(mar=c(2.4, 4.1, 0.5, 2.1))
hist(InsectSprays$count,xlim=c(-4,30), breaks=10,col = "steelblue",freq=F)
lines(density(InsectSprays$count),col="darkred")
boxplot(InsectSprays$count, horizontal=TRUE,  outline=TRUE,ylim=c(-4,30), frame=F, col = "green1",width=
```



The Histogram with boxplot shows that count data is right skewed with two prominent peaks.Therefore it seems the 'count' variable is bimodal.

```r
qqnorm(InsectSprays$count)
qqline(InsectSprays$count)
```



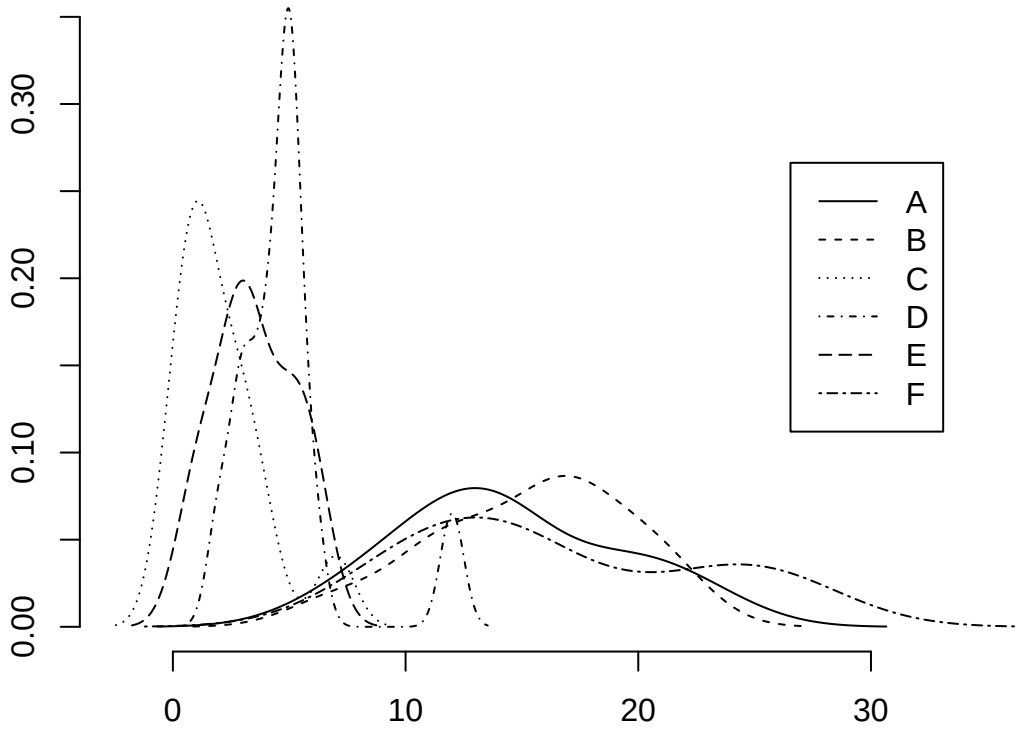
Lets see the distribution of each type of spray variable:-

```
library(UsingR)
```

```
## Warning: package 'UsingR' was built under R version 3.1.1

## Loading required package: MASS
## Loading required package: HistData

## Warning: package 'HistData' was built under R version 3.1.1

## Loading required package: Hmisc

## Warning: package 'Hmisc' was built under R version 3.1.1

## Loading required package: grid
## Loading required package: survival
## Loading required package: splines
## Loading required package: Formula

## Warning: package 'Formula' was built under R version 3.1.1

##
## Attaching package: 'Hmisc'
```

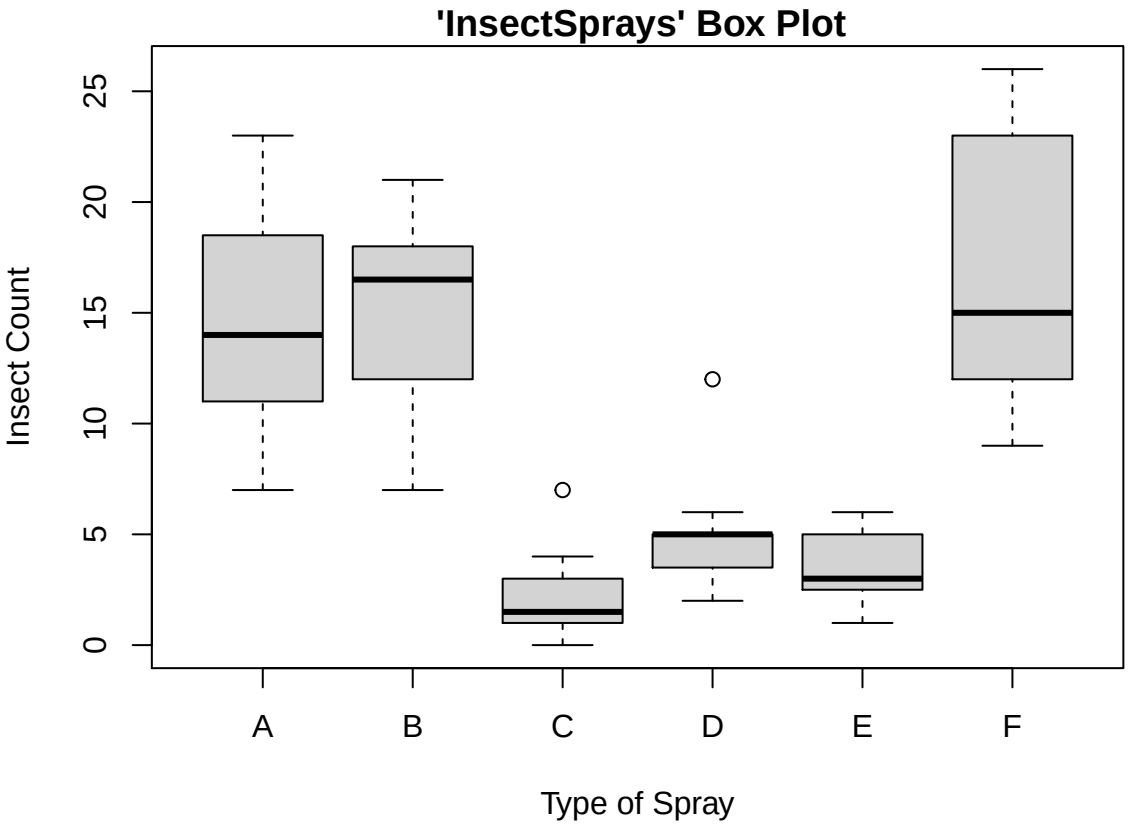
```
##
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
##
## Loading required package: quantreg

## Warning: package 'quantreg' was built under R version 3.1.1

## Loading required package: SparseM

## Warning: package 'SparseM' was built under R version 3.1.1

##
## Attaching package: 'SparseM'
##
## The following object is masked from 'package:base':
##
##     backsolve
##
##
## Attaching package: 'quantreg'
##
## The following object is masked from 'package:Hmisc':
##
##     latex
##
## The following object is masked from 'package:survival':
##
##     untangle.specials
##
##
## Attaching package: 'UsingR'
##
## The following object is masked from 'package:survival':
##
##     cancer
```

```
par(mar=c(4.1, 5.1, 1.0, 2.1))
simple.densityplot(count ~ spray, data =InsectSprays,xlab="InsectSprays$count",main="InsectSprays$spray
```



Another way to visualise the categorical variable is by plotting side by side boxplot :-

```
par(mar=c(5.1, 5.1, 1.2, 2.1))
require(stats); require(graphics)
boxplot(count ~ spray, data = InsectSprays,xlab = "Type of Spray",
        ylab = "Insect Count",main = " 'InsectSprays' Box Plot",
        varwidth = TRUE, col = "lightgray")
```



The boxplot for 'spray'variable suggests that there were less moths on plots sprayed with C, D and E.Also note that the variation in the number of moths is smaller in these treatments