# An Insight into the Effectiveness of Insect Sprays

*Nishant Upadhyay*

*Thursday, October 17, 2014*

This is a dataset available in R.It can be sourced using the following codes:

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

### Introduction:

In the following analysis we analyse an experiment in which 6 different insecticidal sprays are randomly applied to different treatment plots, and the response variable is the number of insects observed per plot(killed by the use of specific spray). This experimental work was carried out using replicated plots, and treatments were randomly assigned to the different plots.

When I started looking for dataset having a numeric and a categorical variable for analysis,i posted the question in stackoverflow and the feedback i got was overwhelming.Finally shortlisted **'InsectSprays'** builtin dataset in R.

### Scope of Inference:

The advantage I see in using this dataset is that it is a randomized study to evaluate the efficacy of insecticides thereby trying to establish a **CAUSAL** connection i.e.Effectiveness of Insect sprays—implying that it is an **EXPERIMENTAL STUDY** having experimental design features of *control,randomize & replicate*.

Here the population of interest is the number i.e. counts of moths effected by using the 6 different types of sprays.Since this is a randomized experiment sampling bias is reduced.Here the comparison will be made between the control group and the treatment group(groups here being plots of land) thus allowing us for causal conclusions.Here the sample size is 72 with adequate no of sprays for inter comparison and hence it seems reasonable for the experiment to be generalized to that population—-Random sampling ensures **'Generalization'** & Random assignment to control & treatment groups ensures that **'causal connections'** can be made.

I have tried to do a detailed analysis whereby some results are same but got from different methods.My intention here is that i should be able to refer this analysis in future also as well as increase my understanding of topic of Inference.

### Data:

The relevant dataset used here is *"InsectSprays"* (inbuilt dataset in R).

Basic Information about the experimental study:

A completely agricultural randomized experiment was conducted to evaluate the effects of different insect sprays on insect prevalence in agricultural plots.

Agricultural plots were assigned completely at random to receive six different insect spray treatments, the first of which is the control. There are 72 plots in the experiments. Each spray was assigned to 12 plots(treatment plots) and the control to 12 plots(control plots). The response variable consists of the counts of insects in the agricultural experimental units (plots) treated with the different insecticides i.e.response variable is the number of insects killed observed per plot.This work was carried out using replicated plots, and treatments were randomly assigned to the different plots.

**cases: 72–counts of insects killed by specific type of spray**

**variable: two–'count' & 'spray'**

Source:

Beall, G., (1942) The Transformation of data from entomological field experiments, Biometrika, 29, 243-262.

### Exploratory data analysis:

Lets first see the basic structure of 'InsectSprays' dataset:

```
head(InsectSprays)
```

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

```
tail(InsectSprays)
```

```
##    count spray
## 67    13     F
## 68    10     F
## 69    26     F
## 70    26     F
## 71    24     F
## 72    13     F
```

```
#To see the structure of the dataset 'InsectSprays':
str(InsectSprays)
```

```
## 'data.frame':    72 obs. of  2 variables:
##  $ count: num  10 7 20 14 14 12 10 23 17 20 ...
##  $ spray: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
#To see the variables in the dataset:
names(InsectSprays)
```

```
## [1] "count" "spray"
```

```
# lets see the levels of the categorical variable"spray"
levels(InsectSprays$spray)
```

```
## [1] "A" "B" "C" "D" "E" "F"
```

```
#To see 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:-

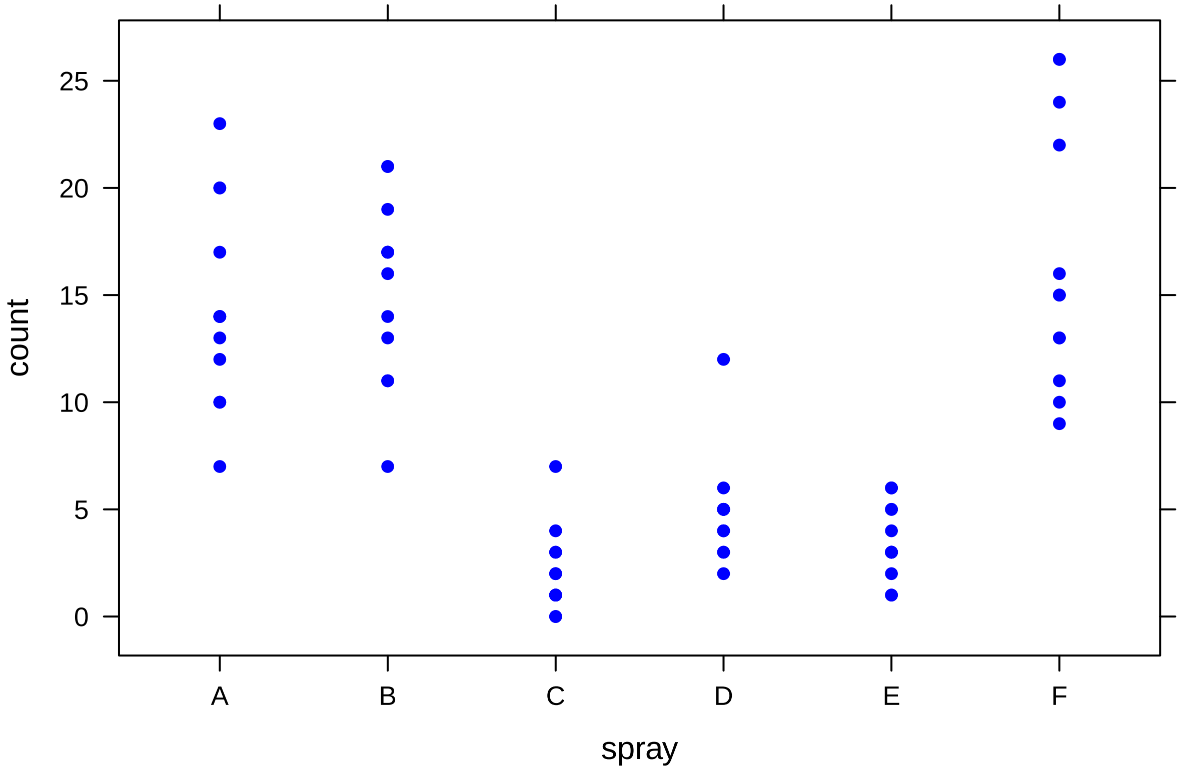
```
by(InsectSprays$count,InsectSprays$spray,summary)
```

```
## InsectSprays$spray: A
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   11.50   14.00   14.50   17.75   23.00
## --------------------------------------------------------
## InsectSprays$spray: B
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   12.50   16.50   15.33   17.50   21.00
## --------------------------------------------------------
## InsectSprays$spray: C
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   1.500   2.083   3.000   7.000
## --------------------------------------------------------
## InsectSprays$spray: D
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   3.750   5.000   4.917   5.000  12.000
## --------------------------------------------------------
## 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.
##    9.00   12.50   15.00   16.67   22.50   26.00
```

Now lets perform some Exploratory data analysis through visualisation of the dataset:

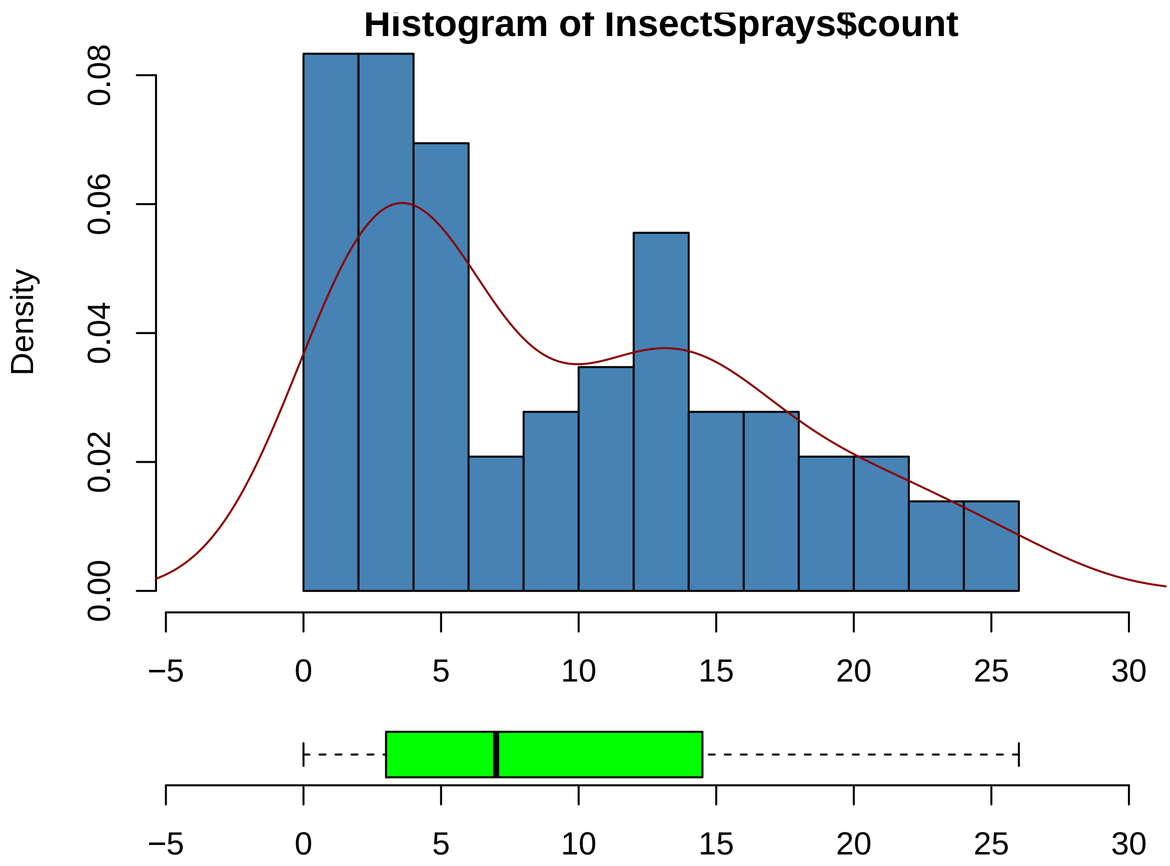
We first plot the response variable'count' vs explanatory variable'spray'

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



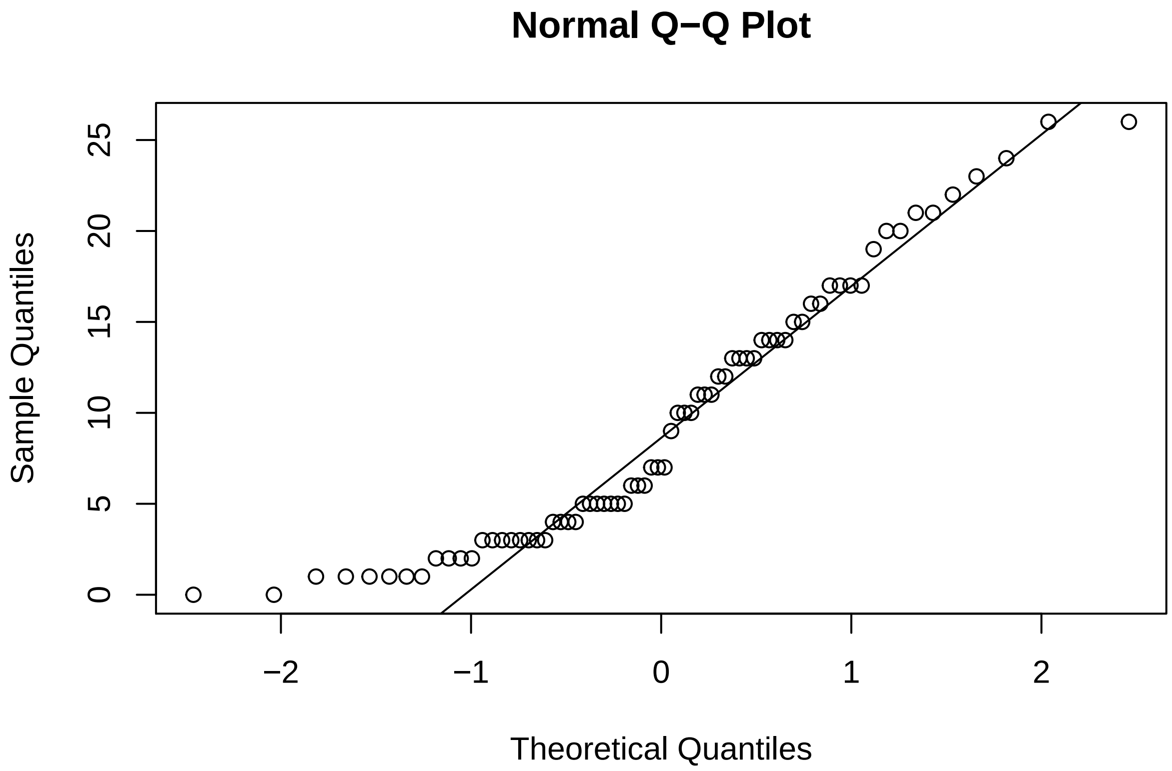
Let’s now see the distribution of numeric variable “count”:-

```
# 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 non-symmetric i.e. it is bimodal.

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



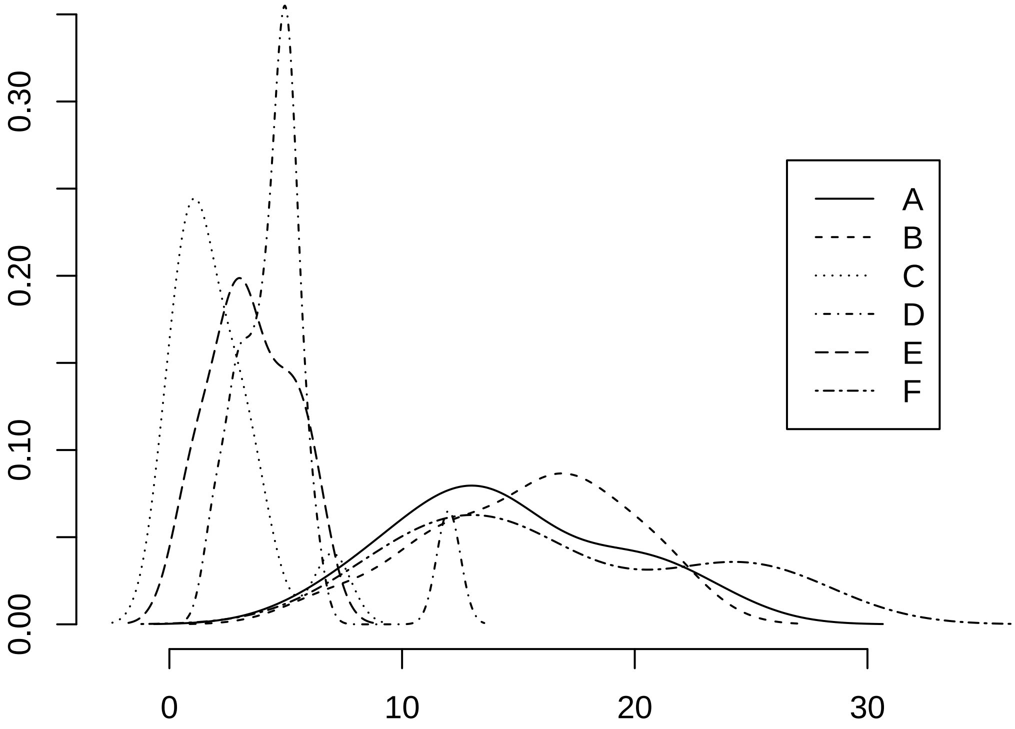
The normal probability plot of the numeric variable'count' shows that not all point lie along the straight line but some points at the lower end are far from the line indicating that the distribution of 'count' is skewed towards the right which is more clearly visible in the above plot(presence of *heavy tails*.

Lets see the distribution of each type of spray variable to visualise symmetry:-

```
library(UsingR)
```

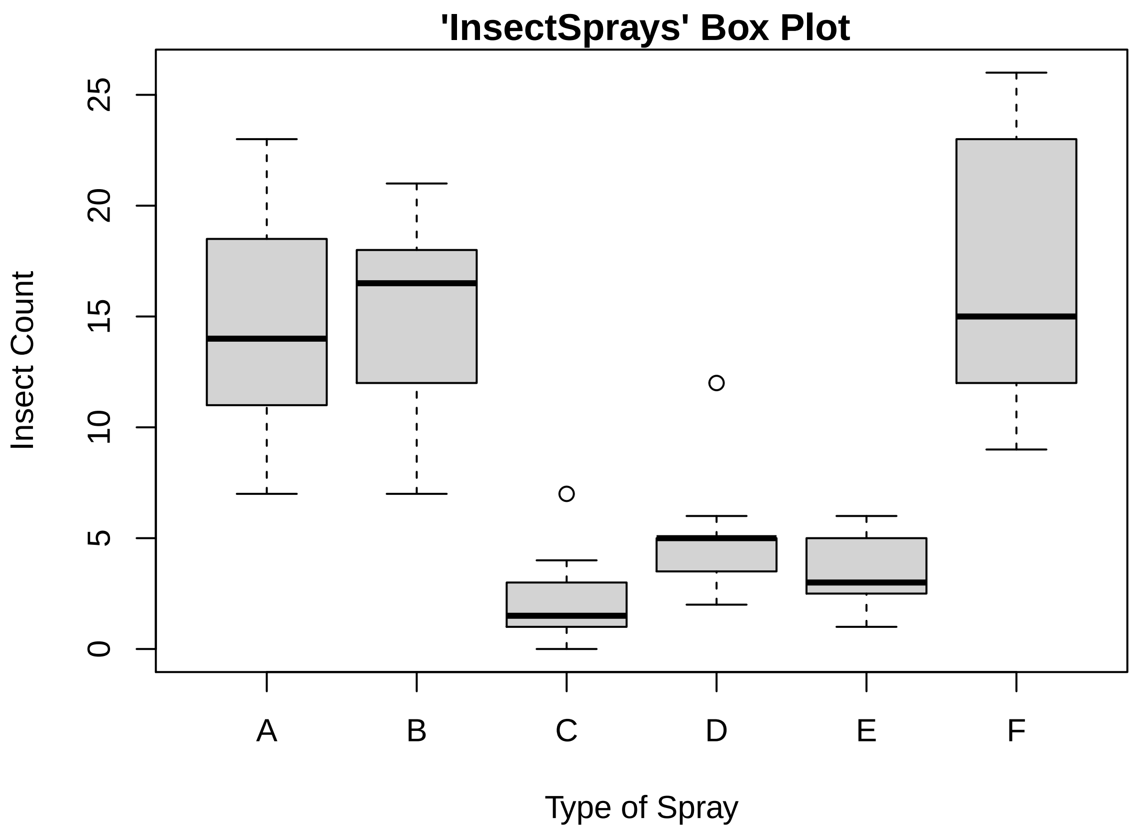
```
## Warning: package 'UsingR' was built under R version 3.1.1

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

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

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

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

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

```
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* :-

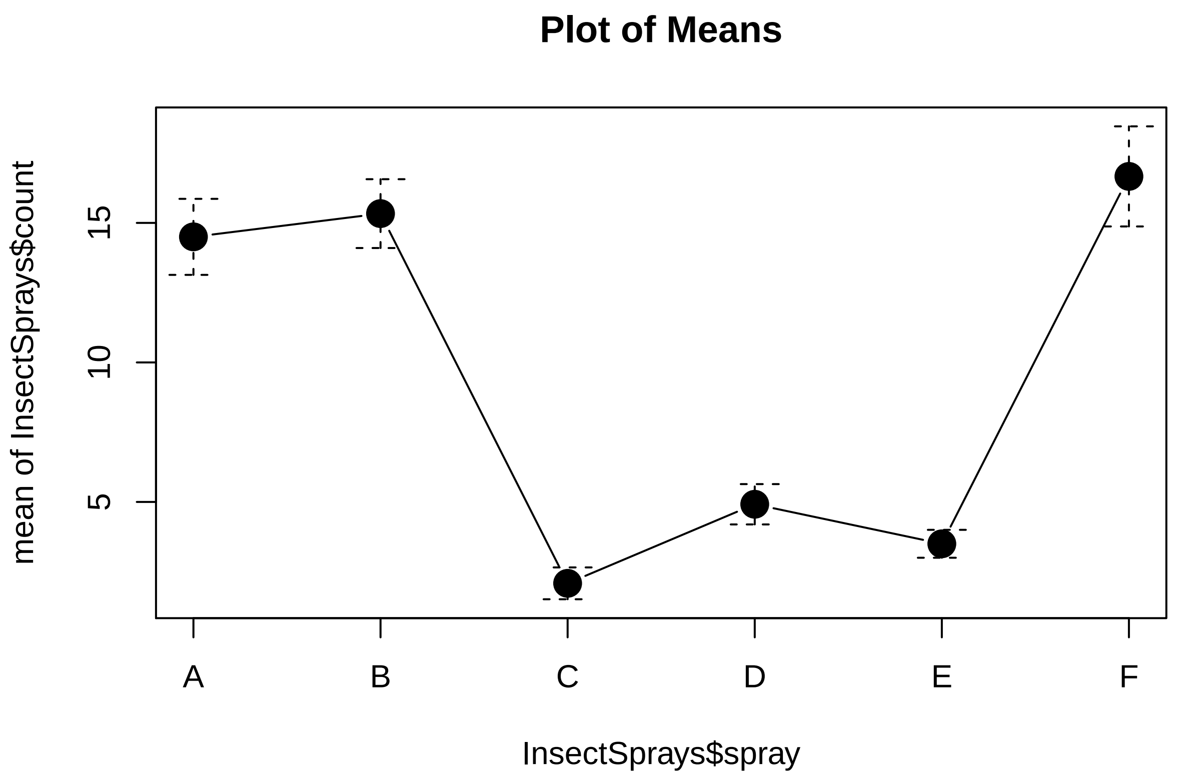
```
par(mar=c(5.1, 5.1, 1.2, 2.1))
require(stats); require(graphics)
boxplot(count ~ spray, data = InsectSprays,xlab = "Type of Spray",
        ylab = "Insect Count",main = " 'InsectSprays' Box Plot",
        varwidth = TRUE, col = "lightgray")
```



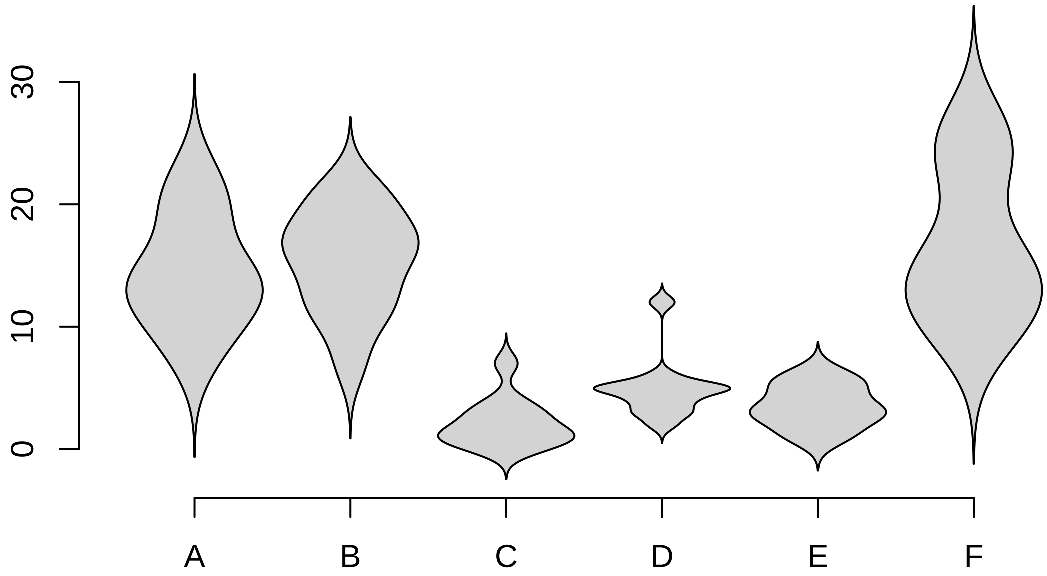
```
library(RcmdrMisc)
```

```
## Loading required package: car
## Loading required package: sandwich
```

```
plotMeans(InsectSprays$count,InsectSprays$spray,Error.Bars="SE")
```



```
library(UsingR)
simple.violinplot(count ~ spray, data = InsectSprays, col = "lightgray")
```



**Conclusions from Exploratory Data Analysis**

The boxplot as well as the Plot of means & violin plot for 'spray'variable suggests that there were less moths killed on plots sprayed with C, D and E implying less efficacy of these sprays.Also note that the variation in the number of moths killed is smaller in these treatments.Numerically the number of insects killed by each spray can be seen as below:

```
# Let's count the number of insects death due to each of the insecticides:
countfreq=c(1:6)
for(i in 1:6){
    countfreq[i]=sum(InsectSprays$count[InsectSprays$spray==LETTERS[i]])}
countfreq
```

```
## [1] 174 184  25  59  42 200
```

```
#The numbers 174, ..., 200 count the number of insects deaths due to
#insecticides A, ..., F. Clearly, the frequencies show that the
#Insecticides A, B, and F are very powerful as compared with C, D, and E.
```

**Inference:**

**Hypothesis** Ho:Spray means are equal Ha: Spray means are not equal

Looking at the structure of the 'InsectSprays' dataset,we know that the categorical variable *'spray'* has more than 3 levels(i.e. 6 levels) and hence to conduct an inferential analysis across many groups, we will apply the **ANOVA** methodology to test whether the means across many groups are equal.Here, since we have *One* *numerical and one categorical variable (with more than 2 levels)*: we conduct hypothesis test only comparing means across several groups with **no defined parameter of interest**, using **ANOVA and pairwise tests (theoretical only)**.

Before proceeding for Hypothesis testing, we conduct the customary check for the conditions for performing ANOVA:

**1: Independence**

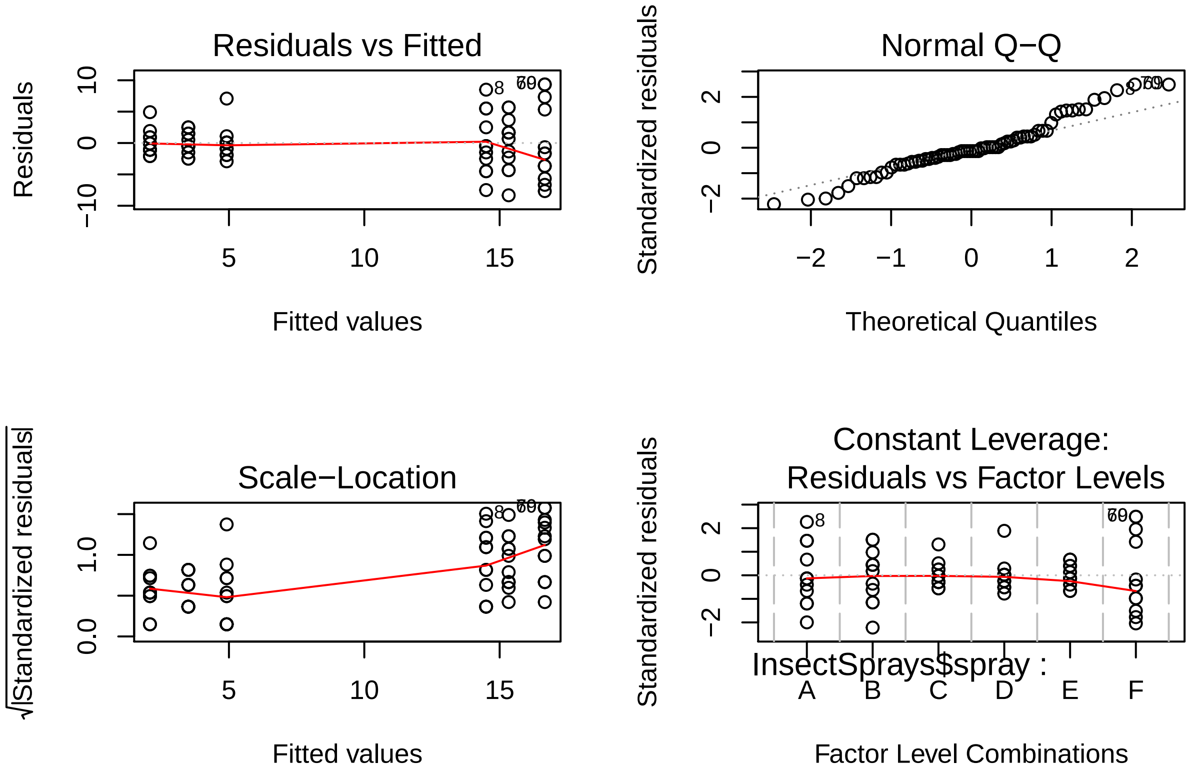
**withingroups**– As this is a completely randomized experiment conducted to evaluate the effects of different insect sprays on insect prevalence in agricultural plots & each sample size is <10% of the respective sample— **HENCE SAMPLE OBS. ARE INDEPENDENT**.

**Between groups**–As this is a sampled data from an experiment involving random sampling and random assignment, we can safely assume the independence among the groups.

**2: Normality and Homogeneity of variance across groups**

Lets perform some diagnostics checks for normality:

```r
lm1<-lm(InsectSprays$count~InsectSprays$spray)
par(mfrow=c(2,2))
plot(lm1)
```



The plots to the left show that variance is increasing with the mean, *a violation of the constant variance assumption*. The top right plot shows that these residuals have longer tails(heavier tails) than we would expect under normality.Hence count data is **NOT Normal**.

Aliter:

```
#Shapiro-Wilk test of normality
shapiro.test(InsectSprays$count)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  InsectSprays$count
## W = 0.9216, p-value = 0.0002525
```

Here,pvalue is less than alpha=0.05, hence *Shapiro-Wilk* test of normality suggests that data is not Normal.

Also the test for homoscedasticity (i.e.contant variance) can be visually checked from side by side boxplot of the count across groups.We can clearly see that variability among the groups **ARE DIFFERENT**.This can also be tested using **BARTLETT** test for homogeneity of variances.

```
# BARTLETT test for homogeneity of variances
bartlett.test(count ~ spray, data = InsectSprays)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  count by spray
## Bartlett's K-squared = 25.9598, df = 5, p-value = 9.085e-05
```

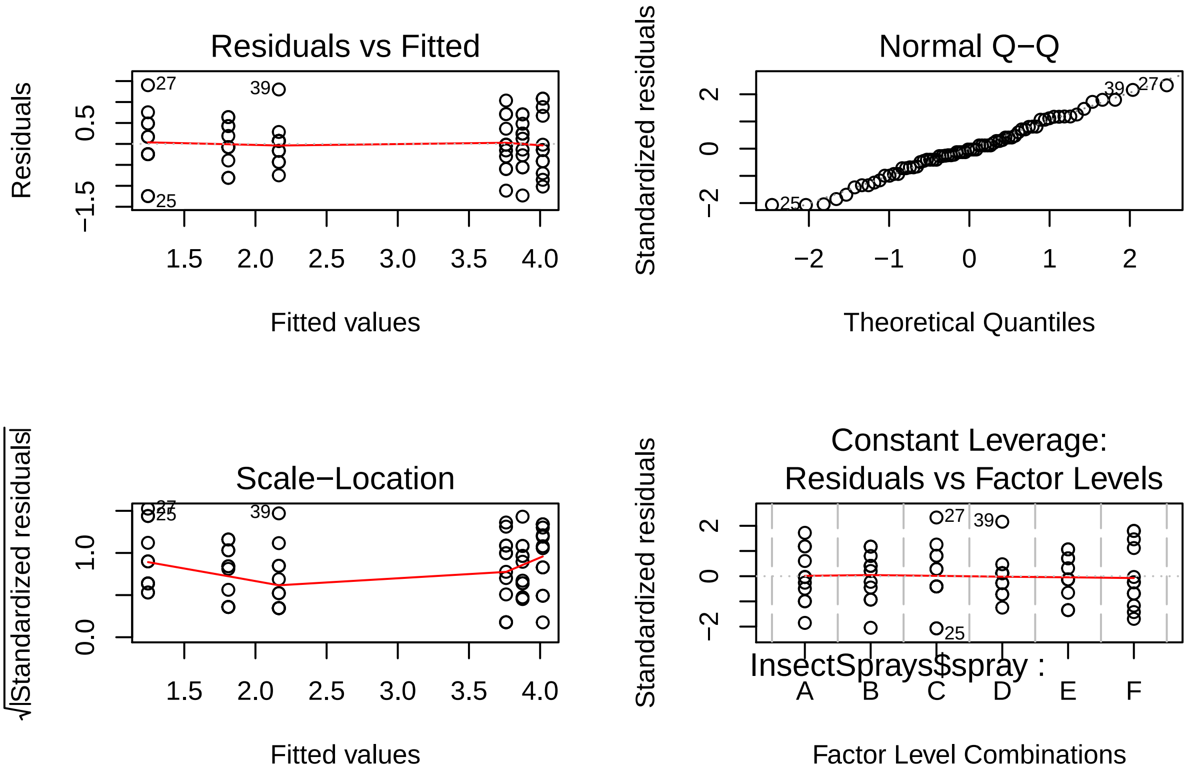
Here,the pvalue is $< 0.05$,hence we reject the null hypothesis in favour of alternate hypothesis which says that atleast one group mean is different.

So we can say that 'InsectSprays' data has observations that are independent but they are **not normal & the variances between the groups are different**.

Moving forward, let's perform Inferential analysis for the 'InsectSprays' dataset.Since the data is NOT normal nor it is having constant variance across groups we transform the Non Normal data.

Therefore,before ANOVA testing, lets transform the'count' data so that it becomes approximately normal.

```
# transforming the count var by taking square root of the'count' variable
lm1<-lm(sqrt(InsectSprays$count)~InsectSprays$spray)
par(mfrow=c(2,2))
plot(lm1)
```



Now we see that data is normal as revealed by the Normal Probability plot.

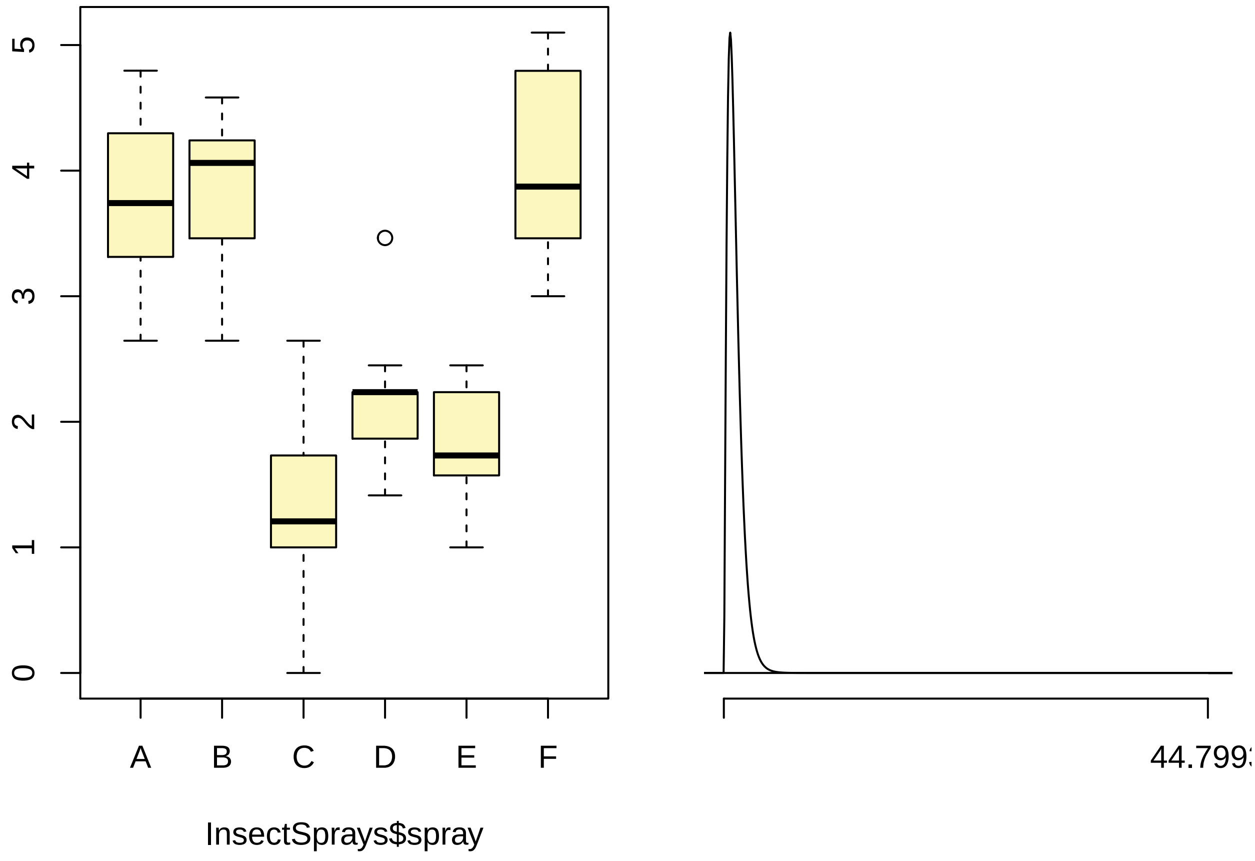
**Now we perform the hypothesis test using ANOVA:**

```
# Load the custom 'inference' function:
load(url("http://s3.amazonaws.com/assets.datacamp.com/course/dasi/inference.Rdata"))

inference(y = sqrt(InsectSprays$count), x = InsectSprays$spray,
          est = "mean",method = "theoretical", type = "ht",
          alternative = "greater")
```

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

```
## Warning: package 'BHH2' was built under R version 3.1.1
```

```
## Response variable: numerical, Explanatory variable: categorical
## ANOVA
## Summary statistics:
## n_A = 12, mean_A = 3.7607, sd_A = 0.6243
## n_B = 12, mean_B = 3.8766, sd_B = 0.5769
## n_C = 12, mean_C = 1.2449, sd_C = 0.763
## n_D = 12, mean_D = 2.1644, sd_D = 0.5033
## n_E = 12, mean_E = 1.8095, sd_E = 0.4964
## n_F = 12, mean_F = 4.0186, sd_F = 0.7513
## H_0: All means are equal.
## H_A: At least one mean is different.
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          5 88.438 17.6876  44.799 < 2.2e-16
## Residuals 66 26.058  0.3948
##
## Pairwise tests: t tests with pooled SD
##        A      B      C      D  E
## B 0.6527     NA     NA     NA NA
## C 0.0000 0.0000     NA     NA NA
## D 0.0000 0.0000 0.0006     NA NA
## E 0.0000 0.0000 0.0312 0.1712 NA
## F 0.3183 0.5818 0.0000 0.0000  0
```



Aliter:using Built in ‘aov’ function in R

```
summary(aov(lm(sqrt(InsectSprays$count)~InsectSprays$spray)))
```

```
##                    Df Sum Sq Mean Sq F value Pr(>F)
## InsectSprays$spray  5  88.44  17.688    44.8 <2e-16 ***
## Residuals          66  26.06   0.395
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see from both test the test statistic *F value= 44.8* is very high indicating that it lies farther in the right tail (which can be seen in the F distribution plot).This gives a very small pvalue ( pvalue (=2e-16)<0.05).This implies that we reject the null hypothesis of equality of means and accept alternate hypothesis that atleast one of mean among the groups is different.

Now that we have established that atleast one of group mean is different,the next question we want to answer is which of the means are different i.e. we have to do multiple comparision between the groups.

Before infering using the pvalues, since pairwise t-test for differences in each possible pair of groups **inflates the Type I error**,we apply the **BONFERRONI** correction which suggests a more stringent significance level.

```
# calculating Bonferroni corrected significance level

alpha<-0.05
k<-6 # no of levels
alpha.mod<-alpha/k
alpha.mod
```

```
## [1] 0.008333333
```

### Method I:

Using the *'inference'* function used above and seeing its output–

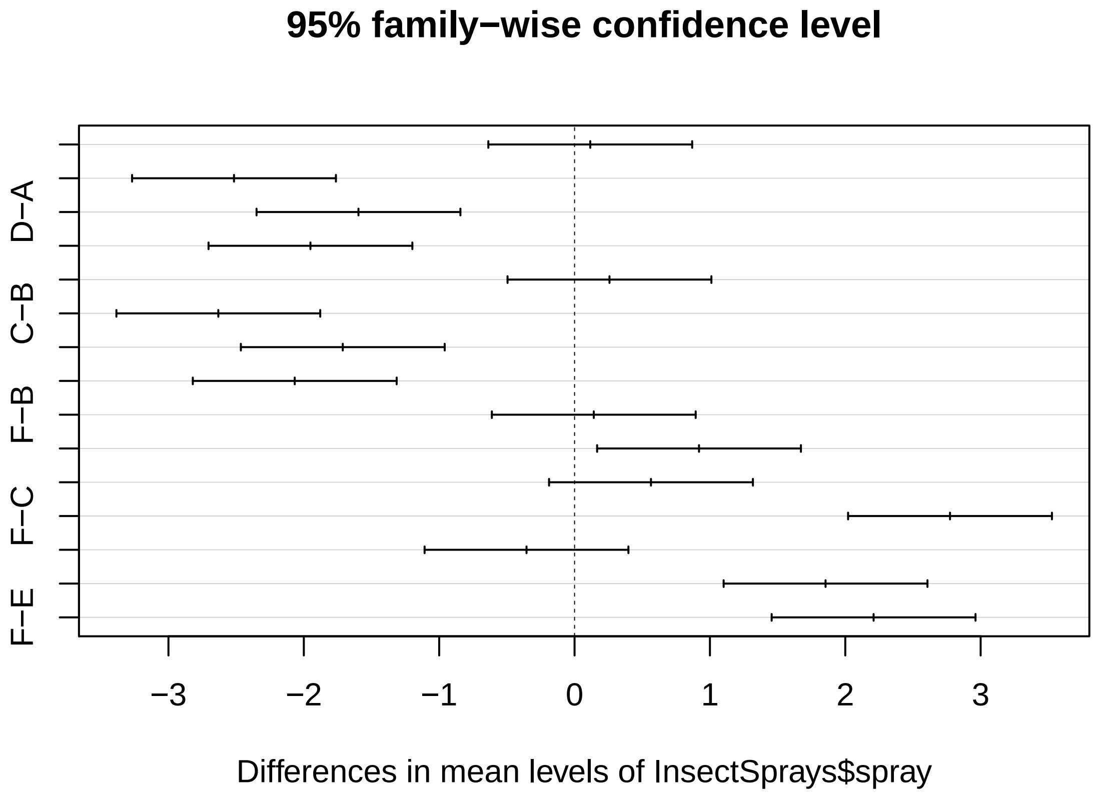
Pairwise tests: t tests with pooled SD A B C D E B 0.6527 NA NA NA NA C 0.0000 0.0000 NA NA NA D 0.0000 0.0000 0.0006 NA NA E 0.0000 0.0000 0.0312 0.1712 NA F 0.3183 0.5818 0.0000 0.0000 0

aliter: **Method II**

```
# TukeyHSD (Tukey Honest Significant Differences) test for multiple
#comparisions
THSD<-TukeyHSD(aov(sqrt(InsectSprays$count)~InsectSprays$spray))
THSD
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = sqrt(InsectSprays$count) ~ InsectSprays$spray)
##
## $`InsectSprays$spray`
##           diff        lwr        upr     p adj
## B-A  0.1159530 -0.6369601  0.8688661 0.9975245
## C-A -2.5158217 -3.2687349 -1.7629086 0.0000000
## D-A -1.5963245 -2.3492377 -0.8434114 0.0000006
## E-A -1.9512174 -2.7041305 -1.1983042 0.0000000
## F-A  0.2579388 -0.4949744  1.0108519 0.9144964
## C-B -2.6317747 -3.3846879 -1.8788616 0.0000000
## D-B -1.7122775 -2.4651907 -0.9593644 0.0000001
## E-B -2.0671704 -2.8200835 -1.3142572 0.0000000
## F-B  0.1419858 -0.6109274  0.8948989 0.9935788
## D-C  0.9194972  0.1665841  1.6724103 0.0080813
## E-C  0.5646043 -0.1883088  1.3175175 0.2512638
## F-C  2.7737605  2.0208474  3.5266736 0.0000000
## E-D -0.3548928 -1.1078060  0.3980203 0.7366389
## F-D  1.8542633  1.1013502  2.6071764 0.0000000
## F-E  2.2091561  1.4562430  2.9620693 0.0000000
```

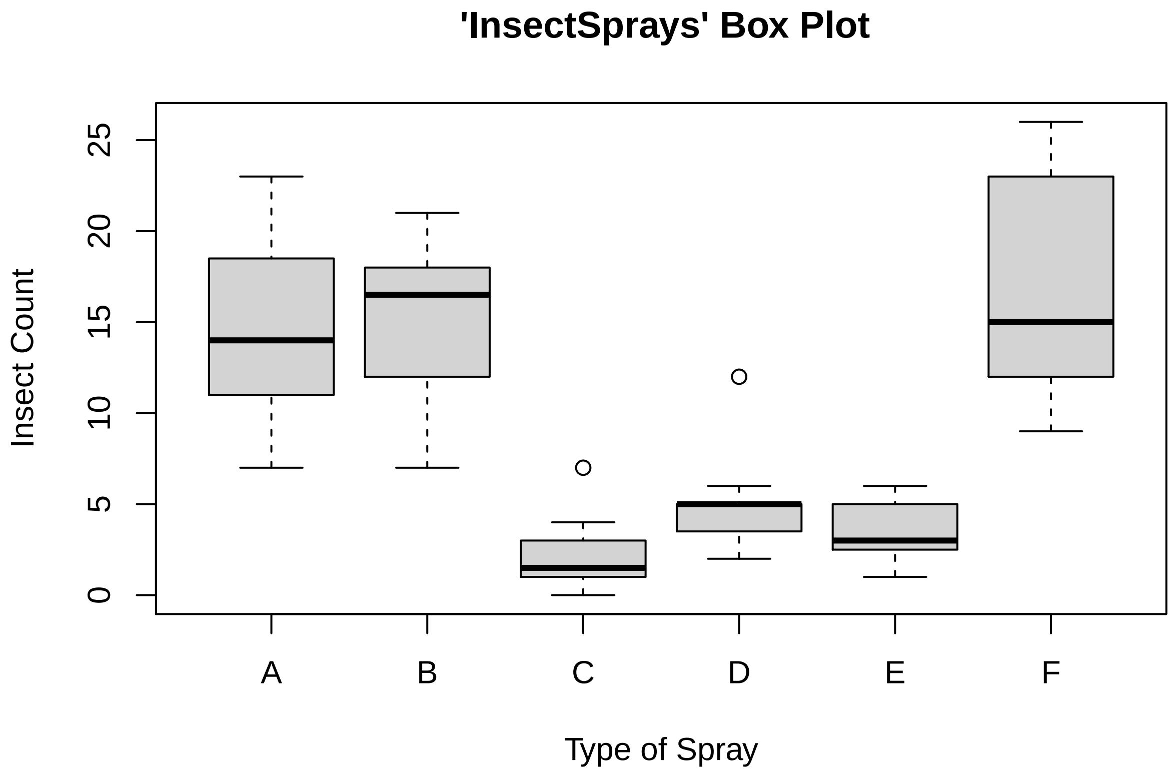
```
plot(THSD)
```



Comparing the pvalues with **Bonferroni adjusted significance level of 0.008333**, we find that sprays C,D & E compared with sprays A,B, & F have pvalues < adjusted alpha indicating that we reject the null hypothesis of mean difference is zero and *accepting* that means of groups are significantly different.

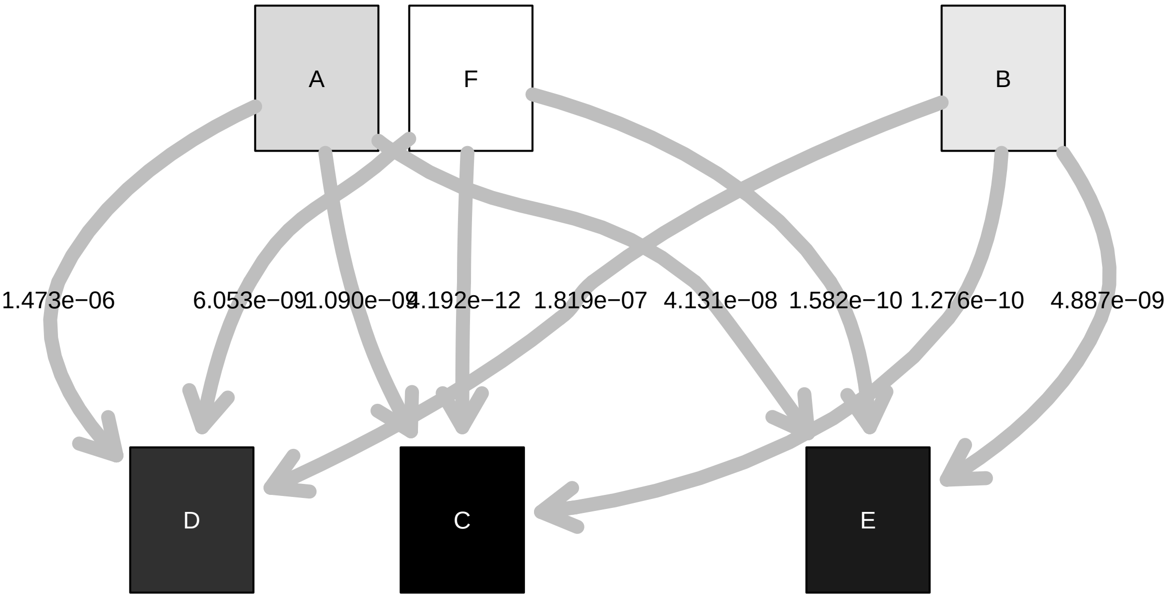
To visualise multiple comparisions, I came across 'paircompviz' package which allows pairwise t tests comparision visualisation through **HASSE diagram**,

```
# Visualization of Multiple Pairwise Comparison Test Results

# let's first load the 'paircompviz' package using the following codes:

library(paircompviz)
require(stats); require(graphics)
boxplot(count ~ spray, data = InsectSprays,xlab = "Type of Spray",
        ylab = "Insect Count",main = " 'InsectSprays' Box Plot",
        varwidth = TRUE, col = "lightgray")
```



```
paircomp(InsectSprays$count, InsectSprays$spray, test="t",compress=FALSE,
         pooled.sd=TRUE,p.adjust.method="bonferroni")
```

**Pairwise comparisons on InsectSprays$count by InsectSprays$spray using t tests with pooled SD and bonferroni p−value adjustment**



Nodes of the graph represent treatments,edges represent statistically significant difference.Pairwise pvalues are also given. Comparing the results with side by side box plot and we can see that sprays C,D,E which are shaded black in Hasse diagram are representing significant difference criteria.

### Conclusion:

In comparing the mean count of insects killed by 6 different sprays (i.e.testing the difference in the effectiveness of the sprays), we determined (through an ANOVA test at the 0.05 significance level) that more than one mean(count of insects killed) differed from others i.e.3 sprays (C,D,E) means are statistically varied from each other when following up the ANOVA test with a series of pairwise T-tests ( co-verified using TukeyHSD test). These result leads us to conclude that the effectiveness is significant across different sprays groups and specifically sprays C,D,E as a group are **less effective** compared to spray group A,B & F.

Inspite of this detailed study ,there are open questions which I could not answer or incorporate here.One such issue was whether transforming the data from non normal to normal was correct? or should i have gone for less powerful Non-parametric tests.Also while using the custom inference function,i could not add the fact/argument that variances were unequal across groups(is there a argument to add unequal variance in 'Inference' function?).

### References:

OpenIntro Statistics text book,2nd Edition_David M Diez, Christopher D Barr & Mine Çetinkaya-Rundel 2012

An Introduction to Statistical Inference and Its Applications with R_Michael W. Trosset, Department of Statistics, Indiana University 2008

AN INTRODUCTION TO R_DEEPAYAN SARKAR 2011

Coursera–Exploratory Data Analysis/Regression/Statistical Inference module notes_John Hopkins

http://en.wikipedia.org/wiki/Tukey's_range_test

http://watson.nci.nih.gov/bioc_mirror/packages/2.13/bioc/html/paircompviz.html

**Appendix:**

```
# Data sample:First twenty obs
head(InsectSprays,30)
```

```
##    count spray
## 1     10     A
## 2      7     A
## 3     20     A
## 4     14     A
## 5     14     A
## 6     12     A
## 7     10     A
## 8     23     A
## 9     17     A
## 10    20     A
## 11    14     A
## 12    13     A
## 13    11     B
## 14    17     B
## 15    21     B
## 16    11     B
## 17    16     B
## 18    14     B
## 19    17     B
## 20    17     B
## 21    19     B
## 22    21     B
## 23     7     B
## 24    13     B
## 25     0     C
## 26     1     C
## 27     7     C
## 28     2     C
## 29     3     C
## 30     1     C
```