ANOVA

##Purpose## ANOVA is used to compare multiple means to see if there are any differences. Are they equal or not?

##Theory## ANOVA is used to compare differences of means among more than 2 groups. It does this by looking at variation in the data and where that variation is found (hence its name). Specifically, ANOVA compares the amount of variation between groups with the amount of variation within groups. It can be used for both observational and experimental studies.

When we take samples from a population, we expect each sample mean to differ simply because we are taking a sample rather than measuring the whole population; this is called sampling error but is often referred to more informally as the effects of “chance”. Thus, we always expect there to be some differences in means among different groups. The question is: is the difference among groups greater than that expected to be caused by chance?

Like other classical statistical tests, we use ANOVA to calculate a test statistic (the F-ratio) with which we can obtain the probability (the P-value) of obtaining the data assuming the null hypothesis. A significant P-value (usually taken as P<0.05) suggests that at least one group mean is significantly different from the others.

Null hypothesis: all population means are equal

Alternative hypothesis: at least one population mean is different from the rest.

###How does ANOVA relate to the t-test?### t-test is used when comparing two groups while ANOVA is used for comparing more than 2 groups. In fact, if you calculate the p-value using ANOVA for 2 groups, you will get the same results as the t-test.

##Assumptions## 1. The dependent variable is continuous.

1. The independent variable is categorical with or more categories.
2. The dependent and independent variables have values for each row of data. (?)
3. Observations in each group are independent.
4. The dependent variable is approximately normally distributed in each group.
5. There is approximate equality of variance in all the groups.
6. We should not have any outliers

###Packages Used###

library(ggplot2)  
library(psych)  
library(car)  
library(readxl)  
library(pwr)  
library(readr)

##Demonstration Data##

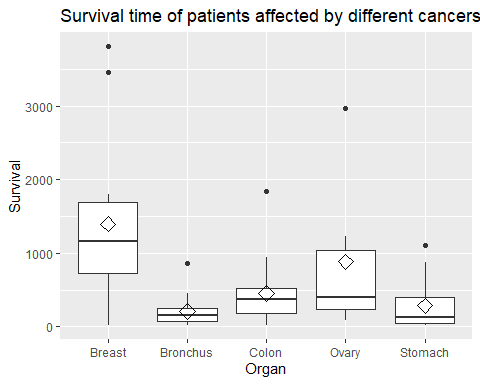
cancer.survival <- read\_csv("C:/Users/leekennedy/Documents/GitHub/Chemistry\_Stats/data/cancer-survival.csv")

###Plot the data### Data should always be visually examined.

cs\_plot <- ggplot(cancer.survival, aes(x = Organ, y = Survival)) +   
 geom\_boxplot() +  
 stat\_summary(fun.y = mean, geom = "point", shape = 23, size = 4) +  
 ggtitle("Survival time of patients affected by different cancers")

## Warning: `fun.y` is deprecated. Use `fun` instead.

cs\_plot



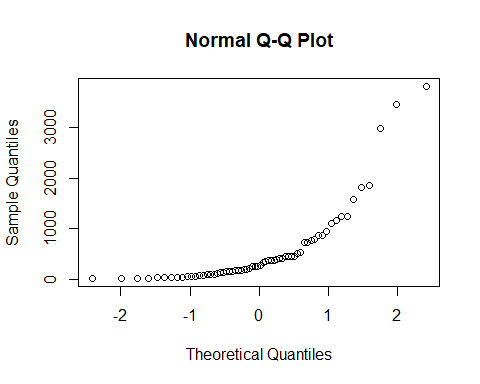
##Testing Assumptions## ###Normality, using Shapiro-Wilk Test### The null-hypothesis of this test is that the population is normally distributed. Thus, if the p-value is less than the chosen alpha level, then the null hypothesis is rejected and there is evidence that the data tested are not from a normally distributed population; in other words, the data are not normal. On the contrary, if the p-value is greater than the chosen alpha level, then the null hypothesis that the data came from a normally distributed population cannot be rejected (e.g., for an alpha level of 0.05, a data set with a p-value of 0.02 rejects the null hypothesis that the data are from a normally distributed population).[2] However, since the test is biased by sample size,[3] the test may be statistically significant from a normal distribution in any large samples. Thus a Q–Q plot is required for verification in addition to the test.

with(cancer.survival, tapply(Survival, Organ, shapiro.test))

## $Breast  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.86857, p-value = 0.07431  
##   
##   
## $Bronchus  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.76596, p-value = 0.0007186  
##   
##   
## $Colon  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.76056, p-value = 0.0006134  
##   
##   
## $Ovary  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.76688, p-value = 0.029  
##   
##   
## $Stomach  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.75473, p-value = 0.002075

###Q-Q Plot### Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight.

qqnorm(cancer.survival$Survival)

 ###Equality of Variance, using Levene’s Test### In statistics, Levene’s test is an inferential statistic used to assess the equality of variances for a variable calculated for two or more groups.[1] Some common statistical procedures assume that variances of the populations from which different samples are drawn are equal. Levene’s test assesses this assumption. It tests the null hypothesis that the population variances are equal (called homogeneity of variance or homoscedasticity). If the resulting p-value of Levene’s test is less than some significance level (typically 0.05), the obtained differences in sample variances are unlikely to have occurred based on random sampling from a population with equal variances. Thus, the null hypothesis of equal variances is rejected and it is concluded that there is a difference between the variances in the population.

leveneTest(Survival~Organ, data = cancer.survival)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 4 4.4524 0.003271 \*\*  
## 59   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##Stop and Reassess## Neither necessary conditions, normality and equal variance is being met. At this point you should stop and assess what to do. Transforming data is an option.

###Transform Data### The boxplot and the Q-Q plot indicates that the relationship is not a normal distribution and so a log transformation is tried:

cancer.survival$log.survival <- log(cancer.survival$Survival)  
  
with(cancer.survival, tapply(log.survival, Organ, shapiro.test))

## $Breast  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.802, p-value = 0.009995  
##   
##   
## $Bronchus  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.98047, p-value = 0.9613  
##   
##   
## $Colon  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.92636, p-value = 0.1891  
##   
##   
## $Ovary  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.983, p-value = 0.9655  
##   
##   
## $Stomach  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.92837, p-value = 0.3245

leveneTest(log.survival~Organ, data = cancer.survival)

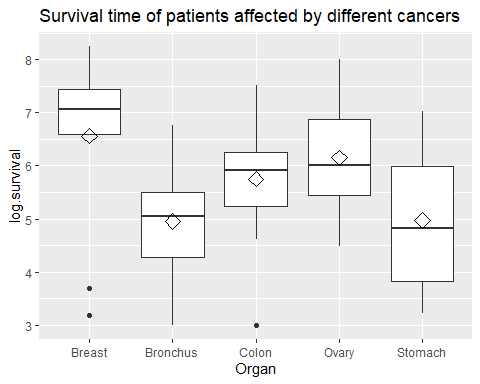
## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 4 0.6685 0.6164  
## 59

Following the transformation, the data appears to be better structured. This is born out in the boxplot and Q-Q plot:

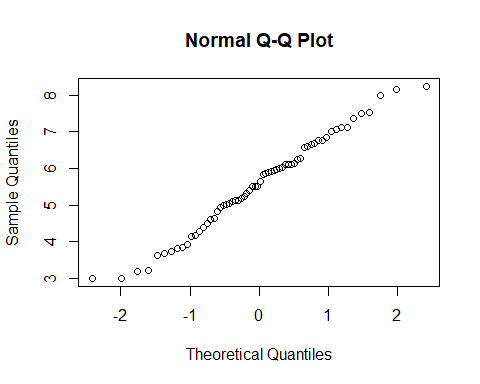
# Diamonds = means (Shape 24)  
  
cs\_plot <- ggplot(cancer.survival, aes(x = Organ, y = log.survival)) +   
 geom\_boxplot() +  
 stat\_summary(fun.y = mean, geom = "point", shape = 23, size = 4) +  
 ggtitle("Survival time of patients affected by different cancers")

## Warning: `fun.y` is deprecated. Use `fun` instead.

cs\_plot



qqnorm(cancer.survival$log.survival)



##Perform the ANOVA Test##

aov1 <- aov(log.survival~Organ, cancer.survival)  
summary(aov1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Organ 4 24.49 6.122 4.286 0.00412 \*\*  
## Residuals 59 84.27 1.428   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The p-value of 0.00412 indicates the Null Hypothesis, Ho, that the means are all the same, is rejected. A Tukey HSD test will assess which one are different.

##Perform the Tukey HSD Comparison## Tukey’s HSD (honest significant difference) test is a single-step multiple comparison procedure and statistical test. It can be used on raw data or in conjunction with an ANOVA (Post-hoc analysis) to find means that are significantly different from each other.

Tukey’s test compares the means of every treatment to the means of every other treatment; that is, it applies simultaneously to the set of all pairwise comparisons and identifies any difference between two means that is greater than the expected standard error.

TukeyHSD(aov1)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = log.survival ~ Organ, data = cancer.survival)  
##   
## $Organ  
## diff lwr upr p adj  
## Bronchus-Breast -1.60543320 -2.906741 -0.3041254 0.0083352  
## Colon-Breast -0.80948110 -2.110789 0.4918267 0.4119156  
## Ovary-Breast -0.40798703 -2.114754 1.2987803 0.9615409  
## Stomach-Breast -1.59068365 -2.968399 -0.2129685 0.0158132  
## Colon-Bronchus 0.79595210 -0.357534 1.9494382 0.3072938  
## Ovary-Bronchus 1.19744617 -0.399483 2.7943753 0.2296079  
## Stomach-Bronchus 0.01474955 -1.224293 1.2537924 0.9999997  
## Ovary-Colon 0.40149407 -1.195435 1.9984232 0.9540004  
## Stomach-Colon -0.78120255 -2.020245 0.4578403 0.3981146  
## Stomach-Ovary -1.18269662 -2.842480 0.4770864 0.2763506

From the above results, there is a significant difference between bronchus and breast and stomach and breast. This can be seen in the boxplot.

##Interpretation##

* The data showed departure from normality and equality of variance.
* Perhaps unequal variance was due to our unbalanced design (we had unequal samples in our groups)
* A log transformation was useful in stabilizing variance.
* Normality was violated in the breast group even after transformation.
* ANOVA is robust to slight deviations from normality
* Differences between groups were statistically significant

##Laboratory Application - Validations## ANOVA looks at the differences between groups and within groups and so it can be used to look at variance within a batch run and between batch runs. This is the repeatability and reproducibility from a validation trial. In this example we have a sample tested six times in the first batch and three times in five subsequent batches. The data is artificial and used sdr = 2 and sdR = 5 in it’s construction.

**NOTE:** We are looking for something slightly different here to the regular ANOVA test. ANOVA looks at data to see if there are differences in the means. In validation trials we expect differences and want to use the ANOVA process to extract some measure of it.

###Data###

val\_data <- read\_excel("C:/Users/leekennedy/Documents/GitHub/Chemistry\_Stats/data/Validation.xlsx")  
round(val\_data,2)

## # A tibble: 6 x 6  
## A B C D E F  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 91.9 99.4 105. 98.0 95.3 107.  
## 2 90 96.0 108. 98.6 94.0 105.  
## 3 91.2 96.1 108. 96.9 98.1 104.  
## 4 91.6 NA NA NA NA NA   
## 5 89 NA NA NA NA NA   
## 6 89.5 NA NA NA NA NA

val\_data2 <- na.omit(stack(val\_data))

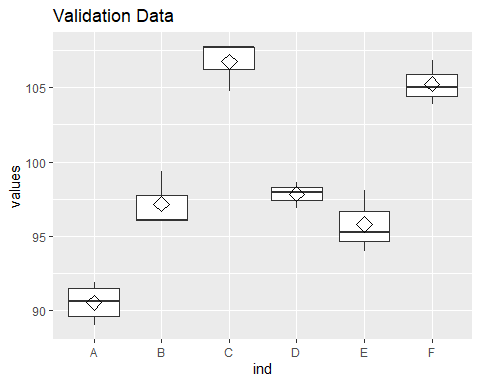
Used *na.omit* to remove NAs from data. *Stack* converts from wide to long format.

###Boxplot### It always pays to view your data:

val\_plot <- ggplot(val\_data2, aes(x = ind, y = values)) +   
 geom\_boxplot() +  
 stat\_summary(fun.y = mean, geom = "point", shape = 23, size = 4) +  
 ggtitle("Validation Data")

## Warning: `fun.y` is deprecated. Use `fun` instead.

val\_plot

 From the boxplot, it is apparent that there is much better within group precision than between group precision.

##Explore the data## ###Shapiro-Wilk Normality Test###

with(val\_data2, tapply(values, ind, shapiro.test))

## $A  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.91786, p-value = 0.4901  
##   
##   
## $B  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.75872, p-value = 0.01934  
##   
##   
## $C  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.76494, p-value = 0.03329  
##   
##   
## $D  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.98467, p-value = 0.7629  
##   
##   
## $E  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.95511, p-value = 0.5922  
##   
##   
## $F  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.98272, p-value = 0.7482

###Levene’s test for Variance###

leveneTest(values~ind, data = val\_data2)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 5 0.1483 0.9775  
## 15

###Interpretation### The variance is good as seen by both the boxplot and Levene’s Test. Not all groups pass the Shapiro-Wilk normality test but they only have three data points. Ultimately we are wanting to measure the variances, not assess them, so the discrepencies are minor.

##Run ANOVA##

val\_aov <- aov(val\_data2$values~val\_data2$ind)  
summary(val\_aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## val\_data2$ind 5 737.1 147.41 62.48 1.66e-09 \*\*\*  
## Residuals 15 35.4 2.36   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The ANOVA confirms the boxplot assessment.

##Extract the Repeatability and Reproducibility standard deviations##

mean.sqr <- summary(val\_aov)[1][[1]][[3]]  
ncount <- as.numeric(length(val\_aov$effects))/as.numeric(length(val\_aov$coefficients))  
sdr <- sqrt(mean.sqr[2])  
interim <- sqrt((mean.sqr[1]-mean.sqr[2])/ncount)  
sdR <- sqrt(sdr^2 + interim^2)  
paste("Repeatability sd = ", round(sdr,2))

## [1] "Repeatability sd = 1.54"

paste("Reproducibility sd = ",round(sdR,2))

## [1] "Reproducibility sd = 6.62"

You will only approach the dummy data’s seed values (2 & 5) by repeated analysis.

## Two Way ANOVA

In two way ANOVA there are three hypotheses of interest as listed below

H0: There is an effect of the first factor on the dependent continuous variable (main effect)  
H0: There is an effect of the second factor variable on the dependent continuous variable (main effect)  
H0: There is a combined effect of the first and second factor variable on the continuous dependent variable (interaction)

The above hypotheses can be extended from two factor variables to N factor variables.

For results of two way ANOVA to be valid there are several assumptions that need to be satisfied. They are listed below.

Observations must be independent within and across groups  
Observations are approximately normally distributed.  
There is equal variance in the observations  
We should not have any outliers especially when our design is unbalanced  
The errors are independent

This example uses a moth trap experiment. The dependent variable is the number of moths in a trap. The independent variables are location and type of lure. There were four locations (top, middle, lower and ground). There were three types of lure (scent, sugar and chemical).

### Data

moth.experiment <- read\_csv("C:/Users/leekennedy/Documents/GitHub/Chemistry\_Stats/data/moth-trap-experiment.csv")

## Parsed with column specification:  
## cols(  
## `number of moths` = col\_double(),  
## location = col\_character(),  
## `type of lure` = col\_character()  
## )

str(moth.experiment)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 60 obs. of 3 variables:  
## $ number of moths: num 32 29 16 18 20 37 40 18 28 36 ...  
## $ location : chr "Top" "Top" "Top" "Top" ...  
## $ type of lure : chr "Chemical" "Chemical" "Chemical" "Chemical" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. `number of moths` = col\_double(),  
## .. location = col\_character(),  
## .. `type of lure` = col\_character()  
## .. )

### Check for balanced data

table(moth.experiment$location, moth.experiment$`type of lure`)

##   
## Chemical Scent Sugar  
## Ground 5 5 5  
## Lower 5 5 5  
## Middle 5 5 5  
## Top 5 5 5

We have equal observations, so the design is balanced.

### Summary Statistics by location

describeBy(moth.experiment$`number of moths`,moth.experiment$location)

##   
## Descriptive statistics by group   
## group: Ground  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 15 19.07 5.09 18 18.85 5.93 12 29 17 0.52 -1.06 1.31  
## ------------------------------------------------------------   
## group: Lower  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 15 33.33 7.5 34 33.77 7.41 17 44 27 -0.6 -0.5 1.94  
## ------------------------------------------------------------   
## group: Middle  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 15 31 9.79 36 31.46 11.86 12 44 32 -0.39 -1.29 2.53  
## ------------------------------------------------------------   
## group: Top  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 15 23.33 7.41 21 23.23 8.9 13 35 22 0.24 -1.63 1.91

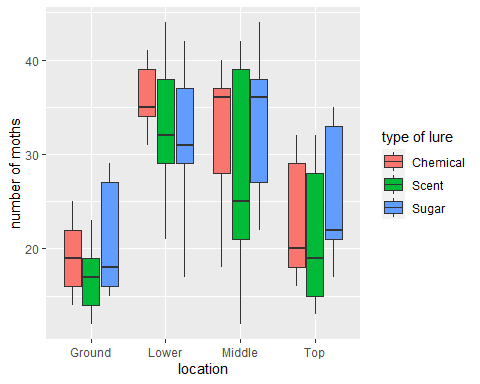
### Summary Statistics by lure

describeBy(moth.experiment$`number of moths`,moth.experiment$`type of lure`)

##   
## Descriptive statistics by group   
## group: Chemical  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 20 27.5 9.06 28.5 27.44 12.6 14 41 27 -0.01 -1.61 2.03  
## ------------------------------------------------------------   
## group: Scent  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 20 24.75 10.29 22 24.06 11.12 12 44 32 0.43 -1.2 2.3  
## ------------------------------------------------------------   
## group: Sugar  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 20 27.8 9.06 28 27.44 11.12 15 44 29 0.14 -1.35 2.03

### Create Boxplots

ggplot(moth.experiment, aes(x = location, y= `number of moths`, fill = `type of lure`)) +  
 geom\_boxplot()



### Check for Normality

shapiro.test(moth.experiment$`number of moths`)

##   
## Shapiro-Wilk normality test  
##   
## data: moth.experiment$`number of moths`  
## W = 0.94533, p-value = 0.009448

This indicates that the data is NOT normally distributed.

###Try a log transformation###

no.of.moth.log = log(moth.experiment$`number of moths`)  
moth.experiment$no.of.moth.log = no.of.moth.log  
shapiro.test(moth.experiment$no.of.moth.log)

##   
## Shapiro-Wilk normality test  
##   
## data: moth.experiment$no.of.moth.log  
## W = 0.94746, p-value = 0.01185

Mmmm…not very useful.

### Check for equality of variances

leveneTest(moth.experiment$`number of moths`~moth.experiment$location\*moth.experiment$`type of lure`)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 11 0.6377 0.7875  
## 48

This indicates that the variance is equivalent across the groups.

### Power of the Test

Our design has 2 factors with 3 and 4 levels, we have 5 observations in each group, our df for the mean squared term is 4\*3(5-1)=48 We choose a medium effect size of 0.25

pwr.f2.test(u=2, v=48, f2=(0.25\*0.25))

##   
## Multiple regression power calculation   
##   
## u = 2  
## v = 48  
## f2 = 0.0625  
## sig.level = 0.05  
## power = 0.3210203

### Atleast! Perform ANOVA

moth.anova = aov(moth.experiment$no.of.moth.log~moth.experiment$location\*moth.experiment$`type of lure`)  
summary(moth.anova)

## Df Sum Sq Mean Sq  
## moth.experiment$location 3 2.966 0.9886  
## moth.experiment$`type of lure` 2 0.262 0.1308  
## moth.experiment$location:moth.experiment$`type of lure` 6 0.196 0.0326  
## Residuals 48 4.828 0.1006  
## F value Pr(>F)   
## moth.experiment$location 9.828 3.64e-05 \*\*\*  
## moth.experiment$`type of lure` 1.301 0.282   
## moth.experiment$location:moth.experiment$`type of lure` 0.324 0.921   
## Residuals   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Location has an effect on number of moths Type of lure does not have an effect on number of moths The combined effect of location and type of lure does not have an effect on number of moths

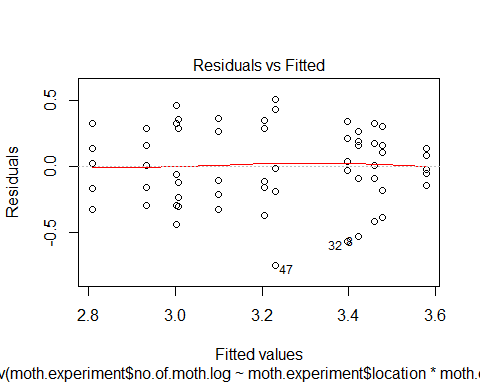
When you have an unbalanced design R does not issue any warnings to correctly analyze an unbalanced design we can use the Anova function in car library we pass results of aov function and specify we would like to use Type III sums of squares

Anova(moth.anova,type = "III")

## Anova Table (Type III tests)  
##   
## Response: moth.experiment$no.of.moth.log  
## Sum Sq Df F value  
## (Intercept) 43.018 1 427.6842  
## moth.experiment$location 1.302 3 4.3144  
## moth.experiment$`type of lure` 0.102 2 0.5054  
## moth.experiment$location:moth.experiment$`type of lure` 0.196 6 0.3245  
## Residuals 4.828 48   
## Pr(>F)   
## (Intercept) < 2.2e-16 \*\*\*  
## moth.experiment$location 0.008988 \*\*   
## moth.experiment$`type of lure` 0.606429   
## moth.experiment$location:moth.experiment$`type of lure` 0.920916   
## Residuals   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

###check for homogeneity of residuals###

plot(moth.anova,1)



Homogeneity assumption is not violated but points 47 and 32 are marked as outliers. Remember our data still had some non normality