Project Part1

Sai Pavan #R11778698,Rishidhar Rao #R11842808, Hemanth Reddy #R11842807

2022-12-04

The experiment was performed using a Statapult, with different type of balls. The Statapult has three factors that could be consider as variables they are nothing but the Pin Elevation, the Bungee Position, and the Release Angle with the Type of Balls being the other factor could be consider as variable.

The Project was performed in three parts.

Part 01: By varying the Type of Balls, and keeping the Pin Elevation and Bungee Position fixed, and Release Angle stretched to 90^o.

Part 02: By considering the Pin Elevation and Release Angle as variables, with a red ball when the Bungee Position is fixed.

Part 03: By considering all factors from the Statapult as variables, with different types of balls.

After determining the necessary sample size, a random data collecting plan is constructed. The gathered data is processed to separate the significant aspects from the insignificant ones, providing us with an assessment of the factors influencing the distance to which the ball is thrown. The ultimate conclusion about the key elements is reached and presented.

Part - 01

Perform a designed experiment to determine the effect of the type of ball on the distance in which the ball is thrown.

The Pin Elevation and Bungee Position should both be at their fourth setting, i.e., highest setting. The Release Angle should be at 90^o, with the arm pulled fully back before releasing. To test this hypothesis, we wish to use a completely randomized design with an α around 0.05.

a) Determine how many samples should be collected to detect a mean difference with a large effect (i.e. 90% of the standard deviation) and a pattern of maximum variability with a probability of 55%.

The sample size determination involves various Arguments:

k: Number of replications of each type of ball.

n: Number of Observations of each type of ball.

f: Effect size (which depends on the variability of the design).

sig.level: The level of significant error we accept.

power: The probability of rejecting the Null hypothesis, if the means differences exceeds the effect.

For the given parameters.

$d = \frac{differencinmean}{sigma}$

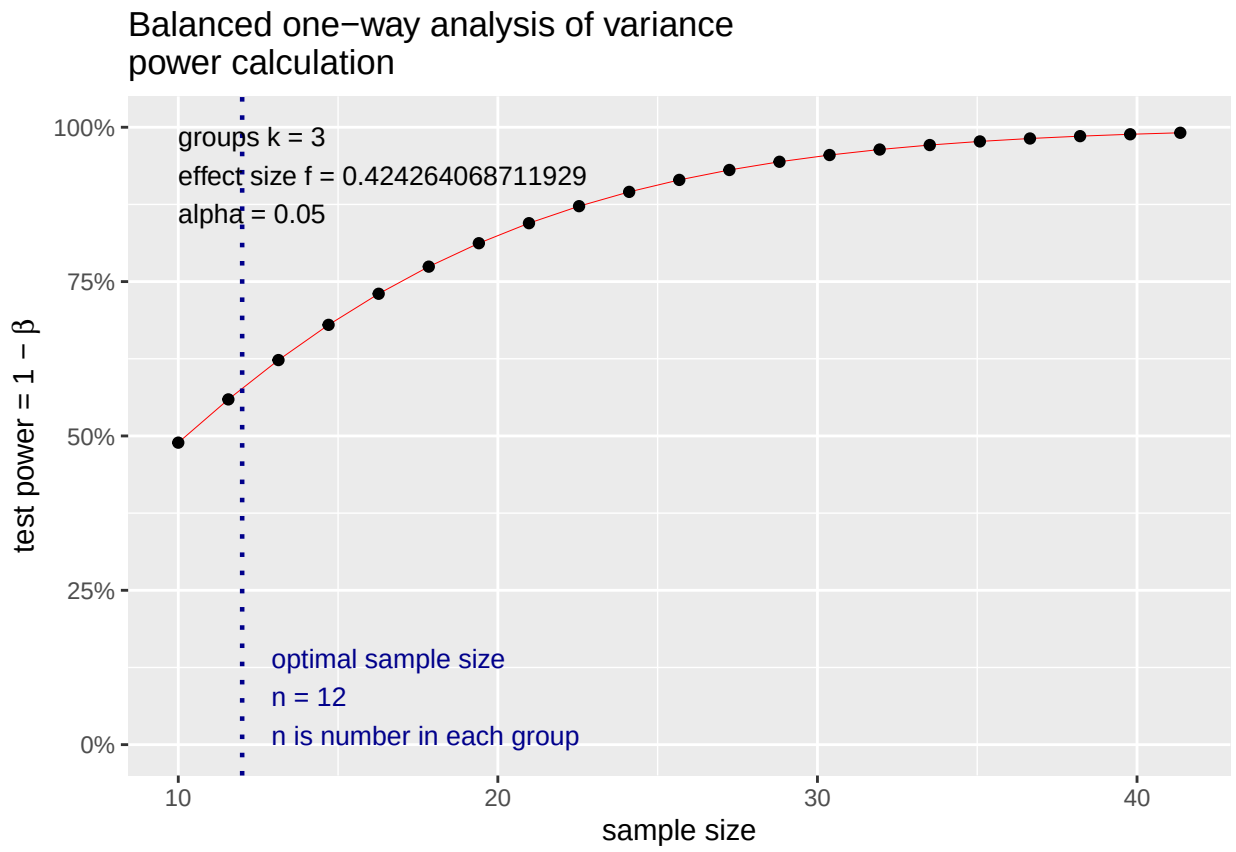
$EffectSize(f) = \frac{d\sqrt{k^2-1}}{2k}$

Sample size determination

```
library(pwr)
pwr.anova.test(k = 3, n = NULL, f = ((0.9*sqrt((3^2)-1))/(2*3)), sig.level = 0.05, power = 0.55)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 3
##               n = 11.35348
##               f = 0.4242641
##       sig.level = 0.05
##           power = 0.55
##
## NOTE: n is number in each group
```

```
plot(pwr.anova.test(k = 3,n=NULL,f=((0.9*sqrt((3^2)-1))/(2*3)),sig.level=0.05,power=0.55))
```



The power test to detect a mean difference with a medium effect, maximum variability and the significance level of 0.05, Gives us the number of samples to collect n = 12 of the each type of ball.

b) Proposing a layout using the number of samples with randomized run order.

The layout generation depends on the type of the design we require, the two major design generators for the Complete Random Design (CRD), Complete Random Block Design (CRBD). Choosing the particular design depends the source of nuisance, if we have any external different sources of nuisance that we need to be considered we choose crbd, but in our experiment we don't have any source of nuisance that effects our results, hence we choose CRD to generate the layout.

Arguments of the layout generation.

Treatment (trt): Number of factors

Replications (r): The number of replications of each type of ball

Seed: Random run code number

Proposed Layout

```
balls<-c("RED","YELLOW","STONE")
library(agricolae)
design<-design.crd(trt = balls, r = 12, seed = 2534356)
design$book
```

```
##    plots  r  balls
## 1    101  1  STONE
## 2    102  1 YELLOW
## 3    103  1    RED
## 4    104  2 YELLOW
## 5    105  3 YELLOW
## 6    106  2    RED
## 7    107  2  STONE
## 8    108  3    RED
## 9    109  3  STONE
## 10   110  4 YELLOW
## 11   111  5 YELLOW
## 12   112  4    RED
## 13   113  5    RED
## 14   114  6    RED
## 15   115  4  STONE
## 16   116  7    RED
## 17   117  5  STONE
## 18   118  8    RED
## 19   119  6  STONE
## 20   120  7  STONE
## 21   121  6 YELLOW
## 22   122  7 YELLOW
## 23   123  8  STONE
## 24   124  9    RED
## 25   125  9  STONE
## 26   126  8 YELLOW
## 27   127  9 YELLOW
## 28   128 10  STONE
## 29   129 11  STONE
## 30   130 10    RED
## 31   131 11    RED
```

```
## 32   132 12    RED
## 33   133 10 YELLOW
## 34   134 11 YELLOW
## 35   135 12 YELLOW
## 36   136 12  STONE
```

Collecting data and record observations on proposed layout in b).

The data is collected according to the pattern of random generated layout in b), following the required conditions of the Pin Elevation and Bungee Position fixed and the Release Angle at 90^o.

Drawing the collected data and Formation of data frames.

```
library(readxl)
dat<-read_excel("C:/Users/Saipa/OneDrive/Desktop/DOE/ProjectData.xlsx")
print(dat)
```

```
## # A tibble: 12 x 3
##    STONE   RED YELLOW
##    <dbl> <dbl>  <dbl>
##  1    44    49     94
##  2    58    53     86
##  3    92    35     49
##  4    60    54     45
##  5    70    64     52
##  6    49    52     68
##  7    37    47     43
##  8    37    42     54
##  9    46    40     44
## 10    97    48     51
## 11    48    45     41
## 12    94    71     41
```

```
library(tidyr)
dat2<-pivot_longer(data = dat, c(RED,YELLOW,STONE))
colnames(dat2)<- c("Type_of_Balls", "Observations")
dat2$Type_of_Balls<-as.factor(dat2$Type_of_Balls)
print(dat2)
```

```
## # A tibble: 36 x 2
##    Type_of_Balls Observations
##    <fct>                <dbl>
##  1 RED                     49
##  2 YELLOW                  94
##  3 STONE                   44
##  4 RED                     53
##  5 YELLOW                  86
##  6 STONE                   58
##  7 RED                     35
##  8 YELLOW                  49
##  9 STONE                   92
## 10 RED                     54
## # ... with 26 more rows
```

d) Performing hypothesis test and check residuals. Be sure to comment and take corrective action if necessary.

Hypothesis to be Tested

1. Null Hypothesis (H_o) : $\mu_1 = \mu_2 = \mu_3$
2. Alternative Hypothesis (H_a) : Atleast one of μ_i differs.

where;

μ_1 = Mean of Red Ball.

μ_2 = Mean of Yellow Ball.

μ_3 = Mean of Stone Ball.

The hypothesis could be tested using Analysis of Varience (anova) or Non Parmetric test (Kruskal-Wallis rank sum test), but the Non Parametric testing reduces the power of the Hypothesis. Hence ideal option for hypothesis testing is anova.

The anova has few of the strong assumptions about the data which has to be fulfilled before the anova is used to test the hypothesis, if this assumptions are not meet the results of the anova will be effects and may lead to wrong conclusion.

Assumptions of ANOVA.

The assumption on normal distribution of observations of each type of ball (This assumption is not strong, could be exempted to certain extent).

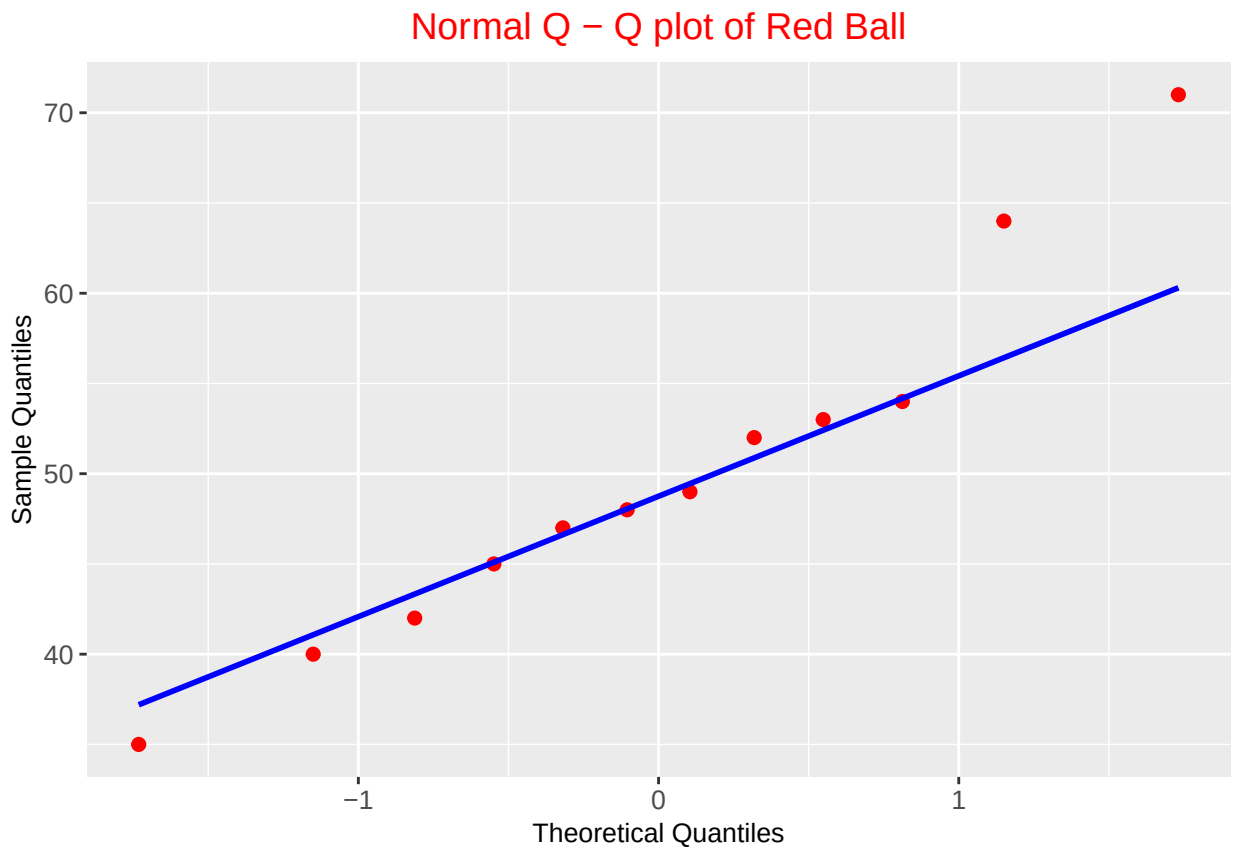
The assumption on constant variance (This is strong assumption, can't be voliated).

Intially choosing ANOVA to test the Hypothesis.

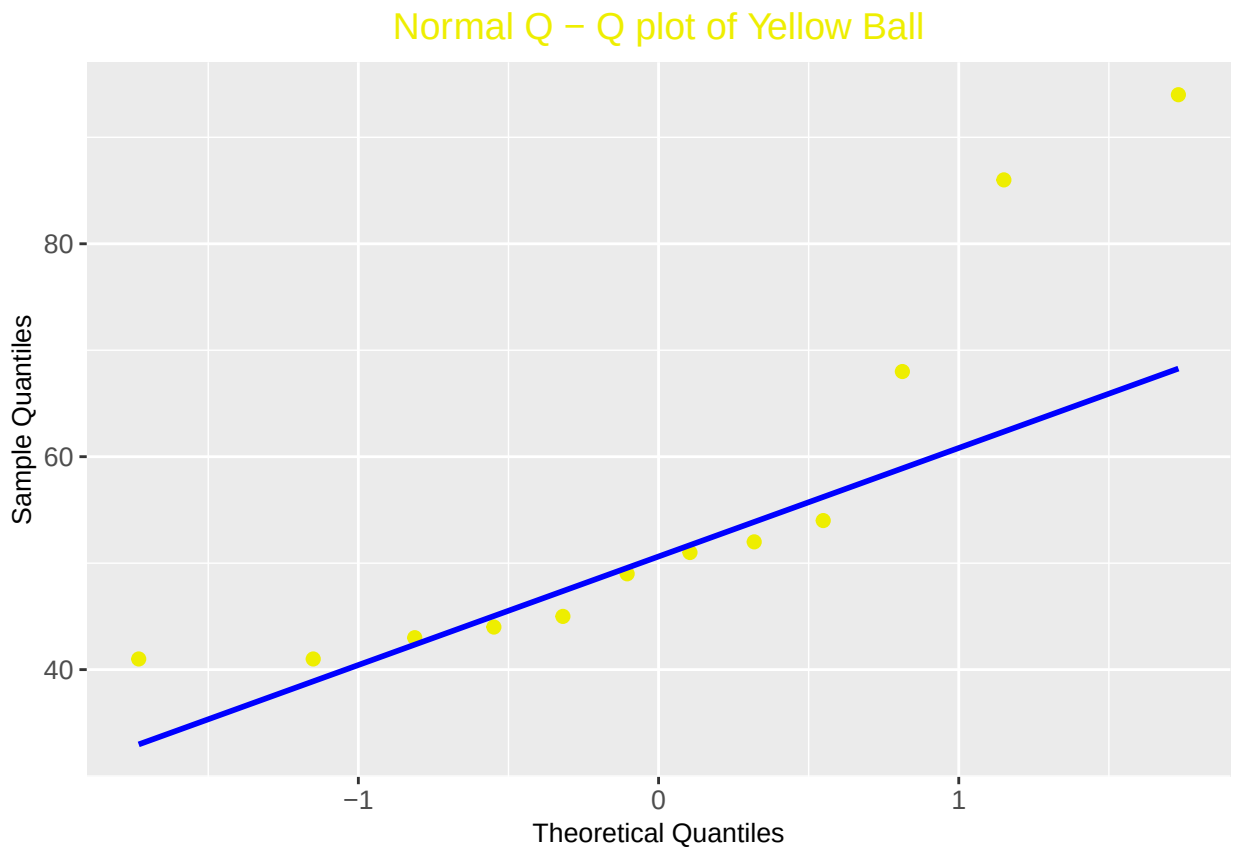
First Checking for Normality Assumption

Normal Probability plots

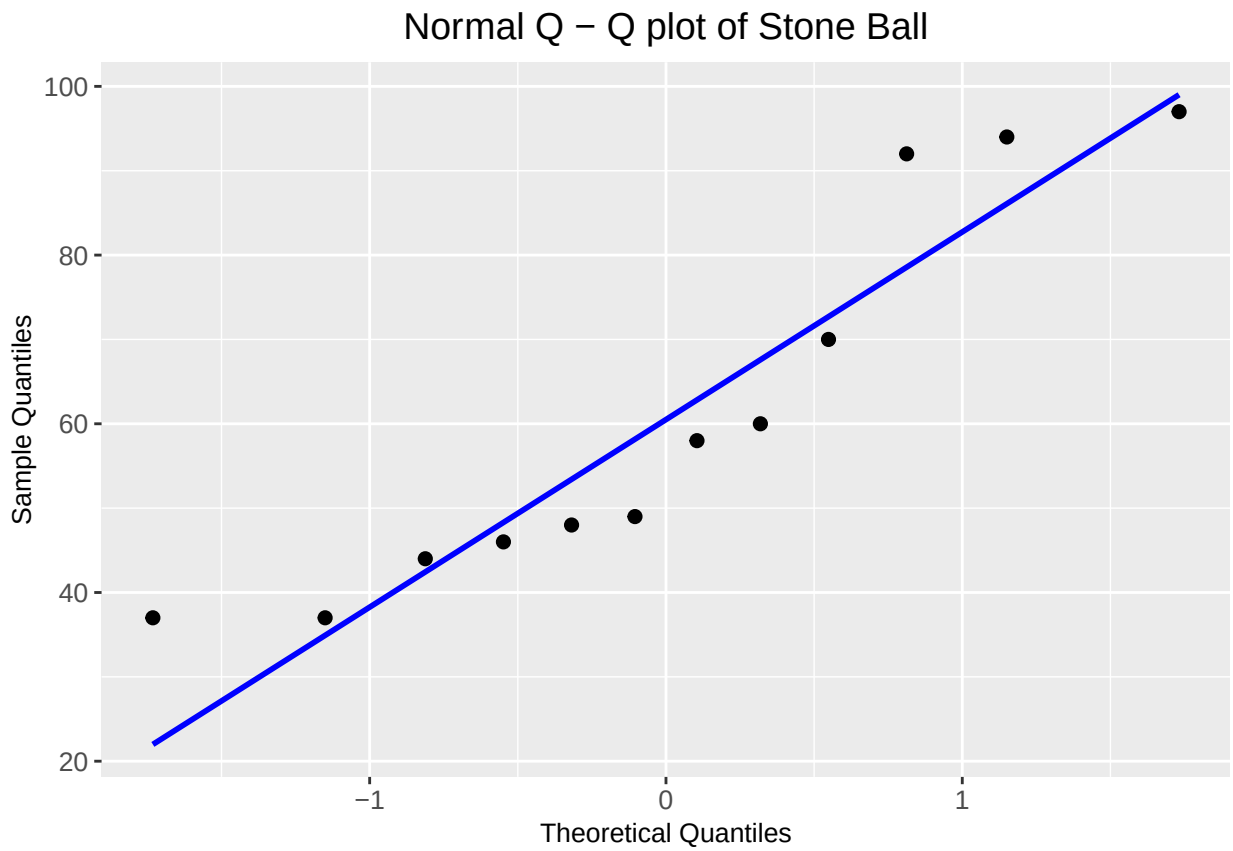
```
library(ggplot2)

ggplot(mapping = aes(sample = dat$RED)) +
  stat_qq(size = 2, col = "red") +
  stat_qq_line(size = 1, col = "blue") +
  xlab("Theoretical Quantiles") + ylab("Sample Quantiles") +
  ggtitle("Normal Q - Q plot of Red Ball", ) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=10),
        axis.title=element_text(size=10),
        plot.title = element_text(size=14, colour = "red" ))
```




```
ggplot(mapping = aes(sample = dat$YELLOW)) +
  stat_qq(size = 2, col = "yellow2") +
  stat_qq_line(size = 1, col = "blue") +
  xlab("Theoretical Quantiles") + ylab("Sample Quantiles") +
  ggtitle("Normal Q - Q plot of Yellow Ball", ) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=10),
        axis.title=element_text(size=10),
        plot.title = element_text(size=14, colour = "yellow2"))
```




```
ggplot(mapping = aes(sample = dat$STONE)) +
  stat_qq(size = 2, col = "black") +
  stat_qq_line(size = 1, col = "blue") +
  xlab("Theoretical Quantiles") + ylab("Sample Quantiles") +
  ggtitle("Normal Q - Q plot of Stone Ball", ) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=10),
        axis.title=element_text(size=10),
        plot.title = element_text(size=14, colour = "black"))
```



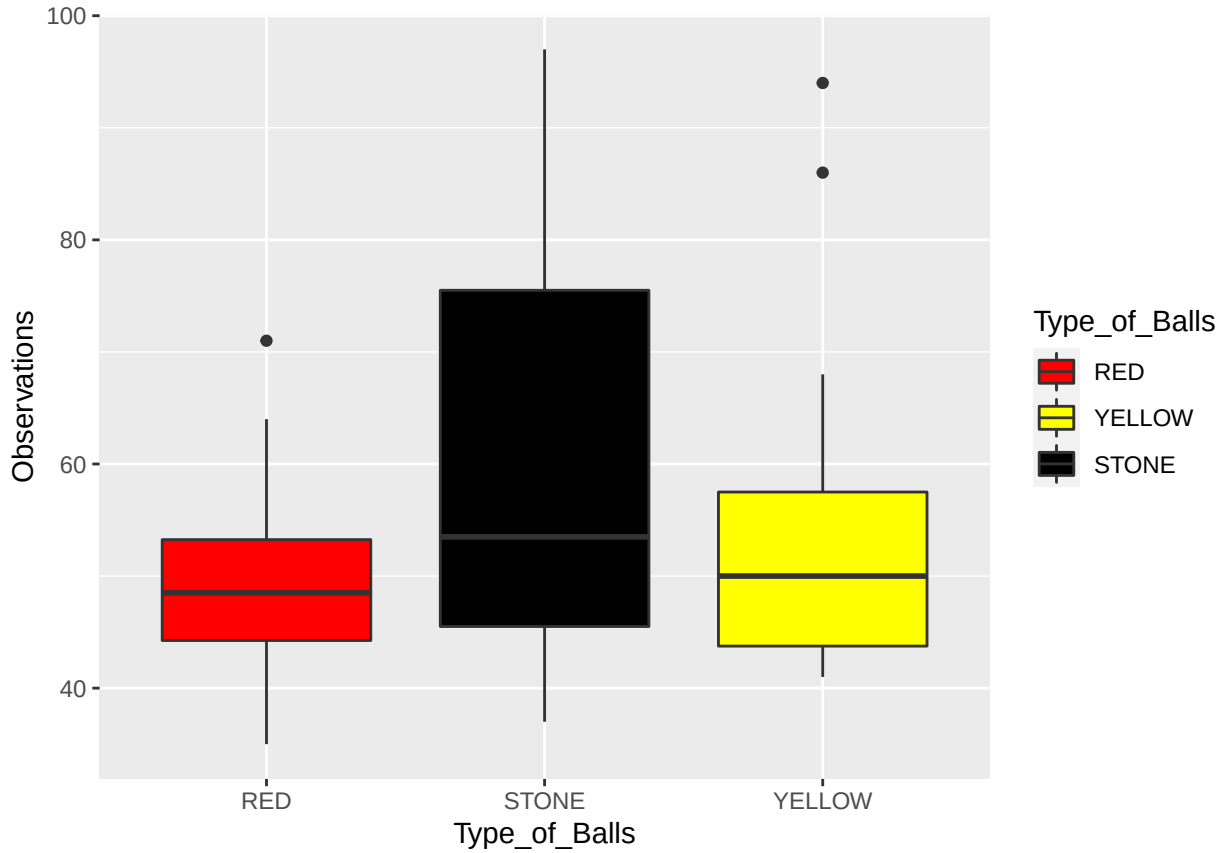
Comments: The Normal Probability plots of the all type of balls looks normally distributed; but Yellow and red balls has tails slightly out drifted. Could be considered the observations are normally distributed.

Checking for constant variance

The boxplot is best and easy way to check for the constant varience, by comparing the median and quartile ranges.

Box plot

```
library(ggplot2)
ggplot(dat2, aes(x = Type_of_Balls, y = Observations, fill = Type_of_Balls)) +
  geom_boxplot() + scale_fill_manual(breaks = dat2$Type_of_Balls,
                    values = c("RED","YELLOW","BLACK"))
```

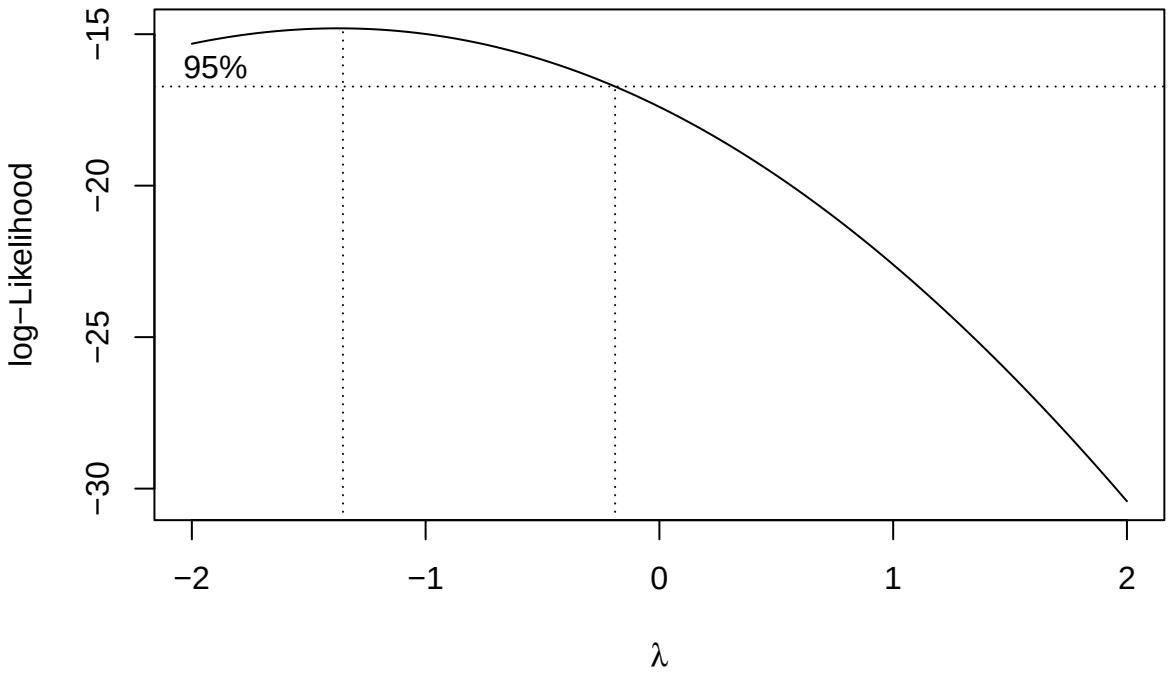


Comments: Clearly there is no constant variance seen. Every ball has different quantile ranges , so obviously we will have different variances across each type of ball.Therefore we can't go with parametric method(Anova) as it has strong assumption on constant variance. But we will try to transform the data and look for constant variance so that there is a chance that we can still go with anova if we got our variance as constant.

The above box plot could not give us the favorable results to use anova for Hypothesis testing, hence we will try transforming the data to stabilize the variance using the Boxcox transformation.

Box cox Transformation

```
library(MASS)
boxcox(dat2$Observations~dat2$Type_of_Balls)
```



The Box cox transformation gives us the λ within the 95% of confidence interval, choosing the λ as -1 at the peak of curve.

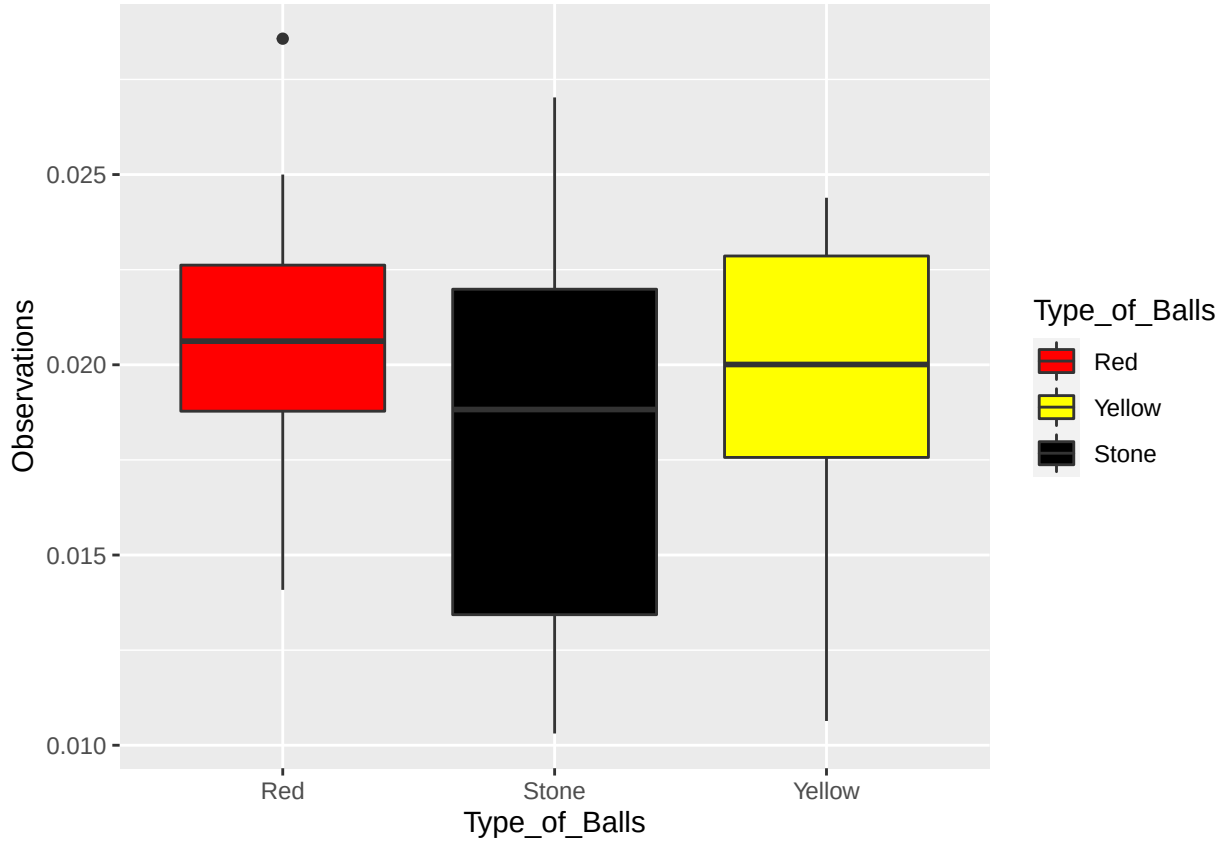
```
Red<-c(dat$RED)^-1
Yellow<-c(dat$YELLOW)^-1
Stone<-c(dat$STONE)^-1

dat3<-cbind.data.frame(Red, Yellow, Stone)

library(tidyr)
dat4<-pivot_longer(data = dat3, c(Red,Yellow,Stone))
colnames(dat4)<- c("Type_of_Balls", "Observations")
dat4$Type_of_Balls<-as.factor(dat4$Type_of_Balls)
print(dat4)
```

```
## # A tibble: 36 x 2
##    Type_of_Balls Observations
##    <fct>                <dbl>
##  1 Red                 0.0204
##  2 Yellow              0.0106
##  3 Stone               0.0227
##  4 Red                 0.0189
##  5 Yellow              0.0116
##  6 Stone               0.0172
##  7 Red                 0.0286
##  8 Yellow              0.0204
```

```
##  9 Stone               0.0109
## 10 Red                 0.0185
## # ... with 26 more rows
```

Boxplot of newly Transformed Data

```
library(ggplot2)
ggplot(dat4, aes(x = Type_of_Balls, y = Observations, fill = Type_of_Balls)) +
  geom_boxplot() + scale_fill_manual(breaks = dat4$Type_of_Balls,
                                     values = c("red","yellow","black"))
```



Comments: Box-cox also didn't helped us(There is no constant variance by transforming the data also) so there fore we can't perform anova we need to go for non- parametric method to check(Conclude) for our hypothesis.

Non-parametric test

```
kruskal.test(dat2$Observations~dat2$Type_of_Balls)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  dat2$Observations by dat2$Type_of_Balls
## Kruskal-Wallis chi-squared = 0.89139, df = 2, p-value = 0.6404
```

Comments: The Kruskal-Wallis rank sum test gives us $p-value = 0.6404$ which is very greater in comparison with significant level $\alpha = 0.05$, Hence we could say, we don't have sufficient strong evidence to **reject the**

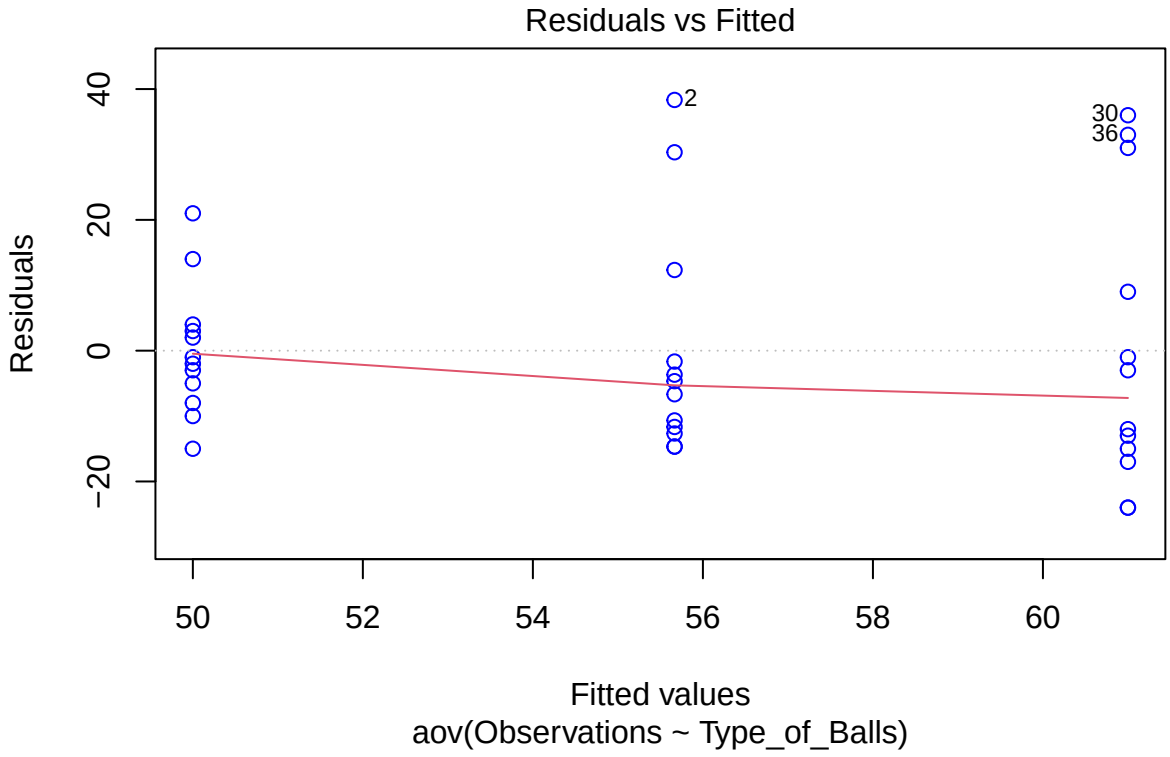
Null Hypothesis.Hence, we need to accept(Null Hypothesis) that means of different types of balls are equal.

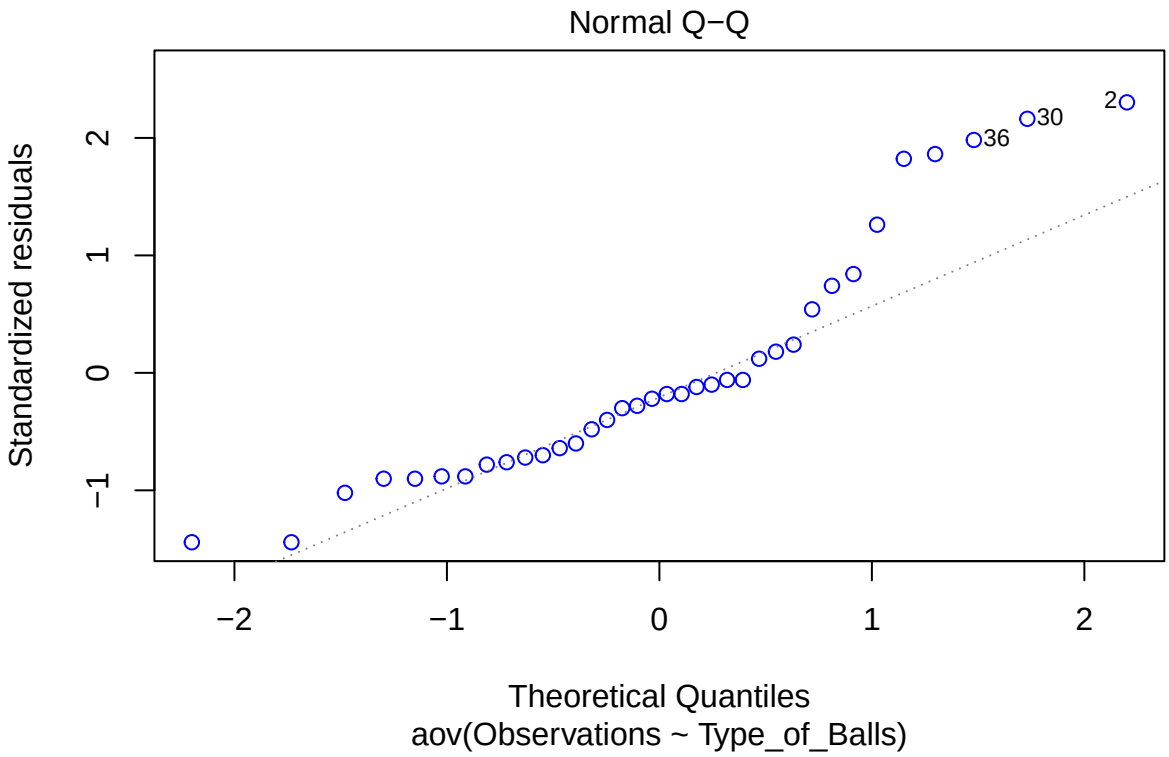
Conculsion: Concluding from the analysis, we could say that the Type of The balls has no significant effect on the distance in which the ball is thrown for the fixed Pin Elevation and Bungee Position, with Release angle stretched to of 90^o.

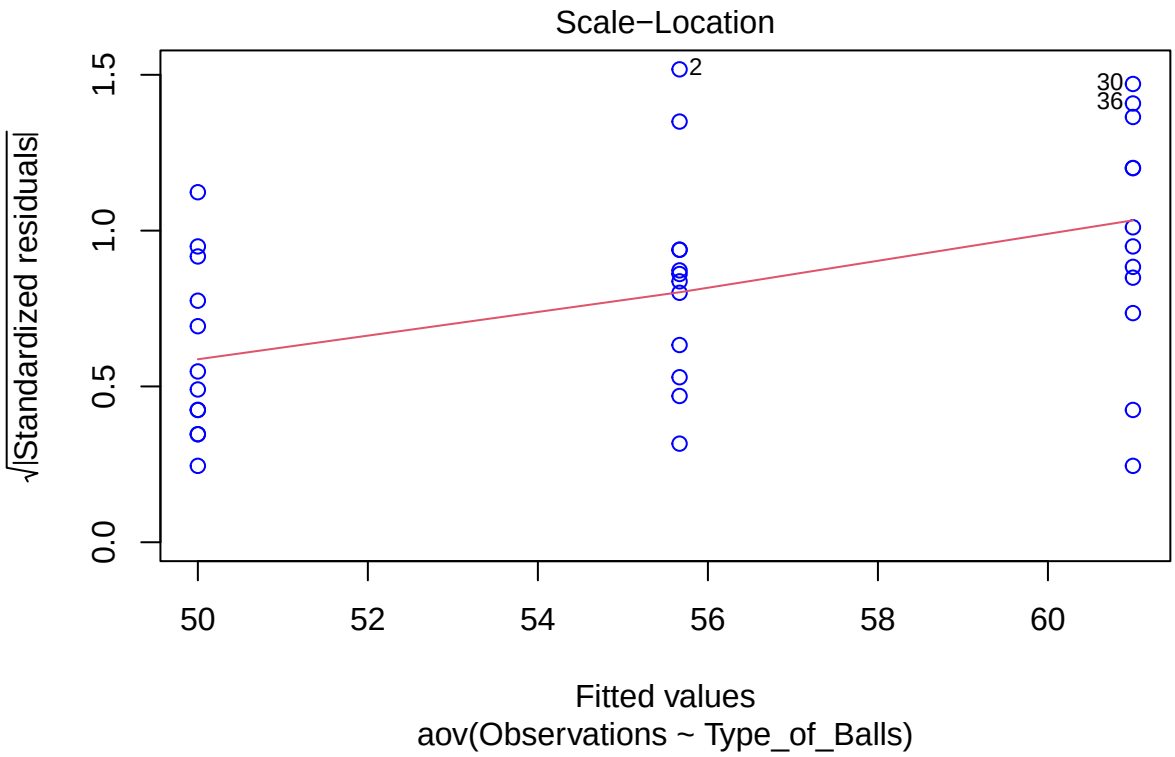
Investigating further about the model adequacy, to check for the outliners using the residuals.

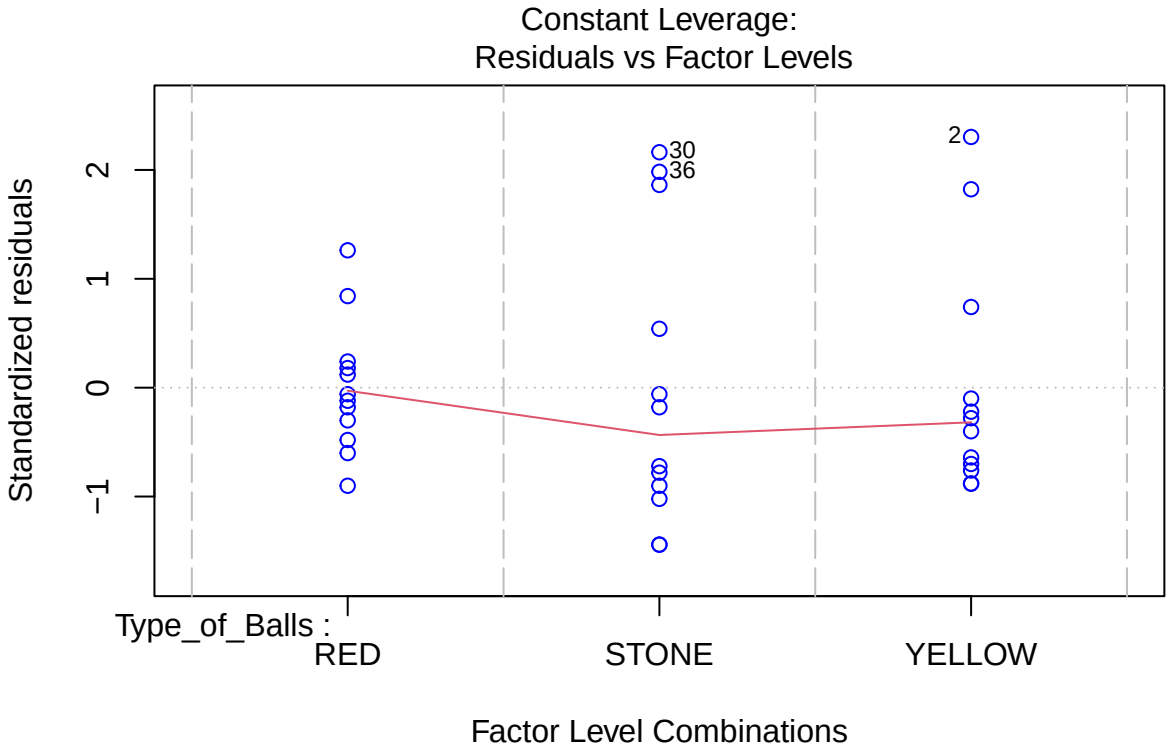
Residual plots

```
aov_model<-aov(Observations~Type_of_Balls, data = dat2)
residules<-resid(aov_model)
plot(aov_model, col = "Blue")
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
aov(Observations ~ Type_of_Balls)
2
30
36
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2
30
36
50
52
54
56
58
60
Fitted values
aov(Observations ~ Type_of_Balls)



Comments: The residuals plots gives us that the model is inadequate with the few of the outlines, but the normal probability plots of the residuals looks almost normally distributed with some points out drifted.

As we failed to reject Null-Hypothesis but we will provide Additional evidence from Tukey's Honest Significant Differences Test. If we can see ZERO in our confidence interval it means that pairs are not differing in mean that they have almost same means.We know that it is Parametric method (we can't perform it on our data) but we are just using it as further evidence that's it.

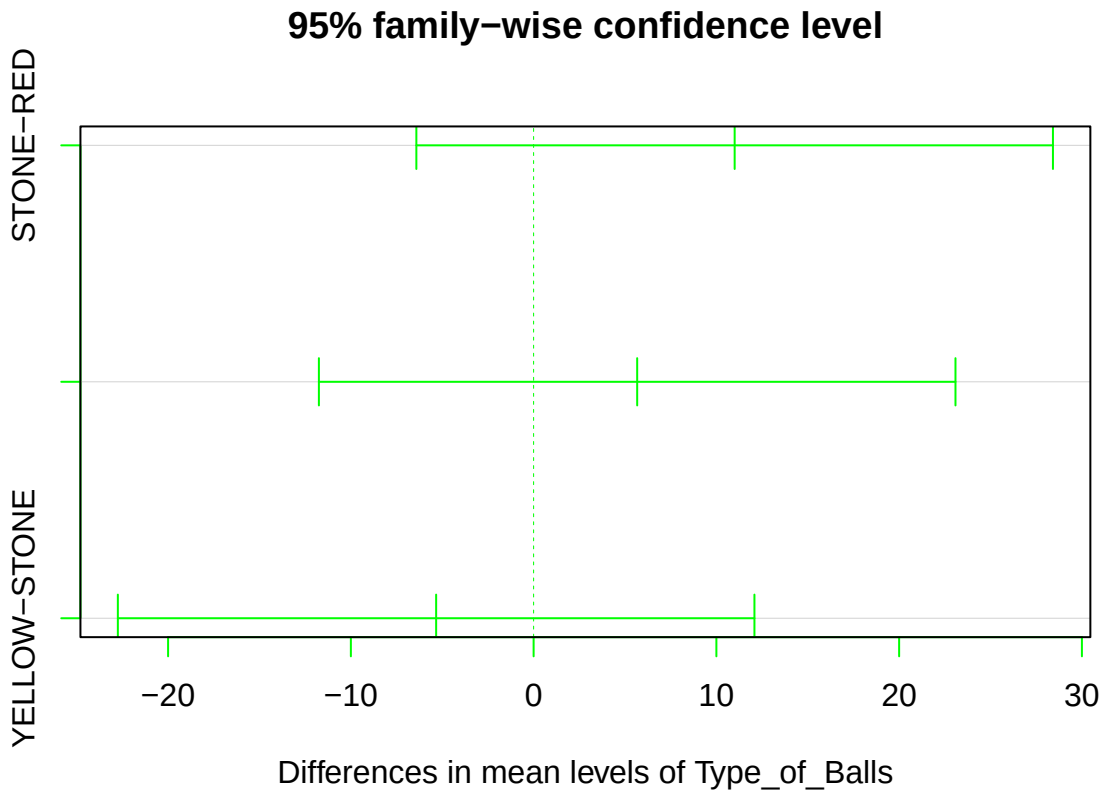
Pairwise comparisons.

Tukey's Honest Significant Differences

```
TukeyHSD(aov_model, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Observations ~ Type_of_Balls, data = dat2)
##
## $Type_of_Balls
##                  diff        lwr      upr     p adj
## STONE-RED    11.000000  -6.416256 28.41626 0.2814358
## YELLOW-RED    5.666667 -11.749589 23.08292 0.7066231
## YELLOW-STONE -5.333333 -22.749589 12.08292 0.7349004
```

```
plot(TukeyHSD(aov_model, conf.level = 0.95), col = "green")
```



Comments: As we assumed we got the result. None of the pairs are differing in the means.As we can conclude it from the above plot zero is in every Confidence interval of the pairs it means that they are not differing in means.

Conculsion: Concluding from the analysis, we could say that the Type of The balls has no significant effect on the distance in which the ball is thrown for the fixed Pin Elevation and Bungee Position, with Release angle stretched to of 90^o.

Source Code

```
# Power Test
library(pwr)
pwr.anova.test(k = 3, n = NULL, f = ((0.9*sqrt((3^2)-1))/(2*3)), sig.level = 0.05, power = 0.55)
plot(pwr.anova.test(k = 3,n=NULL,f=((0.9*sqrt((3^2)-1))/(2*3)),sig.level=0.05,power=0.55))
## The number of samples to collect n = 36(12 from each group).

# Laying out the number of samples from part (a) with randomized run order.
balls<-c("RED","YELLOW","STONE")
library(agricolae)
design<-design.crd(trt = balls, r = 12, seed = 2534356)
design$book

#Formation of data frame for the collected data for analysis
library(readxl)
dat<-read_excel("C:/Users/Saipa/OneDrive/Desktop/DOE/ProjectData.xlsx")
print(dat)
```

```r
library(tidyr)
dat2<-pivot_longer(data = dat, c(RED,YELLOW,STONE))
colnames(dat2)<- c("Type_of_Balls", "Observations")
dat2$Type_of_Balls<-as.factor(dat2$Type_of_Balls)
print(dat2)


# Normal Probability Plots.
library(ggplot2)

ggplot(mapping = aes(sample = dat$RED)) +
  stat_qq(size = 2, col = "red") +
  stat_qq_line(size = 1, col = "blue") +
  xlab("Theoretical Quantiles") + ylab("Sample Quantiles") +
  ggtitle("Normal Q - Q plot of Red Ball", ) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=10),
        axis.title=element_text(size=10),
        plot.title = element_text(size=14, colour = "red" ))

ggplot(mapping = aes(sample = dat$YELLOW)) +
  stat_qq(size = 2, col = "yellow2") +
  stat_qq_line(size = 1, col = "blue") +
  xlab("Theoretical Quantiles") + ylab("Sample Quantiles") +
  ggtitle("Normal Q - Q plot of Yellow Ball", ) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=10),
        axis.title=element_text(size=10),
        plot.title = element_text(size=14, colour = "yellow2"))

ggplot(mapping = aes(sample = dat$STONE)) +
  stat_qq(size = 2, col = "black") +
  stat_qq_line(size = 1, col = "blue") +
  xlab("Theoretical Quantiles") + ylab("Sample Quantiles") +
  ggtitle("Normal Q - Q plot of Stone Ball", ) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=10),
        axis.title=element_text(size=10),
        plot.title = element_text(size=14, colour = "black"))


# Box plot
library(ggplot2)
ggplot(dat2, aes(x = Type_of_Balls, y = Observations, fill = Type_of_Balls)) +
  geom_boxplot() + scale_fill_manual(breaks = dat2$Type_of_Balls,
                    values = c("RED","YELLOW","BLACK"))

# Transforming the data
# Box-cox Transformation
library(MASS)
boxcox(dat2$Observations~dat2$Type_of_Balls)
# $\lambda = -1$
Red<-c(dat$RED)^-1
```

```r
Yellow<-c(dat$YELLOW)^-1
Stone<-c(dat$STONE)^-1

dat3<-cbind.data.frame(Red, Yellow, Stone)

library(tidyr)
dat4<-pivot_longer(data = dat3, c(Red,Yellow,Stone))
colnames(dat4)<- c("Type_of_Balls", "Observations")
dat4$Type_of_Balls<-as.factor(dat4$Type_of_Balls)
print(dat4)


library(ggplot2)
ggplot(dat4, aes(x = Type_of_Balls, y = Observations, fill = Type_of_Balls)) +
  geom_boxplot() + scale_fill_manual(breaks = dat4$Type_of_Balls,
                                     values = c("red","yellow","black"))

## Non parametric test
kruskal.test(dat2$Observations~dat2$Type_of_Balls)

# Residuals plots
aov_model<-aov(Observations~Type_of_Balls, data = dat2)
residules<-resid(aov_model)
plot(aov_model, col = "Blue")

# Tukey's HSD Test
TukeyHSD(aov_model, conf.level = 0.95)
plot(TukeyHSD(aov_model, conf.level = 0.95), col = "green")
```