Problem Set 3 Code

Daniel Crownover

2024-02-05

Question 1

1.1 State the H0 and Ha of your test - 2.5 H_0: the average weight of elephant tusk has stayed the same over time

H_a: the average weight of poached elephant tusks has changed over time.

1.2 Read in Data

```
TuskData <- read.csv("/Users/daniel/Downloads/BIO 591 Coding Files/Problem Set/Problem Set 3/Problem.se
```

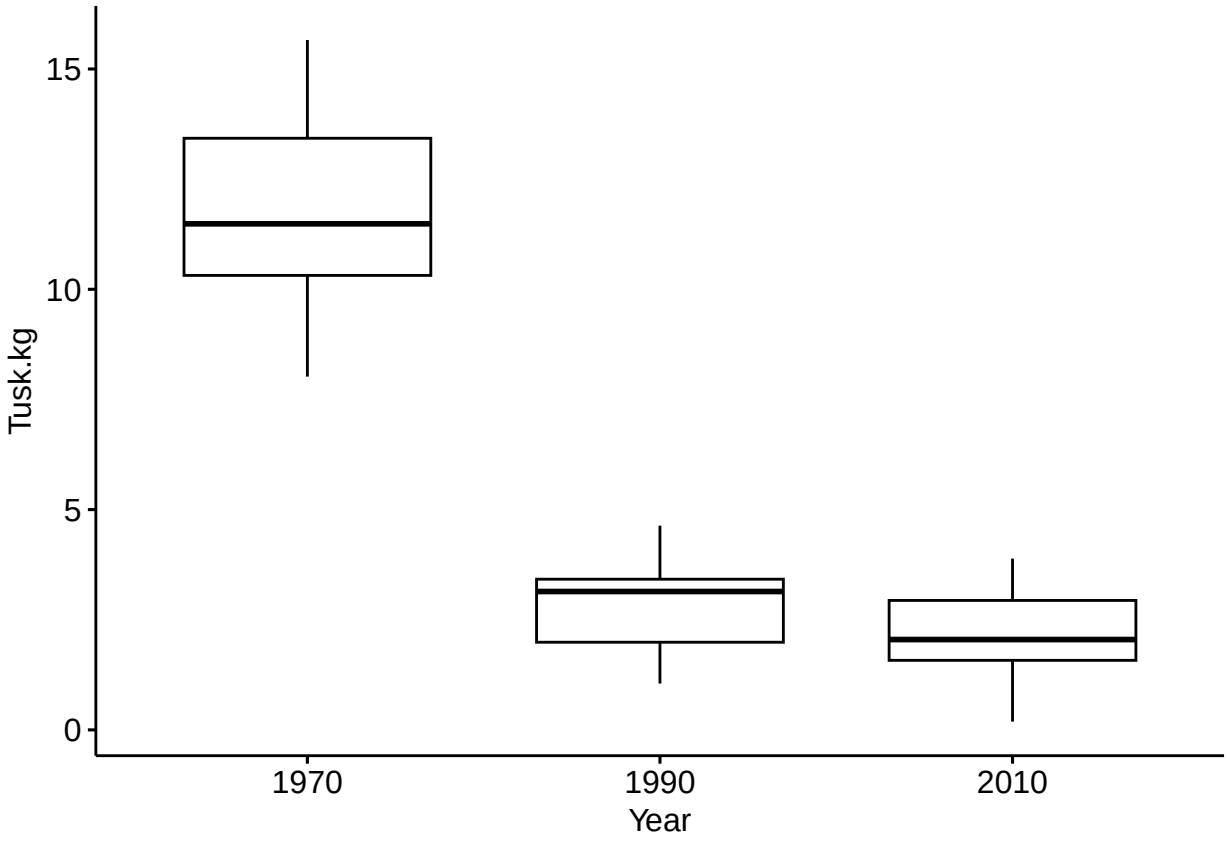
1.3 Summary Statistics of Tusk Data

```
TuskData %>%
  group_by(Year) %>%
  get_summary_stats(Tusk.kg, type = "mean_sd")
```

```
## # A tibble: 3 x 5
##    Year variable     n  mean    sd
##   <int> <fct>    <dbl> <dbl> <dbl>
## 1  1970 Tusk.kg     20 11.7  2.24
## 2  1990 Tusk.kg     20  2.92 1.09
## 3  2010 Tusk.kg     20  2.17 0.927
```

1.4 Create a box plot of Poached Tusk Weight by year:

```
ggboxplot(TuskData, x = "Year", y = "Tusk.kg")
```



1.5 Identify Outliers

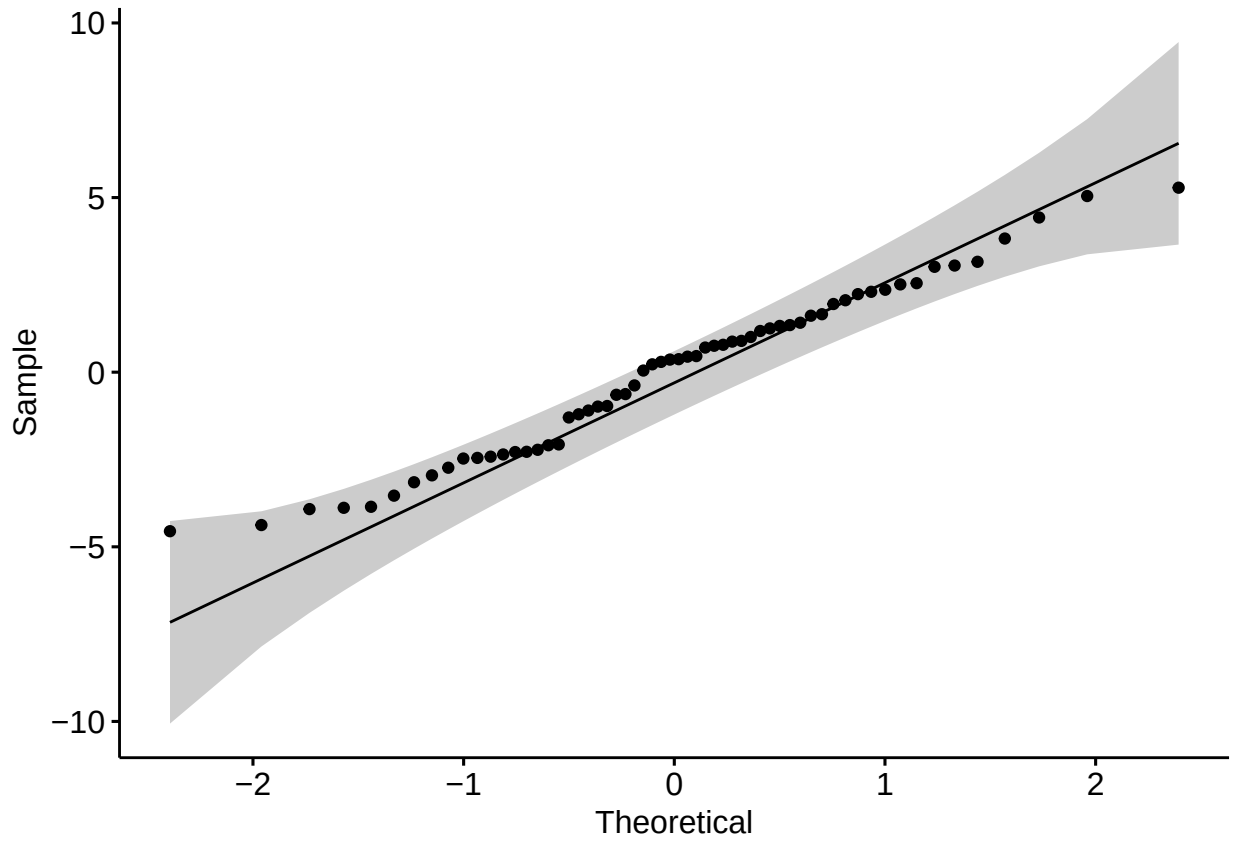
```
TuskData %>%
  group_by(Year) %>%
  identify_outliers(Tusk.kg)
```

```
## [1] Year       Tusk.kg    is.outlier is.extreme
## <0 rows> (or 0-length row.names)
```

There were no extreme outliers. in the situation where you have extreme outliers, this can be due to: 1) data entry errors, measurement errors or unusual values

1.6 Build Linear Model

```
# Build the linear model
model  <- lm(Tusk.kg ~ Year, data = TuskData)
# Create a QQ plot of residuals
ggqqplot(residuals(model))
```




```
# Compute Shapiro-Wilk test of normality
shapiro_test(residuals(model))
```

```
## # A tibble: 1 x 3
##   variable         statistic p.value
##   <chr>                <dbl>   <dbl>
## 1 residuals(model)     0.977   0.323
```

A Shapiro-Wilk test result with a p-value of 0.322 suggests that there is no significant evidence to reject the null hypothesis of normality at the conventional significance level (e.g., 0.05)

OR

With a Shapiro-Wilk test yielding a p-value of 0.33, there isn't compelling evidence to dismiss the null hypothesis, which assumes the data follows a normal distribution, at the standard significance level of 0.05.

1.7 Shapiro Wilks Normality Test per Year

```
TuskData %>%
  group_by(Year) %>%
  shapiro_test(Tusk.kg)
```

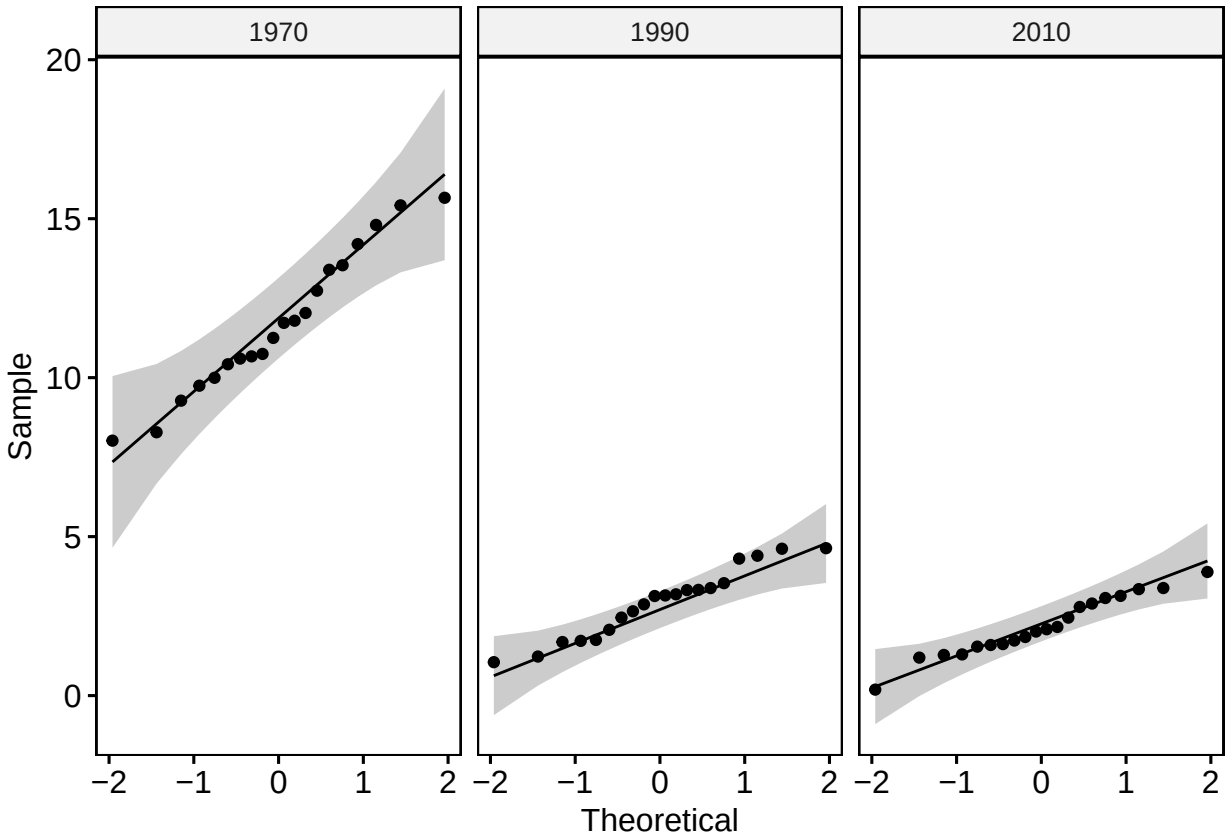
```
## # A tibble: 3 x 4
##    Year variable statistic     p
```

```
##   <int> <chr>         <dbl> <dbl>
## 1  1970 Tusk.kg       0.966 0.667
## 2  1990 Tusk.kg       0.950 0.366
## 3  2010 Tusk.kg       0.971 0.784
```

With a Shapiro-Wilk test yielding a p-value of 0.322, there isn't compelling evidence to dismiss the null hypothesis, which assumes the data follows a normal distribution, at the standard significance level of 0.05.

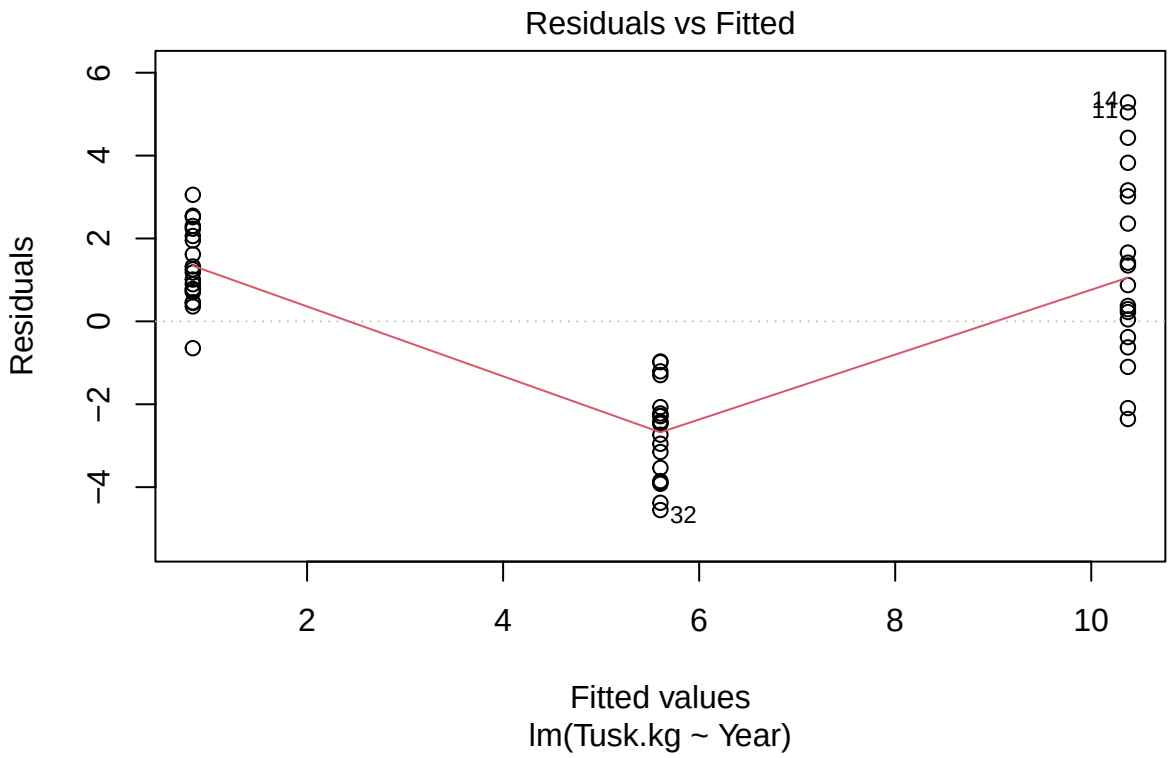
1.8 Creating a QQ plot for each Year:

```
ggqqplot(TuskData, "Tusk.kg", facet.by = "Year")
```



The residuals versus fits plot can be used to check the homogeneity of variances. ## 1.9 Checking for homogeneity of vairances

```
plot(model, 1)
```




```
str(TuskData)
```

```
## 'data.frame':    60 obs. of  2 variables:
##  $ Tusk.kg: num  12.74 12.03 13.54 9.75 11.72 ...
##  $ Year   : int  1970 1970 1970 1970 1970 1970 1970 1970 1970 1970 ...
```

In the plot above, there is evident relationships between residuals and fitted values (the mean of each groups) So, we can't assume the homogeneity of variances. If these where equal this line would be flat

1.9.1 Levene's test to check the homogeneity of variances:

```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:arm':
##
##     logit
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
TuskData$Year <- as.factor(TuskData$Year)
str(TuskData)
```

```
## 'data.frame':    60 obs. of  2 variables:
##  $ Tusk.kg: num  12.74 12.03 13.54 9.75 11.72 ...
##  $ Year   : Factor w/ 3 levels "1970","1990",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
library(rstatix)

TuskData %>% levene_test(Tusk.kg ~ Year)
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic        p
##   <int> <int>     <dbl>    <dbl>
## 1     2    57      8.67 0.000517
```

```
#this is how you do it in base R
summary(aov(Tusk.kg ~ Year, data = TuskData))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Year         2 1125.8   562.9   238.1 <2e-16 ***
## Residuals   57  134.7     2.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.9.2 Describe how you checked the assumptions of your statistical test: All of the preliminary test passed the assumptions of normailty. No significant outliers where detected. The liner model and Shapiro-Wilk test resulted a p-value of 0.322, this suggested that there is no significant evidence to reject the null hypothesis of normality at the conventional significance level (e.g., 0.05). We can assume normality with the Tusk data set and can continue. When testing for homogeneity of variances, we found the plot to be unequal, the Levenes test showed that our p value(0.000517) is well below the critical value of .005, which means that we do have sufficient evidence to reject the null hypothesis(that the variances are equal).

because the classical one-way ANOVA test requires an assumption of equal variances for all groups. With the tusk data, the homogeneity of variance assumption turned out to fail the test: the Levene test was significant.

We used the Welch one-way test as an alternative to the standard one-way ANOVA because this situation is one where the homogeneity of variance can't be assumed (i.e., Levene test is significant).

We used the the Games-Howell post hoc test or pairwise t-tests (with no assumption of equal variances) can be used to compare all possible combinations of group differences.

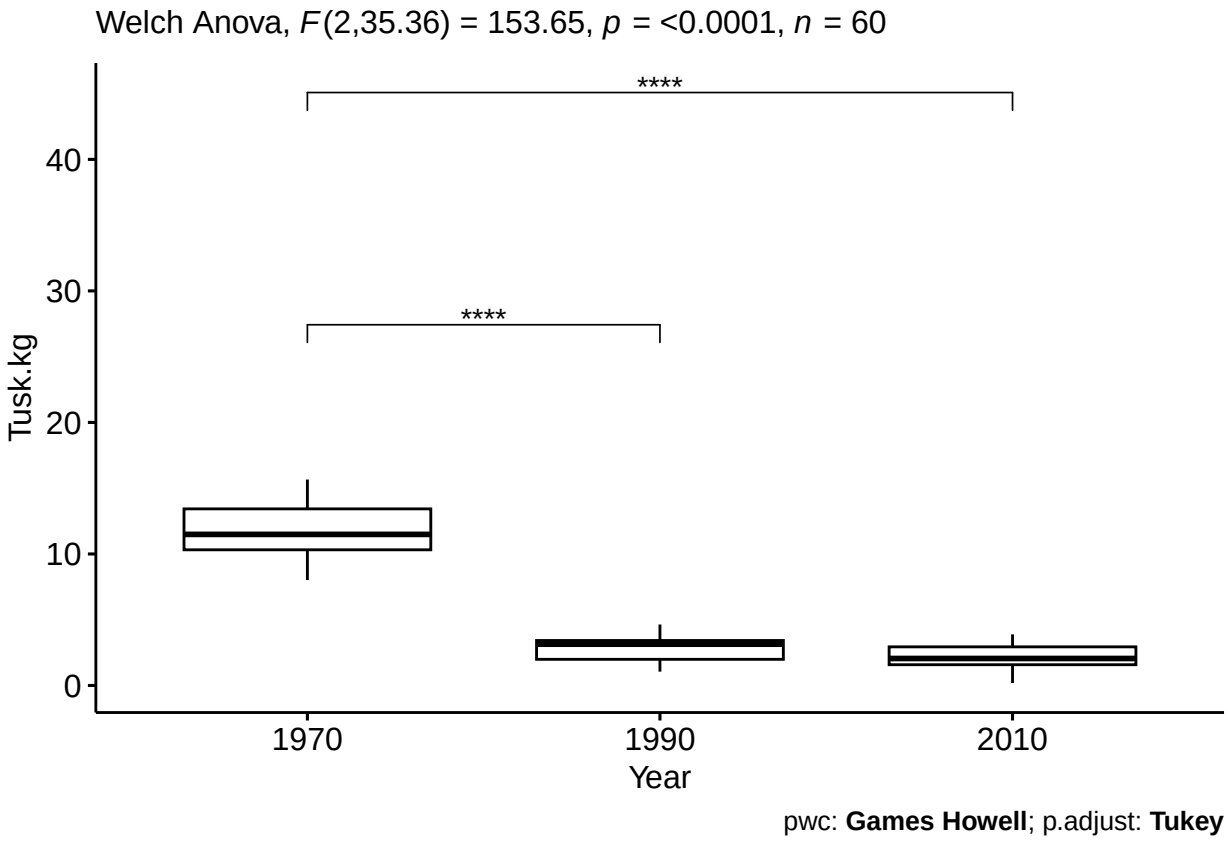
1.9.3 Non Parametric

1.9.4 Welch test –> always use welch when variance are not equal(levene test fails)

```r
TuskData$Year <- as.factor(TuskData$Year)

library(rstatix)
library(ggplot2)

# Welch One way ANOVA test
res.aov2 <- TuskData %>% welch_anova_test(Tusk.kg ~ Year)
# Pairwise comparisons (Games-Howell)
pwc2 <- TuskData %>% games_howell_test(Tusk.kg ~ Year) ## adjusted turkey test
# Visualization: box plots with p-values
pwc2
```

```
## # A tibble: 3 x 8
##   .y.     group1 group2 estimate conf.low conf.high    p.adj p.adj.signif
## * <chr>   <chr>  <chr>     <dbl>    <dbl>     <dbl>    <dbl> <chr>
## 1 Tusk.kg 1970   1990     -8.79    -10.2    -7.41   5.55e-16 ****
## 2 Tusk.kg 1970   2010     -9.54    -10.9    -8.19   4.25e-14 ****
## 3 Tusk.kg 1990   2010     -0.748    -1.53    0.0345 6.3 e- 2 ns
```

```r
pwc2 <- pwc2 %>% add_xy_position(x = "Year", step.increase = 1)
ggboxplot(TuskData, x = "Year", y = "Tusk.kg") +
  stat_pvalue_manual(pwc2, hide.ns = TRUE) +
  labs(
    subtitle = get_test_label(res.aov2, detailed = TRUE),
    caption = get_pwc_label(pwc2)

    )
```



1.9.5 Report the results of the ANOVA and a post-hoc test

from the Tukey test, the F test was able to compute these values across these groups regardless of its unequal variance.

The differences between average tusk weights from 1970 to 1990 (p = 5.55e-16) and from 1970 - 2010(p = 4.25e-14) both had p values that are super statistically significant from between these years. Knowing this, we can reject the null(which is they the weights stay the same) and it is safe to assume that they changed from these time periods. The 1990 - 2010 years (p = 6.30e-02) was not statisticly significant and did not change between those years. The cooresponding box plot shows the significance level(p value) between years represented by the hovering black lines and astrics(showing level of significance) of the mean tusk weights from 1970, 1990 and 2010.

```
pwc3 <- TuskData %>%
  pairwise_t_test(
   Tusk.kg ~ Year, pool.sd = FALSE,
    p.adjust.method = "bonferroni"
    )
pwc3
```

```
## # A tibble: 3 x 10
##   .y.   group1 group2    n1    n2 statistic    df        p    p.adj p.adj.signif
## * <chr> <chr>  <chr>  <int> <int>     <dbl> <dbl>    <dbl>    <dbl> <chr>
## 1 Tusk~ 1970   1990      20    20     15.8   27.5 2.64e-15 7.92e-15 ****
## 2 Tusk~ 1970   2010      20    20     17.6   25.3 1.07e-15 3.21e-15 ****
## 3 Tusk~ 1990   2010      20    20      2.33  37.0 2.5 e- 2 7.5 e- 2 ns
```

This is the bonferroni test and just accounts for type one error, which is fasly rejecting the null. It shows

the same answers with 1970 - 1990 (p = 7.92e-15) and 1970 - 2010 (p = 3.21e-15) both being statisticly significant, which means we can we reject the null.

Question 2

2.0 State H_0 and H_a

H_0_1 The results between each antibiotic is the same H_0_2 The results stayed the same between Lab groups

H_a_1 The results between each antibiotic where different H_a_2 The results between each lab group where different

2.1 Read in data

```
lab <- c(rep(1:4, each = 4))
antibiotic <- rep(c(1:4), 4)
results <- (c(9.3, 9.4, 9.6, 10, 9.4, 9.3, 9.8, 9.9, 9.2,
9.4, 9.5, 9.7, 9.7, 9.6, 10, 10.2))
dlabs <- data.frame(lab = factor(lab),
antibiotic = factor(antibiotic), results)
```

2.2 Renaming

```
dlabs$lab <- factor(dlabs$lab,
                  levels = c(1,2,3,4),
                  labels = c("lab 1", "lab 2", "lab 3", "lab 4"))

dlabs$antibiotic <- factor(dlabs$antibiotic,
                  levels = c(1,2,3,4),
                  labels = c("Antibiotic 1", "Antibiotic 2", "Antibiotic 3", "Antibiotic 4"))

dlabs
```

```
##      lab   antibiotic results
## 1  lab 1 Antibiotic 1     9.3
## 2  lab 1 Antibiotic 2     9.4
## 3  lab 1 Antibiotic 3     9.6
## 4  lab 1 Antibiotic 4    10.0
## 5  lab 2 Antibiotic 1     9.4
## 6  lab 2 Antibiotic 2     9.3
## 7  lab 2 Antibiotic 3     9.8
## 8  lab 2 Antibiotic 4     9.9
## 9  lab 3 Antibiotic 1     9.2
## 10 lab 3 Antibiotic 2     9.4
## 11 lab 3 Antibiotic 3     9.5
## 12 lab 3 Antibiotic 4     9.7
## 13 lab 4 Antibiotic 1     9.7
## 14 lab 4 Antibiotic 2     9.6
```

```
## 15 lab 4 Antibiotic 3    10.0
## 16 lab 4 Antibiotic 4    10.2
```

2.3 Get summary statistics and visualize your data with ggpubr:

```
dlabs %>%
  group_by(antibiotic) %>%
  get_summary_stats(results, type = "mean_sd")
```

```
## # A tibble: 4 x 5
##   antibiotic   variable     n  mean    sd
##   <fct>        <fct>    <dbl> <dbl> <dbl>
## 1 Antibiotic 1 results      4  9.4  0.216
## 2 Antibiotic 2 results      4  9.43 0.126
## 3 Antibiotic 3 results      4  9.72 0.222
## 4 Antibiotic 4 results      4  9.95 0.208
```

2.4 Identify outliers in each cell design:

```
dlabs %>%
  group_by(antibiotic) %>%
  identify_outliers(results)
```

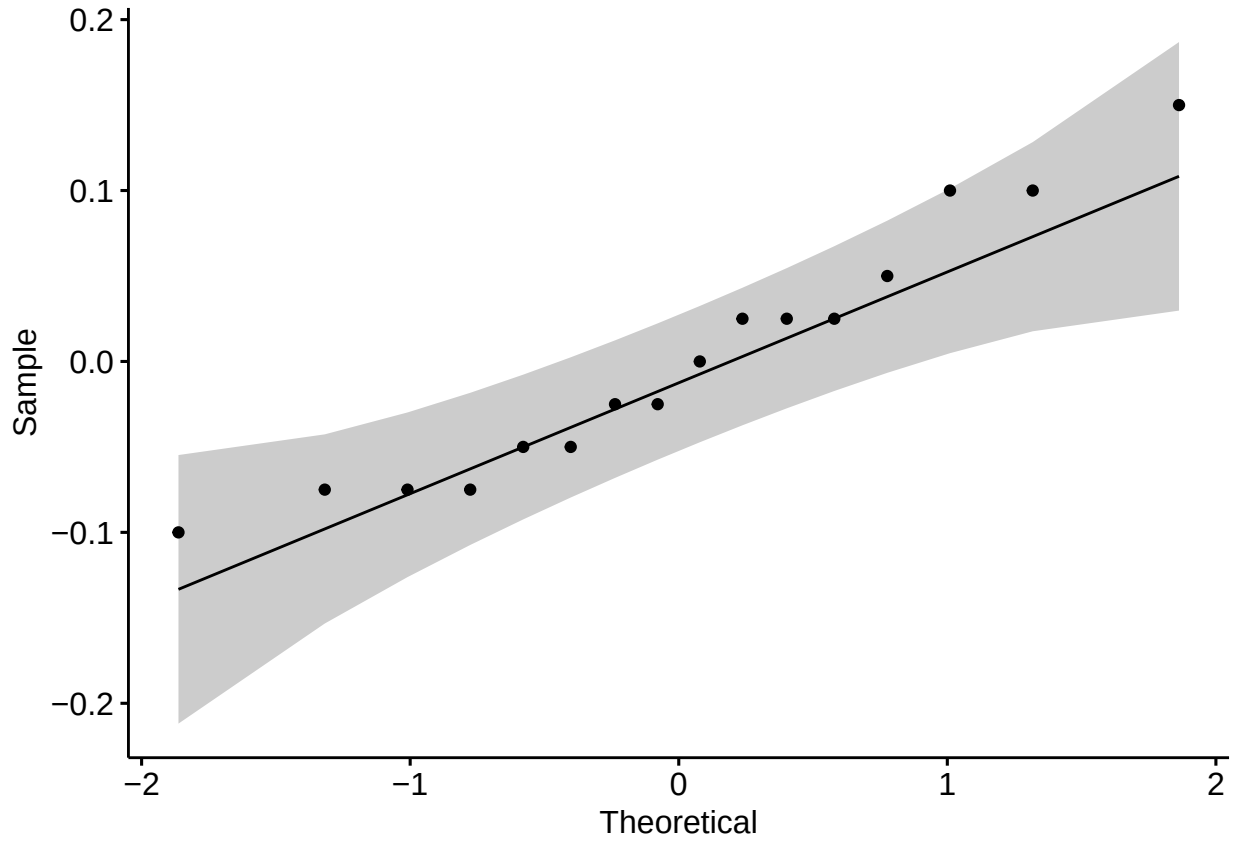
```
## # A tibble: 1 x 5
##   antibiotic   lab   results is.outlier is.extreme
##   <fct>        <fct>   <dbl> <lgl>      <lgl>
## 1 Antibiotic 2 lab 4     9.6 TRUE       FALSE
```

This passed the outlier test

2.5 Normality Assumption

Check normality assumption by analyzing the model residuals. QQ plot and Shapiro-Wilk test of normality are used.

```
# Build the linear model
model  <- lm(results ~ antibiotic + lab,
             data = dlabs)
# Create a QQ plot of residuals
ggqqplot(residuals(model))
```




```
shapiro_test(residuals(model))
```

```
## # A tibble: 1 x 3
##   variable         statistic p.value
##   <chr>                <dbl>   <dbl>
## 1 residuals(model)     0.940   0.344
```

Our value is NOT statistically significant, we fail/ do not have sufficient evidence to reject the null(that its normal) therefore, we can assume normality. In the QQ plot, as all the points fall approximately along the reference line, we can assume normality. This conclusion is supported by the Shapiro-Wilk test. The p-value is also not significant (p = 0.66), so we can assume normality.

2.6 Homogneity of variances

```
library(rstatix)

dlabs %>% levene_test(results ~ antibiotic)
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     3    12     0.581 0.639
```

```
dlabs %>% levene_test(results ~ lab)
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     3    12     0.486 0.698
```

Null Hypothesis: the variances are equal Non-statistically significant value(p = 0.63), we fail to reject null for antibiotics between results – we assume their are equal variances. Using the Levene's test for results between labs, is not significant (p = 0.69). Therefore, we can assume the homogeneity of variances in the different groups in results between labs. Knowing that there is homogeneity among the vairances for each, or pretty much the variances are equal, we are safe to continue with the two way anova(becasue variances are equal).

2.7 Two Way Anova

```
res.aov5 <- dlabs %>% anova_test(results ~ antibiotic + lab)
res.aov5
```

```
## ANOVA Table (type II tests)
##
##       Effect DFn DFd      F        p p<.05   ges
## 1 antibiotic   3   9 30.937 4.52e-05     * 0.912
## 2        lab   3   9 14.437 8.71e-04     * 0.828
```

```
# Pairwise comparisons
pwc3 <- dlabs %>% tukey_hsd(results ~ antibiotic + lab) ## adjusted turkey test
pwc3
```

```
## # A tibble: 12 x 9
##    term       group1       group2 null.value estimate conf.low conf.high   p.adj
##  * <chr>      <chr>        <chr>       <dbl>    <dbl>    <dbl>     <dbl>   <dbl>
##  1 antibiotic Antibiotic 1 Antib~          0   0.0250  -0.183     0.233  9.81e-1
##  2 antibiotic Antibiotic 1 Antib~          0   0.325    0.117     0.533  3.98e-3
##  3 antibiotic Antibiotic 1 Antib~          0   0.550    0.342     0.758  8.3 e-5
##  4 antibiotic Antibiotic 2 Antib~          0   0.300    0.0919    0.508  6.66e-3
##  5 antibiotic Antibiotic 2 Antib~          0   0.525    0.317     0.733  1.2 e-4
##  6 antibiotic Antibiotic 3 Antib~          0   0.225    0.0169    0.433  3.42e-2
##  7 lab        lab 1        lab 2           0   0.0250  -0.183     0.233  9.81e-1
##  8 lab        lab 1        lab 3           0  -0.125   -0.333     0.0831 3.03e-1
##  9 lab        lab 1        lab 4           0   0.300    0.0919    0.508  6.66e-3
## 10 lab        lab 2        lab 3           0  -0.150   -0.358     0.0581 1.82e-1
## 11 lab        lab 2        lab 4           0   0.275    0.0669    0.483  1.13e-2
## 12 lab        lab 3        lab 4           0   0.425    0.217     0.633  6.06e-4
## # i 1 more variable: p.adj.signif <chr>
```

2.8 Report the results of the ANOVA and a post-hoc test

The difference between results of antibiotic testing is statisticly significant between antibiotics 1-3(p = 0.003980), antibiotics 1 and 4(p = 0.000083), antibiotics 2 and 3(p = 0.006660) ,antibiotics 2 and 4(0.000120)

and antibiotics 3 and 4(p = 0.034200). For these antibiotics we reject the null[which is that the results between each antibiotic is the same] with these values and we can assume that the result in antibiotics are different between antibiotic groups. The difference between results of lab 1-4(p = 0.006660), lab 2-4(p = 0.011300) and lab 3-4(p = 0.000606) are statistically significant. This means we are able to reject the null[that the results where/stayed the same between groups] and it is safe to assume that the results are different.

line plot

```
dlabs %>% ggplot(aes(x = antibiotic, results, group = lab, colour = lab)) + geom_line()
```