# Assignment 2 ANOVA

Martin Ottosson

2023-11-01

## Introduction

In the southern parts of the county of Stockholm and in the county of Sörmland there are large populations of ticks in the archipelago. These ticks carry the risk of borrelia transmission which could cause the disease Lyme borreliosis in humans.

Data was collected in a study and samples were taken at three islands (Askö, Torö and Öja) and at each island samples were taken in three different habitats (forest, beach and meadow). Borrelia prevalence was calculated as percentage of ticks infested with borrelia at each sample site. Other measurements were also taken, but they are not part of this exercise.

The aim of this assignment is to see whether there are any differences in means between the islands when it comes to borrelia prevalence and also if there are any differences between habitats or interactions between islands and habitats.

## Method

The methods used in the assignment are one-way and two-way ANOVA tests. One-way ANOVA (Analysis of Variance) tests are appropriate to use when examining differences between more than two groups and in this case the difference between three groups (islands) is being examined. Two-way ANOVA or Factorial Analysis of Variance is appropriate when looking at the difference between groups and there is more than one factor. In this case islands and habitats. With two-way ANOVA we can also look at the interactions between these factors.

The borrelia dataset contains 90 observations Sample size with 30 observations per island separated in 3 habitats with 10 observations in each.

Diagnostic plots will be evaluated in order to check the assumptions of normally distributed residuals and equal spread of residuals between groups. Tables of the ANOVA tests will be presented together with tables for ad hoc tests. Finally, plots will also be shown to allow for an easier interpretation of the results.

## Result

### Part 1. One-way ANOVA

Is there a difference in means of borrelia prevalence between the different islands Askö, Torö and Öja? As described in the introduction, Borrelia prevalence is the percentage of borrelia infested ticks at the time of measurement.

Table 1 - First one-way ANOVA test of borrelia prevalence between islands

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **Island** | 2 | 26,640.1 | 13,320.0 | 42.9 | 0.0 |
| **Residuals** | 87 | 26,988.8 | 310.2 | | |

The one-way ANOVA test shows a high F-value which is clearly significant. The p-value is close to zero.

We can reject the null hypothesis that there is no difference between means and accept the research hypothesis that there is a difference between means of borrelia prevalence among the islands.

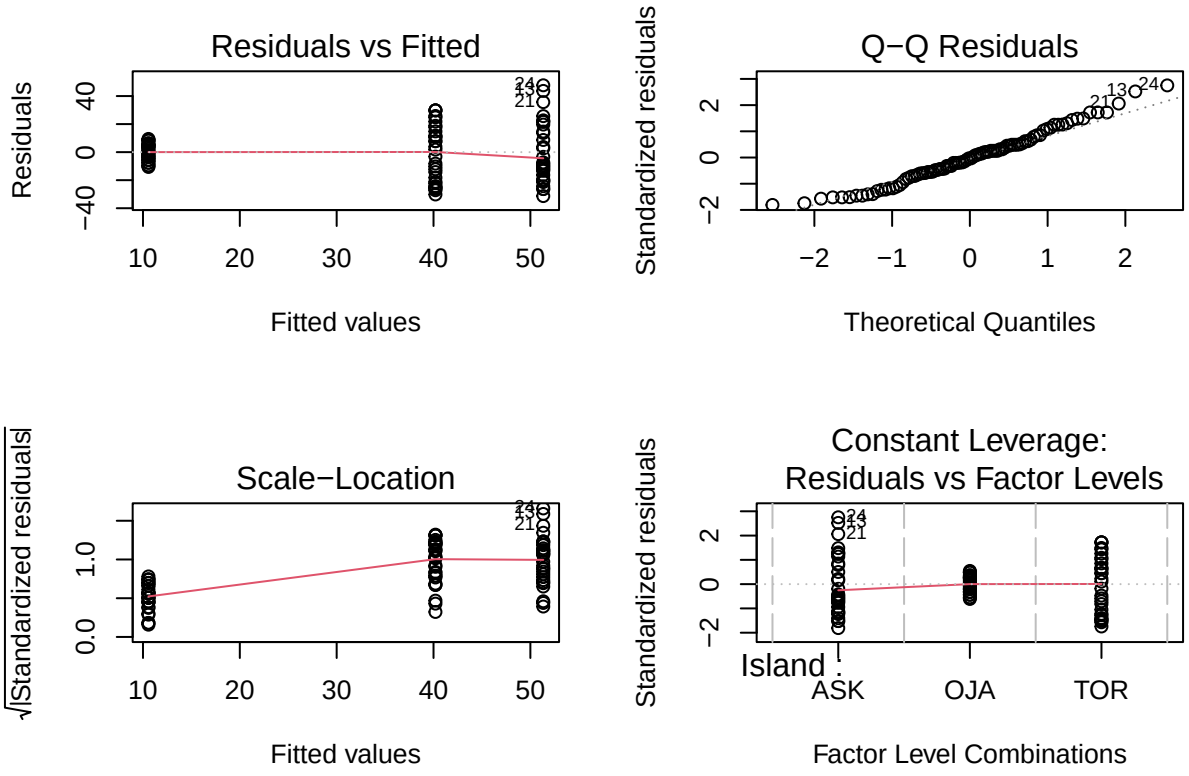



Figure 1 - Diagnostic plots of first ANOVA test

The Residuals vs Fitted plot show the spread of the residuals for the three groups plotted at their mean values. There is a difference of the spread. It is not clear whether the spread of the biggest group is three times bigger than the spread of the smallest group, but a transformation of the data could be used in this case since it could be a sign of heteroscedasticity.

It is also questionable if the residuals are approximately normally distributed since they do not follow the straight line in the Q-Q plot.

Table 2 - Second one-way ANOVA test of arcsine transformed borrelia prevalence between islands

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **Island** | 2 | 4.0 | 2.0 | 47.7 | 0.0 |
| **Residuals** | 87 | 3.6 | 0.0 | | |

The new ANOVA test also is clearly signficant for the F-test. There is a difference of means between the groups.

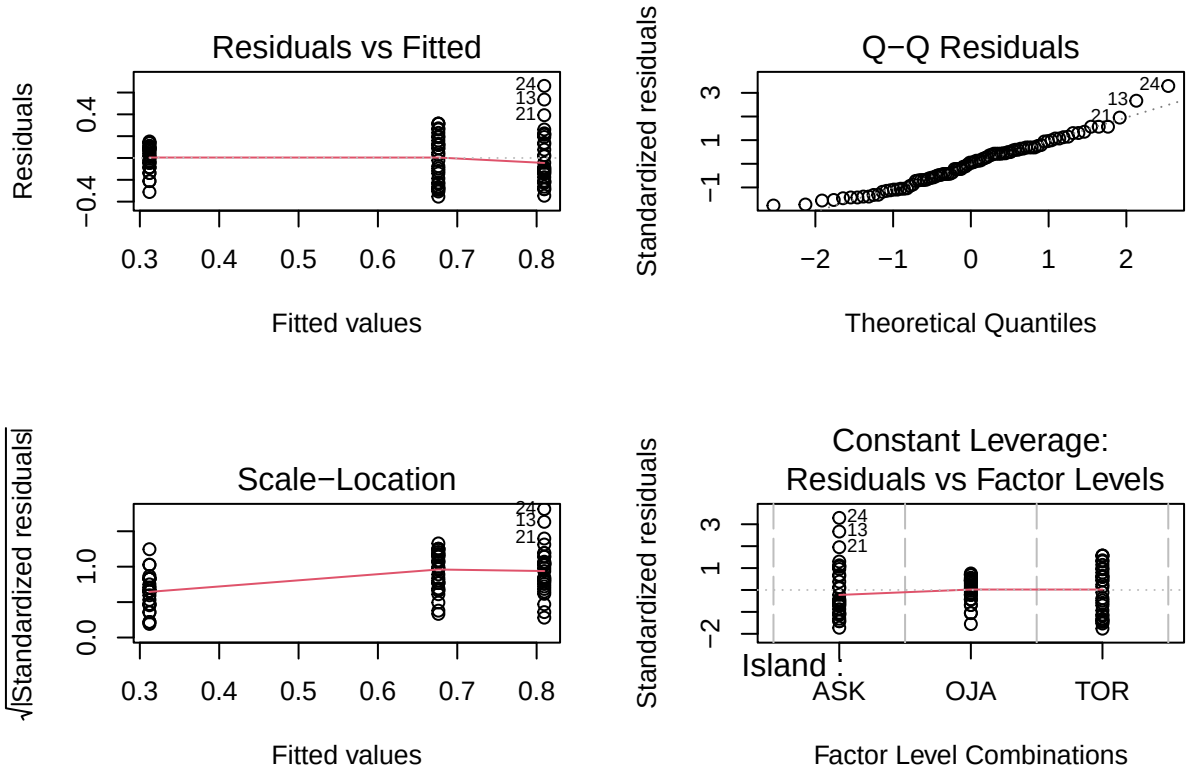



Figure 2 - Diagnostic plots of second ANOVA test

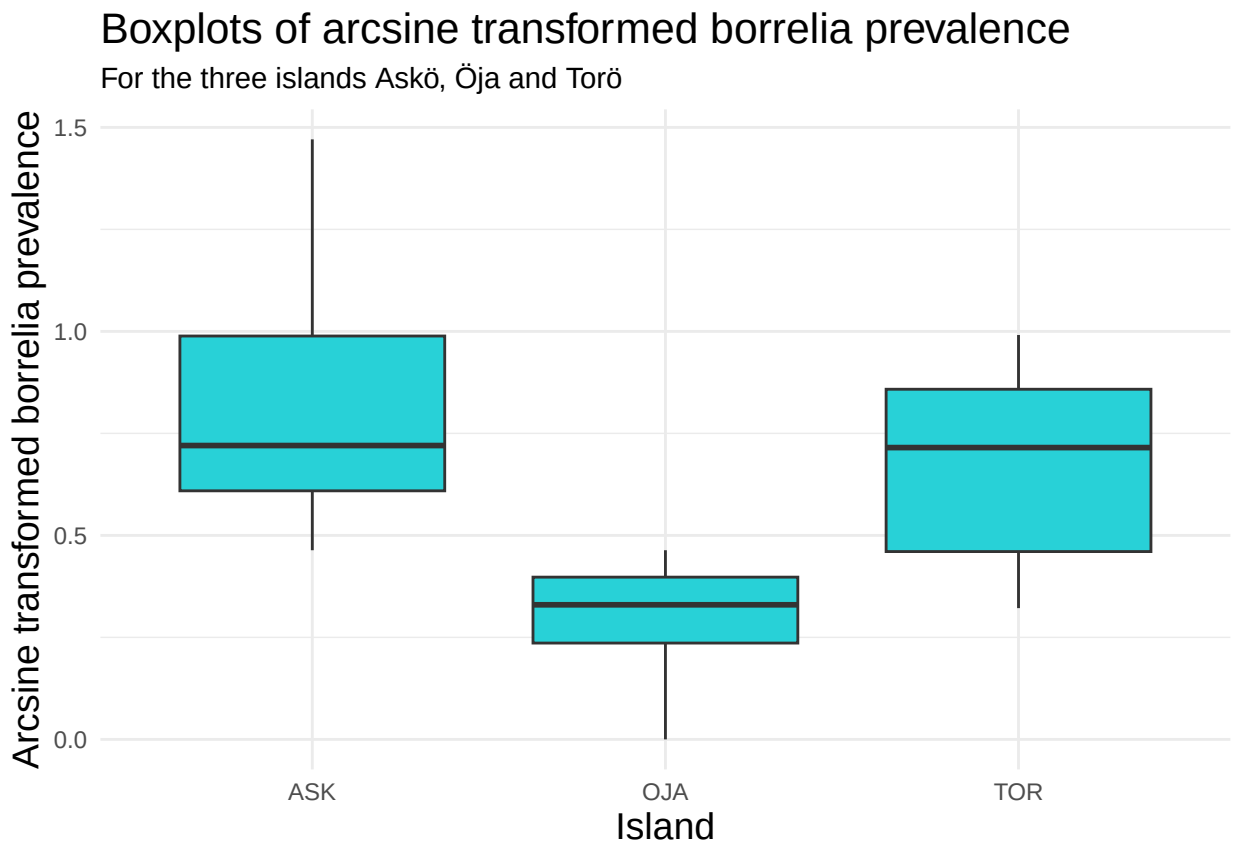
The Residuals vs Fitted plot looks better after the transformation. The difference between the group with most spread of the residuals and the group with the least spread has diminshed. The Q-Q Residuals plot also looks better. Most of the points follow the straight line.

Since we can assume that the assumptions of normal distribution for the residuals and homoscedasticity hold we can perform a post hoc test of pairwise comparisons to show where the differences are.

Table 3 - Tukey HSD post hoc test for second one-way ANOVA test

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = borrelia.prev.arc ~ Island, data = borrelia)
##
## $Island
##               diff        lwr          upr    p adj
## OJA-ASK -0.4975503 -0.6233371 -0.371763458 0.000000
## TOR-ASK -0.1334390 -0.2592258 -0.007652144 0.034963
## TOR-OJA  0.3641113  0.2383245  0.489898153 0.000000
```

The post hoc test is now performed on transformed data and it shows signficant differences of means for all of the pairwise comparisons at the 0.05 level.


Boxplots of arcsine transformed borrelia prevalence
For the three islands Askö, Öja and Torö
Arcsine transformed borrelia prevalence
1.5
1.0
0.5
0.0
ASK
OJA
TOR
Island

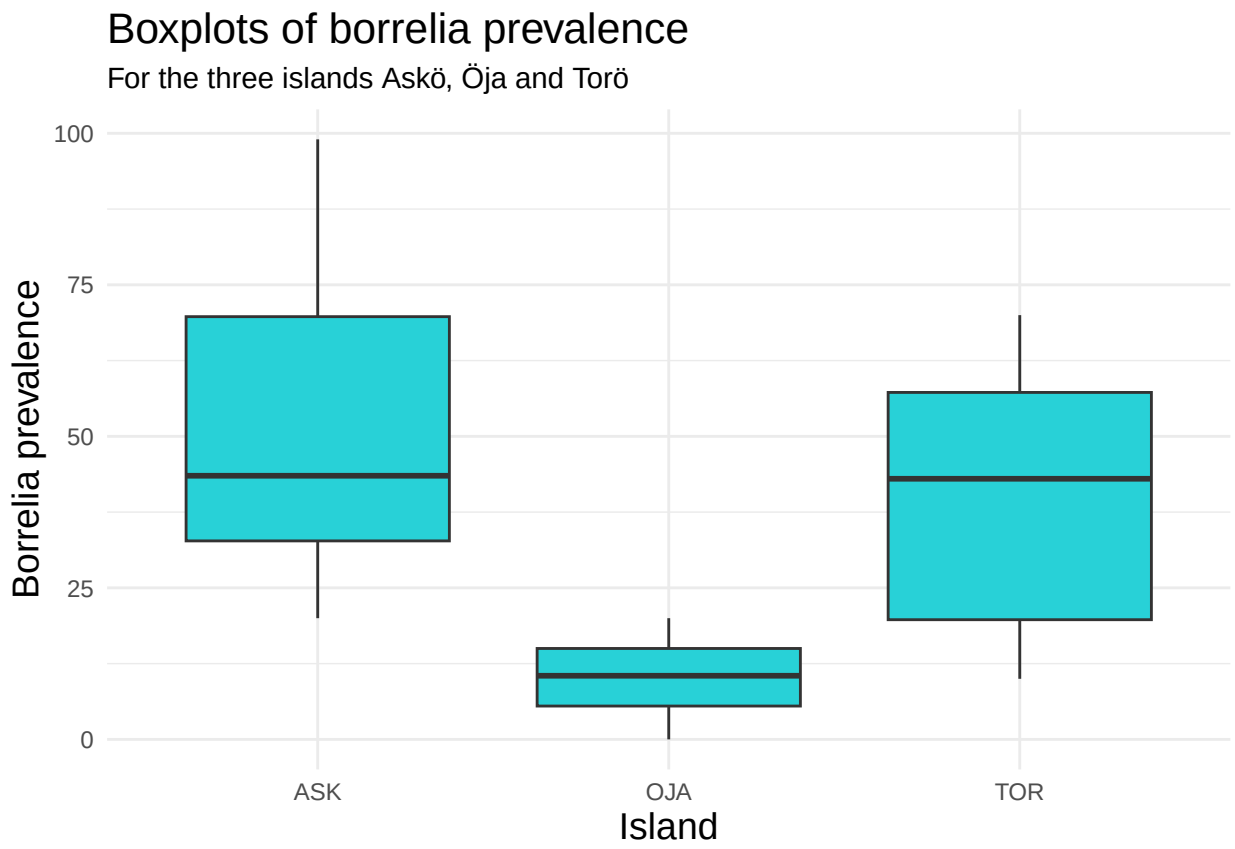



Figure 3 and 4 - Boxplots of borrelia prevalence without and with arcsine transformation

The boxplot shows the differences of borrelia prevalence between the islands and for easier interpretation a boxplot is also included without the arcsine transformation. The smallest difference is between Torö and Askö, but there is still a significant difference with a slightly higher mean for Askö. Both Torö and Askö clearly have a higher proportion of borrelia infested ticks compared to Öja. The Residuals vs Fitted plot for the first ANOVA test showed a mean just above 10% for Öja, around 40% for Torö and above 50% for Askö, which is quite a lot.

### Part 2. Two-way ANOVA with interaction

Is there a difference in means of borrelia prevalence between the different islands Askö, Torö and Öja and the habitat types meadow, beach and forest? Are there any interactions?

First a linear model is created.

Table 4 - Linear model

```
##
## Call:
## lm(formula = Borrelia_prevalence ~ Island * Habitat, data = borrelia)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.000  -9.075  -1.250   8.250  43.000
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                56.500      5.327  10.606  < 2e-16 ***
## IslandOJA                 -47.800      7.534  -6.345  1.2e-08 ***
## IslandTOR                 -24.100      7.534  -3.199  0.00197 **
## HabitatForest              -0.500      7.534  -0.066  0.94725
## HabitatMeadow             -15.000      7.534  -1.991  0.04985 *
## IslandOJA:HabitatForest     3.300     10.655   0.310  0.75757
## IslandTOR:HabitatForest     3.300     10.655   0.310  0.75757
## IslandOJA:HabitatMeadow    17.800     10.655   1.671  0.09865 .
## IslandTOR:HabitatMeadow    35.600     10.655   3.341  0.00126 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.85 on 81 degrees of freedom
## Multiple R-squared:  0.5713, Adjusted R-squared:  0.529
## F-statistic:  13.5 on 8 and 81 DF,  p-value: 3.173e-12
```

The linear model is hard to interpret, but we can use diagnostic plots to check the assumptions.

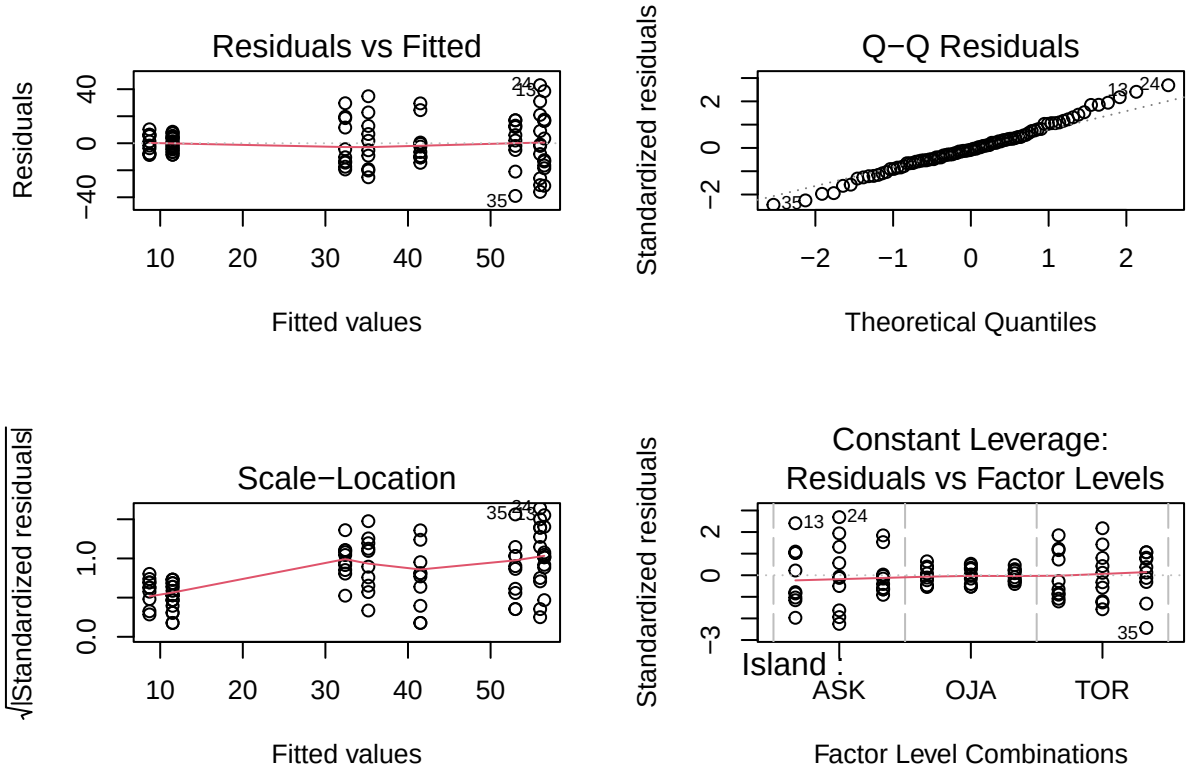



Figure 5 - Diagnostic plots for linear model

The Residuals vs Fitted plots shows spread of dispersions between the groups. The spread could be three times the size for the biggest group compared to the smallest group and therefore heterskedasticity could be a problem. The Q-Q plot also shows that the assumption of approximately normally distributed residuals could be doubted. The points do not follow the straight line strictly.

Since the assumptions are doubtful we will repeat the process again with the arcsine transformation.

Table 5 - Linear model with arcsine transformation

```
##
## Call:
## lm(formula = borrelia.prev.arc ~ Island * Habitat, data = borrelia)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42903 -0.13175 -0.01044  0.12994  0.60322
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.862883   0.062157  13.882  < 2e-16 ***
## IslandOJA                -0.598355   0.087903  -6.807 1.59e-09 ***
## IslandTOR                -0.269465   0.087903  -3.065  0.00295 **
## HabitatForest             0.004529   0.087903   0.052  0.95903
## HabitatMeadow            -0.164228   0.087903  -1.868  0.06534 .
## IslandOJA:HabitatForest   0.063078   0.124314   0.507  0.61325
## IslandTOR:HabitatForest   0.024744   0.124314   0.199  0.84273
## IslandOJA:HabitatMeadow   0.239335   0.124314   1.925  0.05771 .
## IslandTOR:HabitatMeadow   0.383335   0.124314   3.084  0.00280 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1966 on 81 degrees of freedom
## Multiple R-squared:  0.5888, Adjusted R-squared:  0.5482
## F-statistic:  14.5 on 8 and 81 DF,  p-value: 6.41e-13
```

The ouput is hard to interpret as previously stated, but we will look at the diagnostic plots again now with the arcsine transformation of the borrelia prevalence.

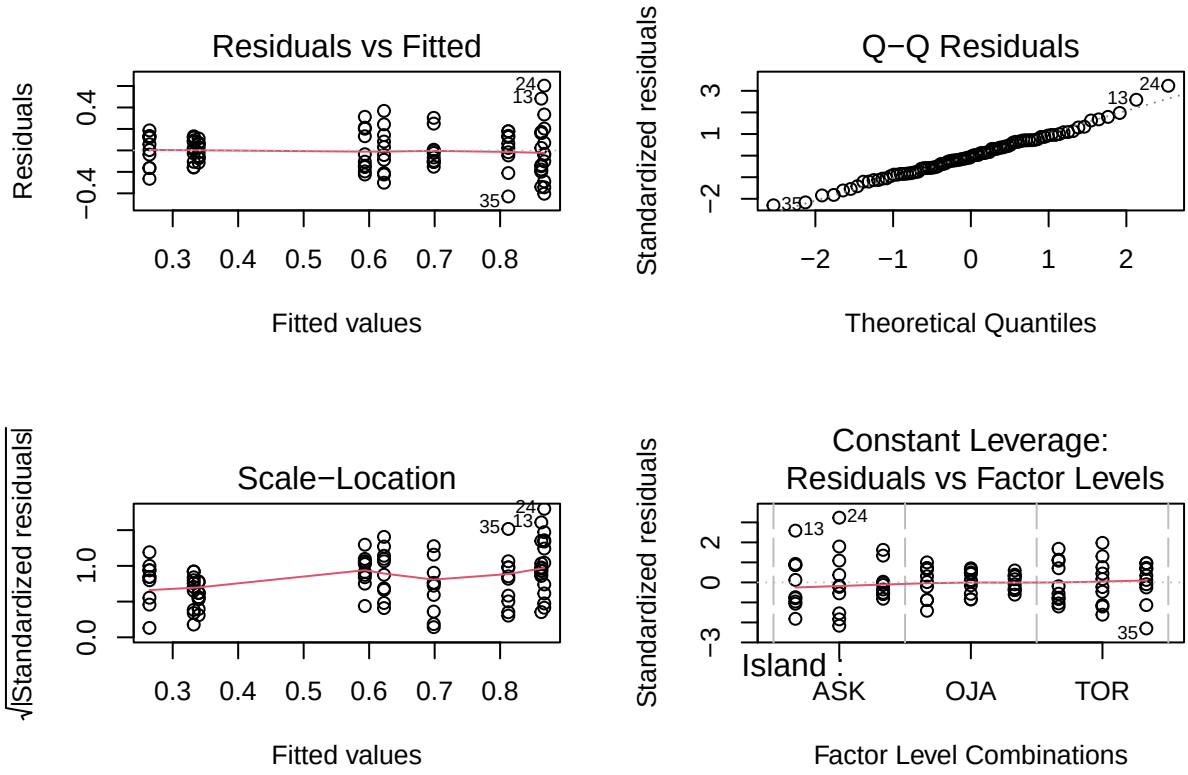



Figure 6 - Diagnostic plots for the linear model with arcsine transformation

After transformation of the borrelia prevalence variable the diagnostic plot looks better with smaller differences in the dispersion of the residuals between the groups and the assumption about normality also seem to hold.

Table 6 - Two-way ANOVA table with arcsine transformed borrelia prevalence as response and island and habitat as factors

```
## Anova Table (Type II tests)
##
## Response: borrelia.prev.arc
##                Sum Sq Df F value    Pr(>F)
## Island         3.9794  2 51.4999 3.703e-15 ***
## Habitat        0.0311  2  0.4026   0.66992
## Island:Habitat 0.4710  4  3.0479   0.02153 *
## Residuals      3.1294 81
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To create a more comprehensible output we can use an Anova table as in table 6. It shows a significant main effect for island and an interaction effect for island and habitat.

As there are significant effects a post hoc test is performed in order to understand where the differences are found.

Table 7 - Post Hoc test for two-way ANOVA

```
##   Tukey multiple comparisons of means
```

```
##     95% family-wise confidence level
##
## Fit: aov(formula = borrelia.prev.arc ~ Island + Habitat + Island * Habitat, data = borrelia)
##
## $Island
##               diff        lwr        upr     p adj
## OJA-ASK -0.4975503 -0.6187207 -0.3763799 0.0000000
## TOR-ASK -0.1334390 -0.2546094 -0.0122686 0.0273537
## TOR-OJA  0.3641113  0.2429409  0.4852817 0.0000000
##
## $Habitat
##                       diff         lwr       upr     p adj
## Forest-Beach  0.033803322 -0.08736706 0.1549737 0.7837818
## Meadow-Beach  0.043328457 -0.07784192 0.1644988 0.6708181
## Meadow-Forest 0.009525135 -0.11164524 0.1306955 0.9807727
##
## $`Island:Habitat`
##                                 diff          lwr          upr     p adj
## OJA:Beach-ASK:Beach   -0.598354621 -0.878520589 -0.318188653 0.0000001
## TOR:Beach-ASK:Beach   -0.269465201 -0.549631169  0.010700767 0.0689076
## ASK:Forest-ASK:Beach   0.004529425 -0.275636543  0.284695393 1.0000000
## OJA:Forest-ASK:Beach  -0.530747199 -0.810913167 -0.250581231 0.0000016
## TOR:Forest-ASK:Beach  -0.240192082 -0.520358050  0.039973886 0.1530401
## ASK:Meadow-ASK:Beach  -0.164228190 -0.444394159  0.115937778 0.6370055
## OJA:Meadow-ASK:Beach  -0.523247835 -0.803413803 -0.243081866 0.0000023
## TOR:Meadow-ASK:Beach  -0.050358427 -0.330524396  0.229807541 0.9996910
## TOR:Beach-OJA:Beach    0.328889420  0.048723452  0.609055388 0.0097913
## ASK:Forest-OJA:Beach   0.602884046  0.322718078  0.883050014 0.0000000
## OJA:Forest-OJA:Beach   0.067607422 -0.212558547  0.347773390 0.9973987
## TOR:Forest-OJA:Beach   0.358162539  0.077996571  0.638328507 0.0032745
## ASK:Meadow-OJA:Beach   0.434126431  0.153960462  0.714292399 0.0001395
## OJA:Meadow-OJA:Beach   0.075106786 -0.205059182  0.355272755 0.9946297
## TOR:Meadow-OJA:Beach   0.547996193  0.267830225  0.828162162 0.0000007
## ASK:Forest-TOR:Beach   0.273994626 -0.006171342  0.554160594 0.0602678
## OJA:Forest-TOR:Beach  -0.261281998 -0.541447966  0.018883970 0.0871744
## TOR:Forest-TOR:Beach   0.029273119 -0.250892849  0.309439087 0.9999951
## ASK:Meadow-TOR:Beach   0.105237011 -0.174928957  0.385402979 0.9548862
## OJA:Meadow-TOR:Beach  -0.253782633 -0.533948602  0.026383335 0.1072700
## TOR:Meadow-TOR:Beach   0.219106774 -0.061059195  0.499272742 0.2511385
## OJA:Forest-ASK:Forest -0.535276624 -0.815442593 -0.255110656 0.0000013
## TOR:Forest-ASK:Forest -0.244721507 -0.524887475  0.035444461 0.1363592
## ASK:Meadow-ASK:Forest -0.168757615 -0.448923584  0.111408353 0.6023434
## OJA:Meadow-ASK:Forest -0.527777260 -0.807943228 -0.247611291 0.0000018
## TOR:Meadow-ASK:Forest -0.054887852 -0.335053821  0.225278116 0.9994166
## TOR:Forest-OJA:Forest  0.290555117  0.010389149  0.570721085 0.0361227
## ASK:Meadow-OJA:Forest  0.366519009  0.086353041  0.646684977 0.0023625
## OJA:Meadow-OJA:Forest  0.007499365 -0.272666603  0.287665333 1.0000000
## TOR:Meadow-OJA:Forest  0.480388772  0.200222804  0.760554740 0.0000172
## ASK:Meadow-TOR:Forest  0.075963892 -0.204202077  0.356129860 0.9942012
## OJA:Meadow-TOR:Forest -0.283055753 -0.563221721 -0.002889784 0.0457375
## TOR:Meadow-TOR:Forest  0.189833655 -0.090332314  0.469999623 0.4415401
## OJA:Meadow-ASK:Meadow -0.359019644 -0.639185612 -0.078853676 0.0031675
## TOR:Meadow-ASK:Meadow  0.113869763 -0.166296205  0.394035731 0.9299072
## TOR:Meadow-OJA:Meadow  0.472889407  0.192723439  0.753055375 0.0000243
```

As table 5 shows there are many comparisons done in the Tukey post hoc test. There seem to be 16 significant effects at the 0.05 level.

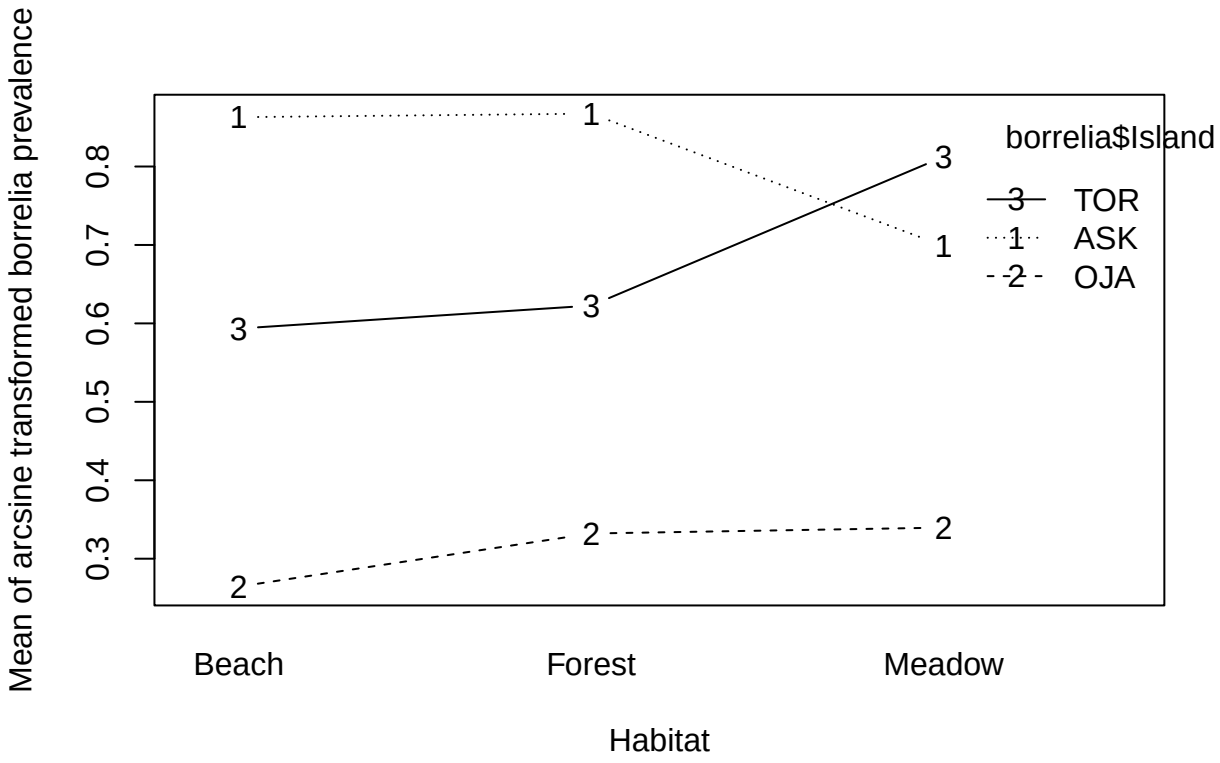



Figure 7 - Interaction plot for two-way ANOVA test

An interaction plot makes it easier to see what is going on. First of all there seem to be clear differences between the islands when it comes to borrelia prevalence on beaches. The Tukey multiple comparisons of means shows a clear significant difference between Askö and Öja. Torö and Öja are also significant, but Torö and Askö comes close to be signficant but has to accept the null hypothesis of no difference at the 0.05 level. Since there are interaction effects which table 5 shows, we should be careful with the main effects and only consider the clear significant comparisons.

In forests there is also a clear difference between Öja and Askö, supported by table 5. Torö and Öja also show a significant difference.

In meadows there are differences between Askö and Öja as well as between Torö and Öja. In meadows Torö takes the first place when it comes to having the highest proportion of infested ticks.

For the interaction effects, with different habits, there are several comparisons with very low p-values as can be seen in table 7. None of the confidence intervals covers 0 in these cases and we can be confident to say there is an actual difference in borrelia prevalence. Some examples of the biggest differences are found between Öja:Forest - Askö:Beach, Askö:Forest - Öja:Beach and Torö:Meadow - Öja:Beach.

## Discussion

This study aimed to find out if there are any differences of borrelia prevalence between three different islands in the Stockholm and Sörmland archipelago. The results from the one-way ANOVA test showed a significant differences between all of the three islands. Askö and Torö have bigger problems with higher proportions of infested ticks which could spread borrelia to humans.

The study also aimed at finding out if there are any differences between habitats and if there are any interaction effects between habitats and islands. Even though we can not say that habitats in itself show a main effect for borrelia prevalence, the two-way ANOVA test showed there is a main effect for island and an interaction effect for island and habitat.

The results showed differences for combinations of habitats and islands, where some of the biggest differences were between Öja:Beach - Askö:Beach, Öja:Forest - Askö:Forest, Öja:Forest - Askö:Beach, Askö:Forest - Öja:Beach and Torö:Meadow - Öja:Beach.

It is commonly known in Sweden to inspect oneself for ticks after a visit in the nature. Especially during summer in the archipelago. The findings of this study could add to current knowledge that Öja might be a safer island to visit. However, one does not need to avoid the other two islands, but when visiting Askö then the meadow could be a better choice for the picknick location and while going to Torö the meadow might be better to avoid. It would be nicer to go to the beach and look at the big waves which Torö is famous for. With that being said, the study could not show a main effect for habits in the two-way ANOVA as already being mentioned, but the interaction plot showed some indications for the habits.

The findings of the study relies on the arcsine transformation of borrelia prevalence in order for the linear model assumptions to hold. Some researchers advise against this kind of transformation (Warton and Hui, 2011) and the interpretation of the results become more difficult with this kind of transformation. This is a limitation for this study.

A strenth of the study is the use of the well proven ANOVA test. With the two-way ANOVA test we can find out about interactions that we could never find with a simpler t-test for differences of means between two groups. This is a strength of the study. There were also some highly significant results and we can be confident in out rejections of the null hypothesis.

Another limitation though would be that there is only 10 observations for each habitat per island. An improvement could be to increase the sample size for each group.

Finally, we can be confident to say that there are differences in the borrelia prevalence between islands in the archipelago and there is an interaction with different habitats. This study makes it easier for visitors in the archipelago to know when extra precaution is needed.

## References

Warton and Hui, 2011. The arcsine is asinine: the analysis of proportions in ecology. Published in Ecology by the Ecological Society of America.

## Appendix

In this appendix the R code from the study is presented.

library(Rcmdr) #this chunk start RCommander: library(ggplot2) #this cunks starts ggplot2: #if you have used other packeges you may need to include them here: library(flextable)
library(xtable) library(esquisse) library(faraway) library(car) library(multcomp)

#include a chunk that opens your indata: borrelia <- read.table("C:/Users/mart0/OneDrive/Dokument/Martin/Södertörns högskola/Statistisk analys och visualisering i R - I/Part 4/Assignment 2/borrelia.txt", header = TRUE, sep = "", dec =",")

head(borrelia)

str(borrelia)

**Table 1**

anova_test1 <- aov(Borrelia_prevalence ~ Island, data = borrelia)

as_flextable(xtable(anova(anova_test1)))

**Figure 1**

par(mfrow=c(2,2)) plot(anova_test1)

**Table 2**

Creates a new variable in the borrelia dataset for arcsine transformed borrelia prevalence: borrelia$borrelia.prev.arc < -asin(sqrt(borrelia$Borrelia_prevalence/100))

New ANOVA test with the arcsine transformed borrelia prevalence as response variable: anova_test2 <- aov(borrelia.prev.arc ~ Island, data = borrelia)

as_flextable(xtable(anova(anova_test2)))

**Figure 2**

par(mfrow=c(2,2)) plot(anova_test2)

**Table 3**

TukeyHSD(anova_test2)

**Figure 3 and 4**

#esquisser()

ggplot(borrelia) + aes(x = Island, y = borrelia.prev.arc) + geom_boxplot(fill = "#28D1D7") + labs(x = "Island", y = "Arcsine transformed borrelia prevalence", title = "Boxplots of arcsine transformed borrelia prevalence", subtitle = "For the three islands Askö, Öja and Torö") + theme_minimal() + theme(plot.title = element_text(size = 16L), axis.title.y = element_text(size = 14L), axis.title.x = element_text(size = 14L))

ggplot(borrelia) + aes(x = Island, y = Borrelia_prevalence) + geom_boxplot(fill = "#28D1D7") + labs(x = "Island", y = "Borrelia prevalence", title = "Boxplots of borrelia prevalence", subtitle = "For the three islands Askö, Öja and Torö") + theme_minimal() + theme(plot.title = element_text(size = 16L), axis.title.y = element_text(size = 14L), axis.title.x = element_text(size = 14L))

### Table 4: lm_borrelia<-lm(Borrelia_prevalence ~ Island * Habitat, borrelia) summary(lm_borrelia)

**Figure 5**

par(mfrow=c(2,2)) #setting to make the following plot to show as a 2x2 multiplot plot(lm_borrelia)

**Table 5**

lm_borrelia2<-lm(borrelia.prev.arc ~ Island * Habitat, borrelia) summary(lm_borrelia2)

**Figure 6**

par(mfrow=c(2,2)) plot(lm_borrelia2)

**Table 6**

Anova(lm_borrelia2)

This is another way of doing the same two-way ANOVA test: Anova_test3 <- aov(borrelia.prev.arc ~ Island + Habitat + Island*Habitat, data = borrelia) summary(Anova_test3)

**Table 7**

TukeyHSD(Anova_test3)

**Figure 7**

interaction.plot(x.factor = borrelia$Habitat, trace.factor = borrelia$Island, response = borrelia$borrelia.prev.arc, fun = mean, type = "b", legend = TRUE, xlab = "Habitat", ylab = "Mean of arcsine transformed borrelia prevalence")