Assignment 2 ANOVA

Yuka Tatsumi

2021-11-05

Introduction

Method

The dataset used in this report is based on a study of ticks that were sampled in three habitats(meadow, forest and beach) in three islands,Askö(ASK), Torö(TOR) and Öja(OJA) in the Stockholm and Sörmland archipelago. The prevalence of Borrelia was measured as % of ticks that were infested by Borrelia at each sample location as well as soli PH and temperature in the study.

Anovas is used for analysis in this report as long as the data meets the assmuption of normality and homoscedascitic because the data in the study is non-binomial data. Before the analysis, normality and homoscedasticity is checked in order to investigate if any data transformation and other measures is required to meet assmuption of normality and homoscedascitic for conducting Anova.

Result

Part 1 One-way Anova

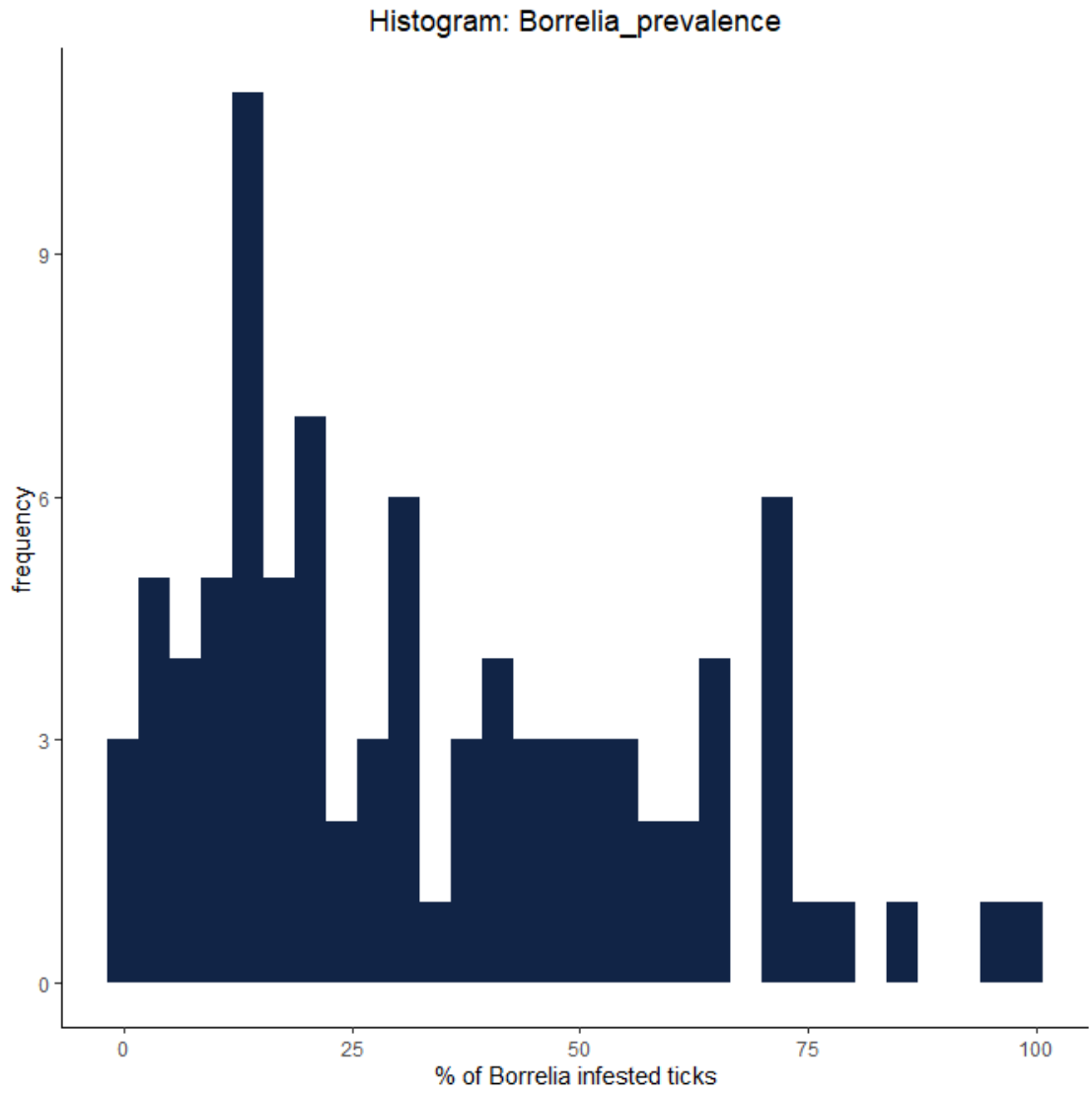
Part 1 in this report aims to answer the question “Are there any differences in the prevalence of Borrelia among the three different islands?” with One-way ANOVA.

Firstly, the normality and homoscedasticity are checked.

According to the histogram below(Figure 1), the data for Borrelia prevalence(% of Borrelia infested ticks) is not normally distributed.

Figure 1

Histogram:Borrelia prevalence(% of Borrelia infested ticks)

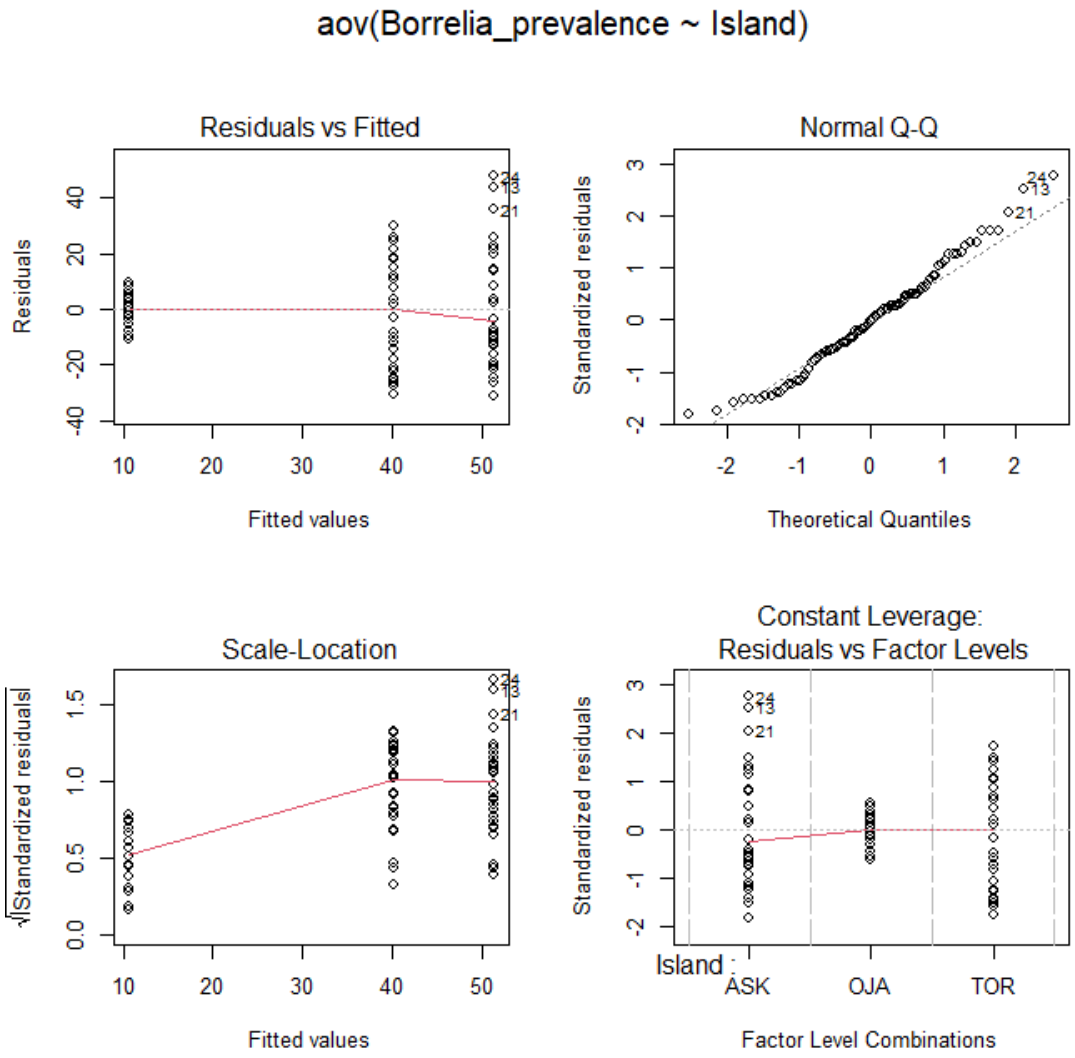


Diagnostic diagrams (basic diagnostic plot)

Normal Q-Q plot in the diagnostic diagrams (Figure 2) shows that there are some deviations from reference line around right and left side of the graph. This indicates that the residual is not completely normally distributed. “Residual vs Fitted values” in the diagnostic diagrams below shows that the biggest spread of the residuals is more than three times than the smallest spread. Therefore, the transformation of the data is required in order to meet the assumption of normality and and homoscedasticity for Anovas.

Figure 2

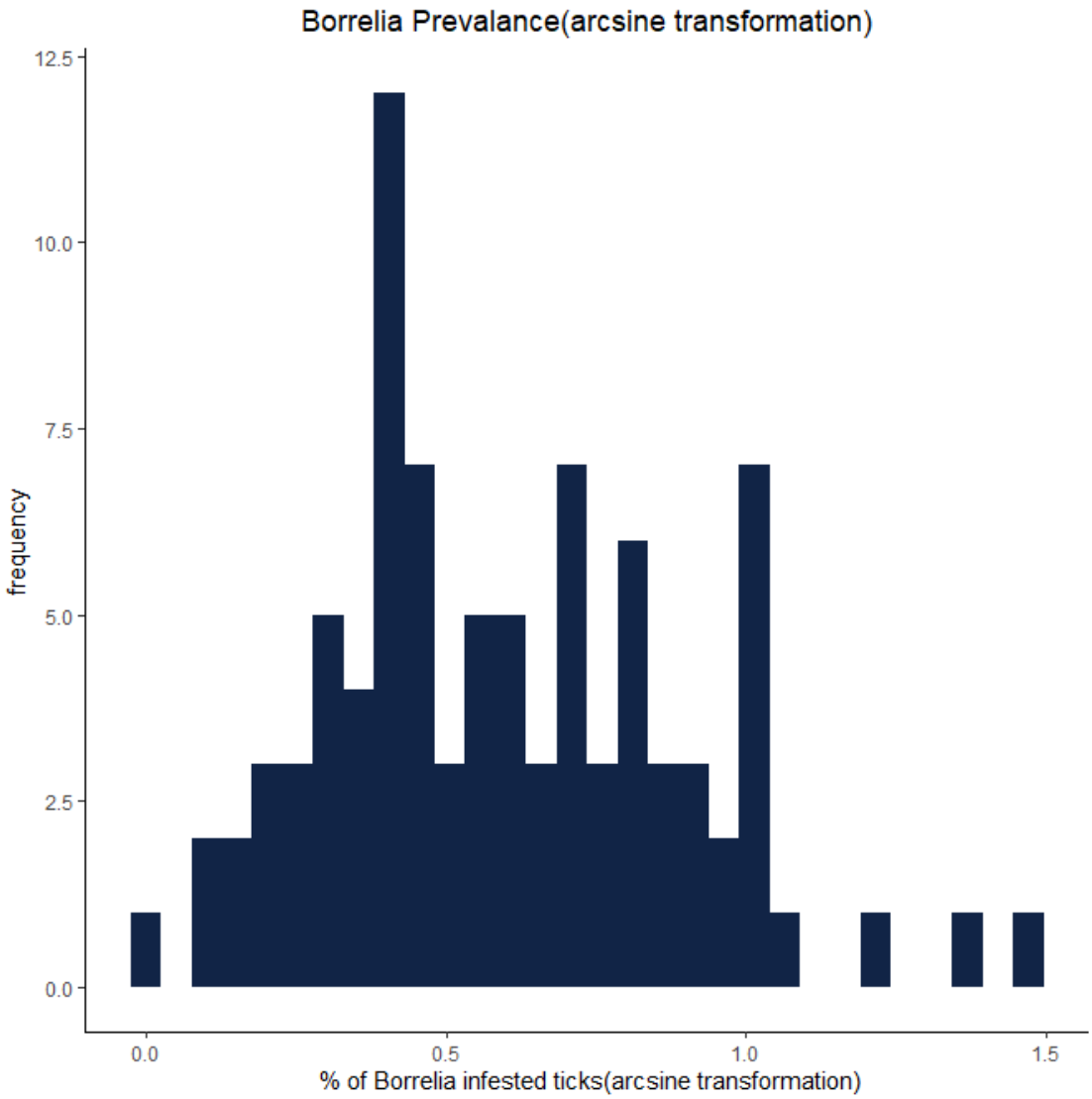
Diagnostic diagrams test (Dependent variable:% of Borrelia infested ticks, Indepent variable: Island)



To meet the assumption of normal distribution, the arcsine transformation is used. After the transformation, the data becomes more normally distributed according to the the histogram(Figure 3).

Figure 3

Histogram: Borrelia Prevalance after arcsine transformation

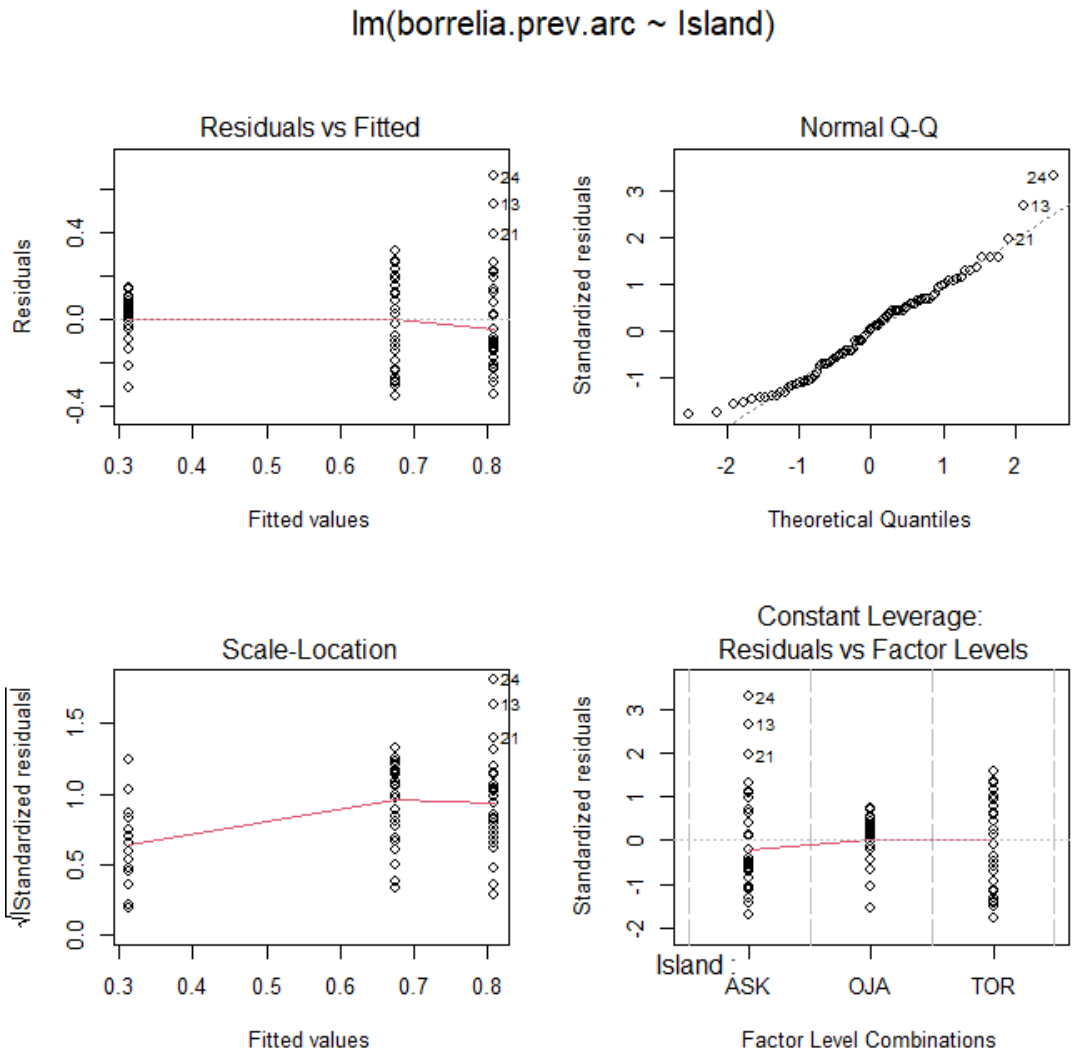


Diagnostic diagrams

After the arcsine transformation, Normal Q-Q plot in diagnostic diagrams (Figure 4) shows that there is less deviation from reference line. This indicates that the residual is more normally distributed. “Residual vs Fitted values” in diagnostic diagrams indicates that the biggest spread of the residuals becomes less than three times than the smallest spread.

Figure 4

Diagnostic diagrams after arcsine transformation Anova one-way test (Dependent variable:Arcsine transformation of % of Borrelia infested ticks, Indepent variable: Island)



The result of Anova one-way test after arcsine transformation in the Table 1 below (F-statistic:47.67, degree of freedom 2 and 87 DF, p-value: 1.051e-14) shows that there are statistically significant differences among 3 different islands.

Table 1

Result of Anova one-way test after after Arcsine transformation (Dependent variable:% of Borrelia infested ticks, Independent variable: Island)

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)

Island	2	3.979	1.99	47.67	1.051e-14
Residuals	87	3.632	0.04174	NA	NA

Linear regression coefficients of Öja is -0.498 (p_value: 0.0000) and the coefficients of Torö is -0.133 (p_value: 0.0132) according to the Table 2 below. It implies that Öja has the lowest risk of a Borrelia transmission and Askö has the highest risk of a Borrelia transmission.

Table 2

Linear Regression Coefficients (Dependent variable:% of Borrelia infested ticks, Independent variable: Island)

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	0.810	0.037	21.706	0.0000	***
IslandOJA	-0.498	0.053	-9.432	0.0000	***
IslandTOR	-0.133	0.053	-2.530	0.0132	*

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05 < '.' < 0.1 < ' ' < 1*

Residual standard error: 0.2043 on 87 degrees of freedom

Multiple R-squared: 0.5229, Adjusted R-squared: 0.5119

F-statistic: 47.67 on 87 and 2 DF, p-value: 0.0000

The results of the post-hoc test (Table 3) indicate that there are statistically significant difference between Öja and Askö (t-value:-9.432, p-value <0.001), Torö and Askö (t-value:-2.53 p-value: 0.035), as well as Torö and Öja(t-value:6.902, p-value: <0.001).

Table 3

Pairwise comparisons(Multiple Comparisons of Means: Tukey Contrasts)

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = borrelia.prev.arc ~ Island, data = Dataset)
##
## Linear Hypotheses:
##                Estimate Std. Error t value Pr(>|t|)
## OJA - ASK == 0 -0.49755    0.05275  -9.432   <1e-04 ***
## TOR - ASK == 0 -0.13344    0.05275  -2.530    0.035 *
## TOR - OJA == 0  0.36411    0.05275   6.902   <1e-04 ***
```

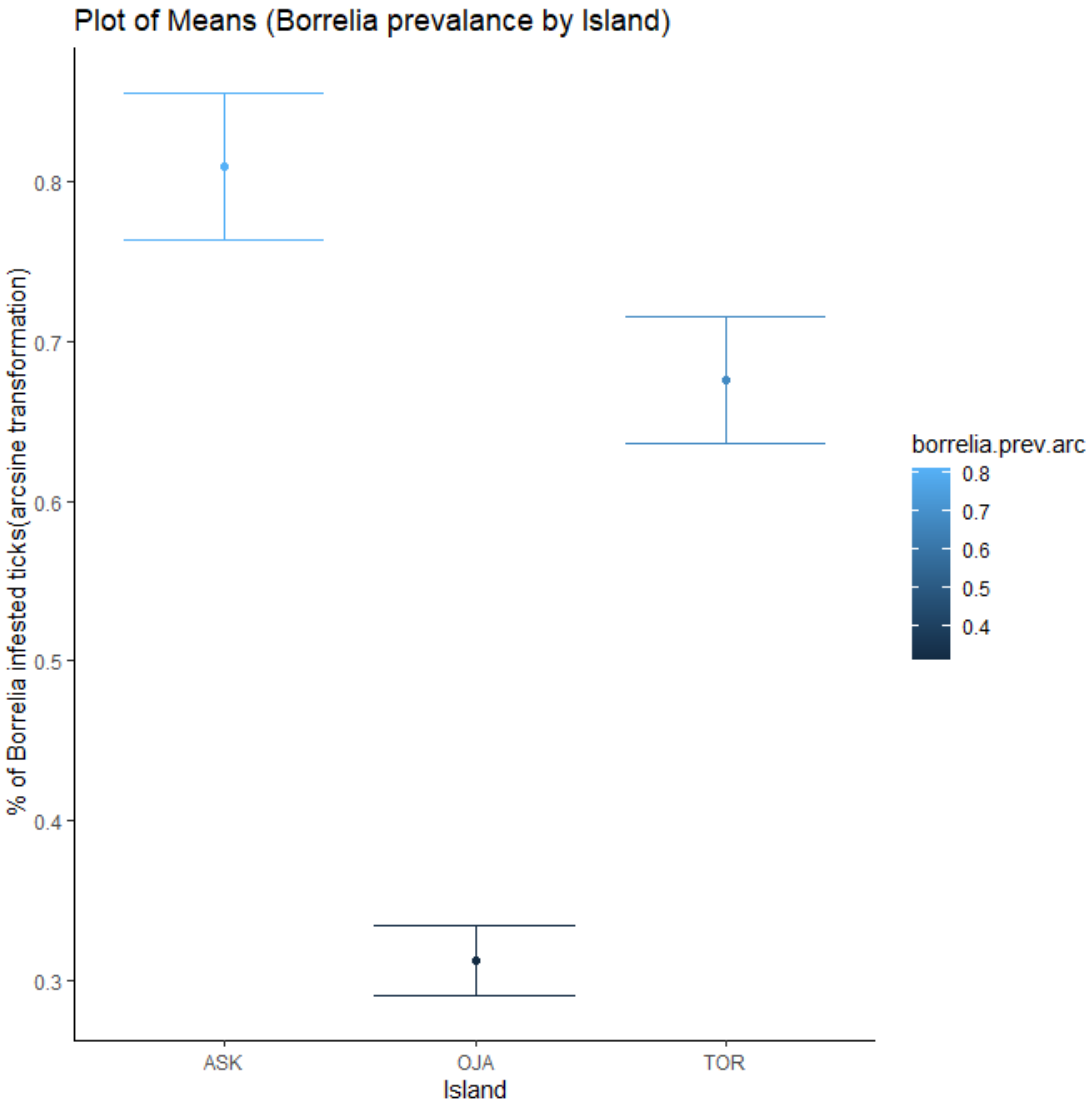
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Plot of Means

Plot of means of Borrelia Prevalence by island (Figure 5) also shows that there are significant difference among different islands.

Figure 5

#####Plot of Means (Borrelia prevalence by Island)



Part 2 Two-way ANOVA interaction

Part 2 in this report aims to answer the question “Does difference in habitat have an influence on the prevalence of Borrelia?” with two-way ANOVA.

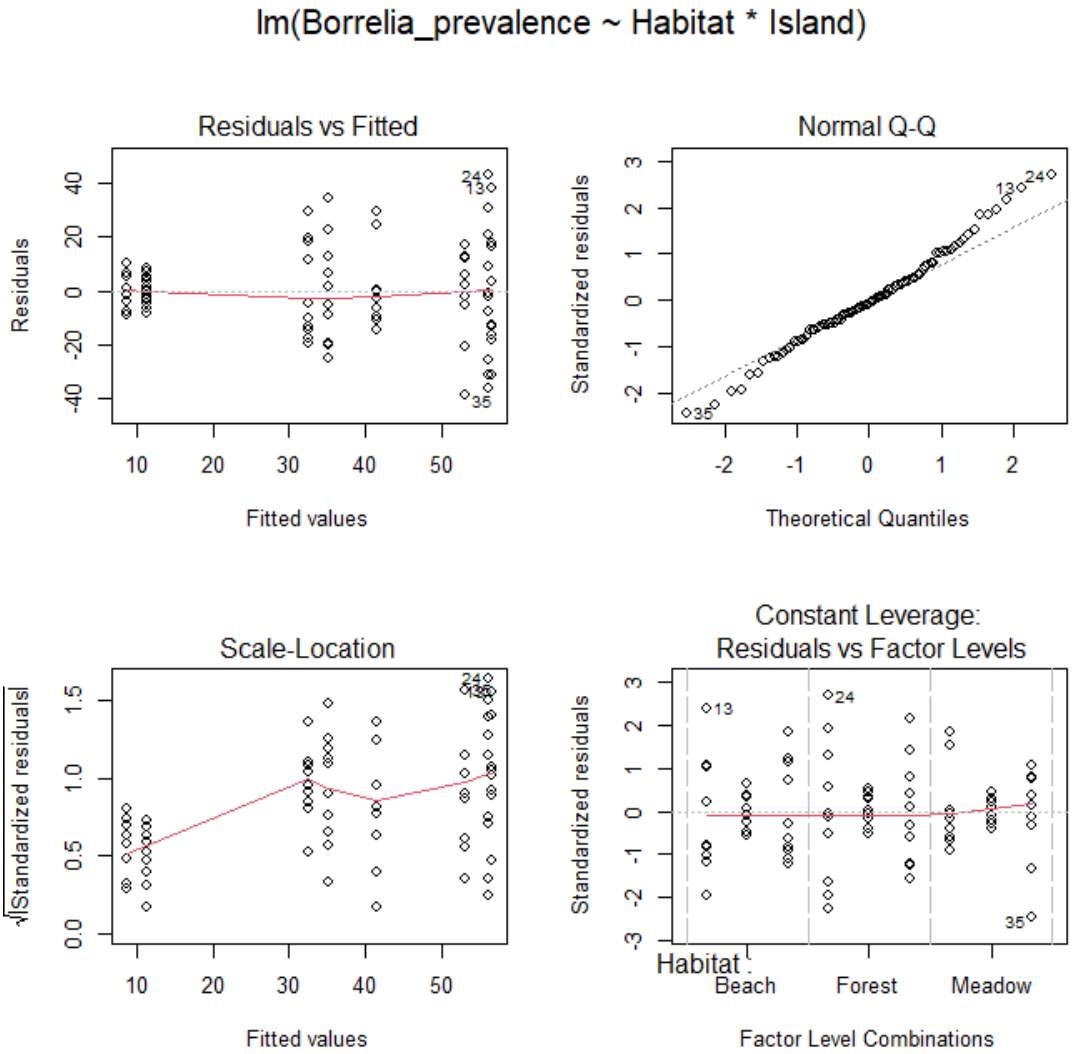
Diagnostic Diagrams

Normal Q-Q plot in the diagnostic diagrams below (Figure 6) shows that there are some deviations from reference line around right and left side of the graph. This indicates that the residual is not completely normally distributed. Residual vs Fitted values graph indicates that the biggest spread of the residuals is more than three times than the smallest spread.

Therefore, the transformation of the data is required in order to meet the assumption of normality and and homoscedasticity for Anova.

Figure 6

Diagnostic Diagrams for two-way ANOVA Model(Dependent variable:% of Borrelia infested ticks, Independent variable: Island, Habitat, Habitat & Island)

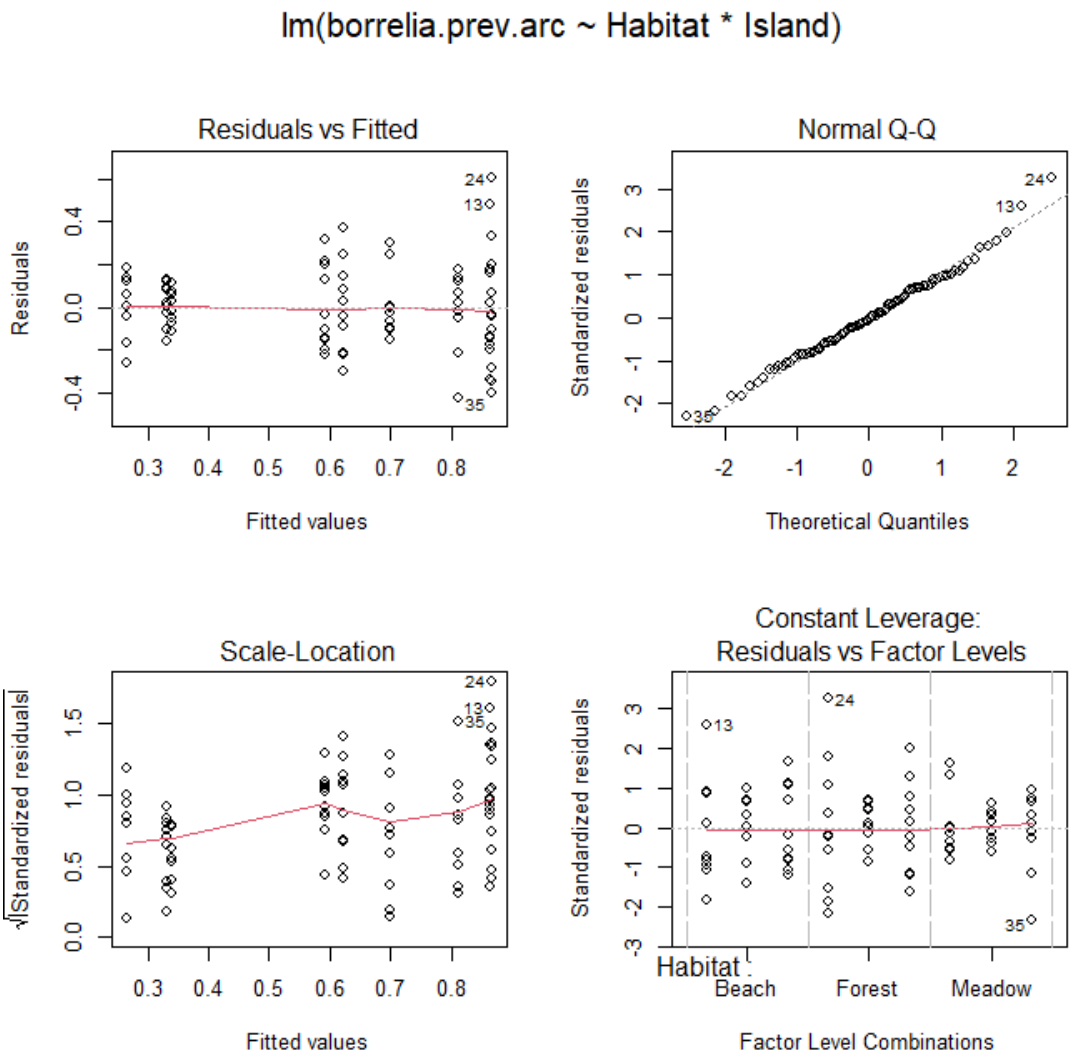


Diagnostic diagrams after arcsine transformation

After the arcsine transformation, normal Q-Q plot shows that there is less deviation from reference line. This indicates that the residual is more normally distributed. Residual vs Fitted values graph indicates that the biggest spread of the residuals is less than three times than the smallest spread.

Figure 7

Diagnostic diagrams after arcsine transformation for two way ANOVA Model(Dependent variable:% of Borrelia infested ticks, Independent variable: Island, Habitat, Habitat & Island)



According to the result of ANOVA test below(Table 4), Island(F value: 51.5, P value:3.703e-15)and combination(interaction) of Habitat and Island(F value:3.0478, P value:0.02153) have a statistically significant effect on prevalence of borrelia.

Table 4

Anova Table (Dependent variable:% of Borrelia infested ticks(arcsine transformation), Independent variable: Habitat, Island, Habitat*Island)

Anova Table (Type II tests)

	Sum Sq	Df	F value	Pr(>F)
Habitat	0.03111	2	0.4026	0.6699
Island	3.979	2	51.5	3.703e-15
Habitat:Island	0.471	4	3.048	0.02153
Residuals	3.129	81	NA	NA

The result of linear regression analysis (F-statistic: 14.5, Degree of freedom: 8 and 81, p-value: 0.0000) in the Table 5 shows that the linear model below is better fitted than the model without independent variable.

According to coefficients in the table below, there are significant difference in the risk of a Borrelia transmission among different island (Island[T.OJA]: t value -6.807, P value :0.00000, Island[T.TOR]: t value-3.065, p value:0.00295). Coefficients for Öja is -0,598 and the one for Torö är -0,269465, which are negative. It means that Askö has the highest risk of Borrelia transmission while Törö has the lowest risk among 3 Islands.

Additionaly, the results of Linear regression analysis below (Habitat Meadow:Island TOR P:0.0028, t value 3.084) in Table 5 indicates that there are also statistically significant interaction effect in the risk of Borrelia by meadows and Torö. Coefficients are positive 0.383, which means that the combination of meadows and Torö will relate to higher risk of borrelia.

Table 5

Linear Regression Model Coefficients (Dependent variable:% of Borrelia infested ticks(arcsine transformation), Independent variable: Island, Habitat, Habitat & Island)

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	0.863	0.062	13.882	0.0000	***
HabitatForest	0.005	0.088	0.052	0.9590	
HabitatMeadow	-0.164	0.088	-1.868	0.0653	.
IslandOJA	-0.598	0.088	-6.807	0.0000	***
IslandTOR	-0.269	0.088	-3.065	0.0030	**
HabitatForest:IslandOJA	0.063	0.124	0.507	0.6132	
HabitatMeadow:IslandOJA	0.239	0.124	1.925	0.0577	.

	Estimate	Standard Error	t value	Pr(>|t|)	
HabitatForest:IslandTOR	0.025	0.124	0.199	0.8427	
HabitatMeadow:IslandTOR	0.383	0.124	3.084	0.0028	**

*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 81 and 8 DF, p-value: 0.0000

Post-hoc test

According to post-hoc test below (Table 6), there are no statistically significant differences between habitat. (Beach-Forest F test:0.4436, P value: 1, Beach-Meadow F test: 0.7289, P value:1, Forest-Meadow F test: 0.0352, P value:1)

Table 6

Result of Post-hoc test(Habitat)

```
## F Test:
## P-value adjustment method: holm
##                   Value Df Sum of Sq      F Pr(>F)
##  Beach-Forest -0.033803  1   0.01714 0.4436      1
##  Beach-Meadow -0.043328  1   0.02816 0.7289      1
## Forest-Meadow -0.009525  1   0.00136 0.0352      1
## Residuals               81   3.12943
```

According to post-hoc test below (Table 7), there are statistically significant differences between ASKÖ and ÖJA (F test: 96.1136 P value:0.000), ASKÖ and TORÖ(F test: 6.9132, P value: 0.01024) as well as ÖJA and TORÖ(F test 51.4730 P value: 0.000).

Table 7

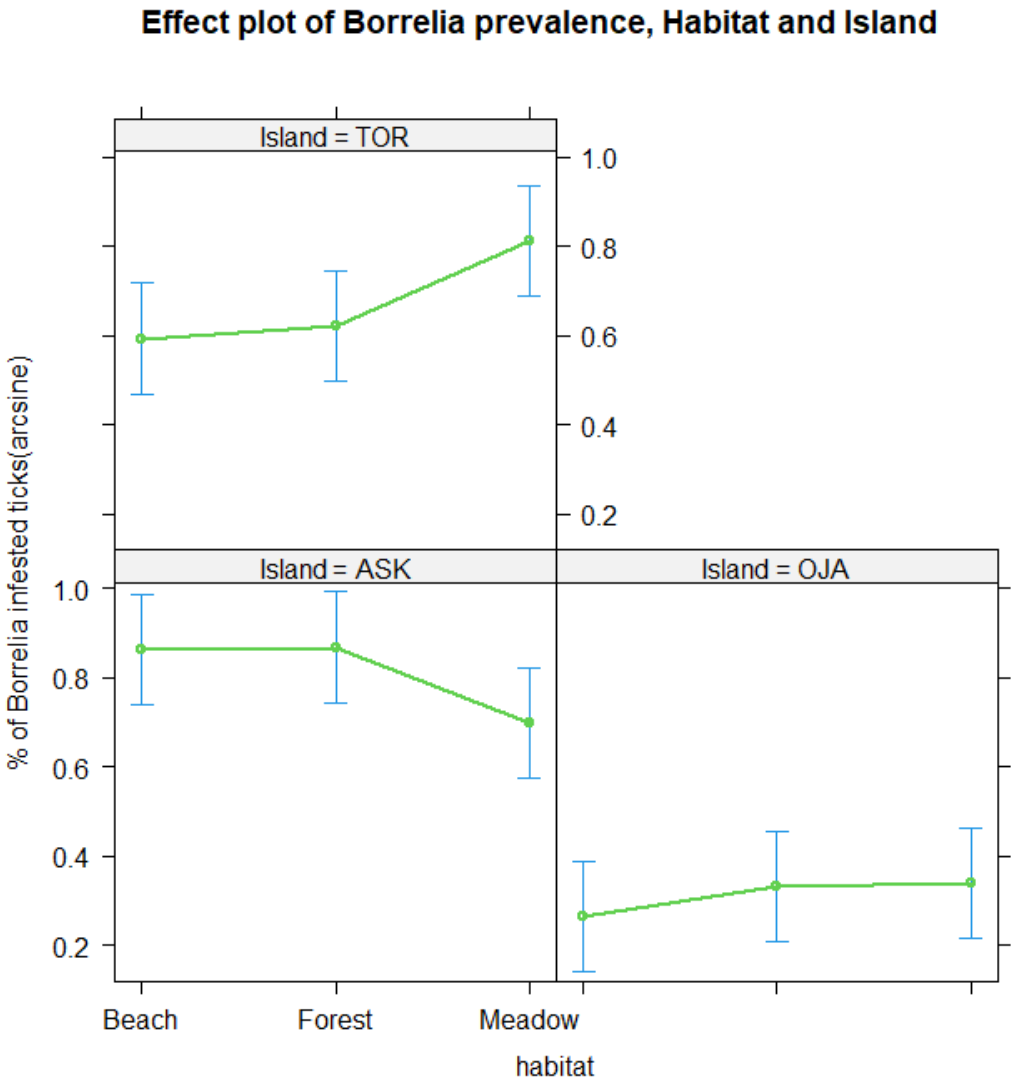
Result of Post-hoc test(Island)

```
## F Test:
## P-value adjustment method: holm
##              Value Df Sum of Sq       F    Pr(>F)
## ASK-OJA    0.49755  1    3.7133 96.1136 6.185e-15 ***
## ASK-TOR    0.13344  1    0.2671  6.9132   0.01024 *
## OJA-TOR   -0.36411  1    1.9887 51.4730 6.198e-10 ***
## Residuals          81    3.1294
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The effect plot below (Figure 8) shows that there are differences between 3 island and Askö has the highest risk of prevalence of Borrelia in Beach and Forest. The graph also indicates that prevalence of Borrelia is higher in Meadow in Torö than Meadow in Askö and Öja.

Figure 8

Effect plot of Borrelia prevalence, Habitat and Island



Discussion

The data in the study indicates that there are differences among 3 islands in prevalence of borrelia and Askö has the highest risk of prevalence of Borrelia. There are no statistically significant differences among habitat in prevalence of Borrelia according to ANOVA test. However, statistically significant interaction effect (habitat and island) was identified. Meadow in Torö has higher risk of prevalence of Borrelia than other habitat in Torö and

than Meadows in other islands. This could be because the meadows in Torö might have particular/different characteristics that is highly correlated to prevalence of Borrelia.

References

Appendix

Import Libary and data

```
library(Rcmdr) #this chunk start RCommander
library(ggplot2) #this cunks starts ggplot2
library(flextable)
library(faraway)
library(car)
library(dplyr)
library(abind, pos=17)
library(e1071, pos=18)

library(mvtnorm, pos=24)
library(survival, pos=24)
library(MASS, pos=24)
library(TH.data, pos=24)
library(multcomp, pos=24)
library(phia)
library(Rmisc)
library(xtable)

#include a chunk that opens your indata
Dataset <-

read.table("C:/Users/y_tat/OneDrive/Documents/R/sh/Assignment2/borrelia.txt",
   header=TRUE, stringsAsFactors=TRUE, sep="\t", na.strings="NA",
  dec=",", strip.white=TRUE)
```

Part 1 One-way Anova

```
AnovaModel.2 <- aov(Borrelia_prevalence ~ Island, data=Dataset)
summary(AnovaModel.2)

local({
  .Pairs <- glht(AnovaModel.2, linfct = mcp(Island = "Tukey"))
  print(summary(.Pairs)) # pairwise tests
  print(confint(.Pairs, level=0.95)) # confidence intervals
  print(cld(.Pairs, level=0.05)) # compact letter display
  old.oma <- par(oma=c(0, 5, 0, 0))
  plot(confint(.Pairs))
  par(old.oma)
})
```

Histogram:Borrelia prevalence(% of Borrelia infested ticks)

```
ggplot(Dataset) +
  aes(x = Borrelia_prevalence) +
```

```
  geom_histogram(bins = 30L, fill = "#112446") +
  labs(x = "% of Borrelia infested ticks", y="frequency", title="Histogram:
Borrelia_prevalence" ) +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))
```

Diagnostic diagrams for Anova one-way test (Dependent variable:% of Borrelia infested ticks, Indepent variable: Island)

```
oldpar <- par(oma=c(0,0,3,0), mfrow=c(2,2))
plot(AnovaModel.2)
par(oldpar)
```

Arcsine transformation of % of Borrelia infested ticks

```
Dataset$borrelia.prev.arc<-asin(sqrt(Dataset$Borrelia_prevalence/100))
```

Histogram: % of Borrelia infested ticks after arcsine transformation

```
ggplot(Dataset) +
  aes(x = borrelia.prev.arc) +
  geom_histogram(bins = 30L, fill = "#112446") +
  labs(x = " % of Borrelia infested ticks(arcsine transformation)",
y="frequency", title="Borrelia Prevalance(arcsine transformation)" ) +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))

lm_island<-lm(borrelia.prev.arc~Island,data=Dataset)
summary(lm_island)
```

Diagnostic diagrams after arcsine transformation

```
oldpar <- par(oma=c(0,0,3,0), mfrow=c(2,2))
plot(lm_island)
par(oldpar)
```

Result of Anova one-way test after after Arcsine transformation

```
Anova_island<-anova(lm_island)
pander(Anova_island)
```

Linear Regression Coefficient (Dependent variable:% of Borrelia infested ticks, Independent variable: Island)

```
lm_island=lm(borrelia.prev.arc~Island,data=Dataset)
lm_island %>% as_flextable()
```

Pairwise comparisons(Multiple Comparisons of Means: Tukey Contrasts)

```
summary(glht(lm_island, linfct=mcp(Island="Tukey")))
```

####Plot of Means (Borrelia prevalence by Island)

```
sum_borrelia_prevalance<-summarySE(Dataset, measurevar="borrelia.prev.arc",
groupvars="Island")
ggplot(sum_borrelia_prevalance, aes(x=Island, y=borrelia.prev.arc,
color=borrelia.prev.arc)) +
  geom_errorbar(aes(ymin=borrelia.prev.arc-se, ymax=borrelia.prev.arc+se),
```

```
width=.8) +
   labs(x = "Island", y = " % of Borrelia infested ticks(arcsine
transformation)",title = "Plot of Means (Borrelia prevalance by Island)")+
  geom_line() +
  geom_point()+
  theme_classic()
```

Plot of Means (Borrelia prevalance by Island)

```
attach(Dataset)
library(gplots)
plotmeans(borrelia.prev.arc ~ Island, connect=list(1:5),
            ccol="black", pch=16, cex.axis=0.95)

detach(Dataset)
```

Part 2 Two-way ANOVA interaction

Two-way ANOVA Model(Dependent variable:% of Borrelia infested ticks, Independent variable: Island, Habitat, Habitat & Island)

```
RegModel.HI <- lm(Borrelia_prevalence~Habitat*Island, data=Dataset)
summary(RegModel.HI)

AnovaModel.HI=Anova(RegModel.HI)
AnovaModel.HI
```

Diagnostic Diagrams for two-way ANOVA Model(Dependent variable:% of Borrelia infested ticks, Independent variable: Island, Habitat, Habitat & Island)

```
oldpar <- par(oma=c(0,0,3,0), mfrow=c(2,2))
plot(RegModel.HI)
par(oldpar)
```

ANOVA Model after arcsine transformation(Dependent variable:% of Borrelia infested ticks, Independent variable: Island, Habitat, Habitat & Island)

```
RegModel.HI_arc <- lm(borrelia.prev.arc~Habitat*Island, data=Dataset)
```

Diagnostic diagrams after arcsine transformation for two way ANOVA Model(Dependent variable:% of Borrelia infested ticks, Independent variable: Island, Habitat, Habitat & Island)

```
oldpar <- par(oma=c(0,0,3,0), mfrow=c(2,2))
plot(RegModel.HI_arc)
par(oldpar)
```

Anova Table after arcsine transformation for two way ANOVA Model(Dependent variable:% of Borrelia infested ticks, Independent variable: Island, Habitat, Habitat & Island)

```
Anova_HI_arc<-Anova(RegModel.HI_arc)
pander(Anova_HI_arc)
```

*Linear Regression Model Analysis of effect on % of Borrelia infested ticks(arcsine transformation) by Habita, Island, Habitat*Island*

```
RegModel.HI_arc <- lm(borrelia.prev.arc~Habitat*Island, data=Dataset)
RegModel.HI_arc %>% as_flextable()
```

Result of Post-hoc test(Habitat)

```
testInteractions(RegModel.HI_arc, pairwise="Habitat", adjustment="holm")
```

Result of Post-hoc test(Island)

```
testInteractions(RegModel.HI_arc, pairwise="Island", adjustment="holm")
```

Effect plot of Borrelia prevalence, Habitat and Island

```
plot(effect(term="Habitat:Island",mod=RegModel.HI_arc,se=TRUE, x.var=
"Habitat"), ylab=" % of Borrelia infested ticks(arcsine)", xlab= "habitat",
main="Effect plot of Borrelia prevalence, Habitat and Island", colors =
c(3,4))
```