“STA106 MIDTERM PROJECT - Spring 2022”

Aditya Mittal & Mitchell Lawrence

Contents

Question I - sparrow.csv

Introduction

Sparrows are birds that live across different continents such as Northern Africa, Europe, the Americas, and much of Asia. These birds migrate seasonally, often travelling to Northern/Central America during the winter where they build new homes. Kent Island, located on the eastern shore in Maryland, often attracts many of these Sparrows who then live there for months.

To better understand the preferable nesting conditions for migrating sparrows, our main focus is to use the data from the experiment and determine whether the size of the sparrows attracted to Kent Island based on nest sizes (control, enlarged, reduced) is statistically significantly different among nesting groups. We are also interested in studying the possible range of sizes of sparrows for the nests that tends to have the largest sparrow. Furthermore, we comparing the differences in weights between control and reduced/enlarged nests to see which group weights differ from one another and if there's statistically significant evidence of variation based on nest groups.

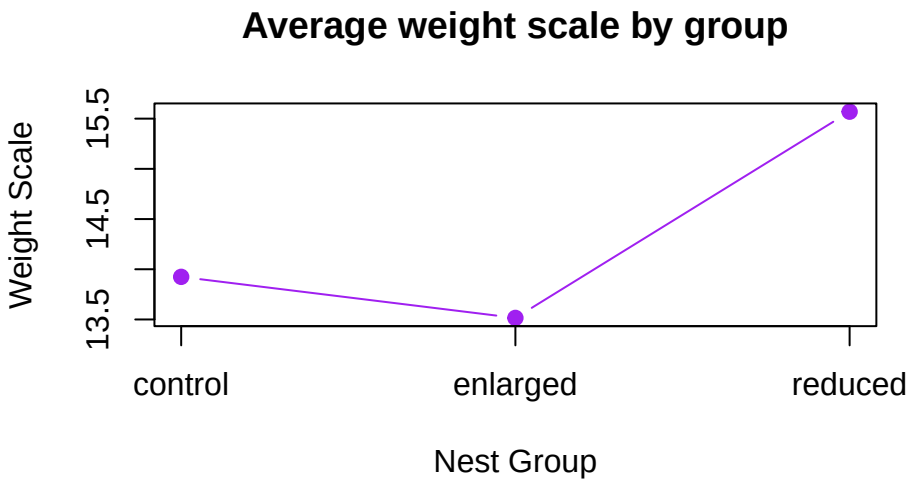
We will use two methods to test for differences in sparrow weight based on grouping:

1. We want to use the Single Factor Anova: F-test for equal means to conduct a hypothesis test to determine if the average weight of the sparrow is different for at least one of the groups.
2. We are creating 3 separate confidence intervals. Our first confidence interval will show the range of weights of the nest that tends of have the largest sparrow. Additionally, we are interested in checking if there is statistically significant evidence of a difference in weights between control and enlarged & reduced nest groups by creating pairwise confidence intervals for differences in means.

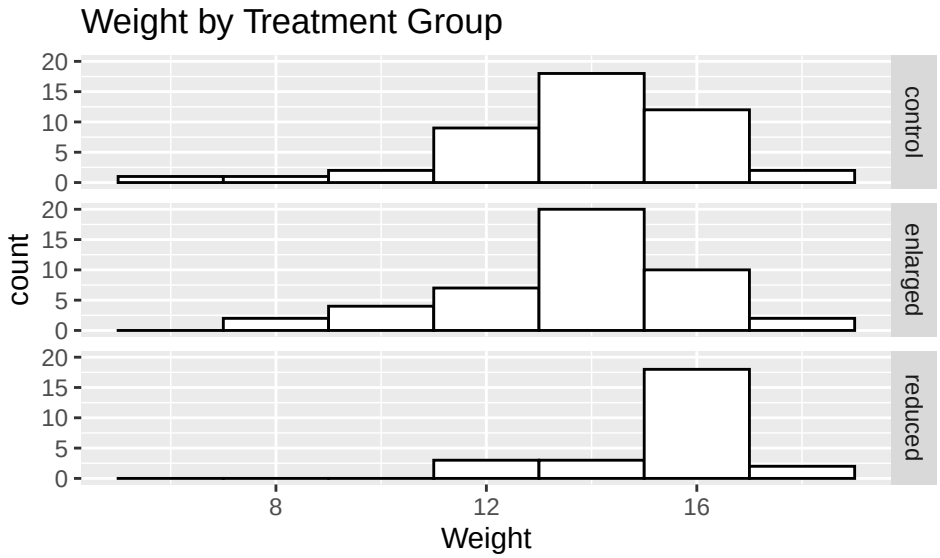
Ultimately, both these methods will help determine whether or not there is statistically significant in the average sizes of sparrows attracted based on nest size. Kent Island can use this information in the future to build different nest sizes in order to attract certain specific types and sizes of sparrows.

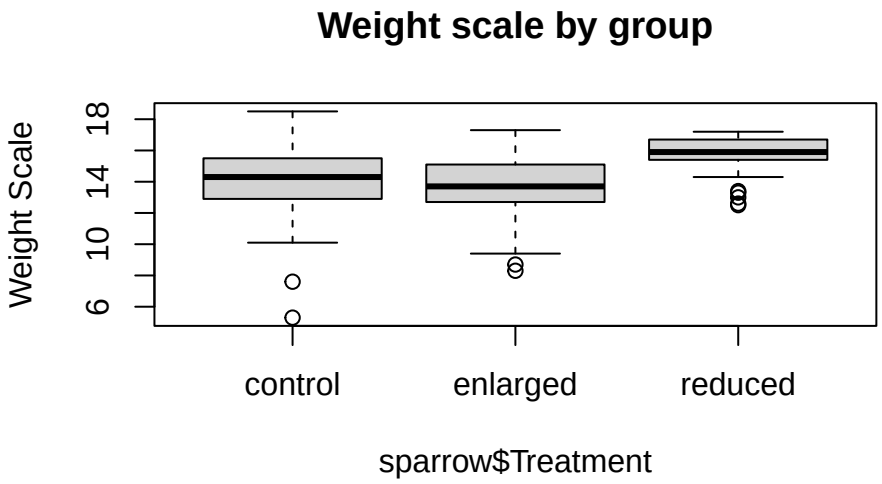
Data Summary

The sparrow dataset contains 116 subjects. There are three treatment nest size categories - control, enlarged, and reduced. To get a more accurate visualization of the data, we create a summary table to showcase the group means, standard deviations, and sample sizes.



Based on our plot of group means, there is a difference between the average weight of sparrow attracted among nesting groups. On average, the reduced size nest group attracted sparrows at about 15.5692 grams whereas the nest group control and enlarged attracted sparrows weighing at 13.9244 and 13.5156 grams, respectively.





The variation in the spread of the data in the histogram and boxplot suggests that the population standard deviations are unequal. Using summary values, we get the sd of sparrow weights in reduced size nest group of 1.4593, whereas the nest group control and enlarged attracted sparrows weighing at 2.4196 and 2.104 grams, respectively. Based on the box plot, the control nest has the widest range of sparrow weights attracted, followed by enlarged nest group and then the narrowest range of data contained in the reduced nesting group. We also note there are are several outliers visible in the boxplot, which we will find and remove using standardized residuals in part III.

It is also important to note that the variance among sample groups are also unequal as the sample standard deviation varies among groups. Additionally, note that the sample sizes for both control and enlarged nests contained 45 sampled sparrows, as opposed to 26 subjects sampled in reduced nest.

Furthermore, the difference in group means and the small standard deviation value of the reduced group may indicate a potential difference of average weight based on nesting groups. We will test this claim further with F-test for equal means and Confidence Intervals.

Diagnostics

To ensure that the ANOVA model is the best fit for our data, we must check if our data follows the needed assumptions.

A general ANOVA model must follow three basic assumptions: 1. All Yij's (each individual sample) were independently, randomly sampled 2. All groups are independent 3. Each error term, Eij follows a normal distributed with mean = 0, equal population variance = $sigma^2$

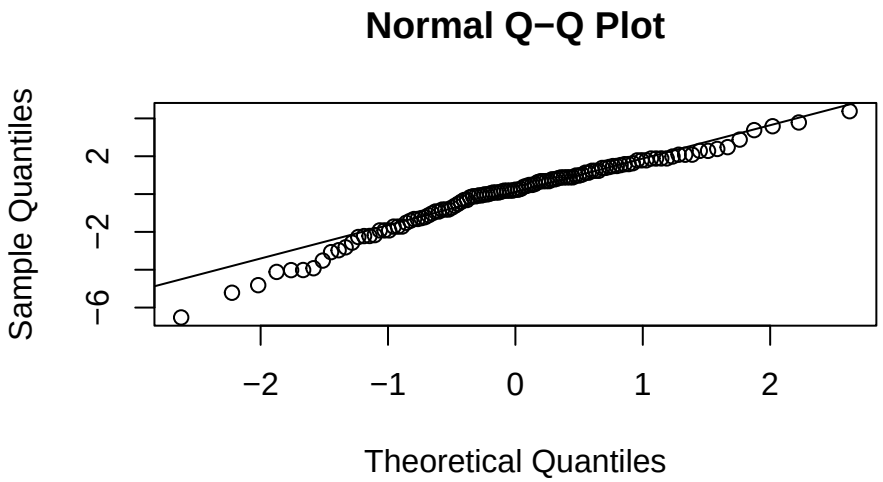
The condition of independence is the most important assumption for the ANOVA model's validity. The first and second condition are met as all sparrows were independently, randomly sampled. Each nesting group (control, reduced, enlarged) is also independent as the weights of sparrows in one nesting group does not provide information about the subjects in other groups. For our third assumption, we need to check if the random error terms follow normality and have equal population variances across groups. First, however, we need to check for and remove potential outliers from our data. Outliers tend increase the overall sample variance within group (MSE), thus lowering the calculated F-statistic value and the chance of rejecting the null hypothesis. Since the box plot revealed that the data contained outliers, we will use two separate methods to find them: Semi-studentized residuals and the Studentized residuals.

Using the Semi-standardized residuals, we are using the assumption that there is equal variance among groups. However, as we noted from earlier that our standard deviations in the data were not equal, we also use the Studentized residuals as it relaxes the assumption that all variances are equal.

$CO.eij = 45$, $CO.rij = 45$. Using both the methods, row 45 of the dataset sparrow contains an outlier. We go ahead and remove this row. Now, we have created a new dataset with 115 entries and no outliers (as tested by the Studentized Residual Method).

After removing the outlier value from the dataset, we can assess the normality of the random error terms using QQ plot. Note, assessing normality based on the QQ plot is subjective thus we also use a more robust method of checking - the Shapiro Wilkis Test.

We begin with assessing the normality QQ Plot:



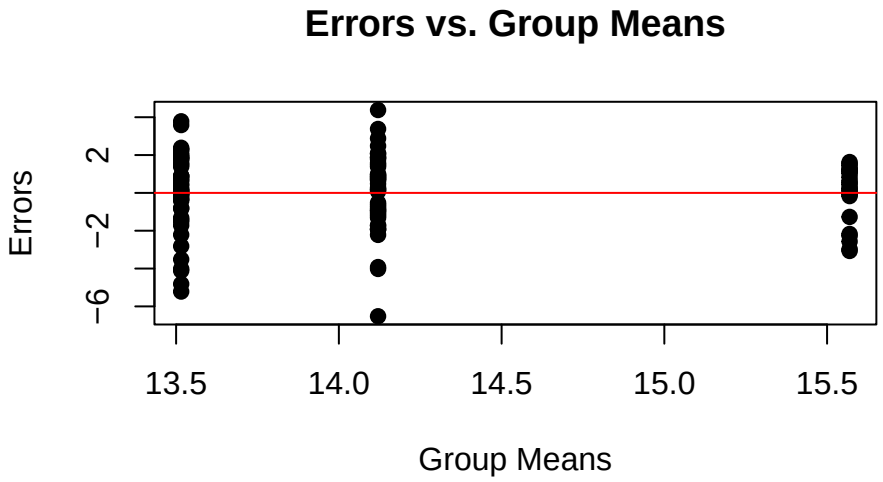
There is slight deviation from the normal line around the end values. Since most quantities are close to the line, we may conclude the data is approximately normal. However, as stated earlier, we will also conduct the Shapiro-Wilkis test to get a more accurate result.

The null and alternate hypothesis for Shapiro-Wilkis test are: $H0$: The data is normally distributed vs. HA: The data is not normally distributed

Since our Shapiro-Wilkis test pvalue = 0.00371448569200948 < common alpha levels (0.01,0.05,0.1), we reject H0 in favor of Ha. We conclude that the the distribution of eij are non-normal. Thus, this condition is violated.

We also want to check homoscedasticity - whether population variances are equal among groups. Similarly, we plot errors vs group to get a approximate graph and also conduct a robust Brown-Forsythe Test to test for constant variance.

We begin with assessing the normality error vs group Plot:



Based on the spread of errors vs group, it appears that variance based on group different because of the unequal vertical spread of the plot. However, this chart may be misleading due to difference in sample sizes among groups; thus, we also conduct Brown-Forsythe Test to get more accurate results.

The hypothesis for Shapiro-Wilkis test are: $H_O : \sigma_C^2 = \sigma_E^2 = \sigma_R^2$ vs. H_A: At least one group variance is unequal.

Since our p-value = 0.2292946 > alpha = 0.05, we fail to reject H0 in favor of Ha. We can conclude that the sample variance among groups are equal and the condition of homoscedasticity is met.

Overall, the third condition is violated that requires the random error terms to be normally distributed. Since the data is non-normal, we may choose to conduct transformation of variables, but that is outside the scope of this study. Despite the non-normality of errors, the single factor ANOVA F-test for equal means is still a robust test to compare average sparrow weight based on nest groups and check if there's variation

Analysis

This section returns the important values from the F-test for equal means and the single factor/pairwise confidence intervals. Interpretations and conclusions based on these values are in the following sections.

The Single Factor ANOVA cell means model is a good model fit for this data since we are interested in finding the group means of each group. The cell means model is: $Yij = \mu_i + \epsilon_{ij}$

where:

Yij = jth value of Y observation in ith group

μ_i = unknown true population mean for ith group

ϵ_{ij} = jth residual for ith group

Despite the normality condition being violated, we can assume normality due to the central limit theorem (our sample size n > 30). Furthermore, the ANOVA model can still used as a robust test in the case of non-normality with little effect on Type I error.

Using the single factor ANOVA F-test for equal means, our null and alternate hypothesis are:

```
## Analysis of Variance Table
##
## Response: Weight
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Treatment   2  70.09  35.045  9.1375 0.0002108 ***
## Residuals 112 429.55   3.835
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : \mu_C = \mu_E = \mu_R$, where μ_i is the average weight of the attracted sparrow based on i-th sparrow group. vs. H_A : At least one of the average weights of sparrow is different (not equal) based on nest group. Using the anova table, we get the calculated test statistic to be 9.1375438. The corresponding p-value for this test is 2.1084938×10^{-4}. Since our pvalue is 2.1084938×10^{-4} < all common alpha levels (0.1,0.05,0.01), we can reject H0 in favor of HA.

The power of test is the probability that we reject H0, given H0 is false. Thus, by avoiding making a type II error, power refers to the probability we make the correct decision. As we rejected H0, we can calculated the power of this test to be 0.9547259.

Furthermore, as we were intersted in finding an interval of sparrow sizes for the nest that tends to have the largest sparrow, we create 3 separate 95% confidence intervals to get possible range of sparrow sizes by each group.

We are 95% confident that the weight of sparrows in nesting group "control" is between (13.2997184, 14.5491705).

We are 95% confident that the weight of sparrows in nesting group "enlarged" is between (12.8908295, 14.1402816).

We are 95% confident that the weight of sparrows in nesting group "reduced" is between (14.7473493, 16.3911123).

Lastly, as we've concluded there's statistical difference in between at-least one group, we can create pairwise confidence interval for difference in means to compare sparrow sizes of control group to enlarged/reduced group.

We are 95% confident that the sparrow weight for control group is less than the sparrow weight for the reduced nest group by (0.612424, 2.6771486)

The 95 % confidence interval for difference in sparrow weights between control group and enlarged group is between (-0.4746072, 1.292385). Since 0 is contained within this interval, we can conclude there is no difference in sparrow size based on nesting group control and enlarged.

Thus, based on hypothesis testing and pairwise CI for difference in means, we can conclude that reduced nesting group has statistically larger sparrow sizes compared to control group and enlarged group.

Interpretation

Using our data from the analysis section, we can make conclusions based on hypothesis testing and confidence intervals.

Since our pvalue = 0.0002108 < alpha = 0.05, we reject H0 in favor of HA. We can conclude that there is a statistically significant difference in average weight of sparrows attracted in at least one nesting group. Since we rejected the null hypothesis, we may have made a type I error (rejecting H0 when in reality it is true). Using alpha = 0.05, there is 5% chance we falsely rejected the null hypothesis that all group means are equal, when in reality they are equal.

The power of the test is 0.9547259, thus there is 95.47% we rejected the null hypothesis that states the average weight of sparrow based on group were equal, when in reality they were not.

The 95% pairwise CI for difference in means helps us find which group had a statistical difference in sparrow sizes as compared to others. Since 0 was included in the pairwise CI comparing control and enlarged nest group, we concluded that there is no statistical evidence of a difference in weights between these groups. However, as we concluded earlier, the sparrow weight for control group is less than the sparrow weight for the reduced nest group by 0.612424 and 2.6771486 grams. Furthermore, since the lower bound of reduced nest group interval is greater than the upper bound for both control and enlarged group intervals, we can suggest this interval contains the sparrow weights for the nest that tends to have largest sparrows. Thus, there is statistical evidence to suggest the reduced nest group attracts larger sparrows as opposed to the other two groups.

Conclusion

Migration of birds during seasonal changes is essential for their survival and changing climates has made this process much more difficult for them. Once we fit our data to follow the assumptions of single factor ANOVA and follow the model $Yij = \mu_i + \epsilon_{ij}$, we used our single factor ANOVA methods we were able to come to conclude that the average sparrow weight was different for at least one group.

Furthermore, we have proven through single factor ANOVA that there is statistical evidence to prove there is a difference in mean Sparrow weight based on their nest size. More specifically, the confidence intervals we constructed indicated that the reduced group had significantly larger sparrows compared to the other two groups. Thus, the results of this report can help guide the migration status of many sparrows in the upcoming year as placing reduced nests will tend to attract larger Sparrows.

Appendix of Code Used

```
knitr::opts_chunk$set(echo = FALSE,
message = FALSE,
warning = FALSE,
fig.width= 5,
fig.height= 3,
fig.align= 'center')
# read dataset sparrow.csv
sparrow <- read.csv("~/Desktop/Year1/Spring Quarter 2022/STA106/Project1/sparrow (1).csv")

# setup basic functions for analysis
# power function
give.me.power = function(ybar,ni,MSE,alpha){
  a = length(ybar)
  nt = sum(ni)
  overall.mean = sum(ni*ybar)/nt
  phi = (1/sqrt(MSE))*sqrt( sum(ni*(ybar - overall.mean)^2)/a)
  phi.star = a *phi^2
  Fc = qf(1-alpha,a-1,nt-a)
  power = 1 - pf(Fc, a-1, nt-a, phi.star)
  return(power)
}

# confidence interval function
give.me.CI = function(ybar,ni,ci,MSE,multiplier){
  if(sum(ci) != 0 & sum(ci !=0 ) != 1){
    return("Error - you did not input a valid contrast")
  } else if(length(ci) != length(ni)){
    return("Error - not enough contrasts given")
  }
  else{
    estimate = sum(ybar*ci)
    SE = sqrt(MSE*sum(ci^2/ni))
    CI = estimate + c(-1,1)*multiplier*SE
    result = c(estimate,CI)
    names(result) = c("Estimate","Lower Bound","Upper Bound")
    return(result)
  }
}

# Part II)
# Summary Statistics
group.means = by(sparrow$Weight,sparrow$Treatment,mean)
group.sds = by(sparrow$Weight,sparrow$Treatment,sd)
group.nis = by(sparrow$Weight,sparrow$Treatment,length)
the.summary = rbind(group.means,group.sds,group.nis)
the.summary = round(the.summary,digits = 4)
colnames(the.summary) = names(group.means)
rownames(the.summary) = c("Means","Std. Dev","Sample Size")

# group means plot/table
plot(group.means,xaxt = "n",pch = 19,col = "purple",xlab = "Nest Group",ylab = "Weight Scale",
     main = "Average weight scale by group",type = "b")
```

```r
axis(1,1:length(group.means),names(group.means))

# histogram
library(ggplot2)
ggplot(sparrow, aes(x = Weight)) + geom_histogram(binwidth = 2,,color = "black",fill = "white") +
  facet_grid(Treatment ~.) +ggtitle("Weight by Treatment Group")

# boxplot
boxplot(sparrow$Weight ~ sparrow$Treatment, main = "Weight scale by group",
        ylab = "Weight Scale")

# Part III)
# finding outliers via Semi-studentized/standardized residuals
the.model = lm(Weight ~ Treatment,data = sparrow)
sparrow$ei = the.model$residuals
nt = nrow(sparrow)
a = length(unique(sparrow$Treatment))
SSE = sum(sparrow$ei^2)
MSE = SSE/(nt-a)
eij.star = the.model$residuals/sqrt(MSE)

alpha = 0.05
t.cutoff= qt(1-alpha/(2*nt), nt-a)
CO.eij = which(abs(eij.star) > t.cutoff)

# finding outliers via studentized/standardized residuals
rij = rstandard(the.model)
CO.rij = which(abs(rij) > t.cutoff)
# remove outlier row
CO1 = c(CO.rij)
outliers = CO1
new.data = sparrow[-outliers,]
new.model = lm(Weight ~ Treatment,data = new.data)

# assessing normality using qq plot
qqnorm(new.model$residuals)
qqline(new.model$residuals)

# assessing normality Shapiro Wilkis test
ei = new.model$residuals
the.SWtest = shapiro.test(ei)

# Assess constant variance
plot(new.model$fitted.values, new.model$residuals, main = "Errors vs. Group Means",
     xlab = "Group Means",ylab = "Errors",pch = 19)
abline(h = 0,col = "red")

# Brown-Forsythe Test
library(car)
the.BFtest = leveneTest(ei~ Treatment, data=new.data, center=median)
p.val = the.BFtest[[3]][1]

# anova table
```

```r
new.model = lm(Weight ~ Treatment,data = new.data)
anova.table = anova(new.model)
anova.table

# get power of test
the.power = give.me.power(group.means,group.nis,MSE,0.05)
#Confidence interval for largest sparrow
t.value = qt(1-0.05/2, sum(group.nis) - length(group.nis))
ci.1 = c(1,0,0)
ci.2 = c(0,1,0)
ci.3 = c(0,0,1)

CI1 = give.me.CI(group.means,group.nis,ci.1,MSE,t.value)
CI2 = give.me.CI(group.means,group.nis,ci.2,MSE,t.value)
CI3 = give.me.CI(group.means,group.nis,ci.3,MSE,t.value)

# pairwise CI
# control - reduced
ci.4 = c(1,0,-1)
# control - enlarged
ci.5 = c(1, -1, 0)

CI4 = give.me.CI(group.means,group.nis,ci.4,MSE,t.value)
CI5 = give.me.CI(group.means,group.nis,ci.5,MSE,t.value)
```