Practicum #2

Mason Deja

8/27/17

EDA for total length

1. The total length for Ruffe captured in the St. Louis Harbor in Lake Superior is slightly left skewed with no obvious outliers (Figure 1). The center is at a mean of 126.25 and the dispersion is 22.21(Table 1). The mean and standard deviation were used because the distribution was not strongly skewed and no outlier was present.

Figure 1- total length for Ruffe captured in the St. Louis Harbor in Lake Superior is slightly left skewed with no obvious outliers

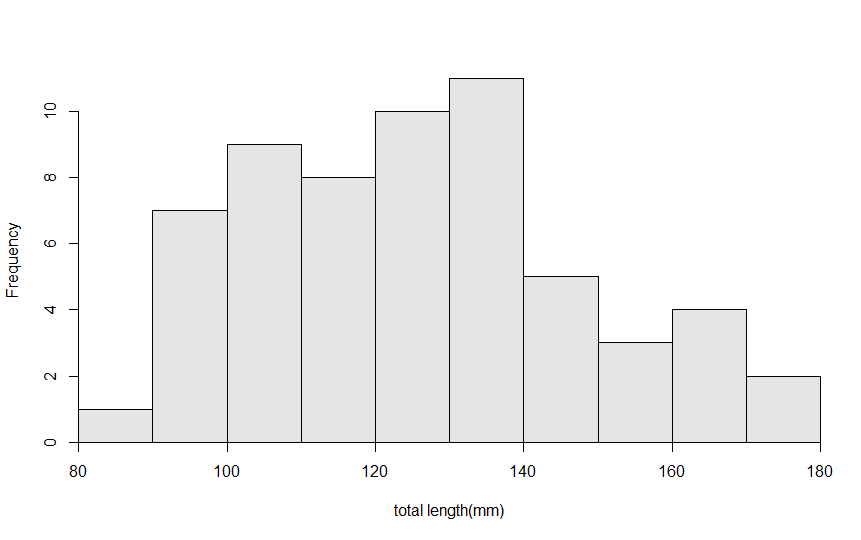


Table 1- Descriptive statistics of total length for Ruffe captured in the St. Louis Harbor in Lake Superior

n mean sd min Q1 median Q3 max

60.000 126.247 22.213 86.700 108.400 125.800 139.700 174.000

EDA for sex

1. The majority of Ruffe captured in the St. Louis Harbor in Lake Superior were mostly females with 42 females and only 18 males captured(Figure 2,Table2).

Figure 4-bar chart showing the frequency of female fish vs male fish captured

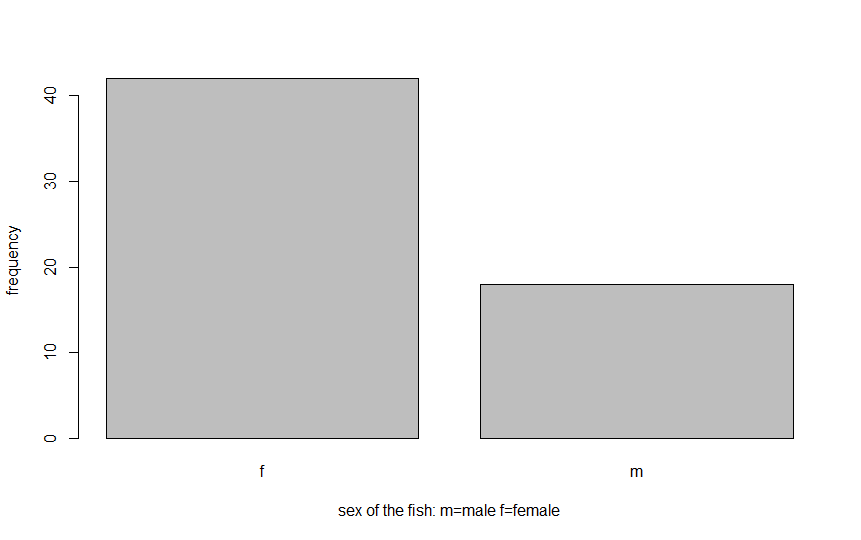


Table 4- showing the amount of males and females captured

Sex

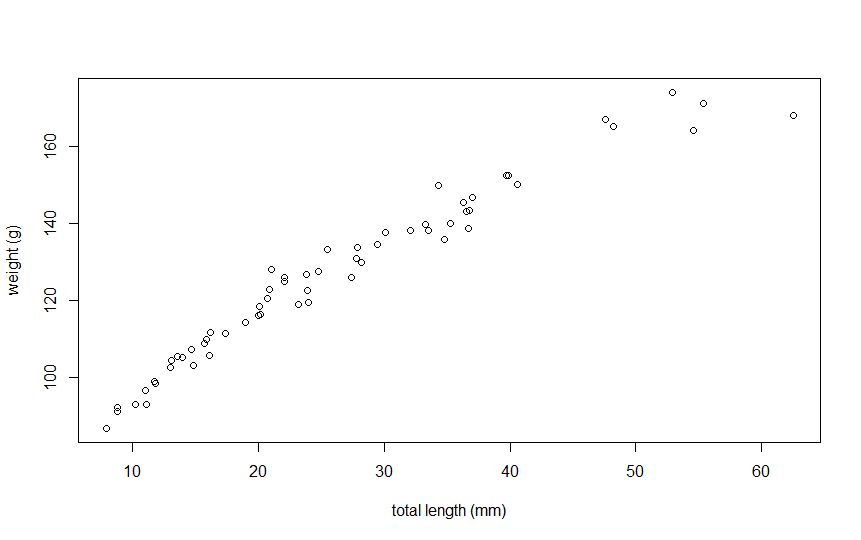
f m

42 18

EDA for the relationship between total length and weight of Ruffe

1. The relationship between weight and total length of Ruffe captured in the St. Louis Harbor in Lake Superior (Figure 3) appears to be linear, positive and a strong correlation of 0.975 with no obvious outliers (very slight curve). The correlation coeﬃcient was used to assess strength because I deemed the relationship to be linear without any outliers.

Figure 3- showing the relationship between weight and total length of Ruffe captured in the St. Louis Harbor in Lake Superior



EDA for location of collection and sex

1. The majority of Ruffe captured in the St. Louis Harbor in Lake Superior at all three locations were females with Whaleback having the highest percentage of females at 73.68% and males at 26.31% (Table 3).

Table 3-percentage table showing the distribution of males and females at three different locations

sex Allouez Interstate Whaleback

f 71.42857 65.00000 73.68421

m 28.57143 35.00000 26.31579

Sum 100.00000 100.00000 100.00000

Linear regression to predict body girth from total length

1. The slope indicates that for every one millimeter increase of total length the body girth increases by 0.616(table 4).
2. The predicted body girth of a Ruffe whose total length is equal to the median total length (125.85) is 78.75.
3. The proportion of the total variability in body girth is explained by knowing the total length is r2=0.867.

Table 4-results of a linear regression predicting body girth from total length of Ruffe

Coefficients:

(Intercept) tl

1.2605 0.6157

Ruffe hypothesis tests

Ruffe from Allouez vs Ruffe from Whaleback

1. A=0.05
2. Ho:UAllouez-UWhaleback=0 vs Ha:UAllouez-UWhaleback>0.
3. A 2 sample t-test is required because (i) a quantitative variable was measured (weight of a Ruffe), (ii) two populations were sampled ( Allouez and Whaleback) and (iii) populations were independent of each other(two locations on lake superior were sampled).
4. The data appears to be from an observational study with the Ruffe being sampled at various locations which implies randomization.
5. (i) 40=40, (ii)individuals are independent as discussed above and (iii) the variances appear to be equal because the leavens tests p-value=0.728 which is greater than a(table 5).
6. Xbar-xbar=27.06-24.63=2.43(Table 6).
7. T=0.498 with 38 degrees of freedom(Table 6).
8. P-value=0.6215(Table 6)
9. DNR Ho.
10. The mean weight of Ruffe from Allouez is different from Ruffe of Whaleback.

Table 5-Levene's Test for Homogeneity of Variance (center = median)

Df F value Pr(>F)

group 1 0.1223 0.7285

38

Table 6-Two Sample t-test with w by loc

t = 0.4978, df = 38, p-value = 0.6215

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-6.534789 10.796444

sample estimates:

mean in group Allouez mean in group Whaleback

27.05714 24.92632

The distribution of Ruffe into the sexes differs among the three locations.

1. A=0.05
2. Ho: “the distribution of Ruffe into the sexes differs among the three locations” vs Ha: “the distribution of Ruffe into the sexes DOES not differ among the three locations”.
3. A Chi square test is required because (i) a categorical variable was recorded (male or female) (ii) and three locations were sampled(Allouez,Interstate and Whaleback).
4. The date appears to be from an observational study with some randomization (Ruffe were collected from different locations around Lake Superior).
5. There are at least five individuals in each cell of the expected table (Table 8).
6. The statistic is the observed frequency table(Table9)
7. X2=0.3813 with 2 degrees of freedom (Table 7).
8. P-value=0.8264(Table 7)
9. DNR Ho because P-value is >A
10. The distribution of Ruffe sexes among the three locations seems to differ.

Table 7-Pearson's Chi-squared test with freq2

X-squared = 0.3813, df = 2, p-value = 0.8264

Table 8-expected frequency table

loc

sex Allouez Interstate Whaleback

f 14.7 14 13.3

m 6.3 6 5.7

Table 9-observed frequency table

sex Allouez Interstate Whaleback

f 15 13 14

m 6 7 5

The mean upper jaw length of female Ruffe is less than 11 mm

1. Cannot continue with 1sample t-test or 2 sample t-test because more than one or two populations of Ruffe was sampled (Allouez,Interstate and Whaleback).

The importance of statistics

Statistics is the science of collecting, organizing and interpreting numerical information. The two major goals of statistics is to summarize large quantities of data in to concise and informative numerical or graphical summaries and the second goal is to make inferences to be made about all individuals and from a few individuals. Statistics is used to conduct research, better understand statistical studies and to be better consumers of statistical information. Statistics is a major part of our everyday life and without the help of statistician’s our modern day life would be in shambles.

Driving in major cities would be a major challenge because without statistics there would be no way to figure out where to place traffic signals properly. Without statistics we would have no proper way to manage our natural resources like game animals, trees and most importantly fish. Without statistics we would not know when to stop the harvest of lake trout on the Chequamegon bay by commercial and sport fishermen.

The tools I learned in this class are related to why I think statistics is important because some day I would like to become a fisheries biologist and to become a fisheries biologist, you must be part statistician. You have to be able to take the data that has been collected all field season and make sense of it. Say if I was a doing one of the first modern age and growth study on the most commercially important fish in Illinois, the Smallmouth Buffalo (*Ictiobus bubalus*). The Smallmouth Buffalo has been on the decline on and off for many years now. Commercial fishing data shows there is no decline but Illinois Department of Natural Resources and Illinois Natural History Survey data shows otherwise. Illinois River Biological Station is in the process of getting this figured out and will need the help of statistics to see if they are seeing less big fish or if young of the year classes are down from the past years.

All in all I feel like I learned a lot from this class and I have a much better understanding of statistics overall and a better understanding of R code in general. I’m not proud I had to take this class twice but I feel like I have a much better grasp on the concepts taught in this class now than I did before.

R code

library(NCStats)

setwd("~/R stuff")

RM<-read.csv("RuffeMorph.csv")

RM<-RM[!RM$miss & !RM$out,]

str(RM)

Summarize(~tl,data=RM,digits=3)

hist(~tl,data=RM,xlab="total length(mm)")

(freq1<-xtabs(~sex, data=RM))

percTable(freq1,digits=1)

barplot(freq1,xlab="sex of the fish: m=male f=female", ylab="frequency")

corr(~tl+w,data=RM)

plot(tl~w,data=RM,ylab="weight (g)",xlab="total length (mm)" )

( freq2 <- xtabs(~sex+loc, data=RM) )

percTable(freq2)

percTable(freq2,margin=1)

percTable(freq2,margin=2)

(bfl<-lm(bg~tl,data=RM))

fitPlot(bfl,ylab="Total length",xlab="Body girth")

rSquared(bfl)

predict(bfl,data.frame(tl=125.85))

str(RM)

Totallength=RM$tl

median(Totallength)

RML<-filterD(RM,loc!="Interstate")

levenesTest(w~loc,data=RML)

t.test(w~loc,data=RML,alt="two.sided",conf.level=0.95,var.equal=TRUE)

( chi <-chisq.test(freq2,correct=FALSE) )

chi$expected

chi$observed