Toronto, Canada

A Report on

Executive Summary of Module 3

Introduction to Data Analytics (ALY6000)

Guided by:

Prof. Mohammad Shafiqul Islam

Submitted by:

Name : Pratikkumar Indravadan Malaviya

NUID : 002963548

Date of submission : 4th February'2022

INTRODUCTION

The given module introduces basic concepts of probability and statistics based on categorical data. Indeed, the probability is a study of randomness, where we need to deal with those statistics which we have never seen before or we may see afterward. Moreover, the dataset elaborates the key information of fishes which is further divided into several species. For example, Black Crappie, Bluegill, Bluntnose Minnow, Iowa Darter, Largemouth Bass, Pumpkinseed, Tadpole Madtom, and Yellow Perch. Furthermore, the provided dataset does include additional information likewise fishID, Species names, the total length of the fish, and fish width. Lastly, on the basis of this key information, there are certain bar graphs are plotted below with the help of R language.

METHODOLOGY

Although, provided inchBio.csv file contains dataset about different fishes where their species name, total length, and fish width become key attributes of fish. As this file is provided additionally, therefore we need to import this csv file manually in R for further calculations and illustrating plots. The R supports the simplest function to read the csv file which is "read.csv(file_path)". By executing predefined functions like 'structure()', 'summary()', and 'data.frame()' on a given csv file will provide some statistics. However, these findings are not enough to plot specified graphs, thus there are determined packages are also mandatory to install prior to illustrating the graphs.

1. **Following an introduction, provide an analysis of descriptive characteristics of the data set provided by your instructor. This includes pertinent statistics including counts, cumulative counts, and frequency, percentages, etc. Include R console screen snippets to support your observations and conclusions. Below is a sample excerpt.**

    - After installing determined packages and loading given inchBio.csv file, now we have 676 records about 8 distinct species of fish. By executing headtail(), we identified top and bottom records of the data.

```
##     netID fishID       species  tl     w  tag scale
## 1      12     16      Bluegill  61   2.9 <NA> FALSE
## 2      12     23      Bluegill  66   4.5 <NA> FALSE
## 3      12     30      Bluegill  70   5.2 <NA> FALSE
## 674   110    863 Black Crappie 307 415.0 1783  TRUE
## 675   129    870 Black Crappie 279 344.0 1789  TRUE
## 676   129    879 Black Crappie 302 397.0 1792  TRUE
```

    - Moving forward to the data where we, simply identified total number of species and each species containing how many records with frequency values .

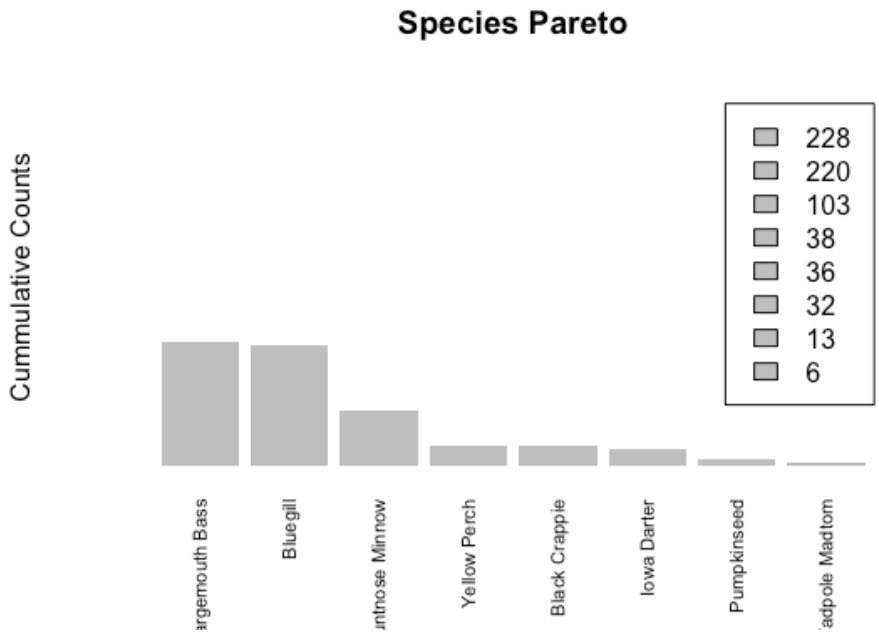
```
##              Var1      Freq
## 1    Black Crappie  5.325444
## 2         Bluegill 32.544379
## 3 Bluntnose Minnow 15.236686
## 4      Iowa Darter  4.733728
## 5  Largemouth Bass 33.727811
## 6      Pumpkinseed  1.923077
## 7   Tadpole Madtom  0.887574
## 8     Yellow Perch  5.621302
```

    - Here, provided data has been separated for the further calculation, where the designated plots are based on the species of the fish and total number of frequencies of each species. In addition we calculated some statistics on the frequency by counts, cumfreq, and cumcounts.

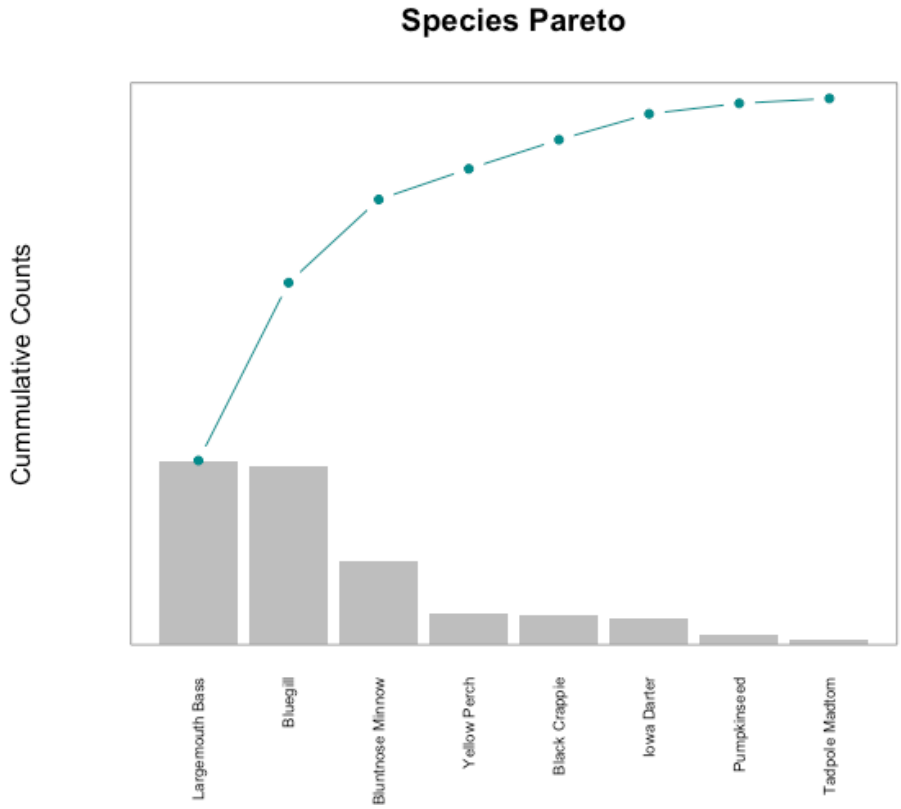
```
##            Species   RelFreq   cumfreq counts cumcounts
## 5  Largemouth Bass 33.727811  33.72781    228       228
## 2         Bluegill 32.544379  66.27219    220       448
```

```
## 3 Bluntnose Minnow 15.236686  81.50888    103       551
## 8     Yellow Perch  5.621302  87.13018     38       589
## 1    Black Crappie  5.325444  92.45562     36       625
## 4      Iowa Darter  4.733728  97.18935     32       657
## 6      Pumpkinseed  1.923077  99.11243     13       670
## 7   Tadpole Madtom  0.887574 100.00000      6       676
```

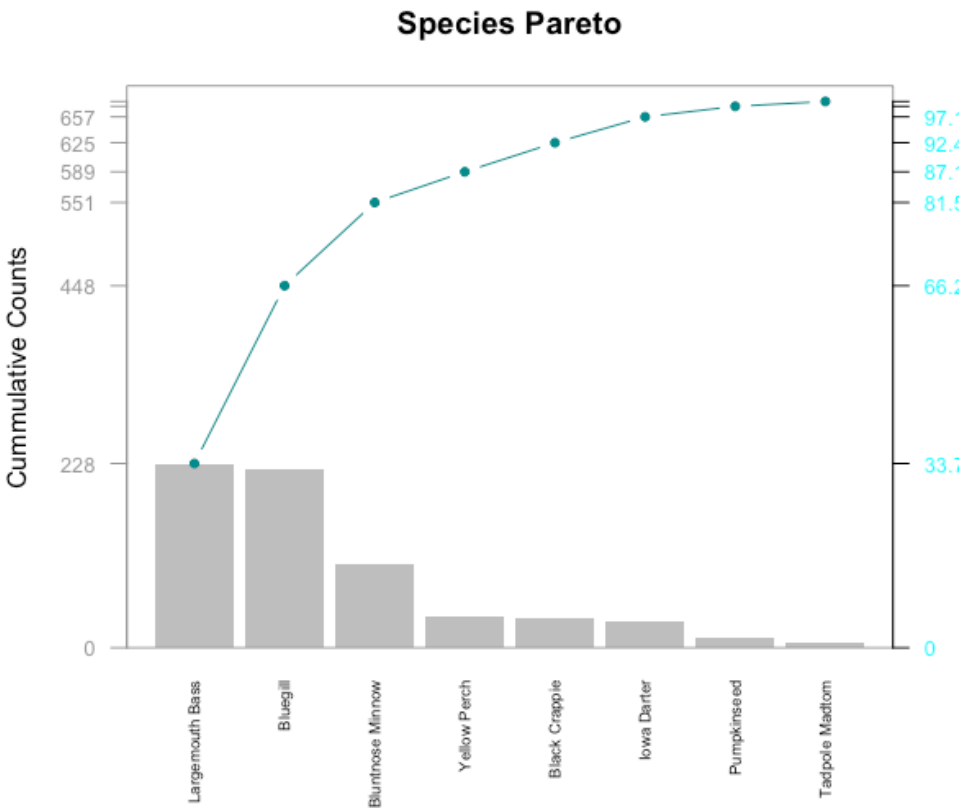
2. **Provide the executive with visualizations (at least 3) in that help them see the key characteristics you want to highlight. They can be boxplots, histograms, frequency and probability distributions, or barplots (bar charts). A pareto plot as illustrated below must be included in this part of your report. Include screen snippets of your plots to support your findings and conclusions. The goal is not only to present your visual results, but also to explain the significance of them.**



- The given barplot defines, that among all species the 'Largemouth Bass' species has highest number of fishes (228). Where, Bluegill species accounted second highest (220) number of fishes.
- However, Tadpole Madtom named species has lowest number of cumulative count (only 6).



- In the given barplot (Species Patero) all species are arranged in descending order, where we are able to figure out their proportion of frequency.
- The line starts with upper most categorical column of ‘Largemouth Bass’, and describes the comparisons between discrete categories of species
- In R, the function creates bar plot bounded by the columns which is depending on the provided parameters.



- In the given bar plot (Species Pareto), one of the axis of plot represents specific categories of fishes which being compared, while others axis represents the measured values.
- The left y-axis indicates the frequency counts of given species, while on the right side the numeric values stands for the parentages.

3. **Finally, provide a clear two to three sentence paragraph summary of the key points that you want the audience to walk away with regarding your analysis. This summary should present accurate analysis and be supported by the data presented in the rest of the report.**
    - The bar plot gives the clear vision on provided dataset which calculates basics of statistics

- Out of total length of species the maximum average length accounted by Largrmoth Bass (298.6) and Tadpole Madtom has minimum average length (36.3) .
- Conversely, the fish width has also major difference, although Tadpole Madtom fish has lowest mean (0.6) , it clearly seem that this fist can fit in human hands also and mostly fishers are using this this species to prey other fishes. Averagely, Black Crapple fish has highest width size (359.5) in compare to others species.
- By considering the width values as per the data set it is likely to be measured in (mm) unit.

CONCLUSION

In this module, we understood the basics of probability, set theory, and statistics. However, the important topic was the difference between mutually exclusive events and independent events, though the probability problems do require several basic prerequisites likewise understanding the possible outcomes of any given problem, on the basis of that we could able to calculate final probability. Moreover, in R we illustrated the Pareto charts based on inchBio.csv file which simply contains the information about the fishes. In addition, we interacted with the new R package so-called 'tidyverse', although the features of that mentioned package is extraordinary. For instance, by loading this package it supports several other packages tidyr, readr, pure so on, so it is great to load one serviceable package instead of a bunch of many more.

BIBLIOGRAPHY

1. Kabacoff, Robert. R In Action: Data Analysis and Graphics with R. Manning, 2015.
2. Tutorialspoint.com. 2022. *R Tutorial*. **https://www.tutorialspoint.com/r/index.htm**
3. R Tutorial: Count With Your Data. (2020, March 21). YouTube. https://www.youtube.com/watch?v=J6ly-bGMTHE.
4. R - Pareto Chart - GeeksforGeeks. (2020, April 30). GeeksforGeeks. https://www.geeksforgeeks.org/r-pareto-chart/.
5. Barplot Function - RDocumentation. (n.d.). barplot function - RDocumentation. https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/barplot.
6. - Robk@statmethods.net, R. K. (n.d.). *Quick-R: Graphical Parameters.* Quick-R: Graphical Parameters. https://www.statmethods.net/advgraphs/parameters.html.
7. - Robk@statmethods.net, R. K. (n.d.). *Quick-R: Axes And Text.* Quick-R: Axes and Text. https://www.statmethods.net/advgraphs/axes.html.

APPENDIX

```
# Malaviye_M3_Project3
# 01 Name & Packages
print('PRATIKKUMAR INDRAVADAN MALAVIYA')
install.packages(c('FSAdata','FSA','dplyr','plotrix','moments','ggplot2',,'magrittr','tidyr
','tidyverse'))
library(FSAdata)
library(magrittr)
library(dplyr)
library(tidyr)
library(plyr)
library(tidyverse)

# 02 inmporting bio.csv
bio = read.csv('/Users/pratik_4511/Desktop/Quarter_1A/M3/inchBio.csv')
print(bio)

# 03 Display the head, tail and structure of <bio>
print(headtail(bio)) #headtail
print(structure(bio)) #structure

# 04 Create an object, <counts>, that counts and lists all the species records,
counts  = table(bio$species)
print(counts)

# 05 #Display just the 8 levels (names) of the species
print(unique(bio$species))

# 06 displays the different species and the number of record
tmp = count(bio$species)
print(tmp)

# 07 <tmp2>, of just the species variable and display the first five records
tmp2 = subset(bio,select = species)
print(head(tmp2,n=5))

# 08 Create a table, <w>, of the species variable. Display the class of w
w = table(bio$species)
print(w)
class(w)

# 09 Convert <w> to a data frame named <t> and display the results
t = data.frame(w)
print(t)

# 10 Extract and display the frequency values from the <t> data frame
t %>% select(Freq)

# 11 Create a table named <cSpec>
cSpec = table(bio$species)
print(cSpec)

#12 <cSpecPct> that displays the species and percentage of records
cSpecPct = prop.table(cSpec)*100
print(cSpecPct)


#13 Convert the table, <cSpecPct>, to a data frame named <u> and confirm that <u> is a data
frame
u = data.frame(cSpecPct)
print(u)

class(u)
#14 barplot of <cSpec>
barplot(cSpec,
        main = 'Fish Count',
        ylab = 'COUNTS',
        col = 'Light Green',
        cex.names = 0.60,
```

```
        las = 2)

#15 barplot of <cSpecPct>
barplot(cSpecPct, ylim = c(0,40), = 'COUNTS', col.lab = 'Light Blue',
       main = 'Fish Relative Frequency')

#16 Rearrange the <u> cSpec Pct data frame in descending order of relative frequency. Save
#the rearranged data frame as the object <d>
d = u[order(-u$Freq),]

print(d)
#17 Rename the <d> columns Var 1 to Species, and Freq to RelFreq
colnames(d) = c('Var1','freq')
colnames(d)[colnames(d) %in% c('Var1','freq')] = c('Species','Relfreq')

print(d)
#18 Add new variables to <d> and call them cumfreq, counts, and cumcounts
counts
t$Freq
tdescending  = t[order(-t$Freq),] # Assign to variable &converting in to descending order
tdescending$Freq
d = d %>%
  mutate(cumfreq = cumsum(d$Relfreq),
         counts = tdescending$Freq,
         cumcounts  = cumsum(tdescending$Freq)
         )

print(d)
#19 Create a parameter variable <def_par> to store parameter variables
def_par = par(no.readonly = TRUE)

#20 barplot <pc>
pc = barplot(d$counts,
             width = 1,
             space = 0.15,
             border = NA,
             axes = F,
             ylim = c(0,3.05*228),
             ylab = "Cummulative Counts",
             names.arg = d$Species,
             las=2,
             cex.names = 0.60,
             main = "Species Pareto",
             d$counts,na.rm=TRUE)

#21 Add a cumulative counts line to the <pc> plot,
pc = barplot(d$counts,
             width = 1,
             space = 0.15,
             border = NA,
             axes = F,
             ylim = c(0,3.05*228),
             ylab = "Cummulative Counts",
             names.arg = d$Species,
             las=2,
             cex.names = 0.60,
             main = "Species Pareto",
             )
lines(pc,
      d$cumcounts,
      type = 'b',
      cex = 0.7,
      pch = 19,
      col = 'cyan4')

#22  Place a grey box around the pareto plot
pc = barplot(d$counts,
```

```
             width = 1,
             space = 0.15,
             border = NA,
             axes = F,
             ylim = c(0,3.05*228),
             ylab = "Cummulative Counts",
             names.arg = d$Species,
             las=2,
             cex.names = 0.60,
             main = "Species Pareto",
             )
lines(pc,
      d$cumcounts,
      type = 'b',
      cex = 0.7,
      pch = 19,
      col = 'cyan4')
box(col= 'grey62')

#23 Add a left side axis
pc = barplot(d$counts,
             width = 1,
             space = 0.15,
             border = NA,
             axes = F,
             ylim = c(0,3.05*228),
             ylab = "Cummulative Counts",
             names.arg = d$Species,
             las=2,
             cex.names = 0.60,
             main = "Species Pareto")
lines(pc,
      d$cumcounts,
      type = 'b',
      cex = 0.7,
      pch = 19,
      col = 'cyan4')
box(col= 'grey62')
axis(side = 2,
     cex.axis = 0.8,
     at = c(0, d$cumcounts),
     las = 1,
     col.axis = 'grey60',
     col = 'grey60')

#24 Add axis details on right side of box
pc = barplot(d$counts,
             width = 1,
             space = 0.15,
             border = NA,
             axes = F,
             ylim = c(0,3.05*228),
             ylab = "Cummulative Counts",
             names.arg = d$Species,
             las=2,
             cex.names = 0.60,
             main = "Species Pareto")
lines(pc,
      d$cumcounts,
      type = 'b',
      cex = 0.7,
      pch = 19,
      col = 'cyan4')
box(col= 'grey62')
axis(side = 2,
     cex.axis = 0.8,
```

```
     at = c(0, d$cumcounts),
     las = 1,
     col.axis = 'grey60',
     col = 'grey60')
axis(side = 4,
     at = c(0,d$cumcounts),
     labels =  c(0,d$cumfreq),
     las = 1,
     col.axis = 'cyan',
     col.lab = 'cyan4',
     cex.axis = 0.8)

#25  finished Species Pareto Plot (without the star watermarks)
pc = barplot(d$counts,
             width = 1,
             space = 0.15,
             border = NA,
             axes = F,
             ylim = c(0,3.05*228),
             ylab = "Cummulative Counts",
             names.arg = d$Species,
             las=2,
             cex.names = 0.60,
             main = "Species Pareto \n Pratik Malaviya")
lines(pc,
      d$cumcounts,
      type = 'b',
      cex = 0.7,
      pch = 19,
      col = 'cyan4')
box(col= 'grey62')
axis(side = 2,
     cex.axis = 0.8,
     at = c(0, d$cumcounts),
     las = 1,
     col.axis = 'grey60',
     col = 'grey60')
axis(side = 4,
     at = c(0,d$cumcounts),
     labels =  c(0,d$cumfreq),
     las = 1,
     col.axis = 'cyan',
     col.lab = 'cyan4',
     cex.axis = 0.8)
```

GITHUB

Username : pratik4511

Repository : https://github.com/pratik4511/malaviya_m3_project-3