Carter Adamson

FISH 558

Lab 12 HW

4/19/2024

*Guidelines:*

* *Include course, lab number, and date at the top of the document; do not write your name.*
* *Number and label the questions and answers clearly! (We should easily be able to find your answers!)*
* *Include all of the requested output (e.g., values, data tables, and plots), not just the code for them. (We will not copy your code into R to see if it works).*
* *Include informative captions for figures and tables. See research articles for examples. [We will take points off if these are not included!].*
* *Submit a Word document unless directed otherwise (no r files or pdfs please).*
* *Include all your code used for the problems.*
* *Answer ALL questions using complete sentences that are clear and informative.*

1. In Lab we calculated a stratified mean of Atlantic Croaker catch-per-unit-effort (CPUE), but we did not explore the data set thoroughly. The data came from a multispecies monitoring trawl survey that is designed to many fish species in Chesapeake Bay. If our goal is to develop the best CPUE time series we can for Atlantic Croaker, we may choose to limit/restrict our dataset to what we think is most representative of the Atlantic Croaker population. For example, Atlantic Croaker may not always be present in large numbers in the Bay, given their seasonal migrations, or they may not regularly inhabit some parts of the bay. Re-evaluate the dataset and develop what might be a better Index for Croaker using the data. Include the following:
   1. **Examine the dataset graphically, and determine whether any data restrictions may be warranted based on Month or Region. Use any graphical approaches you think would be helpful (e.g., coplots, histograms, boxplots, scatterplots, …). Some helpful response values to plot could potentially include CPUE and log(CPUE+1). Include only the most informative plots. (7 pts)**

A graph with numbers and a bar

Description automatically generatedA graph of a number of percent

Description automatically generated

A B

Figure 1: Histograms of catch (in number of individuals) for Atlantic croaker in the Chesapeake Bay. Figure 1A is a histogram of all sets, whereas Figure 1B includes only sets with non-zero catches.

I generated a large group of exploratory plots for this question, and I will be discussing three sets of them to anwser this question. This count histogram is a good way of visualizing the raw data at this intial state to see what we’re dealing with. Figure 1A shows the incredible amount of zeros in this dataset (well over 80%). Even with the zeroes removed, as in Figure 1B, 80% of the catch falls within the smallest bin on the histogram. These graphs tell us that we have zero-inflated data, and that small catches are far more common than large catches.

A graph of different sizes and colors

Description automatically generated with medium confidence

6 7 8

6 7 8

9 10 11

9 10 11

6 7 8

9 10 11

3 4 5

6 7 8

9 10 11

Figure 2: Conditional histogram of the regional distribution of croaker sets in the Chesapeake Bay, broken up into panels by month. Regions are coded by integers from 1 to 5. Months, ranging from March to November, are indicated by the integer in the top right corner of each panel.

The purpose of this conditional histogram was to get an idea of what regions the croaker data are coming from, whether particular regions are favored, and whether the spatial distribution of these sets changes over the months. For most months, we see a minor pattern where Region 4 or 5 has the most data, with Region 1 generally having the least. October is a slight outlier here, with Regions 4 and 5 actually having the least sets. However, the most obvious outlier is August, where all sets are concentrated entirely in Region 5. Both of these months, but August especially, are candidates for removal from the index, since we want to account for factors like movement that may cause variation in the data. Whatever is driving the August results (a migration, for example) should therefore be dealt with before calculating an index.

A graph of different sizes and shapes

Description automatically generated with medium confidence

Figure 3: Boxplots of the effect of five Chesapeake Bay regions on catch-per-unit-effort (CPUE) on the log scale [log(CPUE+1)]. Each panel shows this relationship during a different month.

At first glance, this figure may look a lot like Figure 2, but there are a lot of key differences. Rather than just being histograms, these boxplots have log(CPUE+1) as a response variable. I chose to log transform this data. Since we’ve observed that this dataset is “scrunched up” against zero, the log transformation is useful to compress the outliers and get a better look at these relationships. The boxes show the general distribution of log(CPUE+1) for the different region+month combinations shown. Once again, August stands out as having a distinctly different distribution than the other months, with all catch concentrated in Region 5. The boxplot for Region 5 in this month also indicates higher median CPUE than any other month/region combination. Boxes whose interquartile range and whiskers are “flattened” on zero and don’t have any outliers above indicate that most if not all of the catch data for that region in that month were zero. April has three regions that fit this description, for example, marking it as a possible month to exclude as well.

* 1. **[Extra Credit; 2 pts] – Do some graphical exploration of the data using fraction of tows that are not empty (i.e., tows that had at least 1 Croaker). Describe your findings. [Note, you will need to figure out how to calculate the proportion of non-zero tows, for different factors. For example, how does the proportion of non-zero tows change by Month or Region?]**

**A graph of different sizes and shapes

Description automatically generated with medium confidence**

Figure 4: Multi-panel bar graph showing the effect of region (1-5) and month (panels, March-November) on the probability of non-zero catch.

Something which immediately jumped out at me regarding Figure 4 is its ability to clearly display which region/month combinations never had any non-zero catch. In the analysis of Figure 3, for example, I identified April as a month where three regions had very unpromising boxplots that seemed to indicate a high concentration of zero or near-zero CPUE. This graph corroborates that – though CPUE is not plotted, it shows that during April, regions 1-3 have no probability of nonzero catch. This recurs in August with regions 1-4 and in October with region 1. All three of those months are therefore prime candidates for being removed or subsetted out of the index. March and to a lesser extent November also have conspicuously low catch probilities when compared with other months, indicating that there is some variability there that should be accounted for before calculating the index.

* 1. **Based on your graphical exploration, state your observations of the patterns in the data. Do you think you would be justified in restricting the data set to specific months and/or regions? (You may also support your decision using the scientific literature). (4 pts)**

In my responses to the previous questions, I identified several months as prime candidates of months that ought to be left out of an abundance index. August is the most obvious choice since the happenstance of all catch occurring in region 5 is a strong indicator of an unaccounted-for pattern such as a migration. April is another candidate since all catch occurred in regions 4 and 5 only. Similarly, October did not have any catch in region 1, and March had lower catch than many other months as indicated by low probability of non-zero catch on Figure 4 and “flat boxes” on Figure 3. These are all patterns of variability that should be pulled out of the data to ensure that the index of abundance reflects only the variability associated with the population’s abundance and not any other effects. Therefore, I think it is perfectly reasonably to restrict the data by excluding some or all of those months.

I was curious about whether the life cycle of the Atlantic croaker could explain some of this unexplained variation. I found a report from a Chesapeake Bay Atlantic croaker fishery management plan from 1991, which stated that adult croaker are present in high abundance from May through August (and present in lower numbers in March, April, and October). With that biological context in mind, I think it is reasonable to exclude certain months if they fall outside of that peak and if I have already identified another reason to doubt that month’s inclusion. March, April, October, and November all fit that bill. Even though August falls within that peak, the data for that month are so strikingly distinct from other months that I feel justified excluding those data as well.

Link to the report: [**https://dnr.maryland.gov/fisheries/Documents/fmp/ChesapeakeBay\_AtlanticCroaker\_Spot\_FMP\_12-1991.pdf**](https://dnr.maryland.gov/fisheries/Documents/fmp/ChesapeakeBay_AtlanticCroaker_Spot_FMP_12-1991.pdf)

* 1. **Based on your decisions regarding the Month variable, subset your data set accordingly by removing the unwanted Months. Recalculate the annual time series of stratified mean CPUE for Croaker. Make a plot of this new annual index (with SE). (4 pts)**
  2. **Describe how this new time series differs from the one we calculated in lab. Which is better, and why? (3 pts)**

*[I am answering these two questions together because the plot below helps answer both]*

A graph showing different colors and numbers

Description automatically generated

Figure 5: Time series plot for three indices of Atlantic croaker abundance in the Chesapeake Bay from 2002 to 2014. The solid blue line represents the stratified mean index calculated with a subset of the data as part of this homework assignment. The dotted lines represent the indices calculated as part of Lab 12: the arithmetic mean (pink) and a stratified mean calculated with the entire dataset (green).

The stratified mean index that I calculated for this assignment used a subset of the data by month. Following on from the reasoning described in previous questions, the months excluded were March, April, August, October, and November. The result of this is a much different time series. As seen on Figure 5, the general shape of this population trend is relatively consistent. For the most part, we see peaks and troughs in the same years on each. What has changed is the value of the index (i.e. the new stratified mean indicates generally higher numbers of fish) and the magnitude of variation through the years. Take 2007 and 2008 for example: in 2007, all three trendlines are at a peak but the new index is nearly twice the value of the other two. Then, the three trendlines hit very similar values in 2008. The result is a much more dramatic-looking decline according to the newer index. As for whether this is better, it’s hard to say. Certainly, the steps we’ve taken have excluded some temporal variation that was not being taken into account by the previous indices – we can have some amount of confidence, then, that the new index is more accurately reflective of the genuine abundance trend. However, the graph below should be cause for concern:

A graph with different colored lines

Description automatically generated

Figure 6: Time series plot for three indices of Atlantic croaker abundance in the Chesapeake Bay from 2002 to 2014. Mostly identical to Figure 5, this plot adds error bars as dotted blue vertical lines for just the stratified mean index derived from subsetted data.

Figure 6 shows clearly that the annual stratified means calculated from the subsetted data still all have confidence intervals that include zero and go negative. While this did also happen with the stratified mean calculated from the entire dataset, it still makes me hesitant to describe this index as a major improvement. If we can’t be completely confident that the index of abundance isn’t zero, then what use is it?

* 1. **Explain what a random stratified mean is and how it is different from an arithmetic mean. Include some sort of simple conceptual example (that you make up) to help your explanation. Under what conditions would a stratified mean be much better than an arithmetic mean? (5 pts)**

The arithmetic mean is the “standard” or “common” mean that we are used to, calculated by summing the data and dividing by n (sample size/number of data points summed). The random stratified is slightly more complex, separating data into weighted strata. The arithmetic mean of the data points within a stratum can then be adjusted by the weight and summed to obtain a stratified mean.

A quick example can illustrate this point. Say you are trying to come up with an index of salamander abundance between five interconnected ponds. The ponds can be thought of as strata within a larger wetland, and each is a different size.

Table 1: Table of hypothetical salamander survey data to illustrate arithmetic vs stratified means. This data comes from some hypothetical wetland that is subdivided into five pond strata. Data include salamander catch (number of individuals), area (m2), weights applied to each stratum based on area, and the calculated catch \* weight to be used when finding the stratified mean.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Pond** | **Salamander catch** | **Area** | **Weight** | **Catch \* Weight** |
| 1 | 10 | 10 | 0.1 | 1 |
| 2 | 14 | 10 | 0.1 | 1.4 |
| 3 | 4 | 10 | 0.1 | 0.4 |
| 4 | 20 | 20 | 0.2 | 4 |
| 5 | 100 | 40 | 0.4 | 40 |

Arithmetic means are vulnerable to outliers and may not account for spatial or environmental variability. For example, a survey that usually finds no fish, but hits one particular location with a very high density of fish can have an unrealistically high arithmetic mean. A stratified mean lets the stratum with that high-density location be weighed against other strata, giving it less capability to “pull” the mean towards it. Therefore, the stratified mean can be useful when dealing with spatial variability. In this case, the arithmetic mean may have been biased low by placing the same emphasis on the low catch in pond 3 as the high catch in pond 5, despite the fact that a much larger area was sampled in pond 5. The weights involved in calculating the stratified mean take this into account.

1. Answer these questions (1 pt):
2. **How many hours did you spend on this assignment as a whole?**

6.5 hours

1. **Did you work with anyone else or at least consult with someone? Who? Include an image.**

Solo

1. **Were there any particular things you struggled with in this lab and how did you overcome them?**

I struggled most with the exploratory plotting phase. Trying to get a look at all the data and find the right sort of visualization to identify patterns was challenging because it was easy to feel like I was barking up the wrong tree. I decided to just make a *ton* of visualizations and then whittled it down to just the informative ones afterward, once I had a better sense for the situation.

1. **[558 Students Only] Work on your projects. Provide a short (3-5 sentence) summary of the progress you’ve made and any issues you may be having. (4 pts)**

I have focused on taking the mark-recapture data in the .csv file I downloaded from EDI and convert it into a capture history. With the code and background knowledge from WLDF 578, I am confident (hopefully not overconfident) that once I have established accurate capture histories, I should be able to analyze the data in a fairly straightforward way with Rmark. However, I’ve found the actual process of making capture histories from more than 7,000 observations a tad daunting. I gave a shot at making some code myself to generate encounter histories, to very little success. I also tried to use a technique that Dan Barton showed us in 578, but the capture history making was sort of a “sidebar” and not the complete focus of any single lab, so it’s also a little challenging to go through. I’ve made some progress, but not as much as I’d hoped. But between the resources from that class and expertise from a former colleague, I’m not lost for ideas on how to proceed.

I also have the unique problem of there being two tagging systems represented in this data. Partway through the study, they switched from VIE to PIT tags, which just complicates matters more. I should be able to collapse these into a single “tag number” field, however, since every individual will be identifiable with one of the two techniques, and any double-tagged individuals will be identifiable with either.

**Appendix: CODE**

library(here)

library(ggplot2)

library(tidyverse)

library(lattice)

# QUESTION 1 ####

##a - graphical exploration-----####

#set up data w cpue

croaker <- read.csv(here("Lab 12 - CPUE and Standardization","Croaker catch final.csv"))

croaker$CPUE=croaker$Count/((croaker$TowDist/1000)\*(croaker$NetWidth/(0.3048\*1000)))

croaker$logCPUE1<-log(croaker$CPUE+1)

#explore month and region effects -- less useful plots commented out

histogram(~ Count,data=croaker, main="Histogram of Catch")

histogram(~ Count,data=subset(croaker, CPUE>0), main="Histogram of Non-Zero Catch")

#histogram(~ logCPUE1|as.factor(Month),data=croaker)

#histogram(~ logCPUE1|as.factor(Region),data=croaker)

#plot(CPUE~Month, data=croaker)

#plot(CPUE~Region, data=croaker)

#max(croaker$Month) #11

#min(croaker$Month) #3

#histogram(~Month, data=croaker, breaks=8)

#histogram(~Region, data=croaker)

#histogram(~Month|Region, data=croaker)

histogram(~Region|Month, data=croaker)

#coplot(CPUE~Month|Region, croaker)

croaker %>%

ggplot(aes(x=Month, y=logCPUE1, group=Month)) +

geom\_boxplot() +facet\_wrap(~Region)

croaker %>%

ggplot(aes(x=Region, y=logCPUE1, group=Region)) +

geom\_boxplot() +facet\_wrap(~Month)

#croaker %>% ggplot(aes(x=Month, y=CPUE))+geom\_bar(stat="identity")+facet\_wrap(~Region)

#croaker %>% ggplot(aes(x=Region, y=CPUE))+geom\_bar(stat="identity")+facet\_wrap(~Month)

##b - proportions by month and region-----####

croakprop <- croaker %>%

group\_by(Month, Region) %>%

summarise(ntot=n(), non\_zero= sum(Count != 0)) %>%

mutate(catchprop=non\_zero/ntot)

croakprop %>% ggplot(aes(x=Month, y=catchprop))+geom\_bar(stat="identity")+facet\_wrap(~Region)

croakprop %>% ggplot(aes(x=Region, y=catchprop))+geom\_bar(stat="identity")+facet\_wrap(~Month)

plot(catchprop~Month, data=croakprop)

plot(catchprop~Region, data=croakprop)

croakprop %>%

ggplot(aes(x=Month, y=catchprop, group=Month)) +

geom\_boxplot() +facet\_wrap(~Region)

croakprop %>%

ggplot(aes(x=Region, y=catchprop, group=Region)) +

geom\_boxplot() +facet\_wrap(~Month) #because there is only one data pt for each month/region, the boxplots are the same as bar graphs

##d - subset Month and plot new index-----####

##nominal andstrat mean from lab12-----

croak.mean.nominal = croaker %>% group\_by(Year) %>%

summarize(mean = mean(CPUE),

var = var(CPUE),

sd = sqrt(var),

cv = sqrt(var)/mean,

n = length(CPUE),

se = sd/sqrt(n))

croak.mean = croaker %>% group\_by(Year, StratNum) %>%

summarize(strat.mean = mean(CPUE),

strat.var = var(CPUE),

strat.n = length(CPUE))

W=read.csv(here("Lab 12 - CPUE and Standardization", "CM weights\_new.csv"))

croak.mean = left\_join(croak.mean, W, by="StratNum")

croak.mean <- croak.mean %>%

mutate(mean.calc = strat.mean \* Weight,

var.calc = Weight^2\*(TotalN-strat.n)/(TotalN-1)\*strat.var)

croak.index = croak.mean %>% group\_by(Year) %>%

summarise(index = sum(mean.calc),

var = sum(var.calc),

se = sqrt(var),

cv = sqrt(var)/index,

LCI = index-1.96\*se,

UCI = index+1.96\*se)

##strat mean with fewer months ----

#include months 5,6,7,9

croak.subset <- filter(croaker,Month>4)

croak.subset <- filter(croak.subset, Month<10)

croak.subset <- filter(croak.subset, Month!=8)

croak.mean.strat <- croak.subset %>% group\_by(Year, StratNum) %>%

summarize(strat.mean = mean(CPUE),

strat.var = var(CPUE),

strat.n = length(CPUE),

.groups="keep")

croak.mean.strat = left\_join(croak.mean.strat, W, by="StratNum")

croak.mean.strat <- croak.mean.strat %>%

mutate(mean.calc = strat.mean \* Weight,

var.calc = Weight^2\*(TotalN-strat.n)/(TotalN-1)\*strat.var)

croak.index.2 = croak.mean.strat %>% group\_by(Year) %>%

summarise(index = sum(mean.calc),

var = sum(var.calc),

se = sqrt(var),

cv = sqrt(var)/index,

LCI = index-1.96\*se,

UCI = index+1.96\*se)

croak.index.2

croak.indecies <- data.frame(croak.mean.nominal, croak.index, croak.index.2)

#make a ggplot with all three indecies

plot.mean <- ggplot(data=croak.indecies, aes(x=Year, y=mean, colour="Arithmetic Mean"))+

geom\_line(linetype="dashed")+ theme\_bw()+labs(y="Croaker index")+

geom\_line(data=croak.indecies, aes(x=Year, y=index, colour = "Stratified Mean (all months)"), linetype="dashed")+

geom\_line(data=croak.indecies, aes(x=Year, y=index.1, colour="Stratified Mean (subset)"), linetype="solid") +

geom\_errorbar(aes(ymin=LCI.1, ymax=UCI.1, color="Stratified Mean (subset)"), width=0.2, linetype="dotted") #error bars if wanted

plot.mean