FISH 6004. Assignment 1

Due October 5, 2021

Marks: 25% of course total

Please copy R scripts and outputs into a file (e.g. MS Word). Organize and label your results clearly. Use captions for any tables or figures you create. You will be marked for clarity of presentation as well as correct results. Email your completed assignment to me, noel.cadigan@mi.mun.ca

1. Compute catch-at-age for 3Ps cod using the catch-at-length and age-at-length (ALK) information provided. Note that the ALK data is in the age_length_data.csv and the catch-at-length is in catch_length.csv. Provide barplots of catch-at-length and catch-at-age. Plot the distribution of age-at-length. (10 marks)
2. Compute a catch-at-age time series for the 3Ps cod 1959 to 2016 from the simulated catch-at-length time series. The catch-at-length time series is in catch_length_3Ps.csv. Use the ALK computed from the previous problem. Provide a proportion-at-age plot. Plot the average length and age of a fish in the catch over time. Generate a 'spay' plot for catch-at-age. (20 marks)
3. Consider the stratified random survey depicted in the attached EXCEL spreadsheet (survey design.xls). Strata are numbered in the top right-hand corner, and the total number of sampling units (Nh) are indicated in the top left-hand corner of each stratum. The sampling units selected in the survey are colored red. The data associated with the red sampling units are provided in the file 'survey.dat'. Estimate the population average catch. Provide an estimate of the variance of this estimate. Please show all workings (*20 marks*).
4. For the survey data in 3), identify which stratum was sampled the most (i.e. highest sampling fraction) and which stratum was sampled the least. What would happen to the variance of your estimate of the population average catch if you added 3 more samples to the least-sampled strata, and took away 3 samples from the most-sampled strata (i.e. total sample size for all strata remains the same). Hints. The population mean and variance for stratum *h* are defined as $\bar{R}_h = N_h^{-1} \sum_{i=1}^{N_h} R_{hi}$ and $S_h^2 = \sum_{i=1}^{N_h} (R_{hi} - \bar{R}_h)^2/(N_h - 1)$. To answer this question, treat the sample variances from the data collected as the true values for S_h^2, and estimate what happens to $Var(\hat{\bar{R}})$ when the survey design is changed (*15 marks*).
5. It is well-known that when the cost of a sample is the same for all samples, then the optimal allocation of survey effort among strata for a stratified design is

$$n_h = n\left(\frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}\right)$$

   Estimate the optimal sampling design based on the survey data in question 4, and round sample sizes. Keep the total sample size the same. Estimate $Var(\hat{\bar{R}})$ based on the optimal survey design you construct. How much smaller is this variance compared to the variance you estimated in question 3 (*15 marks*).

6. Consider the following stratified population of size N=32 members and responses given by the following R code

```
set.seed(1234)
Nh = c(15,10,7)
N=sum(Nh)
stratum = rep(1:3,times=Nh)
y.pop = rpois(N,stratum*5)

pop = data.frame(y=y.pop,stratum=stratum)
pop$istrat = as.numeric(as.factor(pop$stratum))
```

```
> pop
     y stratum istrat
1    2       1      1
2    6       1      1
3    5       1      1
4    6       1      1
5    7       1      1
6    6       1      1
7    1       1      1
8    3       1      1
9    6       1      1
10   5       1      1
11   6       1      1
12   5       1      1
13   4       1      1
14   8       1      1
15   4       1      1
16  13       2      2
17   8       2      2
18   7       2      2
19   6       2      2
20  12       2      2
21  14       2      2
22 100       2      2
23   7       2      2
24   7       2      2
25  10       2      2
26  13       3      3
27  13       3      3
28  16       3      3
29  12       3      3
30  13       3      3
31  15       3      3
32  15       3      3
```

i. For the stratified random sampling design with `n_h = c(5,3,2)`, how many possible sample values of $\hat{\bar{R}} = N^{-1}\sum_{h=1}^{H} N_h \bar{r}_h$ could be obtained?

ii. Write some R code to compute all of these possible values.

iii. Verify that $\hat{\bar{R}}$ is an unbiased estimator of $\bar{R} = N^{-1}\sum_{h=1}^{H}\sum_{i=1}^{N_h} R_{hi}$ for the above population data.

iv. Verify that $\mathrm{Var}\left(\hat{\bar{R}}\right) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 \frac{1-f_h}{n_h}\frac{\sum_{i=1}^{N_h}(R_{hi}-\bar{R}_h)^2}{N_h-1}$ gives the correct result (i.e. compare with the variance of $\hat{\bar{R}}$ from all possible samples).

v. Plot a histogram of all possible values of $\hat{\bar{R}}$, and plot a normal qqplot of these values. Does the distribution of $\hat{\bar{R}}$ seem like a Normal distribution?

vi. Replace the largest value of y with y=100. What effect does this have on $\mathrm{E}\left(\hat{\bar{R}}\right)$ and $\mathrm{Var}\left(\hat{\bar{R}}\right)$. Give the percent change from the values you obtained in iii) and iv).

(*20 marks*) *hint: split(pop, list(pop$stratum))*. Use *lapply()* and *combinations()* to get all possible samples in each stratum, and use *expand.grid()* to get all possible sets of strata means.