QSCI 381 AB - AUT 22
Homework #8- Vinsensius

The PDF file of the code is attached at the end of the HW

Number 1

In this question we will be using the Cuckoo dataset

```
cuckoo <- read.csv(file="cuckoo_eggs.csv")
```

Cuckoos are known to lay their eggs in the nests of other (host) birds. The eggs are then adopted and hatched by the host birds. The data give the lengths of cuckoo eggs (mm) found in nests of various other bird species.

(1a) Create a box-whisker plot of Cuckoo egg length (y axis) as a function of species (x axis), labelling the axes appropriately and giving the plot a meaningful title (4 pts)

(1b) Fill in the table below by calculating the corresponding sample sizes, means and standard deviations of egg length for each host bird species. Round your answers to 2 decimal places (7 pts)

(1c) You wish to analyse whether mean egg lengths differ among host bird species using analysis of variance. State the null and alternate hypothesis associated with this ANOVA test (2 pts)

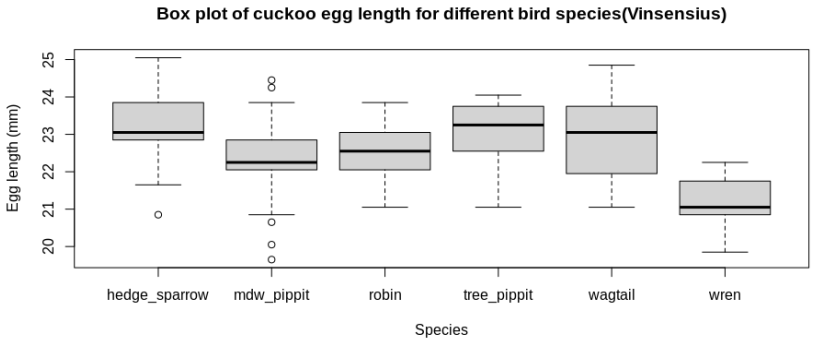
(1d) By filling in the ANOVA table below, test the null hypothesis stated in (3c) at a significance level of $\alpha = 0.05$. Based on the ANOVA test, make a decision regarding whether to reject, or fail to reject the null hypothesis. Round all numeric values to 3 decimal places (8 pts)

(1e) Based on the mean values calculated in (a), which species do you think is driving the significance of this test result (2 pts)

(1f) Load the DescTools package and run a post-hoc Scheffe test on this ANOVA, and report on which species were significantly different from one another at a significance level of 0.05. In your answer, refer to the difference in mean values and p-values associated with that difference to back up your answer (6 pts)

(a) The code is

```
cuckoo_dat <- read.csv(file="cuckoo_eggs.csv")
boxplot(cuckoo_dat$Length ~ cuckoo_dat$Bird, main="Box plot
↪ of cuckoo egg length for different bird
↪ species(Vinsensius)", xlab="Species", ylab="Egg length
↪ (mm)")
```



(b) The code is

```
N_vec =tapply(cuckoo_dat$Length,cuckoo_dat$Bird,length)
sd_vec =  tapply(cuckoo_dat$Length,cuckoo_dat$Bird,sd)
mean_vec = tapply(cuckoo_dat$Length,cuckoo_dat$Bird,mean)
sd_total = sd(cuckoo_dat$Length)
N_tot = NROW(cuckoo_dat)
mean_total = mean(cuckoo_dat$Length)
```

Species	Sample size(n)	Sample mean ($\bar{x}$)	Sample SD (s)
Hedge Sparrow	14	23.12	1.07
Meadow Pipit	45	22.30	0.92
Robin	16	22.58	0.68
Tree Pipit	15	23.09	0.90
Wagtail	15	22.90	1.07
Wren	15	21.13	0.74
All	120	22.46	1.07

The total mean should be the mean of the whole group, similar to the sd

(c) The null hypothesis (H_0) is that the mean of all egg length of all bird species are equal to each other, $\mu_1 = \ldots = \mu_6$. The alternate hypothesis (H_A) is that at least one of the group mean is not equal to other groups.

(d) The code is

```
summary(aov(Length~Bird,cuckoo_dat))
Fcrit = qf(0.95,5,114)
```

Variation	Df	Sum Squares	Mean Squares	F	F crit	p-val
Between species	5	42.940	8.588	10.388	2.294	3.152e-08
within species	114	94.248	0.827			
Total	119	137.188	9.415	10.388	2.294	3.152e-08

(e) Hedge sparrow and Wren have maximum and minimum mean respectively. So, i expect either or both species to affect the result.

(f) The code is

```
cuckoo_anova = aov(Length~Bird,cuckoo_dat)
ScheffeTest(cuckoo_anova)
```

Species difference	difference value	p-val
wren - hedge_sparrow	-1.99	1.1e-05
wren- mdw_pippit	-1.17	0.0037
wren-robin	-1.45	0.0026
wren-tree_pippit	-1.96	1.0e-05
wren-wagtail	-1.77	9.8e-05

Based on the combination of species difference that are statistically significant(p-val<0.05), Wren occurs in all of the combination difference, so Wren is statistically different from other species.

Number 2

On your way to a trailhead for a hike in bear country, a well-meaning logger tells you that 50% of fatal black bear attacks are perpetrated by mother bears defending their cubs, and all other attacks are more or less random (equally likely to be lone females or males). Being a naturally skeptical, curious, and resourceful person, you dig up some data on fatal black bear attacks in North America since 1900, and tally the results as follows:

Type of Black Bear	Number of Documented Attacks
Male	55
Female	3
Female with cubs	5

Create a vector of the bear attack data by entering the data manually: attacks <- c(55, 3, 5)

(2a) What type of test will you need to conduct to evaluate the logger's claim, and what are the null and alternative hypotheses? (3 pts)

(2b) Calculate and present the expected frequencies from this data, based on the expectation that 50% of attacks will be from Females with Cubs, and 25% each for lone males, lone females. (4 pts)

(2c) Given the information calculated in (b) does this data meet the requirements for running the test you stated in (a)? (2 pts)

(2d) Calculate and present the test statistic for running the test on this data (4 pts)

(2e) Test the null hypothesis associated with this test at a significance level $\alpha = 0.01$, stating the result of the test and relevant test values (4 pts)

(2f) Given the result you found in (e), provide a written interpretation of whether the initial claim was supported or denied by the data (3 pts)

(a) The test that will be done is chi-squared test for goodness of fit because we have the expected probability for the attacks.

The null hypothesis is that the observed frequencies are the same as expected frequencies, i.e. the attacks by female bears with cubs is 50% of the total attacks while

the attacks by either female or male bears are 25% of the total attacks. The claim is represented by the null hypothesis.

The alternate hypothesis is that at the least one of observed frequencies of attack types is not the same as the expected frequencies.

(b) The code is

```
attacks <- c(55, 3, 5)
type_of_bear = c("Male","Female","Female-cubs")
bear_attack_df = data.frame(type_of_bear,attacks)
bear_attack_df$expected_freq = c(0.25, 0.25,0.5)*sum(attacks)
```

Type of Black Bear	Number of Documented Attacks	expected freq
Male	55	15.75
Female	3	15.75
Female with cubs	5	31.50

(c) Given that all the expected frequencies>5, we can apply chi-squared test for goodness-of-fit.

(d) The code is

```
chisq.test(x=bear_attack_df$attacks,p=c(0.25, 0.25,0.5))
```

The test statistic $\chi^2 = 130$.

(e) The code is

```
chisq.test(x=bear_attack_df$attacks,p=c(0.25, 0.25,0.5))
```

The p-val calculated is 2.2e-16. Since p-val $< \alpha$, we reject the null hypothesis that 50% of bear attacks are by female bear with cubs.

(f) We reject the null hypothesis means that the claim about 50% of the bear attacks are by female bear with cubs is rejected. This means that the claim was denied by the data.

Number 3

The North Pacific undergoes decadal scale variability in ocean temperatures (known as the Pacific Decadal Oscillation). In 2014, the Northeast Pacific underwent a shift from predominantly cooler conditions that had been present since the early 2000's to a very warm phase that prevailed from 2014 to 2020. Seabirds are particularly sensitive to changes in ocean temperatures, as changes in ocean temperatures alter food-web structure, ultimately changing the quality and/or quality of seabird prey (fish, krill). We will look at a dataset of seabird mortality records from southern Washington to identify whether the taxonomic composition of seabird mortality is independent of temperature regime by comparing records from 2007-2009 (cold regime) to records from 2014-2016 (warm regime).

Load the seabird mortality dataset available on canvas

```
birds <- read.csv(file="seabird_mortality.csv")
```

Each row is a record of a dead seabird found on a beach, and the column bird.id corresponds to the taxonomic group that bird belonged to.

(3a) Create a contingency table of the data using `table()` with columns corresponding to bird taxonomic ID and rows corresponding to temperature regime (cold, warm). Enter/show your contingency table below (4 pts)

(3b) Upon first glance, what is your initial impression of the differences between warm and cold periods (2 pts)

(3c) Calculate the row and column sums of your contingency table, and then use this to calculate a table of expected frequencies if taxonomic composition was independent of temperature regime. Include the expected frequency table below (6 pts)

(3d) Which type of test would you use to test whether composition is independent

of temperature regime, and what are the requirements for that test? Does the data meet those requirements? (4 pts)

(3e) Run the appropriate test to test whether composition is independent of temperature regime at a significance level $\alpha = 0.01$. In your answer include: the test statistic, critical value of the test statistic, and the p-value of that test. (5 pts)

(3f) Based on your answer in (e), do you think that taxonomic composition is independent of temperature regime for this data? (2 pts)

(a) The code is

```
birds <- read.csv(file="seabird_mortality.csv")
birds_table = table(birds$temp.regime,birds$bird.id)
```

The contingency table is

	Auklet	Cormorant	Fulmar	Gull	Murre	Puffin	Shearwater
cold	38	126	923	572	813	109	88
warm	911	68	500	444	1418	256	147

(b) The Cormorant, Fulmar, and Gull have higher mortality during the cold weather, while the other four species have higher mortality during warm weather. However, there are more birds in warm temperature regime so we do not know whether the percentages is the same between the two temperature regime.

(c) The code for

```
birds_table = table(birds$temp.regime,birds$bird.id)
birds_df_mortality =
↪  as.data.frame.matrix(table(birds$temp.regime,birds$bird.id))
birds_df_mortality = cbind(birds_df_mortality, total =
↪  rowSums(birds_df_mortality))
birds_df_mortality = rbind(birds_df_mortality, total =
↪  colSums(birds_df_mortality))

birds_chisq = chisq.test(birds_table)
expected_birds_table = addmargins(birds_chisq$expected)
```

The contingency table with sum added is

	Auklet	Cormorant	Fulmar	Gull	Murre	Puffin	Shearwater	total
cold	38	126	923	572	813	109	88	2699
warm	911	68	500	444	1418	256	147	3744
total	949	194	1423	1016	2231	365	235	6413

The expected table is

	Auklet	Cormorant	Fulmar	Gull	Murre	Puffin	Shearwater	total
cold	395	81	592	423	929	152	98	2669
warm	554	113	831	593	1302	213.09	137.2	3744
total	949.00	194.00	1423.00	1016.00	2231.00	365	235	6413

(d) The test type is chi-squared test of independence. The requirement is that the expected frequencies ≥ 5, which the data has met.

(e) The code for

```
birds_chisq = chisq.test(birds_table)
chisq_crit3 = qchisq(0.99,6)
```

The test statistic $\chi^2 = 1050$. The p-val is 2e-16. The critical value is $\chi^2_{\text{crit}} = 16.8$.

(f) Since the p-val$< \alpha$, we reject the null hypothesis that taxonomic composition is independent of temperature regime and this means that there might be some correlation between the two variables.