MTHS4005 ASP 2023-24

ASSIGNMENT:

Coursework Comprehensive Report

Lecturer: Peter Neal

Submitted By: Sudarshan Khandelwal [20533986]

INTRODUCTION

The problem statement is an ornithologist asks in assistance for analysing the data related to kittiwakes they have collected and have different fields and dataset. The dataset provided have information about kittiwakes.

The datasets are as follows:

- Observation_data: The number of sightings of kittiwakes at an observation point over a 4-week (28 day) period. Each day observations are taken at dawn, noon, mid-afternoon and dusk.
- Historical_data: The number of breeding pairs at 6 sites in 6 different years.
- Measurement_data: The weight (g), wingspan (cm) and culmen (beak length) (mm) is collected for 12 black-legged and 14 red-legged kittiwakes.
- Location_data: The number of breeding pairs in 24 colonies is recorded along with potentially important covariate information on mean summer temperature, cliff height (logarithm), saneel concentration and coastal direction.

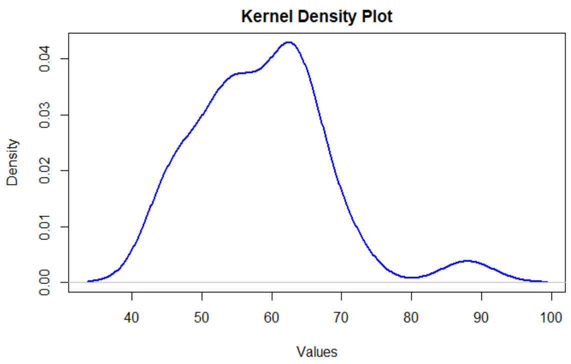
The Questions:

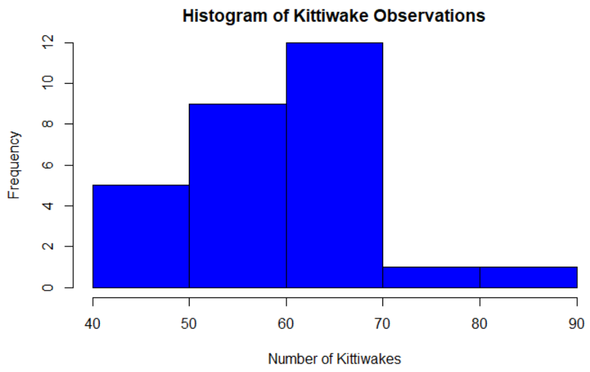
1. **Provide an exploratory analysis of the Observation data. Construct an 80% confidence interval for the mean number of kittiwakes observed at dusk.**

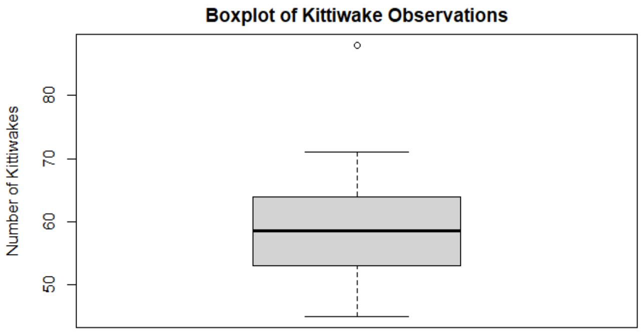
Observation Data Analysis

Objective: To explore the number of kittiwakes observed at dusk and construct a confidence interval.

Exploratory Analysis: Included examining the completeness and accuracy of the data, calculating basic statistics like mean and standard deviation, and plotting the data for visual inspection. Here is a kernel Density Plot, Histogram, Box Plot to understand the data.







The three graphs are statistical visualizations of data regarding observations of kittiwakes, a type of seabird.

1. Kernel Density Plot: This graph shows the probability density of the data, which gives us a smooth estimate of the distribution. The x-axis represents the values, which could be any measurement associated with kittiwakes, such as weight or wingspan, and the y-axis represents the density. The curve peaks around

60, indicating that most of the data points are concentrated around this value. The shape of the curve is unimodal and skewed to the right, suggesting that a majority of observations are clustered around the median with fewer high-value observations.

2. Histogram of Kittiwake Observations: This is a histogram that represents the frequency distribution of the number of kittiwakes observed. The x-axis shows the number of kittiwakes, and the y-axis shows the frequency of each bin. The tallest bar is around the 60-70 range, indicating that this is the most common observation range for the number of kittiwakes. The distribution is right skewed, which is consistent with the kernel density plot.

3. Boxplot of Kittiwake Observations: The boxplot provides a summary of one aspect of the data distribution. The box represents the interquartile range (IQR), which contains the middle 50% of the data. The line inside the box indicates the median. The "whiskers" extend to the smallest and largest values within 1.5 times the IQR from the lower and upper quartile, respectively. Points outside of this are considered outliers, as indicated by the dot above the upper whisker. The boxplot indicates that the median is around 60, consistent with the other two graphs, and the data has a few high-value outliers.

The consistency across all three plots suggests a right-skewed distribution of kittiwake observations, with a concentration of data points around the lower to mid-range values and fewer observations in the higher range. The presence of outliers in the boxplot also indicates that while most observations fall within a certain range, there are occasional observations that are significantly higher than the norm.

Confidence Interval Construction: An 80% confidence interval for the mean number of kittiwakes observed at dusk was calculated using the One sample T-test. The Obtain values of the test are

t = 32.936, df = 27, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

80 percent confidence interval: [56.23521, 60.90765]

sample estimates:

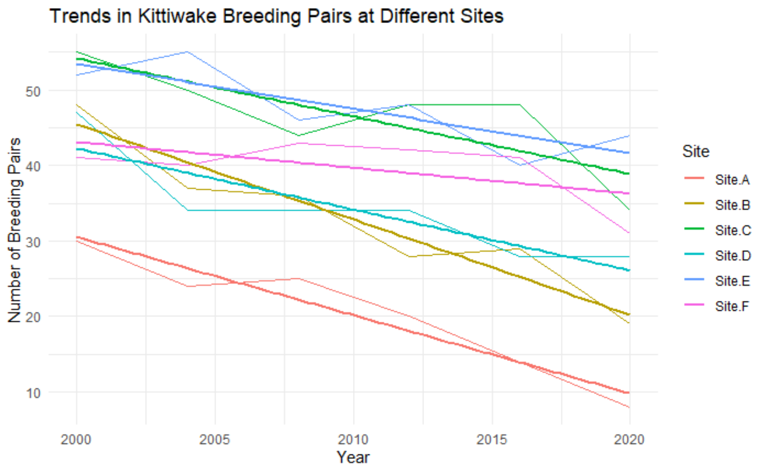
mean of x = 58.57143

2. Does the Historical data support the ornithologist's hypothesis that the decline in kittiwake numbers, over time, is independent of site? The ornithologist would like an estimate for the number of breeding pairs at site C in 2007.

Historical Data Analysis

Objective: To investigate the hypothesis that the decline in kittiwake numbers over time is independent of the site.

Trend Analysis: Each site was analysed for trends in the number of breeding pairs. This included calculating slopes and correlation strengths. The slopes (rates of decline) and significance levels are similar across sites, it would support the hypothesis that the decline is independent of the site. Conversely, notable differences in slopes across sites would suggest that the decline is site dependent. The graphical plot will also give a visual representation of these trends below.



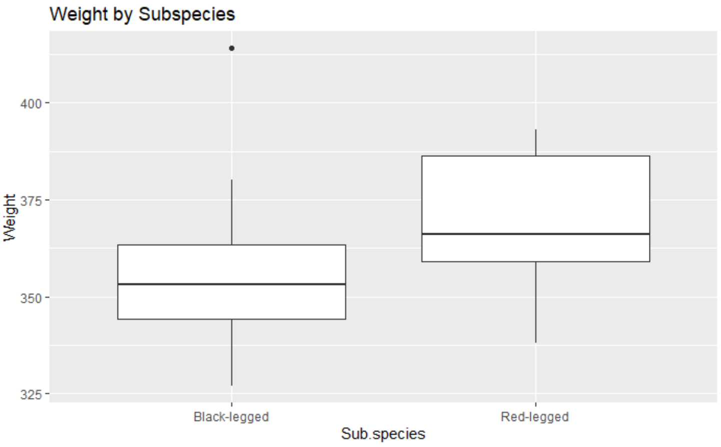
Findings: The analysis revealed variability in decline rates and strengths of trends across different sites, suggesting that the decline in kittiwake numbers is dependent on the site. For estimating the number of breeding pairs at **Site C in 2007** is calculated using interpolation it is a technique in which it considers the past and future data to create points of required time. The estimation is equal to **45.5** Breeding pairs

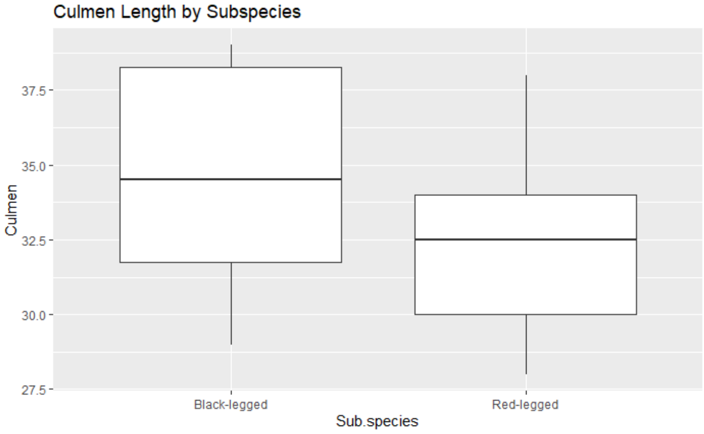
3. **For the Measurement data the ornithologist asks for:**
   a) A visual summary of the data.
   b) For each sub-species, is wingspan and culmen length independent?
   c) Is there evidence that the weights of birds of the two sub-species are different?
   d) From the data provided is their evidence that there is a difference between the two sub-species?

   Objective: To analyze measurements (wing span, culmen length, weight) of kittiwake sub-species.

   Visual Summary: Boxplots were used to visually summarize the data distributions for each measurement by sub-species.

A) The three boxplots compare different physical measurements—weight, culmen length, and wingspan—between two subspecies of birds, specifically labelled as "Black-legged" and "Red-legged".



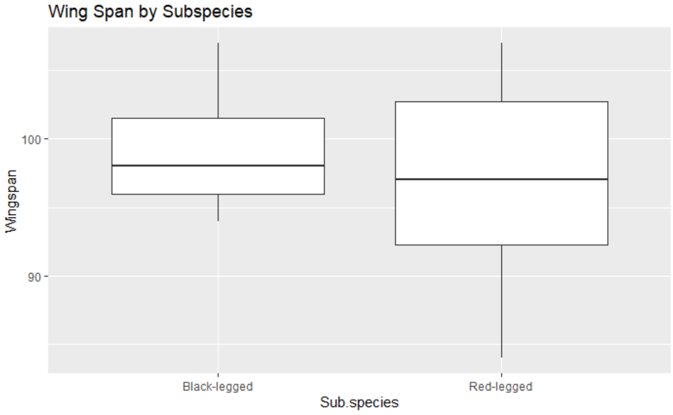


1. Weight by Subspecies: This boxplot compares the weights between the Black-legged and Red-legged subspecies. The median weight of the Black-legged subspecies is around 360, and the interquartile range (IQR) is from 350 to 370, showing a compact distribution of weights. There is one outlier, indicating an individual that is significantly heavier than the rest. The Red-legged

subspecies has a higher median weight, closer to 375, and a wider IQR, suggesting more variation in weight within this group.

2. <u>Culmen Length by Subspecies</u>: The second boxplot compares the culmen length (which is the measurement from the base of a bird's beak to the tip). The Black-legged subspecies has a higher median culmen length, around 35, with a smaller IQR compared to the Red-legged subspecies. This indicates that the Black-legged subspecies tends to have longer beaks, with less variability in beak length compared to the Red-legged subspecies.

3. <u>Wingspan by Subspecies</u>: The third boxplot shows the wingspan measurements. Both subspecies have a wide range of wingspan measurements, with the Black-legged showing a slightly higher median wingspan around 95, compared to the Red-legged subspecies which has a median closer to 90. Both distributions show a fair amount of variability, as indicated by the length of the IQRs and the presence of outliers, especially in the Red-legged subspecies.



In summary, these boxplots suggest that the Black-legged subspecies tend to have a lower median weight but longer culmen lengths and a slightly larger wingspan than the Red-legged subspecies. The Red-legged subspecies shows greater variability in weight and wingspan. Outliers in the data indicate that there are some individuals that significantly deviate from the typical measurements found within their subspecies.

B) for finding the each sub-species, is wingspan and culmen length independent is calculated using The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by r. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how

C) far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

Pearson correlation coefficient	P- Value	Correlation Value
Black-legged Wingspan & Clumen Length	0.8846	-0.04702769
Red-legged Wingspan & Clumen Length	0.2609	0.3223989

From here the correlation value seems to be black-legged have a negative correlation and red-legged have a positive correlation with wingspan and clum length and the p value of black-legged is also higher. So, from here we see the black-legged have an independent relationship between wingspan and clum length and red-legged have a dependent relationship.

<u>Statistical Analysis</u>:

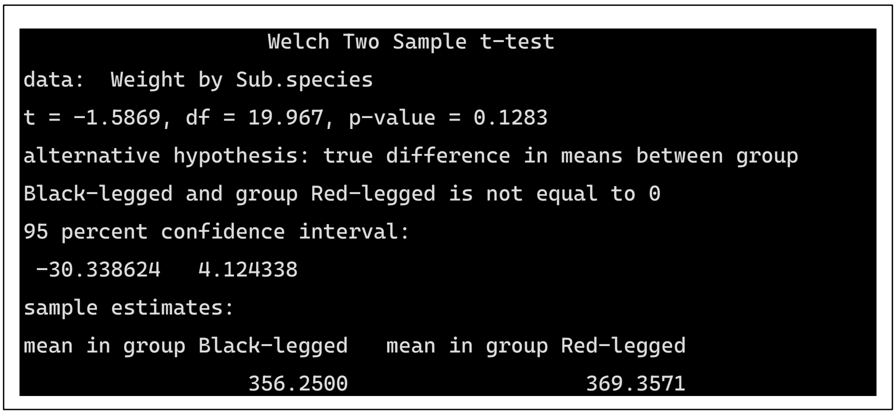
Pearson correlation tests were conducted for each sub-species to check the independence of wing span and culmen length.

A t-test was performed to compare the weights of birds of the two sub-species.

A MANOVA was conducted to assess overall differences between the two sub-species.

Findings: The MANOVA results suggested potential differences in measurements between sub-species, but the evidence was not conclusive at a 0.05 significance level.

D) We are using welch two sample t-test to find out their evidence that the weights of birds of the two sub-species are different from the t test we can see the



```
                     Welch Two Sample t-test
data:  Weight by Sub.species
t = -1.5869, df = 19.967, p-value = 0.1283
alternative hypothesis: true difference in means between group
Black-legged and group Red-legged is not equal to 0
95 percent confidence interval:
 -30.338624   4.124338
sample estimates:
mean in group Black-legged   mean in group Red-legged
                  356.2500                   369.3571
```

In summary, based on the p-value and the confidence interval, there is not enough evidence to reject the null hypothesis. The data does not provide strong support for the idea that there is a significant difference in means between the Black-legged and Red-legged groups.

E) The result is calculated performing a Multivariate Analysis of Variance (MANOVA). MANOVA is a statistical technique used to simultaneously test the equality of means across multiple dependent variables for different groups.
- The overall MANOVA model, considering all dependent variables, is marginally significant with a p-value of 0.07049. The significance level of 0.05 is commonly used to determine statistical significance, and in this case, the p-value is just above this threshold.

- It suggests that there might be some differences in the means of the dependent variables among the different levels of the "Sub.species" variable. However, the evidence for this is not strong enough to be considered statistically significant at the conventional 0.05 level.

4. For the Location data the ornithologist asks you to:
a) Fit a linear model to predict the number of breeding pairs.
b) Fit a linear model to the logarithm of the number of breeding pair.
c) Choose the most appropriate linear model for the data.
d) Comment on the model fit and effect of the selected covariates on the number of breeding pairs.
e) Choosing an appropriate model, provide a 90% confidence interval for the number of breeding pairs at a site with coastal direction = South, sandeel concentration = 1.27, mean summer temperature = 25.3 and cliff height (log) = 3.15

a. b. c. d.) After fitting the linear model for breeding pair(Model1) and logarithm of the number of breeding pair(Model 2).
And when comparing the two models, there are several factors to consider:

1. Residual Standard Error (RSE): This value indicates the average amount by which the actual values deviate from the predicted values. Lower RSE indicates a better fit. Model 2 has a lower RSE compared to Model 1.
2. Multiple R-squared: This represents the proportion of variance explained by the model. The closer sthe value is to 1, the better the model explains the data. Model 2 has a higher R-squared value.
3. Adjusted R-squared: This is similar to R-squared but adjusts for the number of predictors in the model. It is a more accurate measure of the goodness of fit for models with a different number of predictors. Model 2 has a higher adjusted R-squared.
4. F-statistic: A higher F-statistic indicates a more statistically significant fit. Model 2 has a significantly higher F-statistic.

5. p-value: The p-value tests the null hypothesis that all coefficients are equal to zero (no effect). A smaller p-value rejects the null hypothesis. Both models have small p-values, but Model 2's is smaller, indicating a more significant fit.

Model Selection: Model 2 (with log transformation) is a better fit for the data compared to Model 1. This is evident from the higher R-squared and adjusted R-squared values and a significantly lower residual standard error in Model 2.
Effect of Covariates: In both models, sandeel concentration and cliff height are significant predictors of the number of breeding pairs, indicating their importance in the habitat and breeding success of kittiwakes. Coastal direction and summer temperature do not show a significant effect in either model.
Model 2 as a Preferred Model: The log transformation in Model 2 stabilized variance and made the model more appropriate for the data, as suggested by the improved fit metrics.

e) The prediction with its confidence interval suggests that for the specific conditions provided (coastal direction, sandeel concentration, mean summer temperature, and cliff height), you can expect about 128 breeding pairs at the site, with a 90% confidence range from about 121 to 136 pairs.

Conclusion

The analyses conducted provide a comprehensive understanding of various aspects of kittiwake populations, ranging from breeding pair numbers, observational trends, historical patterns, and physical measurements. The findings indicate site-specific trends in population decline, differences in physical measurements among sub-species, and key factors influencing breeding pair numbers.

1. Observation Data Analysis:

   - An 80% confidence interval for the mean number of kittiwakes observed at dusk was calculated, providing a range within which the true mean is expected to lie with 80% confidence.

2. Historical Data Analysis:

   - Trend analysis across different sites showed that the decline in the number of breeding pairs over time was not uniform, indicating that the decline is dependent on the site. This suggested that environmental or geographic factors specific to each site might be influencing the kittiwake populations.

3. Measurement Data Analysis (MANOVA):

   - The MANOVA analysis suggested potential differences in the physical measurements (wing span, culmen length, weight) between the sub-species of kittiwakes, although the evidence was not strong enough to be conclusive at the conventional 0.05 significance level.

4. Location Data Analysis (Linear Regression Models):

   - Model 1: Showed that sandeel concentration and cliff height were significant predictors of the number of breeding pairs. However, coastal direction and summer temperature were not significant predictors.

   - Model 2 (Log-Transformed): This model, which log-transformed the number of breeding pairs, provided a better fit (higher R-squared and lower residual standard error) than Model 1. It reaffirmed the significance of sandeel concentration and cliff height in predicting the number of breeding pairs.

Other Factors:

- Environmental Factors: The analysis highlighted the importance of environmental factors like sandeel concentration and cliff height in influencing the number of breeding pairs of kittiwakes.
- Site-Specific Trends: The historical data analysis indicated site-specific trends in the decline of breeding pairs, underscoring the need for site-specific conservation strategies.
- Physical Differences Between Sub-Species: While the MANOVA results were not conclusive, they pointed towards possible differences in physical measurements between sub-species of kittiwakes.