# Applied Statistics 161.111 S3 2022

# Assignment 2

**Due date:** Friday 3rd February 2023

**Total marks: 46** **Assessment value:** 17%

## Background

A picture containing text, clipart

Description automatically generatedThis assignment uses the same dataset as Assignment 1. The population of interest is the Pygoscelis penguins nesting on the islands within the Palmer Archipelago. The sample data came from a study in which researchers randomly selected three typical islands in the Palmer Archipelago. They took measurements on all penguins found nesting on these three islands during three breeding seasons.

## The data in the Excel file *penguins\_2022\_S3.xlsx* includes information on species (Adélie, Gentoo and Chinstrap), sex (female or male, determined from a blood test), Bill length and Bill depth (both measured in mm), which season the data were collected in and which island the penguin was found on.

## Import this data into RStudio and call it *penguins*.

## Use the data to answer the following questions in the spaces provided. You can re-size the answer spaces.

## Use RStudio and incorporate the output into your answers.

## When you are finished your assignment, save it as a pdf and upload to Stream Assignment 2 Dropbox.

**Part A: Comparing means: Analysis of bill length according to sex [25 marks]**

## Many species of penguins show sexual size dimorphism. This is the condition where the two sexes of the same species have differences in characteristics such as size, weight and markings. We are interested in whether Pygoscelis penguins nesting on the islands within the Palmer Archipelago show sexual size dimorphism.

A1: Use RStudio to draw a side-by-side boxplot of bill lengths versus sex for the penguins in the sample. [3 marks]

|  |
| --- |
|  |

A2: Use RStudio to calculate numerical summaries for bill lengths for each sex. Fill in the table with the values rounded appropriately. [4 marks]

|  |  |  |
| --- | --- | --- |
| Sex | female | male |
| Minimum | 32.1 | 34.1 |
| Lower Quartile | 37.62 | 40.98 |
| Median | 42.75 | 46.75 |
| Mean | 42.06 | 45.76 |
| Upper Quartile | 46.18 | 50.23 |
| Maximum | 58.0 | 59.6 |
| Sample size | 170 | 172 |

A3: What do the plots and numerical summaries tell you about lengths of bills of male and female penguins **in the sample**? [4 marks]

|  |
| --- |
| The sample data provides information on the lengths of bills of male and female penguins. The summary statistics and the box plots show the following:   * The minimum length of bills for female penguins is 32.1 and for male penguins is 34.1. * The lower quartile for female penguins is 37.62 and for male penguins is 40.98, indicating that 25% of the bills of female penguins are shorter than 37.62 cm and 25% of the bills of male penguins are shorter than 40.98 cm. * The median length of bills for female penguins is 42.75 and for male penguins is 46.75, indicating that 50% of the bills are less than these values. * The mean length of bills for female penguins is 42.06 and for male penguins is 45.76. * The upper quartile for female penguins is 46.18 and for male penguins is 50.23, indicating that 75% of the bills of female penguins are shorter than 46.18 cm and 75% of the bills of male penguins are shorter than 50.23 cm. * The maximum length of bills for female penguins is 58 cm and for male penguins is 59.6 cm.   Overall, the data suggests that the length of bills for male penguins is greater than that for female penguins in the sample, with a larger spread (range) in bill length for male penguins as well. |

A4: Do a two-sample t-test to determine if there is any evidence that on average there is a difference in bill length for male and female penguins **in the population**.

* 1. Step 1: Write the hypotheses. [2 marks]

|  |
| --- |
| The null hypothesis is: H0: There is no significant difference in the mean bill length between male and female penguins in the population.  The alternative hypothesis is: Ha: There is a significant difference in the mean bill length between male and female penguins in the population. |

* 1. Use RStudio to do the two-sample t-test. [1 mark]

|  |
| --- |
| # Load data  bill\_lengths <- data.frame(sex = c(rep("female", 170), rep("male", 172)),  length = c(penguins$`Bill length`[penguins$Sex == "female"],  penguins$`Bill length`[penguins$Sex == "male"]))  # Perform two-sample t-test  t.test(length ~ sex, data = bill\_lengths) |

1. Step 2: State the value of the test statistic. [1 mark]

|  |
| --- |
| The value of the test statistic is -6.6698. |

1. Step 3: State the statistical decision with reason. [1 mark]

|  |
| --- |
| The statistical decision based on the results of the Welch Two Sample t-test is to reject the null hypothesis. This is because the p-value of 1.05e-10 is less than the commonly used significance level of 0.05, indicating strong evidence against the null hypothesis and in favor of the alternative hypothesis. The negative t-value of -6.6698 also supports this decision, as it suggests a significant difference in mean bill length between male and female penguins |

1. Step 4: Write your conclusion. [2 marks]

|  |
| --- |
| Based on the results of the Welch Two Sample t-test, we can conclude that there is evidence to suggest that there is a significant difference in mean bill length between male and female penguins in the population. The t-statistic was -6.6698, and the p-value was 1.05e-10, which is less than 0.05, the commonly used significance level. This indicates that we can reject the null hypothesis that there is no difference in mean bill length between the two groups.  The 95% confidence interval for the difference in means between the two groups ranges from -4.800105 to -2.613671, which does not contain 0, further supporting the conclusion that the means are significantly different.  In conclusion, we can say that male penguins tend to have longer bill lengths than female penguins in the population, on average. |

1. Step 5: Check the conditions are met. [1 mark]

The issue of lack of representativeness was discussed in Assignment 1. Discuss whether the normality condition is met.

|  |
| --- |
| Based on the results of the Shapiro-Wilk normality test for both "female" and "male" bill lengths, the p-value is less than 0.05, which indicates that the data does not come from a normally distributed population. As a result, the normality condition is not met and further investigation may be necessary before conducting the two-sample t-test. |

1. Write a sentence to interpret the confidence interval. Explain how it adds to your conclusion. [4 marks]

|  |
| --- |
| The lower bound of the interval (-4.800105) is less than 0, indicating that we are 95% confident that the mean bill length for female penguins is lower than the mean bill length for male penguins. |

A5: Explain why a two-sample t-test is better than a t-test of differences for this context. [1 mark]

|  |
| --- |
| In the case of this data, we want to compare the average bill length between female and male penguins, which requires a comparison of means between two separate groups, making a two-sample t-test more appropriate. Additionally, the two-sample t-test also accounts for potential unequal variances between the two groups, which can affect the validity of the test results in cases where the variances are significantly different. |

A6: In the following breeding season, a conservation worker found two penguins nesting on Biscoe Island. She took blood samples to identify sex and found one was male and one was female. She had read that bill length can help with identifying sex of penguins. She measured their bill lengths. The female had a bill length that was 2mm longer than the male. Does this contradict analysis in A4? Explain. [1 mark]

|  |
| --- |
| The finding of one female penguin with a bill length 2mm longer than a male penguin does not contradict the analysis in A4. This is because the analysis in A4 was based on the mean bill lengths of groups of male and female penguins and the t-test was used to determine if there was a significant difference in the mean bill lengths between the two groups. The finding of one female penguin with a longer bill length than one male penguin is just a single observation and cannot be used to contradict the results of the t-test, which was based on a larger sample size of male and female penguins. It's important to remember that while individual observations may not match the group's mean, the overall trend and differences between the groups can still be significant. |

**Part B: Exploratory Data Analysis of species and season. [9 marks]**

We are interested in investigating if the prevalence of the species depends on which season the data was collected in.

B1: Use RStudio to produce a **table of counts** for species in the sample for each season. Put species as the rows and seasons as the columns. [2 marks]

|  |
| --- |
| # A tibble: 3 × 4  Species season1 season2 season3  <chr> <dbl> <dbl> <dbl>  1 Adelie 49 50 52  2 Chinstrap 26 18 24  3 Gentoo 34 46 43 |

B2: Use RStudio to produce a table of penguin species as a **proportion** of the penguins in each season**.** Note: the columns of your table should sum to 1 [2 marks]

season1 season2 season3

Adelie 0.3245033 0.3311258 0.3443709

Chinstrap 0.3823529 0.2647059 0.3529412

Gentoo 0.2764228 0.3739837 0.3495935

B3: Use RStudio to produce a **stacked bar plot** of the distributionof penguin species as a proportion of penguins each season. [2 marks]

|  |
| --- |
|  |

B4: What do the tables and plot tell you about the distribution of penguin species in the three seasons? [3 marks]

Species A was the most dominant in all three seasons, followed by species B and C being the least dominant in terms of propotion.

**Part C: Inferential Analysis of species and season. [12 marks]**

C1: Do a Chi-squared test of species and season.

a. Step 1: Write the hypotheses. [2 marks]

|  |
| --- |
| H0: There is no association between species and season.  Ha: There is an association between species and season. |

b. Use RStudio to do the Chi-Squared test. Include the RStudio output here. [1 mark]

|  |
| --- |
| I used to methods for the Chi ;  1.Using counts: Pearson's Chi-squared test data: counts X-squared = 3.1543, df = 4, p-value = 0.5323   1. Using a contingency table: Pearson's Chi-squared test data: contingency\_table X-squared = 215.43, df = 163, p-value = 0.003719   The most accurate one was the counts data and results. |

c. Step 2: State the value of the test statistic. [1 mark]

|  |
| --- |
| The value of the test statistic for Pearson's Chi-squared test with data: counts is 3.1543 |

1. Step 3: State the statistical decision with reason. [1 mark]

|  |
| --- |
| The statistical decision is to accept the null hypothesis since the p-value (0.5323) is greater than the significance level (0.05). This means that there is no statistically significant difference between the observed and expected frequencies, and the null hypothesis cannot be rejected. |

1. Step 4: Write your conclusion. [1 mark]

|  |
| --- |
| Based on the results of the Pearson's Chi-squared test, there is no statistically significant difference between the observed and expected frequencies, and the null hypothesis cannot be rejected. |

f. Step 5: Check the conditions are met. [2 marks]

The issue of lack of representativeness was discussed in Assignment 1. What is the other condition? Discuss whether this condition is met.

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | The other condition for Pearson's Chi-squared test is that the expected frequency in each cell should be at least 5. This condition is met in this case since all of the expected frequencies are greater than 5. | | |  | | --- | |  | | |

g. Use RStudio to calculate the residuals. [1 mark]

|  |
| --- |
| season1 season2 season3  Adelie 0.12602491 -0.04698410 -0.07462713  Chinstrap 0.92956756 -0.98019606 0.06972957  Gentoo -0.83080115 0.78086881 0.03083969 |

1. Do the residuals add to your conclusion? Explain. [1 mark]

|  |
| --- |
| Yes, the residuals add to the conclusion that there is no statistically significant difference between the observed and expected frequencies. This is because the residuals indicate that the observed frequencies are close to the expected frequencies, and therefore the null hypothesis cannot be rejected. |

C2 Do the results and conclusions from part C contradict what you found part B? Explain. [2 marks]

|  |
| --- |
| No, the results and conclusions from Part C do not contradict what was found in Part B. In Part B, we found that Adelie penguins had the highest prevalence in season 1, Chinstrap penguins had the highest prevalence in season 2, and Gentoo penguins had the highest prevalence in season 3. This is consistent with the results from Part C, where we found that Adelie penguins had the highest mean prevalence, followed by Gentoo and then Chinstrap. |

+ + + + + + + +