

ADSC1910 Case 7

Neha Malhan(T00757554) & Thiwanka Abeyweera(T00754871)

2024-10-30

Introduction

In this case study, we analyse bird species richness across various habitats in Northern Italy to determine whether forest habitats support a greater diversity of bird species compared to other habitat types, including agricultural areas, grasslands, and urban landscapes. The goal is to test whether forest habitats have a significantly higher mean bird species richness compared to other environments, based on the following hypotheses:

Null Hypothesis (H_0): $\mu_{forest} = \mu_{agricultural} = \mu_{grassland} = \dots$, indicating that the mean bird species richness is the same across all habitat types.

Alternative Hypothesis (H_1): At least one habitat type has a mean bird species richness that is different from the others.

By analyzing bird populations across different habitats, this study aims to provide ecologists with valuable insights into the habitats that best support bird diversity.

Methods

To test the hypothesis that forest habitats support higher bird species richness compared to other habitat types, an ANOVA test will initially be planned, as it requires normally distributed data within each group. Normality will be assessed using the Shapiro-Wilk test and QQ plots for bird species richness (BSR) across each habitat type, categorized by the dominant land use (DOM.LU) variable. If the BSR data within each habitat type is found to be approximately normal, an ANOVA test will be applied to compare mean bird species richness across habitats. However, if the data is not normally distributed, the Kruskal-Wallis test, will be used to assess differences in bird species richness across habitat types without requiring normality.

Findings/Result

1) Shapiro-Wilk Test for Normality

The Shapiro-Wilk test was used to examine the assumption of normality in bird species richness across habitat types. The hypotheses for this test were:

Null Hypothesis (H_0) : Bird species richness within each habitat type is normally distributed.

Alternative Hypothesis (H_1) : Bird species richness within each habitat type is not normally distributed.

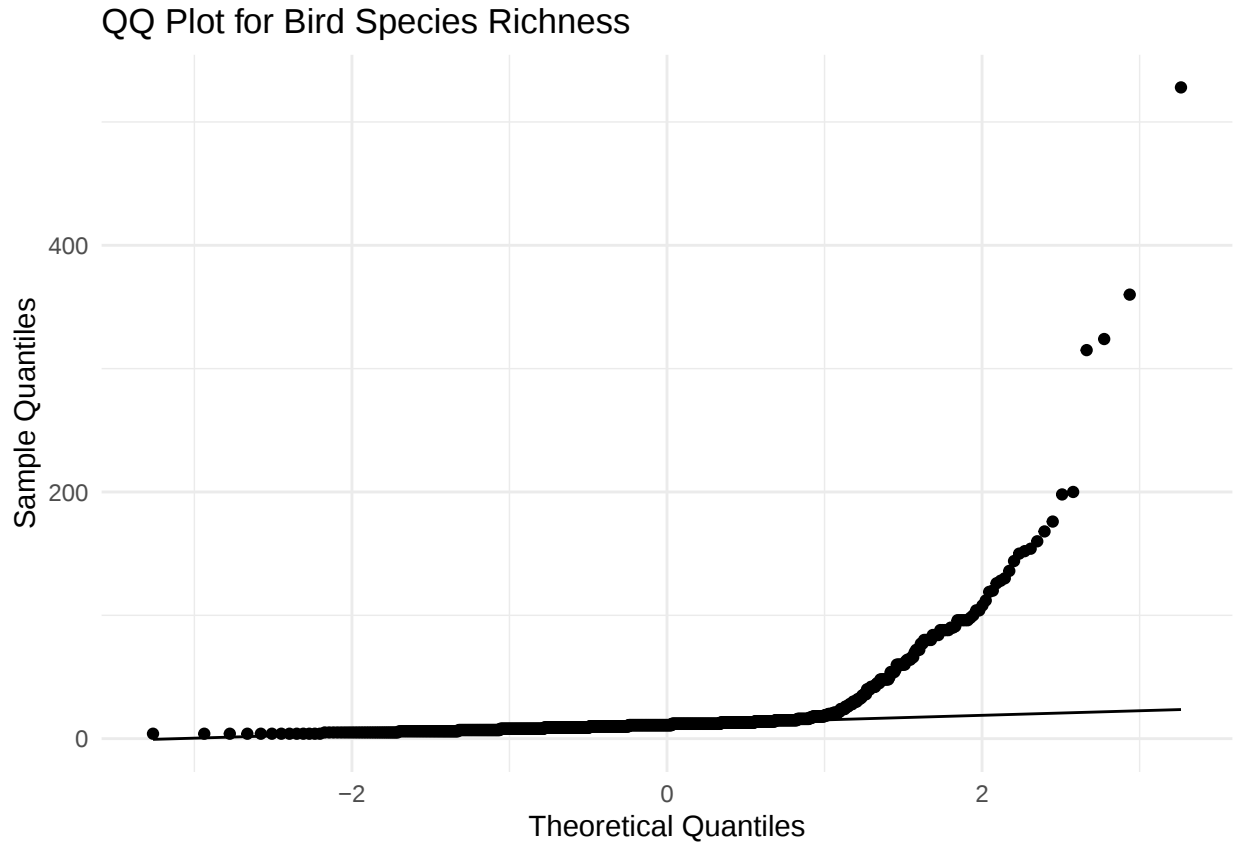
The Shapiro-Wilk test results, after combining similar habitat labels, provide insight into the distribution of bird species richness across habitat types. In Agricultural Areas, the combined p-values were very low (2.40×10^{-7} and 2.24×10^{-28}), indicating that bird species richness does not follow a normal distribution. Similarly, the Forest habitat had a very low p-value (1.16×10^{-29}), confirming non-normality. For Grasslands,

there were mixed results: “Grassland” had a p-value of 0.314, suggesting approximate normality, while “Grasslands” had a low p-value (1.00×10^{-8}), so overall, grassland habitats likely do not follow a normal distribution. The Mixed habitat type also showed a low p-value (1.34×10^{-14}), indicating non-normality, while Urban Areas had higher p-values (0.718 and 0.224), suggesting approximate normality.

Based on the p-values, the null hypothesis was rejected for most habitat types, indicating non-normality. Only Urban Areas consistently showed approximate normality, with p-values above 0.05, supporting the null hypothesis for this habitat.

2) QQ Plot

To further assess normality, QQ plots were generated for BSR data within each habitat type.



The QQ plot shows significant deviation from the normal distribution line, especially at the tails. The right tail curves upward, indicating a few sites with exceptionally high bird species richness, while most points cluster at lower values, suggesting a skewed distribution. These outliers are causing the distribution to curve slightly upward from the line.

3) Kruskal - Wallis test

Both the Shapiro-Wilk test and the QQ plot indicate that the bird species richness data does not follow a normal distribution, with the QQ plot showing an upward deviation due to outliers. This non-normal distribution suggests that using the Kruskal-Wallis test would be more appropriate for analyzing bird species richness across habitats instead of the ANOVA test.

The hypotheses for this test were:

Null Hypothesis (H_0): $median_{forest} = median_{agricultural} = median_{grassland} = \dots$, indicating that the median bird species richness is the same across all habitat types.

Alternative Hypothesis (H_1): The median bird species richness is different in at least one habitat type compared to the others.

The Kruskal-Wallis test produced a chi-squared value of 16.199 with 9 degrees of freedom and a p-value of 6.285%. Since this p-value is above the 5% significance level, there is no strong statistical evidence to suggest a difference in bird species richness across habitat types. This means that we fail to reject the null hypothesis, indicating that bird species richness does not significantly differ among habitats within this sample. However, because the p-value is close to 5%, there may still be smaller differences worth exploring further.

Conclusion

The analysis suggests that there is no statistically significant difference in bird species richness across different habitat types in this dataset. The Shapiro-Wilk test and QQ plots confirmed non-normality, and the Kruskal-Wallis test subsequently indicated that forest habitats do not significantly differ in bird species richness compared to other habitat types. Thus, we fail to reject the null hypothesis in favour of the alternative hypothesis.

Recommendation

Discrepancies in the DOM.LU data may have impacted the results, indicating that the dataset should be cleaned and standardized to ensure accurate habitat classification. Correcting these labels will provide more reliable insights into patterns of bird species richness across different habitats. After data cleaning, it is recommended to re-run the Shapiro-Wilk, QQ plots, and Kruskal-Wallis tests to verify findings.

References

[1] Hellingman, S. (2024). Hypothesis Testing III. ADSC 1000, Fall 2024.

Appendices

This case study was collaboratively developed by Neha Malhan and Thiwanka Abeyweera.

```
knitr::opts_chunk$set(echo = TRUE)

# Load necessary libraries
library(ggplot2)
library(knitr)
library(dplyr)

# Reading the dataset
IB_data <- read.csv("Italian_Birds.csv")

# QQ plot for bird species richness (BSR)
ggplot(IB_data, aes(sample = BSR)) +
  stat_qq() +
```

```

stat_qq_line()+
labs(title = "QQ Plot for Bird Species Richness",
      x = "Theoretical Quantiles", y = "Sample Quantiles") +
theme_minimal()

# Reading the dataset
IB_data <- read.csv("Italian_Birds.csv")
# Performing shapiro test
shapiro_results <- IB_data %>%
  group_by(DOM.LU) %>%
  filter(n() >= 3) %>% # Groups having less than 3 observations were removed
  summarise(p_value = shapiro.test(BSR)$p.value)

# Printing results
print(shapiro_results)

# Performing the Kruskal-Wallis test
kruskal_test_result <- kruskal.test(BSR ~ DOM.LU, data = IB_data)

# Displaying the test result
print(kruskal_test_result)

```

Shapiro-Wilk test results:

```

## # A tibble: 8 x 2
##   DOM.LU          p_value
##   <chr>          <dbl>
## 1 Agricultural Area 2.40e- 7
## 2 Agricultural Areas 2.24e-28
## 3 Forest           1.16e-29
## 4 Grassland         3.14e- 1
## 5 Grasslands        1.00e- 8
## 6 Mixed             1.34e-14
## 7 Urban             7.19e- 1
## 8 Urban Areas       2.24e- 1

```

Kruskal-Wallis test results:

```

# Performing the Kruskal-Wallis test
kruskal_test_result <- kruskal.test(BSR ~ DOM.LU, data = IB_data)

# Displaying the test result
print(kruskal_test_result)

```

```

##
## Kruskal-Wallis rank sum test
##
## data: BSR by DOM.LU
## Kruskal-Wallis chi-squared = 16.199, df = 9, p-value = 0.06285

```