# Univariate Analysis and Basic R Programming

## Table 1: Summary Statistics for BD5_Species with Winsorized Mean

| Variable | Min Value | 1st Quartile | Median | Mean | 3rd Quartile | Max Value | Winsorized_Mean (20%) |
|---|---|---|---|---|---|---|---|
| Bird | 0.2415 | 0.8462 | 0.9038 | 0.8872 | 0.9570 | 1.1720 | 0.9007 |
| Bryophyte | 0.3941 | 0.6886 | 0.7993 | 0.7866 | 0.8855 | 1.1746 | 0.7889 |
| Butterflies | 0.3167 | 0.7926 | 0.8863 | 0.8746 | 0.9677 | 1.3944 | 0.8788 |
| Hoverflies | 0.1235 | 0.5696 | 0.6957 | 0.6795 | 0.8063 | 1.1453 | 0.6876 |
| Isopods | 0.0462 | 0.3917 | 0.5394 | 0.5500 | 0.7612 | 1.2577 | 0.5489 |

## Table 2: Pairwise Correlation Coefficients for all Pairs of Variables within BD5_Species

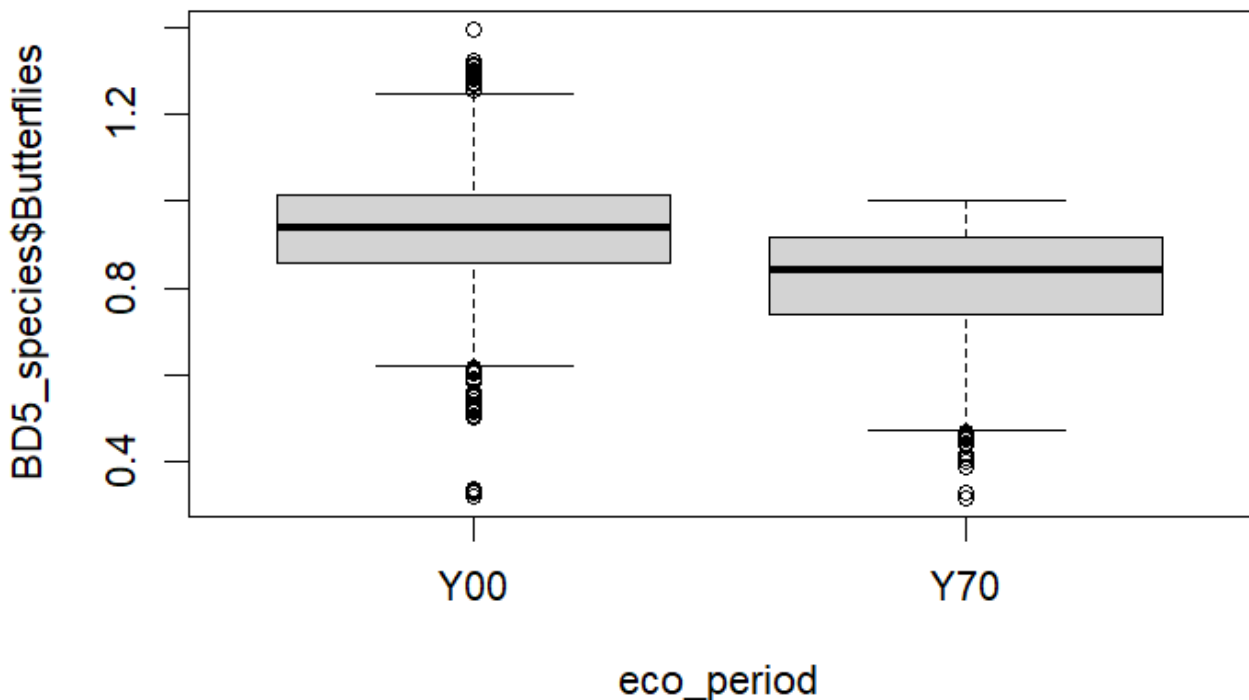| | Bird | Bryophyte | Butterflies | Hoverflies | Isopods |
|---|---|---|---|---|---|
| Bird | 1.0000 | -0.0851 | 0.3423 | 0.4352 | 0.1271 |
| Bryophyte | -0.0851 | 1.0000 | 0.1930 | 0.1322 | -0.0375 |
| Butterflies | 0.3423 | 0.1931 | 1.0000 | 0.2797 | -0.0510 |
| Hoverflies | 0.4352 | 0.1322 | 0.2797 | 1.0000 | 0.3824 |
| Isopods | 0.1271 | -0.0375 | -0.0510 | 0.3824 | 1.0000 |



Figure 1: Boxplot illustrating the biodiversity of bees during two periods: 1970-1990 and 2000-2013

From the summary statistic table (Table 1) above:

- The range (max-min) in the Butterflies variable is **1.0777 (1.3944 - 0.3167**), suggesting substantial variability.

- For all variables, the median is between the first quartile (Q1) and the third quartile (Q3), indicating a somewhat symmetrical distribution.

- Winsorized mean (20%) values are provided, which means that the top and bottom 10% of extreme values are replaced with the 10th and 90th percentiles, respectively.

- From the summary statistic table (Table 1) above, minor differences between the mean and the winsorized mean suggest less influence of extreme values.

- The fact that the median is close to the mean suggests a relatively symmetrical distribution for each variable, with no pronounced skewness.

From the correlation coefficient table (Table 2) above:

- There is a positive correlation between Butterflies and birds (**0.3423**) and a stronger positive correlation between Hoverflies and birds (**0.4352**).

- Hoverflies also show a positive correlation with Butterflies (**0.2797**). There is a negative correlation between Bird and Bryophyte (**-0.0851**) and between Bryophyte and Isopods (**-0.0375**).

- Isopods generally have weak correlations with the other variables, indicating that the other species might not strongly influence their abundance or presence.

It can be concluded that:

- The positive correlations between Birds and both Butterflies and Hoverflies suggest some association or co-occurrence between these species. This could be due to shared habitat preferences or ecological interactions.

- The positive correlation between Bird and Butterflies, as well as Hoverflies, may indicate that changes in the biodiversity of one group are moderately associated with changes in the others.

- Isopods have weaker associations with the other species, suggesting that factors influencing their presence are more independent of the dynamics affecting birds, butterflies, bryophytes, or hoverflies.

- The positive correlations suggest potential ecological relationships or dependencies between certain species. For example, an increase in bird biodiversity might be associated with increased butterfly or hoverfly biodiversity.

From the boxplot (Figure 1) above:

- All statistical measures (max, third quartile, median, first quartile, and min) show higher values in the latter period, suggesting a consistent or improving trend in butterfly biodiversity rather than a sporadic occurrence.

- Higher biodiversity in butterflies might indicate a healthier environment, as butterflies are sensitive to changes in habitat and environmental conditions. It could suggest that the ecosystem is supporting a diverse range of species. This could result from conservation efforts, habitat restoration, changes in environment, climate, or land use over time, favouring butterfly populations.

- If a positive trend in butterfly biodiversity is desired, it may suggest that conservation measures implemented between 1973-1990 and 2000-2013 have had a positive impact.

# Hypothesis Testing

**Hypothesis Test 1(Kolmogorov-Smirnov – k-test):**
$H_0$: The distributions of eco_status_5 is the same between Y70 and Y00.
$H_1$: The distributions of eco_status_5 is different between Y70 and Y00.

The test is performed on the variable eco_status_5 grouped by the variable period. The **D** (k-test) value is 0.18864. The p-value is less than < 2.2e-16 which is an extremely small value. The alternative hypothesis is two-sided, indicating that the test is checking for any kind of difference in the distributions of eco_status_5 between the two periods. The small p-value (at < 0.05) suggests that there is convincing evidence to conclude that the distributions differ significantly. Hence, the null hypothesis is rejected, and the alternative hypothesis is accepted.

**Hypothesis Test 2 (t-test):**
$H_0$: There is no significant difference between the means of BD11_species and BD5_species.
$H_1$: There is a significant difference between the means of BD11_species BD5_species.

*Table 3: T-test Results for BD5 Biodiversity Measure Between the mean of BD11_species and BD5_species*

| | Mean | t | df | p-value | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| ecologicalStatus | 0.7153582 | | | | Lower | Upper |
| eco_status_5 | 0.7555614 | -20.675* | 10262 | < 2.2e-16 | -0.0440 | -0.0364 |

The t-statistic is calculated as -20.675. The 95 percent confidence interval for the mean difference is reported as (-0.04401471, -0.03639155). This interval suggests that we can be 95 percent confident that the true difference in means falls within this range. The p-value is extremely small, reported as < 2.2e-16, indicating that the probability of observing such extreme differences in means by random chance is virtually zero. With such a small p-value, we reject the null hypothesis. Evidence suggests that the ecological status means of the two groups (BD11_species and BD5_species) are significantly different. Therefore, the null hypothesis (H0) is rejected based on the small p-value, indicating a highly significant result, and the alternative hypothesis (H1), which states that there is a significant difference in means between the two groups is accepted.

## Contingency Table

*Table 4a: Contingency Table (BD11up against BD5up)*

| | | | BD5_Species | | |
|---|---|---|---|---|---|
| | | | Down | Up | Total |
| | | Down | 1297 | 341 | 1638 |
| | | Up | 552 | 450 | 1002 |
| BD11_Species | | Total | 1849 | 791 | 2640 |

*Table 4b: BD11up against BD5up independent model Table.*

| | Down | Up |
|---|---|---|
| Down | 804.73 | 833.27 |
| Up | 1044.27 | -42.27 |

The likelihood ratio (G-test) and chi-squared test of independence, both yielding extremely small p-values (< 2.2e-16 and 2.764e-39, respectively), provide robust evidence against independence between the two factors in BD5_species and BD11_species. **G (69.21)** and **Chisq (171.96)** emphasize a significant association between BD11up and BD5up, reinforcing the rejection of the null hypothesis. The analysis, based on 2640 cases, suggests a meaningful relationship between the factors.

*Table 5: Displaying Odds Ratio, Sensitivity, Specificity, and Youden's Index*

| Parameters | Values |
|---|---|
| Odds_ratio | 3.100695 |
| Sensitivity | 0.7014602 |
| Specificity | 0.5689001 |
| Youden's_index | 0.2703604 |

From the table above (Table 5), the odds ratio of **3.100695** signifies a 3.1 times higher likelihood of the event happening. With a sensitivity of **0.7014602**, the model accurately identifies positive instances around **70.1%** of the time. Specificity at **0.5689001** indicates room for improvement in identifying negative instances. Youden's Index of **0.2703604** reflects a moderate balance between sensitivity and specificity. Overall, the model demonstrates potential, but enhancements may be considered for improved performance, especially in correctly identifying negative cases.

## Simple Linear Regression

*Table 6: Showing Results of Simple Linear Regression of BD1 and each of BD5.*

|  | Intercept | Slope | p-Value |
|---|---|---|---|
| Bird | 0.05045 | 0.62739 | < 2.2e-16*** |
| Bryophytes | 0.75924 | -0.19347 | < 2.2e-16*** |
| Butterflies | 0.70428 | -0.11116 | <1.48e-07*** |
| Hoverflies | 0.19755 | 0.60265 | < 2.2e-16*** |
| Isopods | 0.373025 | 0.425556 | < 2.2e-16*** |

*** denotes highly significant

2. For each BD5_Species, the simple linear regression analysis revealed significant and meaningful relationships between BD1 (response variable) and BD5 (predictor variable). The estimated slopes ($\beta_1$) were as follows: Bird (0.62739, p < 2.2e-16), Bryophytes (-0.19347, p < 2.2e-16), Butterflies (-0.11116, p = 1.48e-07), Hoverflies (0.60265, p < 2.2e-16), and Isopods (0.425556, p < 2.2e-16). In all cases, **the estimated slopes indicate that the associations are statistically significant.**

## Multiple Linear Regression

*Table 7a: Displaying Results of Multiple Linear Regression of Initial MLR*

| Predictor | Coefficient | Std. Error | t value | Pr(>|t|) |
|---|---|---|---|---|
| Intercept | 0.26704 | 0.02634 | 10.138 | <2e-16*** |
| Bird | 0.33562 | 0.02567 | 13.077 | <2e-16*** |
| Bryophytes | -0.17775 | 0.01851 | -9.601 | <2e-16*** |
| Butterflies | -0.32344 | 0.01858 | -17.406 | <2e-16*** |
| Hoverflies | 0.49476 | 0.01626 | 30.419 | <2e-16*** |
| Isopods | 0.23414 | 0.01190 | 19.672 | <2e-16*** |

*** denotes highly significant

Residual Standard Error: 0.169
Multiple R-squared: 0.3825
Adjusted R-squared: 0.3819
F-statistic: 653.3 on 5 and 5274 DF, p-value: < 2.2e-16

1. For the Initial MLR, the AIC Value is **-3783.29**

*Table 7b: Feature Selection Table Showing AIC Values and p-Values for Different Regression Models.*

| Model of Predictor | Features | AIC Value | p-value |
|---|---|---|---|
| Initial MLR | Bird + Bryophytes + Butterflies + Hoverflies + Isopods | -3783.29 | < 2.2e-16 |
| MLR1 | Bird + Bryophytes + Butterflies + Hoverflies | -3411.416 | < 2.2e-16 |
| MLR2 | Bird + Bryophytes + Butterflies + Isopods | -2931.801 | < 2.2e-16 |
| MLR3 | Bird + Bryophytes + Hoverflies + Isopods | -3490.382 | < 2.2e-16 |
| MLR4 | Bird + Butterflies + Hoverflies + Isopods | -3693.798 | < 2.2e-16 |
| MLR5 | Bryophytes + Butterflies + Hoverflies + Isopods | -3616.816 | < 2.2e-16 |

2. In the feature selection process, multiple linear regression (MLR) models was performed, progressively simplifying from an initial model with Bird, Bryophytes, Butterflies, Hoverflies, and Isopods, resulting in subsequent models (MLR1 to MLR5) with reduced AIC values and consistently low p-values, **indicative of improved model fit and statistical significance**, ultimately **justifying the removal of some variables** and yielding a more concise and interpretable set of predictors (Bird, Bryophytes, Butterflies, Hoverflies, Isopods) that collectively explain the variability in the response variable.

3. The **interaction term between Bird and Bryophytes** (-1.11482) is the lowest which mean it is the best. This model yields an AIC value of **-3832.356**, suggesting its superior fit compared to alternative models.

4. The linear regression model, trained on the Y70 subset and evaluated on the Y00 subset, demonstrated promising performance with a relatively low Mean Squared Error (MSE) of **0.0293** on the training set. However, when applied to the Y00 validation set, the higher MSE of **0.0871** suggests potential challenges in generalizing to new, unseen data, indicating the need for careful consideration and possible refinement of the model to enhance its overall predictive capability across different temporal periods.
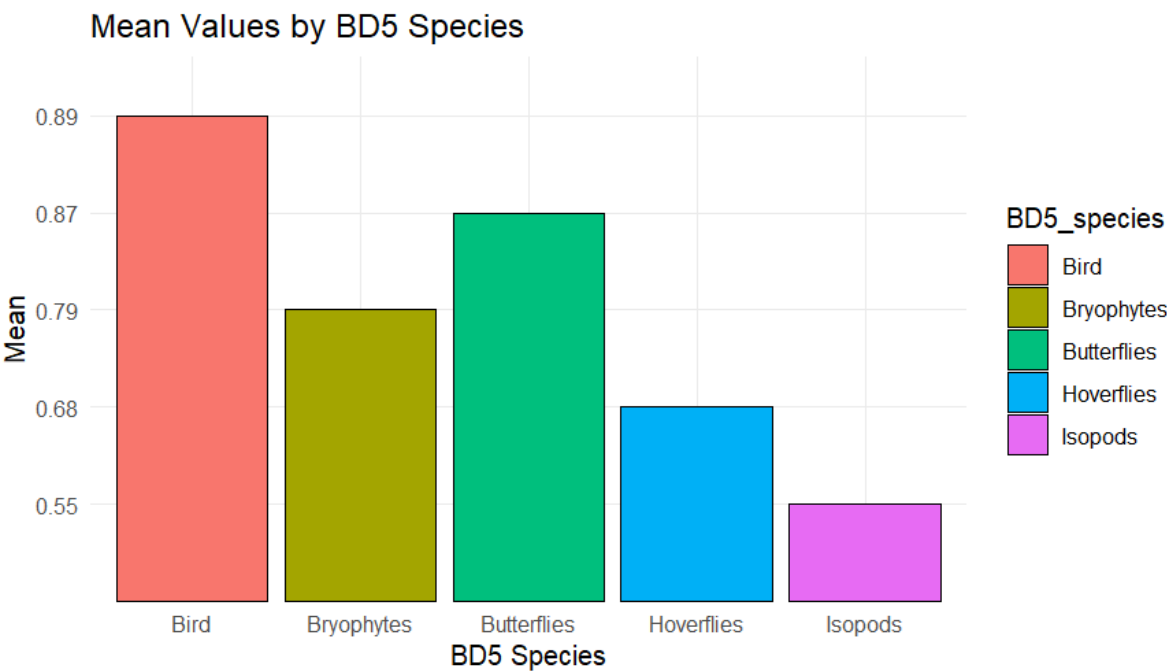
# Open Analysis



Figure 2: Histogram showing the distribution of mean values for the BD5 Species.

Here are some insights from fig 2:

**Bird Species:**
High average bird activity with low variability and a negatively skewed distribution. This indicates higher representation and a consistently favourable environment for birds, possibly with specific areas supporting exceptionally high bird activity.

**Bryophytes Species:**
Moderate presence of bryophytes with moderate variability and a slightly skewed distribution to the left. It suggests diverse ecological conditions supporting varied bryophyte growth, with a tendency for slightly lower presence in some areas within Britain.

**Butterflies Species:**
Relatively high presence of butterflies with moderate variability and a moderately negatively skewed distribution. This indicates favourable conditions for butterflies, with diverse habitats supporting different species and a tendency for slightly lower presence in some areas.

**Hoverflies Species:**
Moderate presence of hoverflies with relatively high variability and a moderately negatively skewed distribution. It suggests varying ecological conditions suitable for different hoverfly species, with a tendency for slightly lower presence in some areas.

**Isopods Species:**
Moderate presence of isopods with relatively high variability and a nearly symmetric distribution. It indicates a moderate presence of isopods with diverse ecological conditions supporting different species, showing consistent presence without a strong skew to either side.

Open Analysis 2:

*Table 8: Pairwise Correlation Coefficients for all Pairs of Variables BD5_Species and Easting and Northing*

|  | Bird | Bryophyte | Butterflies | Hoverflies | Isopods | Easting | Northing |
|---|---|---|---|---|---|---|---|
| **Bird** | 1.00 | -0.44 | 0.44 | 0.64 | 0.20 | 0.55 | -0.79 |
| **Bryophyte** | -0.44 | 1.00 | 0.21 | -0.09 | -0.27 | -0.91 | 0.48 |
| **Butterflies** | 0.44 | 0.21 | 1.00 | 0.40 | -0.17 | -0.17 | -0.38 |
| **Hoverflies** | 0.64 | 0.09 | 0.40 | 1.00 | 0.54 | 0.23 | -0.79 |
| **Isopods** | 0.20 | -0.27 | -0.17 | 0.54 | 1.00 | 0.26 | -0.52 |
| **Easting** | 0.55 | -0.91 | -0.17 | 0.23 | 0.26 | 1.00 | -0.67 |
| **Northing** | -0.79 | 0.48 | -0.38 | -0.79 | -0.67 | -0.67 | 1.00 |

From table 8, the following analysis could be drawn:
- **Bird and Hoverflies** show a strong positive correlation of **0.64**. This suggests that locations with high bird activity also tend to have high hoverfly activity.
- **Easting and Northing** have a positive correlation of **0.55**, indicating a spatial relationship.
- Bird and Northing have a strong negative correlation of **-0.79**, suggesting an inverse relationship. Areas with high bird activity may have lower northing values.
- **Bryophytes and Easting** show a strong negative correlation of **-0.91**, indicating an inverse spatial relationship.

- The strong negative correlation between **Bryophytes and Easting** suggests an inverse spatial relationship. This could imply that certain areas with higher Bryophytes are spatially located differently.
- The positive correlation between **Easting and Northing (0.55)** indicates a spatial trend, highlighting how ecological variables may vary across geographical coordinates.
- The positive correlation between **Bird and Hoverflies (0.64)** suggests a potential ecological association. Certain ecological conditions favourable for birds may also support hoverflies.
- The negative correlation between **Bird and Northing (-0.79)** indicates a spatial dynamic. Understanding this relationship can help in spatially targeted conservation efforts.
- Species with strong negative correlations (e.g., **Bryophytes and Easting**) may exhibit multicollinearity, impacting the stability of regression models.