Caitlin Forsyth
T00687692
BIOL4001-Biostatistics
Aditya Sharma
July 30, 2024

1. Summarize the data in ChickadeeData.csv numerically.

   The csv file "ChickadeeData.csv" contains 12 data variables, 7 numerical variables: EscShi, Entero, PathRich, WingChord, BirdWeight, TailLen, and TarsusLen and 5 factorial variables: Site, Habitat, Source, CommRich, and BirdSex.

   The variable 'EscShi' has a minimum value of 0, and a maximum value of 6520. The mean of 'EscShi' is 577.85, the median is 112, the standard deviation is 1199.91, the variance is 1439797, and the coefficient of variance is 2.08. There is an interquantile range of 435.75, and at the 25% and 75% quantiles the values are 23.50 and 459.25 respectively.

   The variable 'Entero' has a minimum value of 0, and a maximum value of 687. The mean of 'Entero' is 92.74, the median is 20, the standard deviation is 176.42, the variance is 31123.85, and the coefficient of variance is 1.90. There is an interquantile range of 39.75, and at the 25% and 75% quantiles the values are 9.00 and 48.75 respectively.

   The variable 'PathRich' has a minimum value of 11, and a maximum value of 88. The mean of 'PathRich' is 44.30, the median is 41, the standard deviation is 13.91, the variance is 193.52, and the coefficient of variance is 0.31. There is an interquantile range of 17.75, and at the 25% and 75% quantiles the values are 36.00 and 53.75 respectively.

   The 'nest' source has N/A values for the variables WingChord, BirdWeight, TailLen, TarsusLen, and BirdSex. Subsequently, these 'nest' variables were filtered to only include the 'feathers' data as to not have N/A values included.

   The variable 'WingChord' has a minimum value of 60, and a maximum value of 69. The mean of 'WingChord' is 64.78, the median is 65, the standard deviation is 2.25, the variance is 5.06, and the coefficient of variance is 0.03. There is an interquantile range of 4.00, and at the 25% and 75% quantiles the values are 63 and 67 respectively.

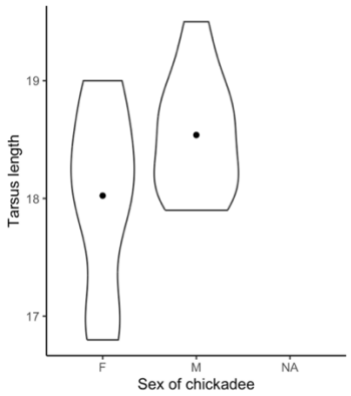
   The variable 'BirdWeight' has a minimum value of 10, and a maximum value of 17. The mean of 'BirdWeight' is 12, the median is 11, the standard deviation is 1.80, the variance is 3.24, and the coefficient of variance is 0.15. There is an interquantile range of 1, and at the 25% and 75% quantiles the values are 11 and 12 respectively.

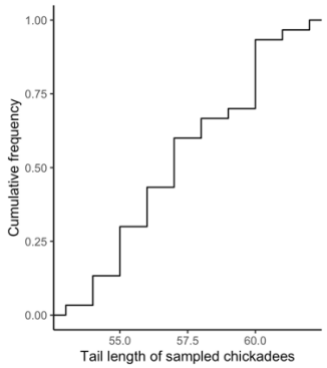
The variable ‘TailLen’ has a minimum value of 53, and a maximum value of 62. The mean of ‘TailLen’ is 57.23, the median is 57, the standard deviation is 2.46, the variance is 6.05, and the coefficient of variance is 0.04. There is an interquantile range of 5.00, and at the 25% and 75% quantiles the values are 55 and 60 respectively.

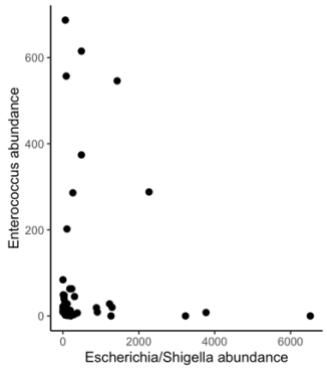
The variable ‘TarsusLen’ has a minimum value of 16.8, and a maximum value of 19.5. The mean of ‘TarsusLen’ is 18.25, the median is 18.25, the standard deviation is 0.69, the variance is 0.47, and the coefficient of variance is 0.04. There is an interquantile range of 0.80, and at the 25% and 75% quantiles the values are 17.9 and 18.7 respectively.

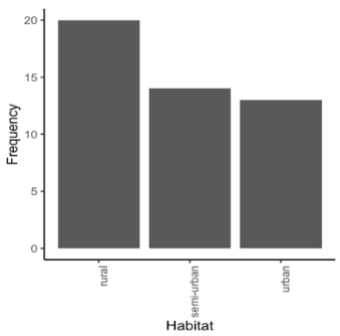
2. Summarize the data graphically.

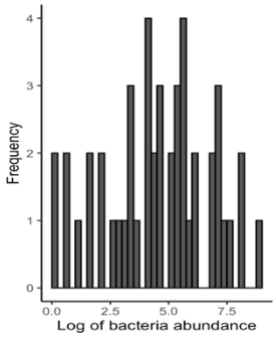
Please see below.

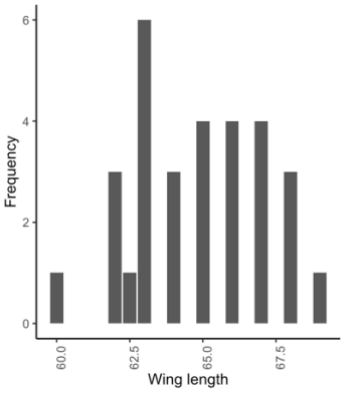


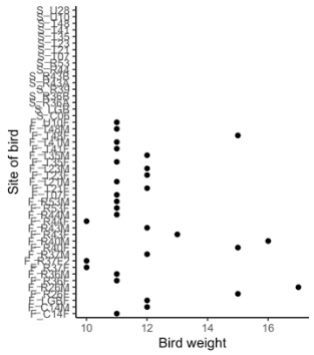


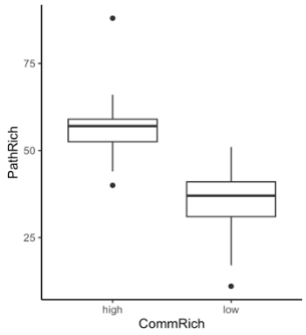


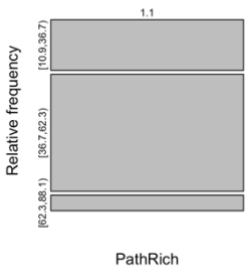


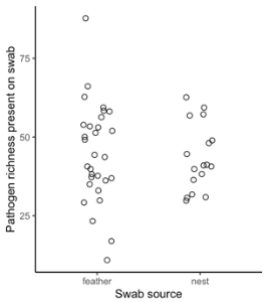


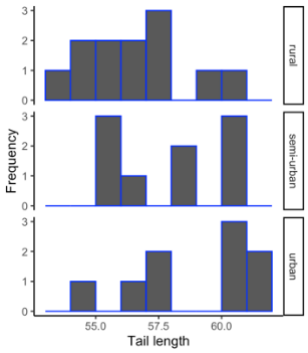












3. Use a two-sample *t*-test to determine whether the mean abundance of bacteria identified as genus *Escherichia/Shigella* differs significantly between nests and feathers. [Consider transforming variable **`EscShi`** and use R function **`t.test`**. Be careful when applying the natural logarithm transformation to data containing zeros since ln(0) is undefined. The usual way around this is to calculate ln(y+1) instead of ln(y).]

H0= There are no significant differences between the mean abundance of bacteria identified as genus *Escherichia/Shigella* in nest and feather 'Source' categories (p-value > 0.05)

HA= There are significant differences between the mean abundance of bacteria identified as genus *Escherichia/Shigella* in nest and feather 'Source' categories (p-value < 0.05).

The two-sample t-test returns a t-value (test statistic) equal to 4.0756 at 45 degrees of freedom, with a p-value equal to 0.0001842. Since the p-value is less than 0.05, we reject the null hypothesis, and accept the alternative hypothesis that there is a significant difference between the mean abundance of bacteria identified as genus *Escherichia/Shigella* in nest and feather 'Source' categories.

4. Use a Mann-Whitney U-test (Wilcoxon rank sum test) to determine whether the population distribution of the abundance of bacteria identified as genus *Escherichia/Shigella* differs significantly between nests and feathers. [Use R function **`wilcox.test`**.]

H0= There are no significant differences between the population distribution of the abundance of bacteria identified as genus *Escherichia/Shigella* in nest and feather 'Source' categories (p-value > 0.05).

HA= There are significant differences between population distribution of the abundance of bacteria identified as genus *Escherichia/Shigella* in nest and feather 'Source' categories (p-value < 0.05).

The Mann-Whitney U-test returned a W-value (test statistic) of 405.5, and a p-value of 0.0008952. Since the p-value is less than 0.05, we reject the null hypothesis, and accept the alternative hypothesis that there are significant differences between population

distribution of the abundance of bacteria identified as genus *Escherichia/Shigella* in nest and feather 'Source' categories.

5. Use a two-sample *t*-test to determine whether the mean abundance of bacteria identified as genus *Enterococcus* differs significantly between nests and feathers. [Consider transforming variable **Entero** and use R function **t.test**.]

H0= There are no significant differences between the mean abundance of bacteria identified as genus *Enterococcus* in nest and feather 'Source' categories (p-value > 0.05)

HA= There are significant differences between the mean abundance of bacteria identified as genus *Enterococcus* in nest and feather 'Source' categories (p-value < 0.05).

The two-sample t-test returns a t-value (test statistic) equal to - 2.6195 at 45 degrees of freedom, with a p-value equal to 0.01196. Since the p-value is less than 0.05, we reject the null hypothesis, and accept the alternative hypothesis that there is a significant difference between the mean abundance of bacteria identified as genus *Enterococcus* in nest and feather 'Source' categories.

6. Use a Mann-Whitney U-test (Wilcoxon rank sum test) to determine whether the population distribution of the abundance of bacteria identified as genus *Enterococcus* differs significantly between nests and feathers. [Use R function **wilcox.test**.]

H0= There are no significant differences between the population distribution of the abundance of bacteria identified as genus *Enterococcus* in nest and feather 'Source' categories (p-value > 0.05).

HA= There are significant differences between population distribution of the abundance of bacteria identified as genus *Enterococcus* in nest and feather 'Source' categories (p-value < 0.05).

The Mann-Whitney U-test returned a W-value (test statistic) of 148, and a p-value of 0.01826. Since the p-value is less than 0.05, we reject the null hypothesis, and accept the alternative hypothesis that there are significant differences between population distribution of the abundance of bacteria identified as genus *Enterococcus* in nest and feather 'Source' categories.

7. Use analysis of variance to determine whether mean pathogen richness is related to habitat. [Use R functions **lm** and **anova**.]

H_0: $\mu_1 = \mu_2 = \mu_3$ (p-value > 0.05).

H_A: At least one μ_i is different from the others (p-value < 0.05).

The F-value (test statistic) is 5.2745, and based on 2 and 44 degrees of freedom, the computed p-value was 0.008844. Since the p-value < 0.05, we reject the null hypothesis and conclude that there is sufficient evidence that at least one μ_i is different from the others.

8. Use a Kruskal-Wallis test to determine whether mean pathogen richness is related to habitat. [Use R function **`kruskal.test`**.]

H0: The distributions are the same for all three densities (p-value >0.05).
HA: The distributions differ among the densities. (p-value <0.05).

The Kruskal-Wallis test returned a chi value (test statistic) of 13.587 at 2 degrees of freedom, and a p-value of 0.001121. Since the p-value is less than 0.05, we can reject the null hypothesis and accept the alternative that the distributions differ among the densities (mean pathogen richness and habitats).

9. Use the Tukey-Kramer method to which habitats differ with respect to mean pathogen richness. [Use R functions **`emmeans`** and **`contrast`**.]

H_0: $\mu_1 = \mu_2 = \mu_3$ (p-value > 0.05).

H_A: At least one μ_i is different from the others (p-value < 0.05).

The Tukey-Kramer method returned t-ratios (test statistics) of -2.036, -3.136, and -1.0599 for rural- (semi-urban), rural-urban and (semi-urban)-urban, respectfully. Each of these variables had 44 degrees of freedom, and varying p-values. For the rural- (semi-urban) and (semi-urban)-urban variables, the p-values were 0.1155 and 0.5445 respectfully. For these comparisons, we would accept the null hypothesis that there are no significant differences between these habitats in terms of pathogen richness (p-value > 0.05). For the rural – urban variables however, there is a p-value of 0.0084, meaning we reject the null hypothesis and accept the alternative that at least one mean is different from the others.

10. Use contingency table analysis to determine whether community richness on feathers is independent of mountain chickadees sex? [Use R function **`filter`** in the **`dplyr`** package to subset the data frame to retain only feather swabs, then use R functions **`table`** and **`chisq.test`**.]

H0= There are no significant differences between community richness on feathers and chickadee sexes (p-value > 0.05).

HA= There are significant differences between the community richness of feathers and chickadee sexes (p-value < 0.05).

The contingency table analysis returned a chi-squared value (test statistic) of 0.5333. Using one degree of freedom, the p-value was calculated as 0.4652. This is above the significance level of 0.05, which means that based on this data, we accept the null hypothesis that there are no significant differences between community richness on feathers and chickadee sexes (p-value > 0.05).

11. Use simple linear regression to determine which of **`WingChord`**, **`BirdWeight`**, **`TailLen`**, or **`TarsusLen`** is most linearly associated with pathogen richness on feathers. [Use R functions **`lm`** and **`summary`**.]

    H0= There are no significant differences between the chosen predictor and pathogen richness on feathers (p-value > 0.05).

    HA= There are significant differences between the chosen predictor and pathogen richness on feathers (p-value < 0.05).

For the WingChord (predictor) variable on the PathRich (response) variable of feathers, the estimated linear regression equation is: PathRich= -38.829+ 1.29 (WingChord). The f-value (test statistic) is 0.9984 and the p-value for this regression analysis is 0.326.

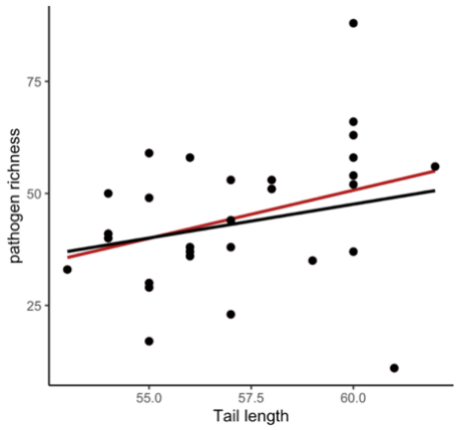
For the BirdWeight (predictor) variable on the PathRich (response) variable of feathers, the estimated linear regression equation is: PathRich= 32.8943+ 0.989(BirdWeight). The f-value (test statistic) is 0.3676 and the p-value for this regression analysis is 0.5492.

For the TailLen (predictor) variable on the PathRich (response) variable of feathers, the estimated linear regression equation is: PathRich= -78.153+ 2.148 (TailLen). The f-value (test statistic) is 3.599 and the p-value for this regression analysis is 0.06816.

For the TarsusLen (predictor) variable on the PathRich (response) variable of feathers, the estimated linear regression equation is: PathRich= 145.14 – 5.501 (TarsusLen). The f-value (test statistic) is 1.746 and the p-value for this regression analysis is 0.1971.

Each of these tests used 1 degree of freedom as the numerator, and 28 degrees of freedom as the denominator. All p-values were above the significance level of 0.05, so the F-value was used to determine best fit. The valid f-value at the corresponding degrees of freedom was 3.34, and the TailLen predictor had an F-distribution above this accepted value at 3.599.

12. For the simple linear regression model selected in Question 11, investigate whether there are any outliers. An outlier in the context of linear regression is an observation with an extreme difference between the observed response value and the predicted response value from the model. Use the **`filter`** function to create a new data frame that excludes the most extreme outlier and refit the simple linear regression model. Report the results for the model fit to the data frame excluding the outlier and use this data frame to answer Questions 13-18.



H0= There are no significant differences between the chosen predictor and response variables (p-value > 0.05).

HA= There are significant differences between the chosen predictor and response variables (p-value < 0.05).

From the above graph it was easy to visualize the outlier lying above a pathogen richness of 80. The filter function was utilized to remove the outlier, and a new regression analysis was calculated. The new regression analysis equation is:
PathRich= -43.019 +1.510(TailLen). The regression was calculated on 1 and 27 degrees of freedom, has an f-value of 2.155 and a p-value of 0.1537. Since the p-value is greater than the significance level of 0.05, we must accept the null hypothesis that there are no differences between the predictor (TailLen) and the response variables (PathRich on feathers).

13. Find the best multiple linear regression model for predicting pathogen richness on feathers from two predictors out of **`WingChord`**, **`BirdWeight`**, **`TailLen`**, and **`TarsusLen`**. [Use R functions **`lm`** and **`summary`**.]

H0= There are no significant differences between the chosen predictor and pathogen richness on feathers (p-value > 0.05).

HA= There are significant differences between the chosen predictor and pathogen richness on feathers (p-value < 0.05).

For the WingChord (predictor) variable on the PathRich (response) variable of feathers, the estimated linear regression equation is: PathRich= -38.829+ 1.29 (WingChord). The f-value (test statistic) is 0.9984 and the p-value for this regression analysis is 0.326.

For the BirdWeight (predictor) variable on the PathRich (response) variable of feathers, the estimated linear regression equation is: PathRich= 32.8943+ 0.989(BirdWeight). The f-value (test statistic) is 0.3676 and the p-value for this regression analysis is 0.5492.

For the TailLen (predictor) variable on the PathRich (response) variable of feathers, the estimated linear regression equation is: PathRich= -78.153+ 2.148 (TailLen). The f-value (test statistic) is 3.599 and the p-value for this regression analysis is 0.06816.

For the TarsusLen (predictor) variable on the PathRich (response) variable of feathers, the estimated linear regression equation is: PathRich= 145.14 – 5.501 (TarsusLen). The f-value (test statistic) is 1.746 and the p-value for this regression analysis is 0.1971.

Each of these tests used 1 degree of freedom as the numerator, and 28 degrees of freedom as the denominator. All p-values were above the significance level of 0.05, and ideally we would accept the null hypothesis, but to make a determination the F-value was used to determine best fit. The valid f-value at the corresponding degrees of freedom was 3.34, and the TailLen predictor had an F-distribution above this accepted value at 3.599. The second

predictor that had the highest f-value and closest p-value to an accepted value was the TarsusLen predictor.

14. Use a two-sample *t*-test to determine whether mean pathogen richness on feathers differs significantly between male and female birds. [Use R function **`t.test`**.]

H0= There are no significant differences between mean pathogen richness on feathers and chickadee sexes (p-value > 0.05).

HA= There are significant differences between mean pathogen richness on feathers and chickadee sexes (p-value < 0.05).

The two-sample t-test returns a t-value (test statistic) equal to – 0.9893 at 28 degrees of freedom, with a p-value equal to 0.331. Since the p-value is greater than 0.05, we accept the null hypothesis and conclude there is no significant difference between the mean pathogen richness of either chickadee sex.

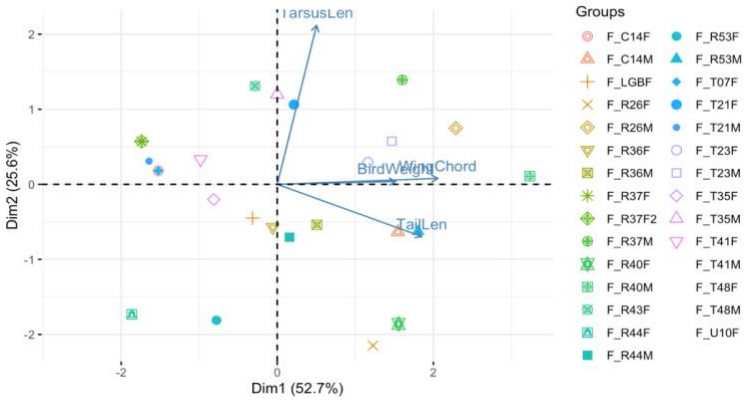
15. Starting from the simple linear regression model in Question 12, use analysis of covariance to investigate whether the linear association in the model differs for male and female birds. [Use R functions **`lm`** and **`summary`**.]

H0= There are no significant differences between mean pathogen richness on feathers and chickadee sexes (p-value > 0.05).

HA= There are significant differences between mean pathogen richness on feathers and chickadee sexes (p-value < 0.05).

The type III ANCOVA test returns a t-value (test statistic) equal to 1.05 at 2 and 26 degrees of freedom, with a p-value equal to 0.3643. Since the p-value is greater than 0.05, we accept the null hypothesis, and conclude there is no significant difference between the mean pathogen richness of either chickadee sex.

16. Apply principal component analysis to the variables **`WingChord`**, **`BirdWeight`**, **`TailLen`**, and **`TarsusLen`**. [Use R functions **`scale`**, **`prcomp`**, and **`summary`**. Also, summarize the principal component loadings and use R function **`fviz_pca_ind`** with argument **`habillage`** to visualize the results and colour the observations by bird sex.]



17. Fit a simple linear regression model with response variable **`PathRich`** and predictor variable equal to the first principal component from Question 16. Compare this model with the simple linear regression model in Question 12. [Use R functions **`lm`** and **`summary`**.]

H0= There are no significant differences between the chosen predictor and response variables (p-value > 0.05).

HA= There are significant differences between the chosen predictor and response variables (p-value < 0.05).

The regression from Q12 was calculated on 1 and 27 degrees of freedom, has an f-value of 2.155 and a p-value of 0.1537. The regression from Q17 was calculated on 1 and 27 degrees of freedom, has an f-value of 0.4894 and a p-value of 0.4902. Since the p-value is greater than the significance level of 0.05, we must accept the null hypothesis that there are no differences between the predictor (TailLen) and the response variables (PathRich on feathers).

18. Fit a multiple linear regression model with response variable **`PathRich`** and predictor variables equal to the first two principal components from Question 16. Compare this model with the multiple linear regression model in Question 13. [Use R functions **`lm`** and **`summary`**.]

    H0= There are no significant differences between the chosen predictor and response variables (p-value > 0.05).

    HA= There are significant differences between the chosen predictor and response variables (p-value < 0.05).

    The regression from Q13 for BirdWeight (predictor) variable on the PathRich (response) variable of feathers, the f-value is 0.3676 and the p-value for this regression analysis is 0.5492.

    The regression from Q18 for BirdWeight (predictor) variable on the PathRich (response) variable of feathers, the f-value is 0.4894 and the p-value for this regression analysis is 0.4902

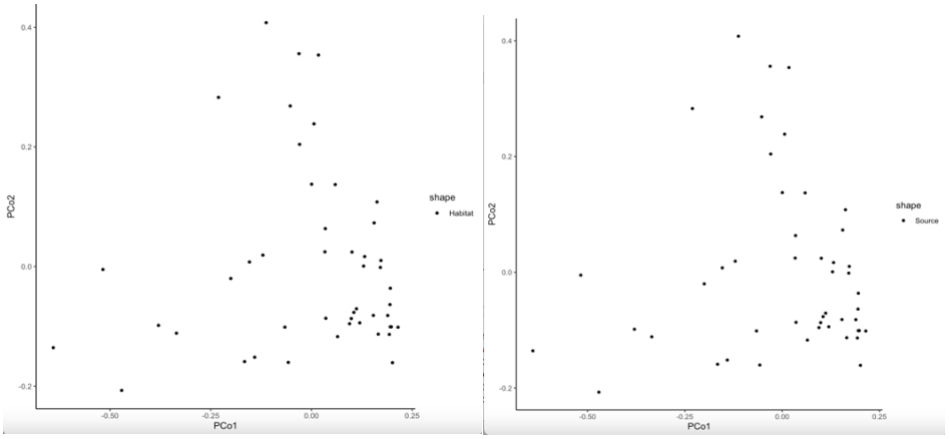
The regression from Q13 for TarsusLen (predictor) variable on the PathRich (response) variable of feathers, the f-value is 1.746 and the p-value for this regression analysis is 0.1971.

The regression from Q18 for TarsusLen (predictor) variable on the PathRich (response) variable of feathers, the f-value is 3.644 and the p-value for this regression analysis is 0.06694.

Since all p-values are above the significance level of 0.05, we must accept the null hypothesis that H0= There are no significant differences between the chosen predictor and response variables (p-value > 0.05).

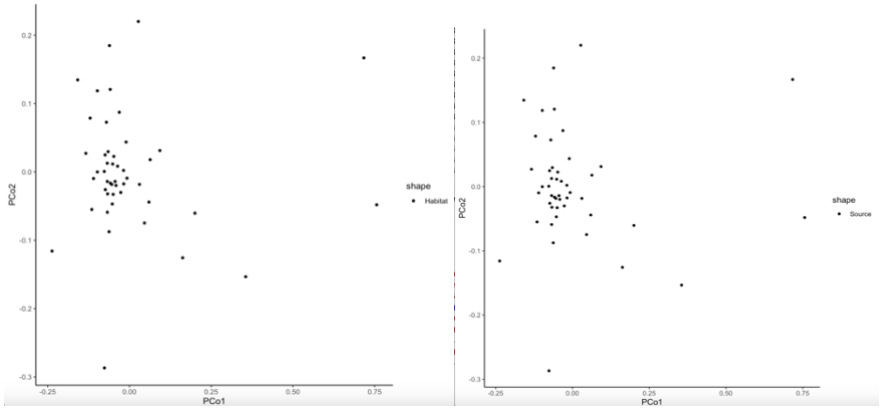
19. Apply metric multidimensional scaling (MDS) to the dissimilarities data in the **ChickadeeDissimilarities.csv** file. [Use R function **cmdscale** to perform the metric MDS and use function **ggplot** to create a scatterplot that projects the data onto the first two principal coordinates. Then use the **shape** aesthetic to first mark the points by **Habitat** and then by **Source**.] Does the composition of the microbial communities appear to be related to **Habitat** or **Source**?

The microbial communities did not change when related to 'Habitat' or 'Source'. Please see below.



20. Apply nonmetric multidimensional scaling (NMDS) to the dissimilarities data in the **ChickadeeDissimilarities.csv** file. [Use R function **metaMDS** to perform the NMDS and use function **ggplot** to create a scatterplot that projects the data onto the first two NMDS axes. Then use the **shape** aesthetic to first mark the points by **Habitat** and then by **Source**.] Does the composition of the microbial communities appear to be related to **Habitat** or **Source**?

The microbial communities did not change when related to 'Habitat' or 'Source'. Please see below.


PCo2
PCo1
shape
Habitat
PCo2
PCo1
shape
Source