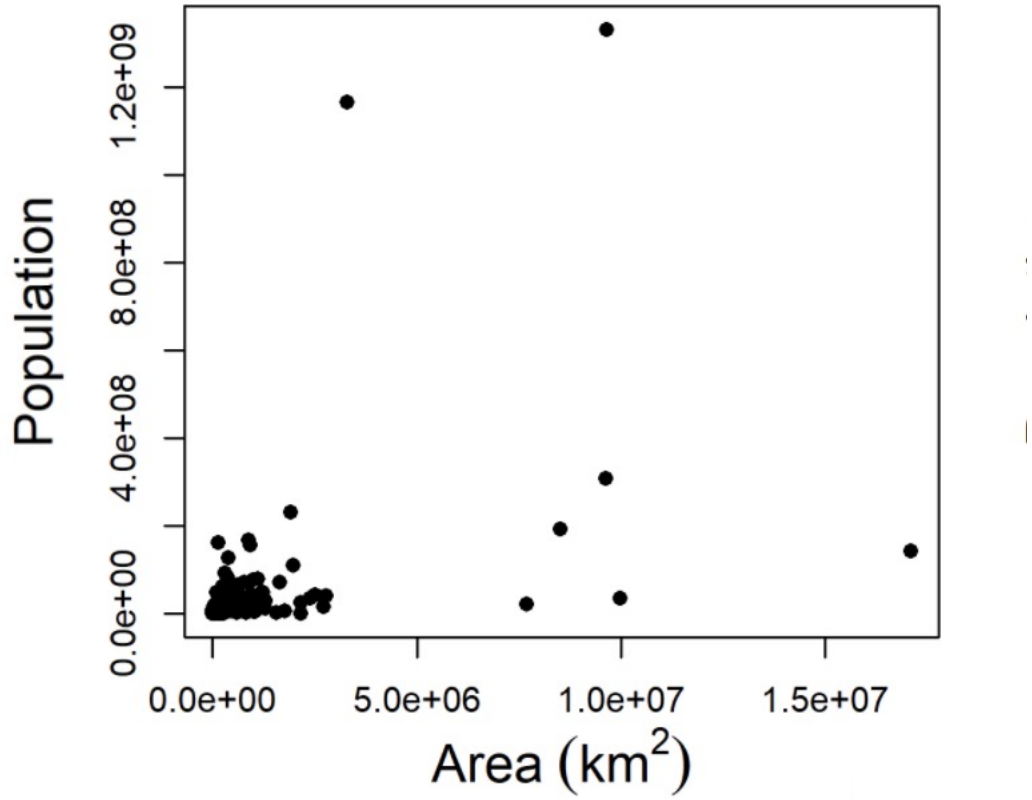
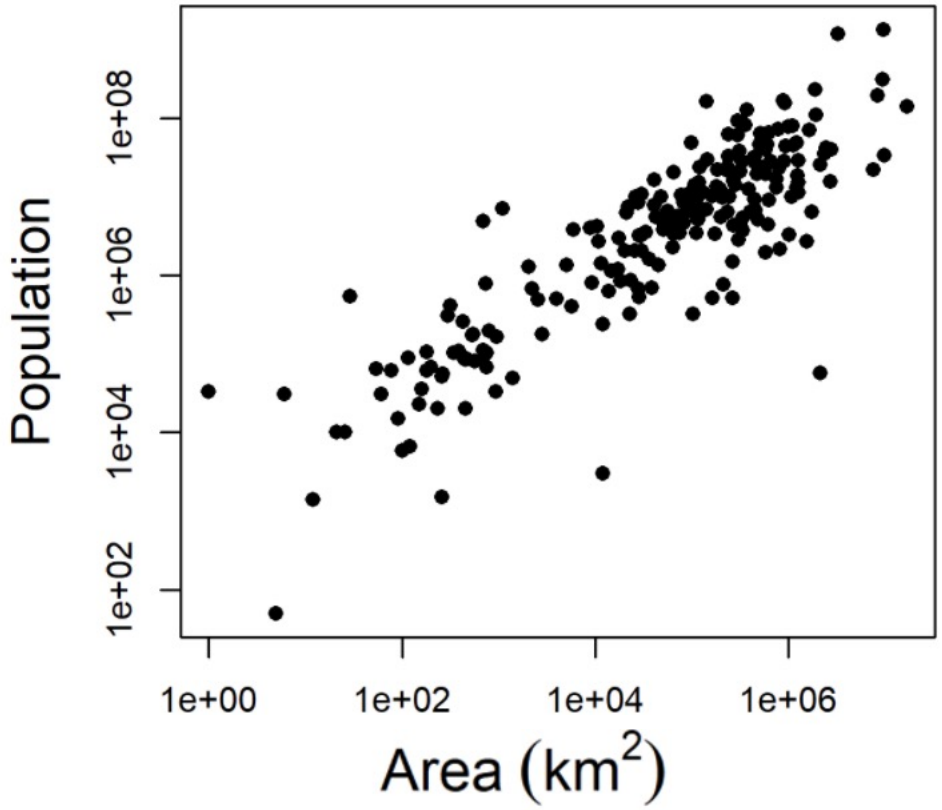
PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

- Data Transformations
- MANOVA (Multivariate Analysis of Variance)

Raw Data – no obvious relationship between variables. A few locations with large populations dominate the relationship



Log Transformed Data – clear positive, linear relationship between the two variables



Data Transformations

- Mathematical functions that are applied to all observations of a given variable
- Can be applied to specific or all variables in a dataset if desired
- Usually simple algebraic functions
- Often used for linearizing data

Data Transformations

- Mathematical functions that are applied to all observations of a given variable
- Can be applied to specific or all variables in a dataset if desired
- Usually simple algebraic functions
- Often used for linearizing data

Two main types:

- Linear. Such as adding constant or multiplying by constant. They do not change results of statistical tests. Preserve the relationships between variables
- Non-linear. Such as taking the square root. Results of statistical tests will differ. Change the relationships between variables

Data Transformations

- Mathematical functions that are applied to all observations of a given variable
- Can be applied to specific or all variables in a dataset if desired
- Usually simple algebraic functions
- Often used for linearizing data

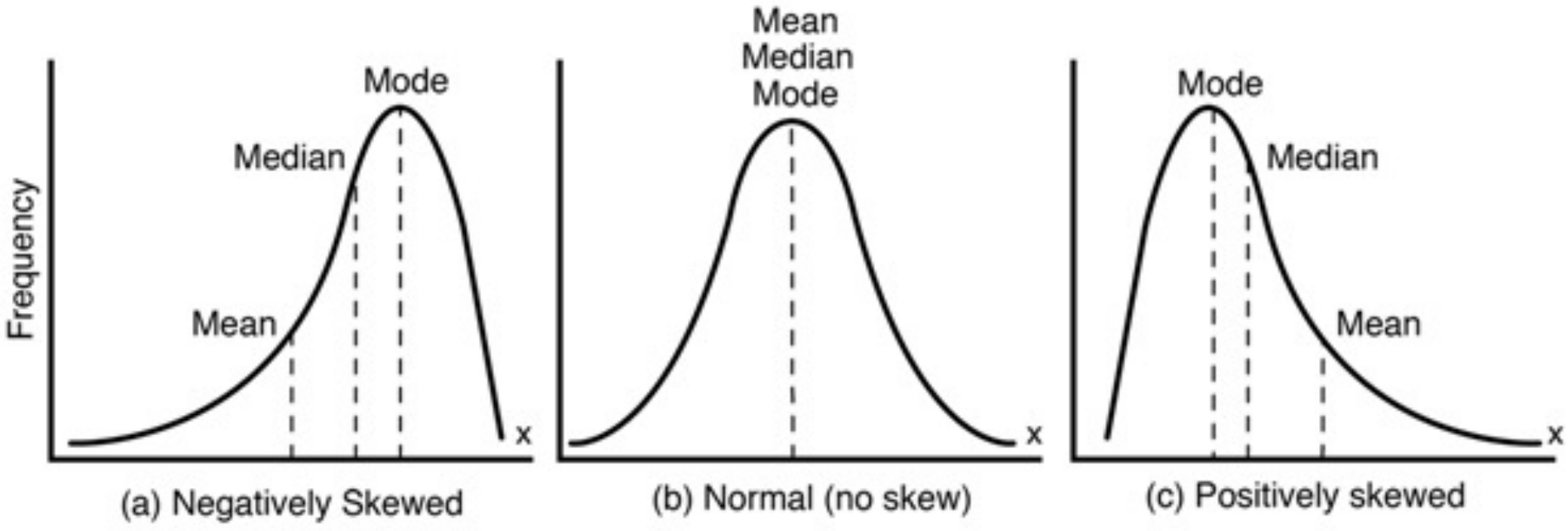
Two main types:

- Linear. Such as adding constant or multiplying by constant. They do not change results of statistical tests. Preserve the relationships between variables
- Non-linear. Such as taking the square root. Results of statistical tests will differ. Change the relationships between variables

It is important to check that a transformation:

a) Didn’t make the distribution of the variable worse
b) Didn’t generate outliers

Distributions of Variables



Also called left-skewed data

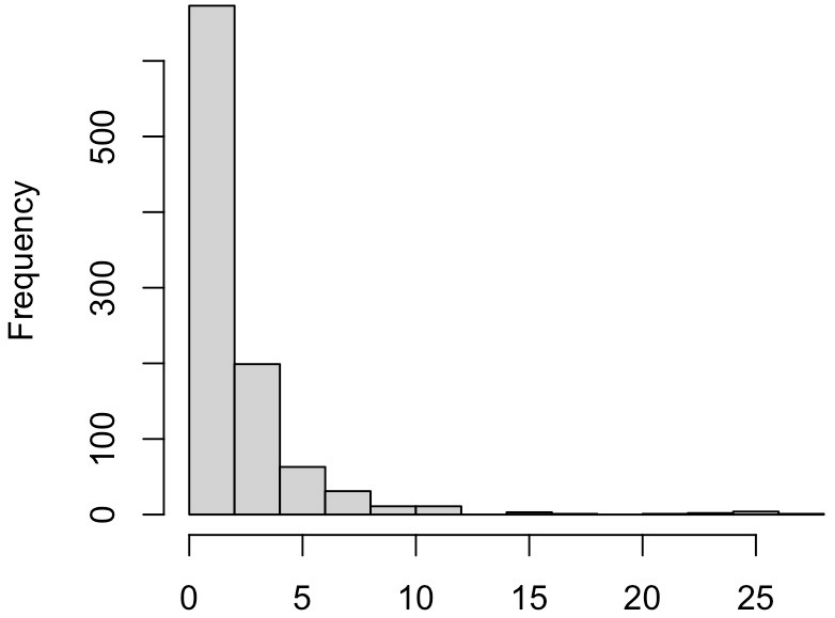
Also called right-skewed data

Logarithmic Transformation

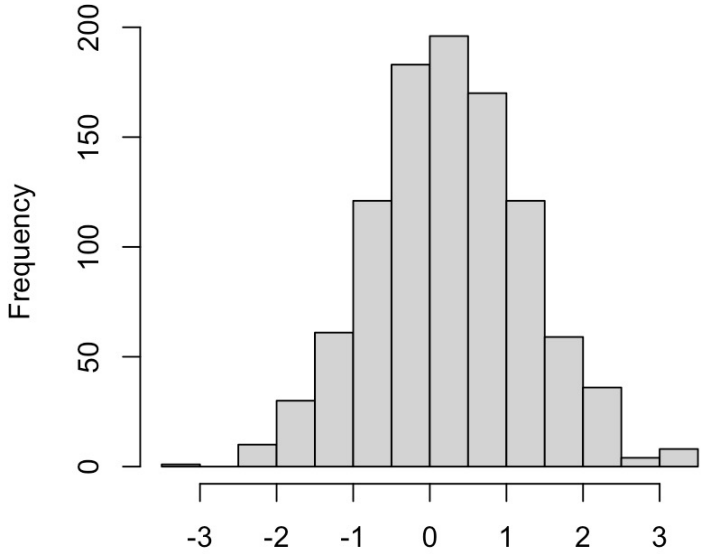
$$y' = \log(y)$$

Replaces the value of each observation with its logarithm

Raw data with extreme positive skew



Same data after log transformation



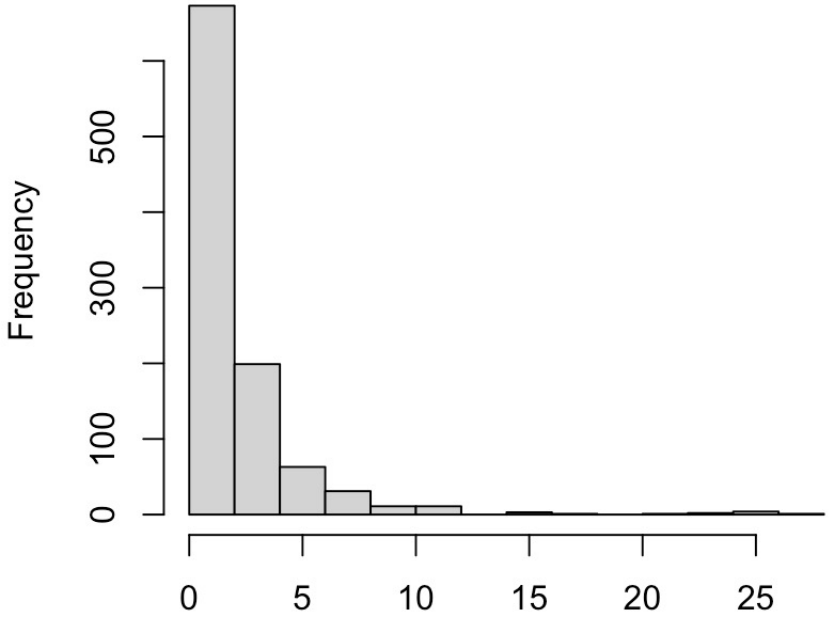
Logarithmic Transformation

$$y' = \log(y)$$

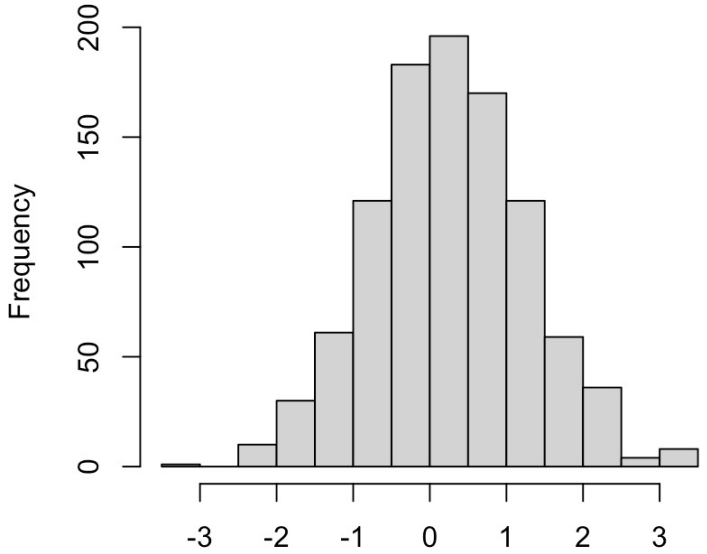
Replaces the value of each observation with its logarithm

- Reduces the range of a variable
- Often used to equalize variances for data in which the mean and variance are correlated
  - Aka variables with larger means also have larger variances
- Makes positively skewed data more symmetrical
- Produces similar but more extreme results than a square root transformation
- Log of 0 is not defined, so need to add 1
  - log1p() in R

Raw data with extreme positive skew



Same data after log transformation

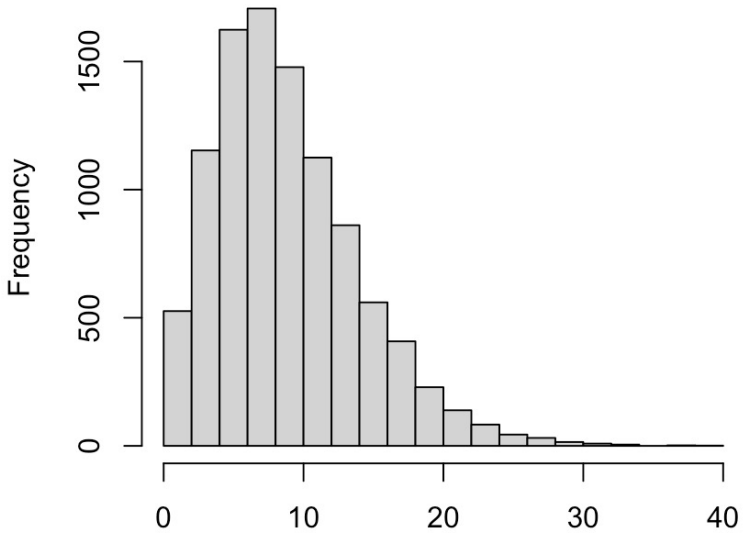


Square Root Transformation

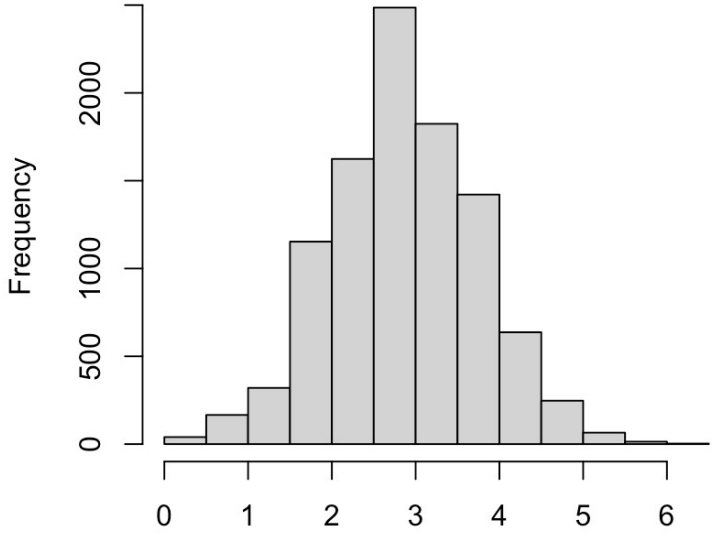
$$y' = \sqrt{y}$$

Replaces each value with its square root

Raw data with positive skew



Same data after square root transformation



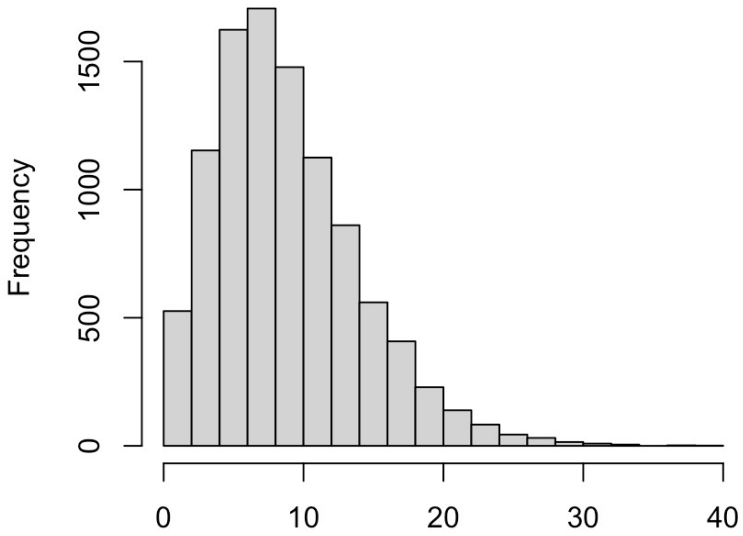
Square Root Transformation

$$y' = \sqrt{y}$$

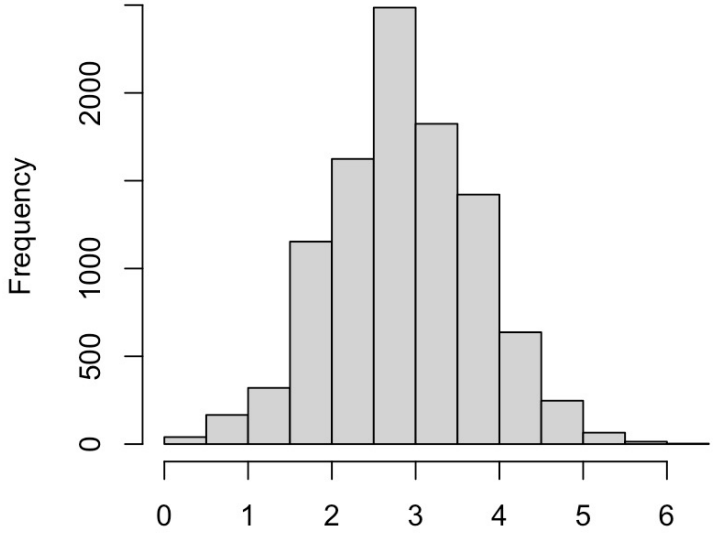
Replaces each value with its square root

- Reduces the range of the data
- Compresses large values more than small ones
- Useful in transforming variables with a small proportion of large values which distort the overall distribution
- Often used on data with a positively skewed distribution

Raw data with positive skew



Same data after square root transformation

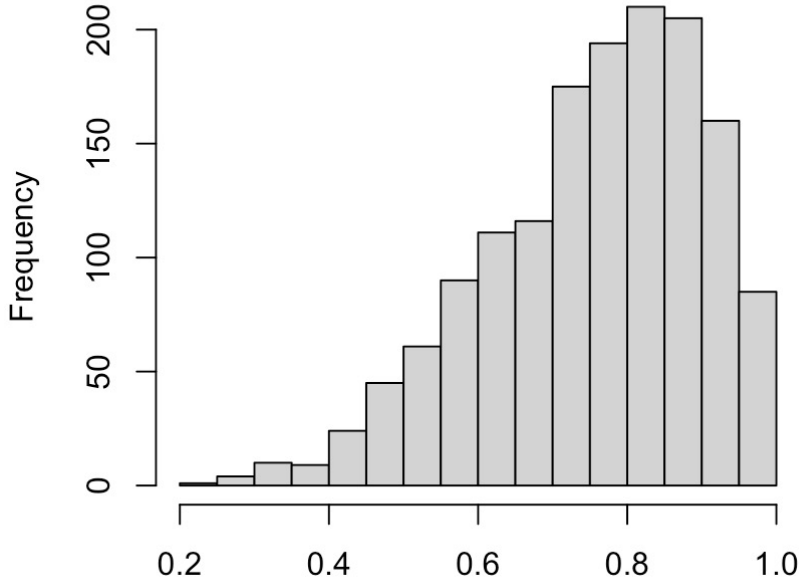


Power Transformation

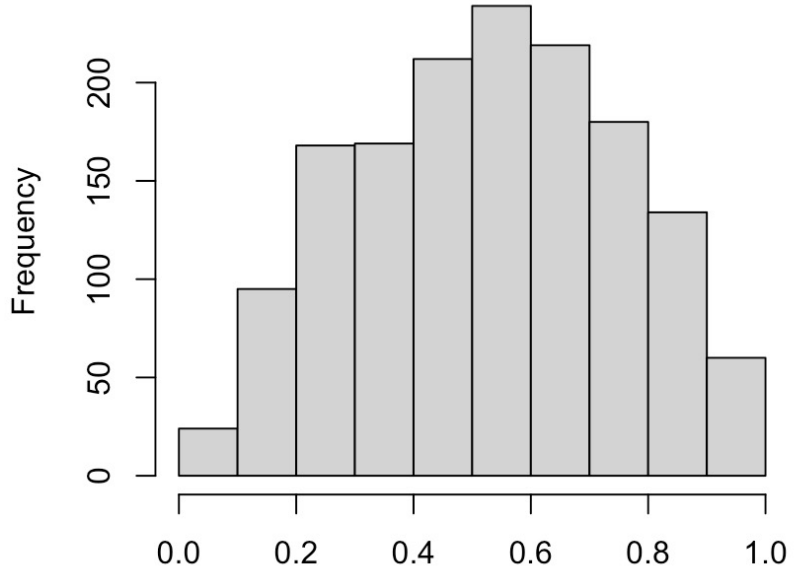
$$y' = y^p$$

- Replaces the value of each observation with the value raised to a certain power

Raw data with negative skew



Same data after y^2 transformation

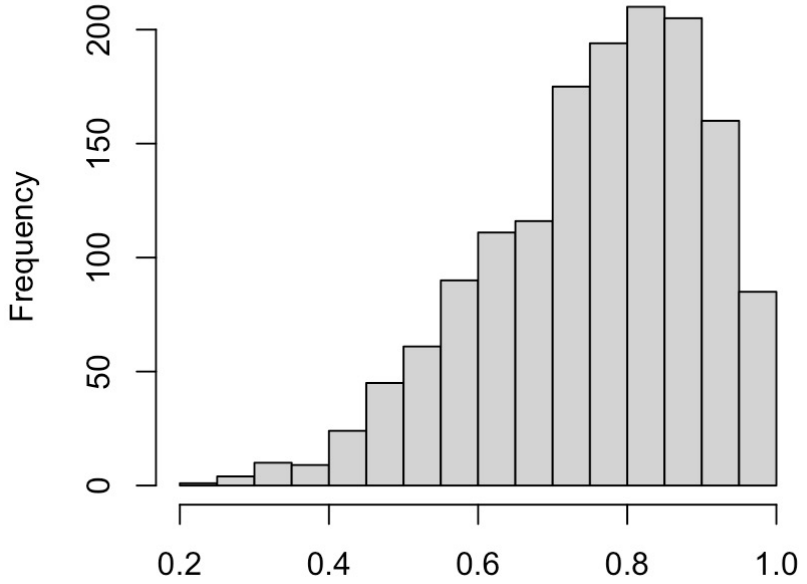


Power Transformation

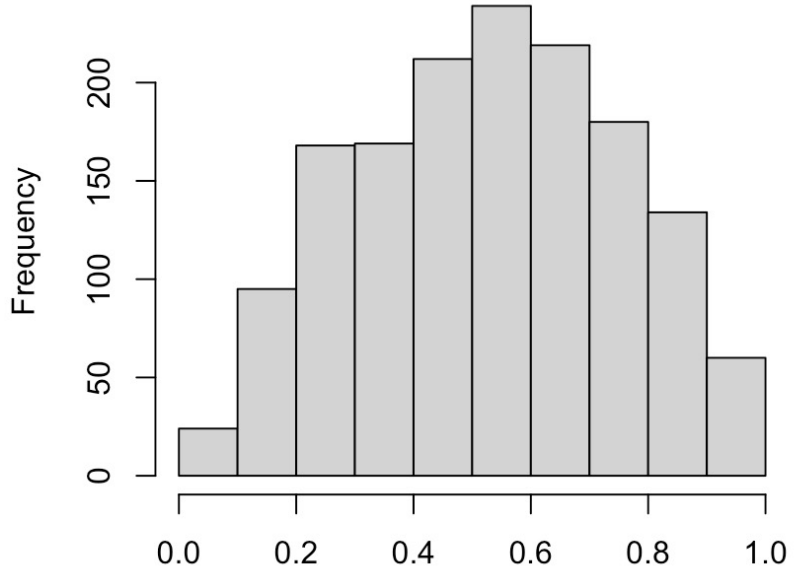
$$y' = y^p$$

- Replaces the value of each observation with the value raised to a certain power
- This is a family of transformations since *p* can vary in value
  - P = 0.5 is square root transformation
- *p* =2 and *p* = 3 often suitable for transforming negatively skewed data

Raw data with negative skew



Same data after y^2 transformation



Others

Arcsine, arsine-square root or angular transformation

$y' = arcsin(y)$ or $y' = arcsin(\sqrt{y})$

Replaces the value of each observation with the arcsine of the square root of the value

Primarily used for percentages or proportions

Not everyone is in favor of this transformation

See: Warton and Hui. 2011. The arcsine is asinine. Ecology. 92:3-10

Reciprocal transformation

$$y' = \frac{1}{y}$$

Replaces the value of each observation with its reciprocal

Suitable for data recorded as rates

e.g. number of offspring per female, height/body mass ratio

Standardization

- Methods that change the data using the data itself
- Applied to all columns or rows in a data set
- Applied when variables differ in units, variability or scale to remove differences in relative weights (importance) of individual variables
- Do not change the skew of the data

Standardization sensu stricto

- Also called Z-scores
- Variables have a mean of zero and standard deviation of one after standardization
- Variables become unitless
- Important to use when the variables have different scales or units of measurement

scale() function

Standardization

- Methods that change the data using the data itself
- Applied to all columns or rows in a data set
- Applied when variables differ in units, variability or scale to remove differences in relative weights (importance) of individual variables
- Do not change the skew of the data

Ranging/Normalizing

- Changes the range of a variable
- Max is 1 and the min is 0
- Often achieved by dividing each value by the maximum for the variable
- Brings all variables to a common scale

Centering

- Changes each variable to have a mean of 0. Achieved by subtracting the mean
- Removes differences in scale between variables
- Does not expand or contract the distribution of values for a variable

Standardization sensu stricto

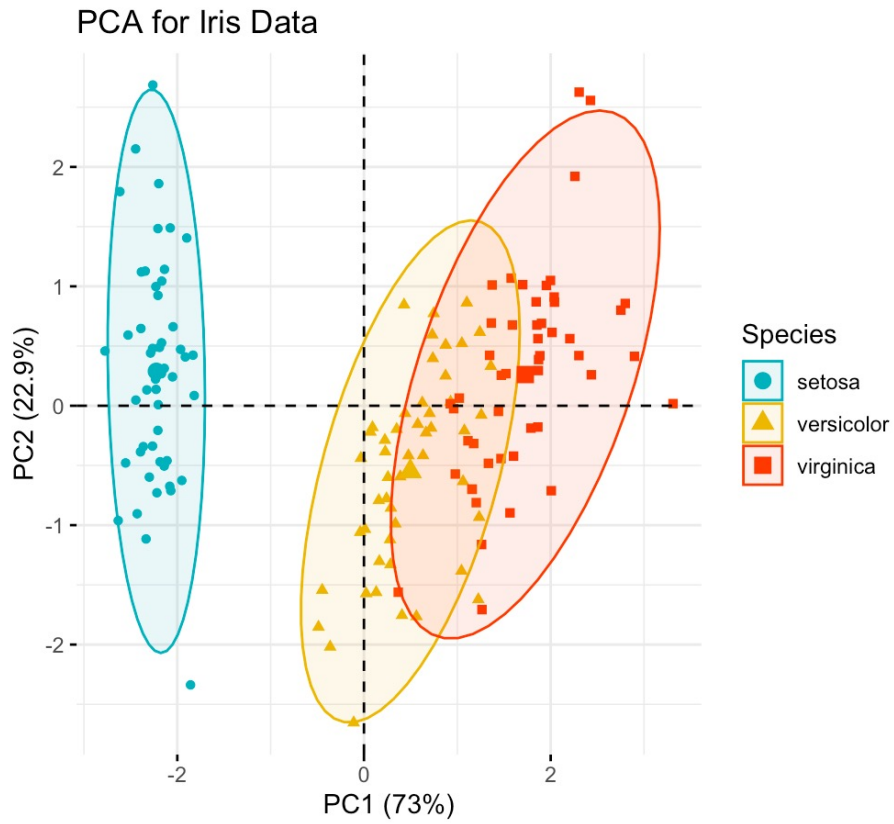
- Also called Z-scores
- Variables have a mean of zero and standard deviation of one after standardization
- Variables become unitless
- Important to use when the variables have different scales or units of measurement

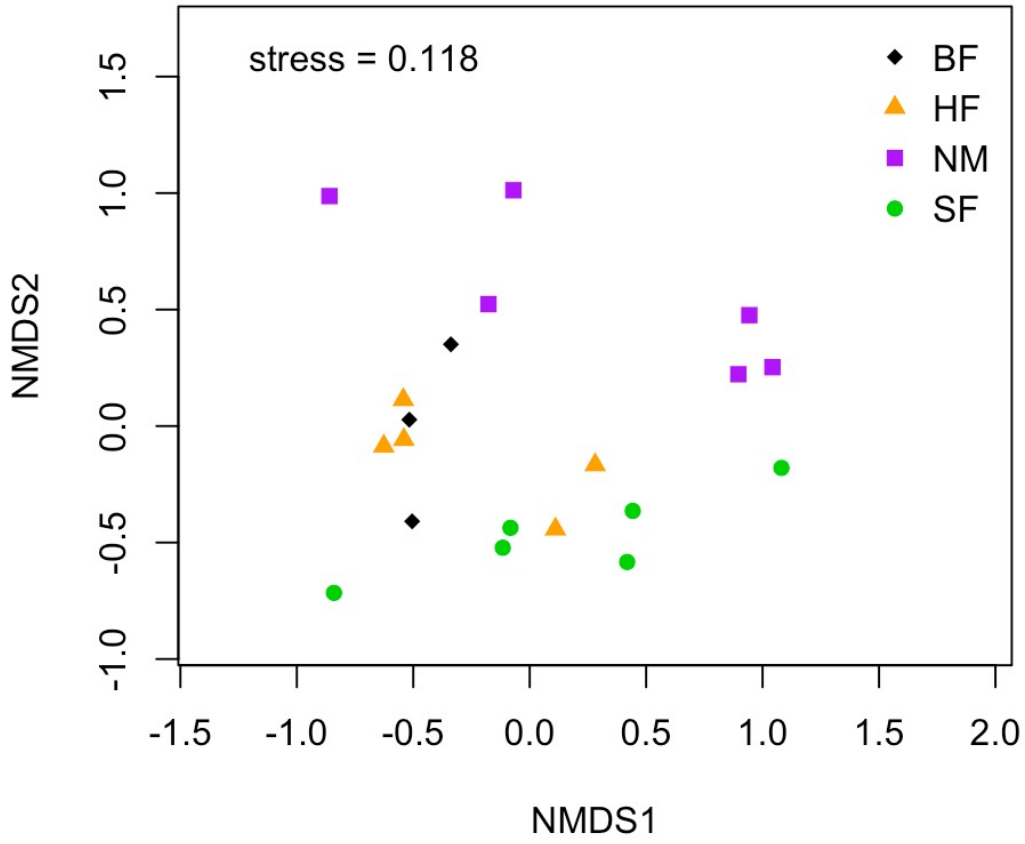
scale() function

PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

- **MANOVA (Multivariate Analysis of Variance)**

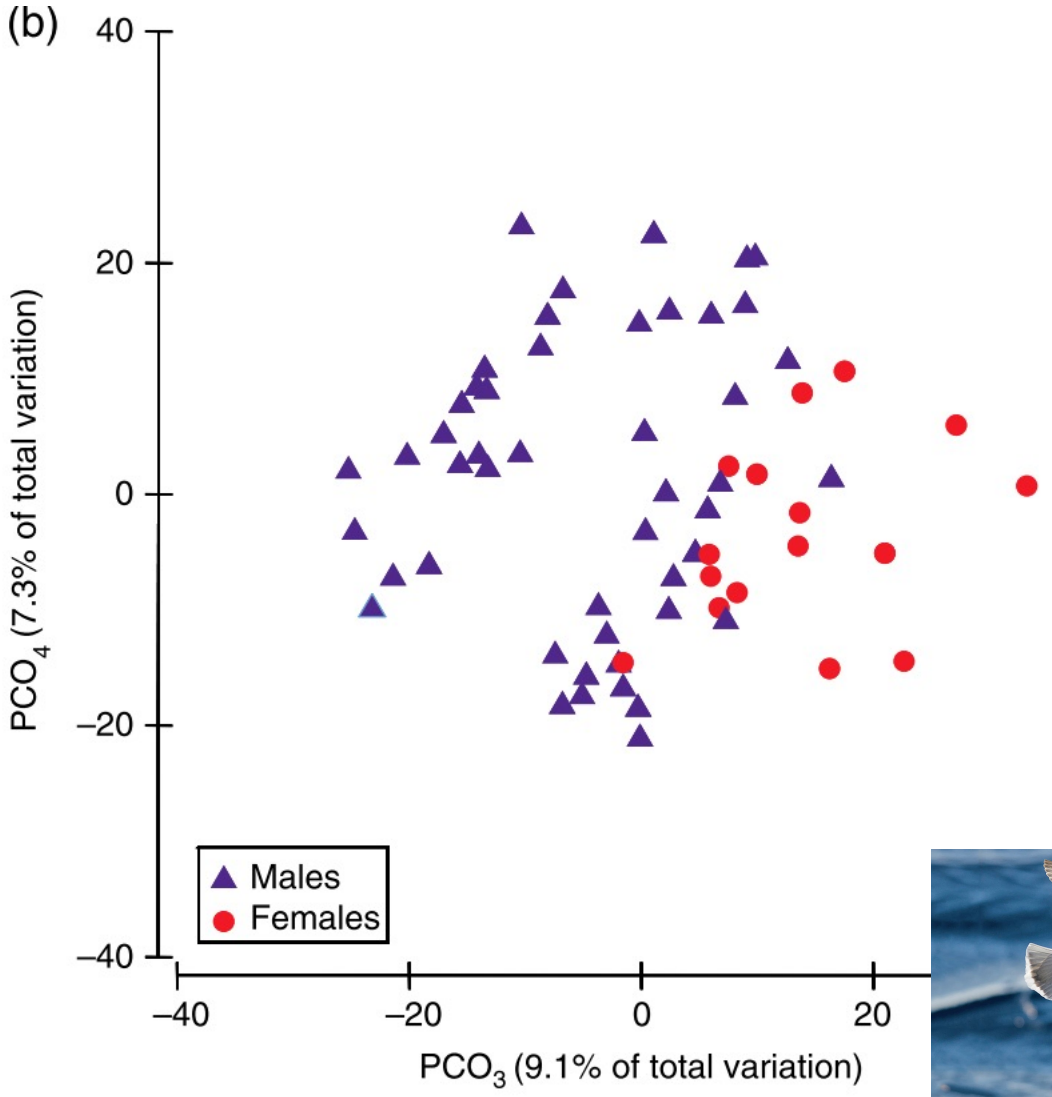
Ordination of Multivariate Data Can Reveal Patterns





While groups can appear distinct in ordination plots, theses types of analysis are unsupervised and are not testing a hypothesis

Subtle patterns are sometimes visible beyond axis 1 and 2



PCO4

PCO3

Differences between males and females only visible when looking at PCO3 and PCO4 (the 3rd and 4th axis of a PCoA)

Notice how little of the total variation in the data are explained by these two axes despite the pattern being quite clear

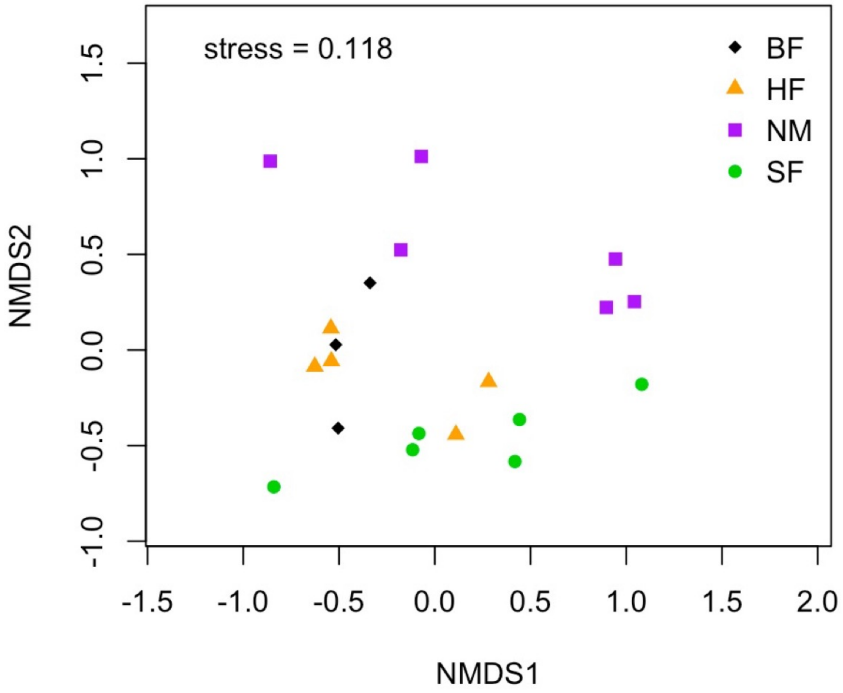
We need inferential methods to test whether these patterns are significant

- Do females occupy a different location in multivariate space than males?

Figure from Mardon at al. 2011. J Avian Biol. 42:29-38

How do we test if groups are different?

MANOVA!



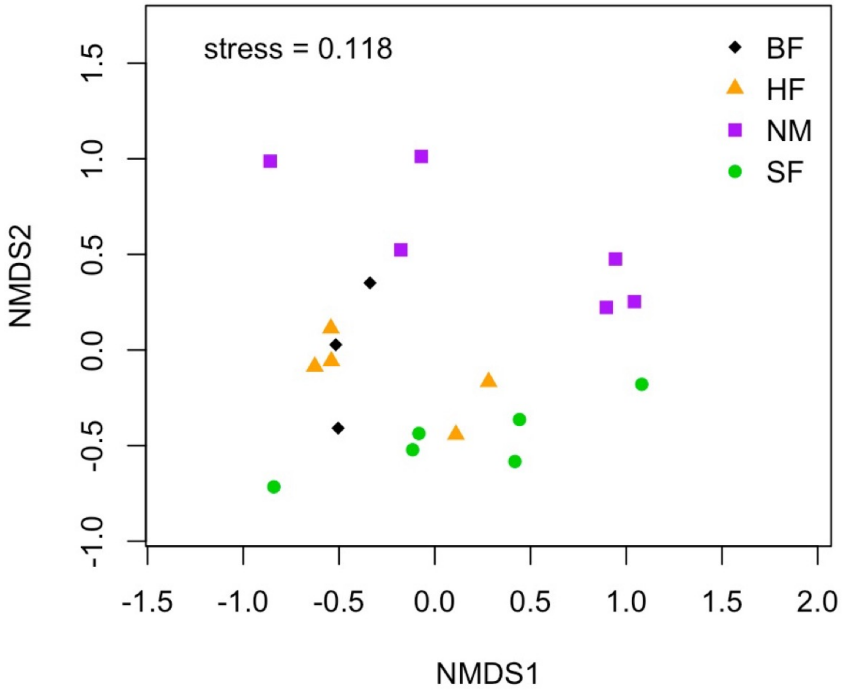
Linear regression (ANOVA)

lm(response ~ predictors)

Single response variable

How do we test if groups are different?

MANOVA!



Linear regression (ANOVA)

lm(response ~ predictors)

Single response variable

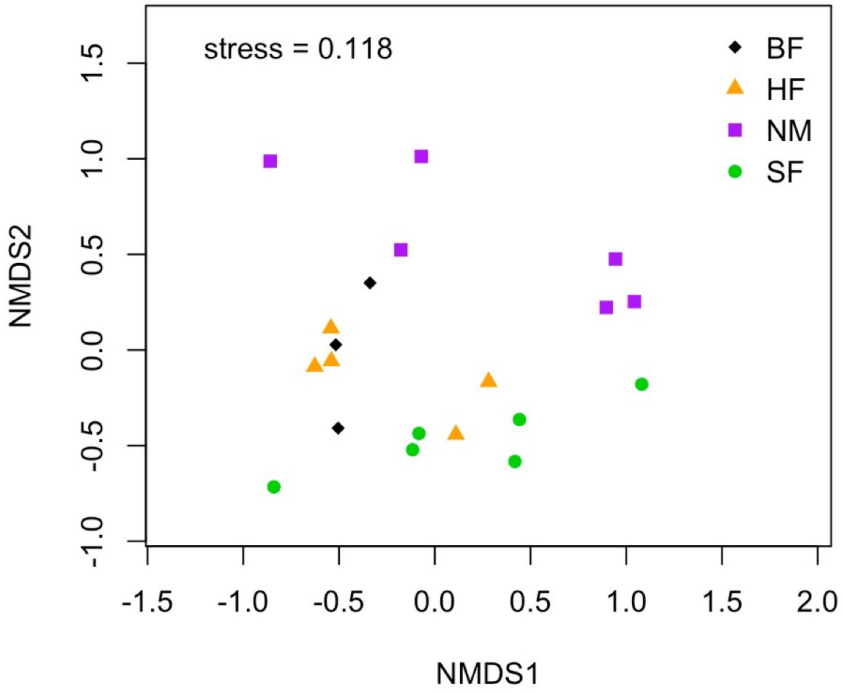
MANOVA

manova(**responses** ~ predictors)

Multiple response variables

How do we test if groups are different?

MANOVA!



Linear regression (ANOVA)

lm(response ~ predictors)

Single response variable

MANOVA

manova(response**s** ~ predictors)

Multiple response variables

PERMANOVA (distance based)

adonis2(responses ~ predictors)

Analysis of variance using distance matrices

Multiple response variables

MANOVA

Multivariate Analysis of Variance

- Extension of ANOVA to analyze data with more than 2 dependent variables
- Reduces Type I error that would occur by performing multiple, separate ANOVAs on the dependent variables
- Instead of comparing group means, we are now comparing group centroids in multivariate space

MANOVA

Multivariate Analysis of Variance

- Extension of ANOVA to analyze data with more than 2 dependent variables
- Reduces Type I error that would occur by performing multiple, separate ANOVAs on the dependent variables
- Instead of comparing group means, we are now comparing group centroids in multivariate space

H_0: the multivariate means of all groups are equal

H_A: at least one pair of groups have different multivariate means

The MANOVA alone won't tell you which pairs of groups differ, so post-hoc pairwise comparisons are also needed

Assumptions of MANOVA

Some carry over from univariate ANOVA:

- Independent observations/samples
- The variance of the different groups is equal and normally distributed (homogeneity of variance)

Assumptions of MANOVA

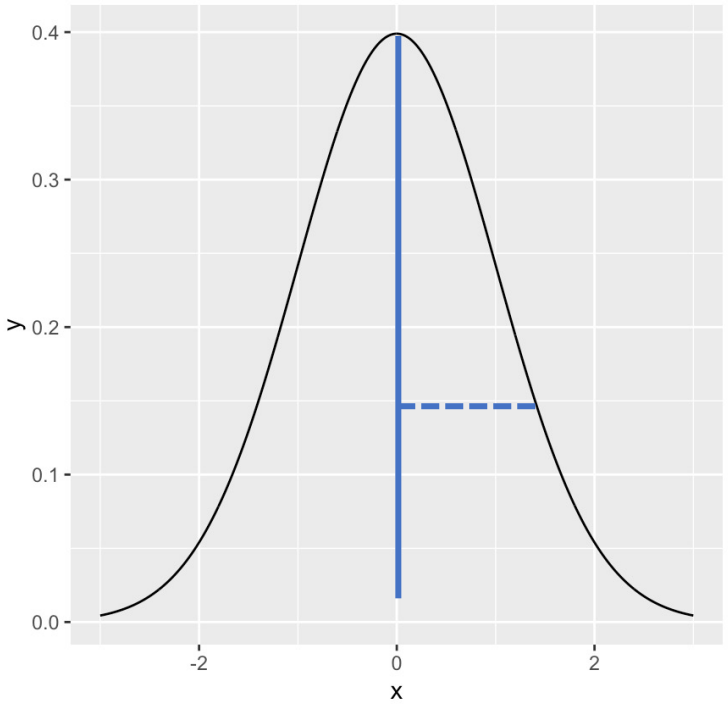
Some carry over from univariate ANOVA:

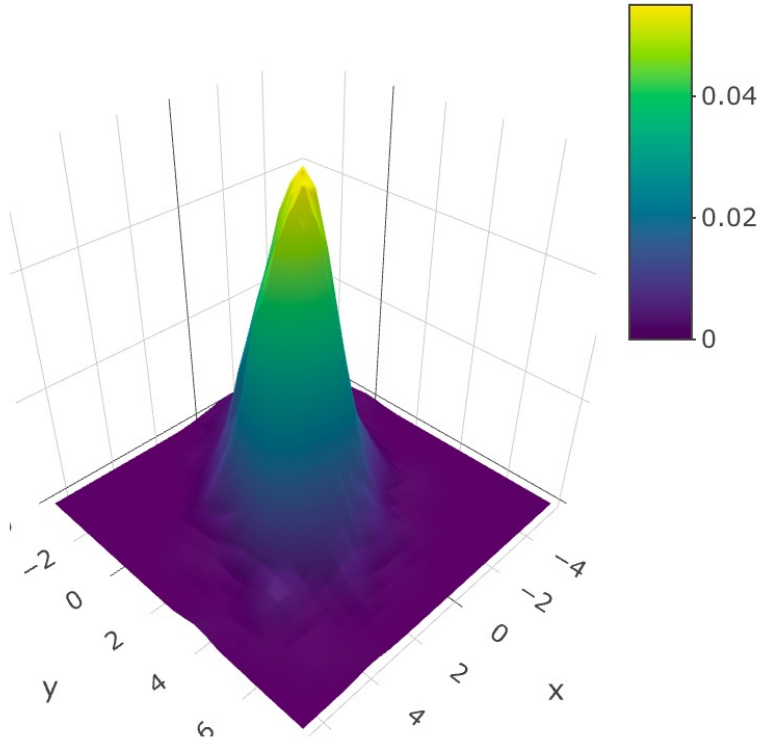
- Independent observations/samples
- The variance of the different groups is equal and normally distributed (homogeneity of variance)

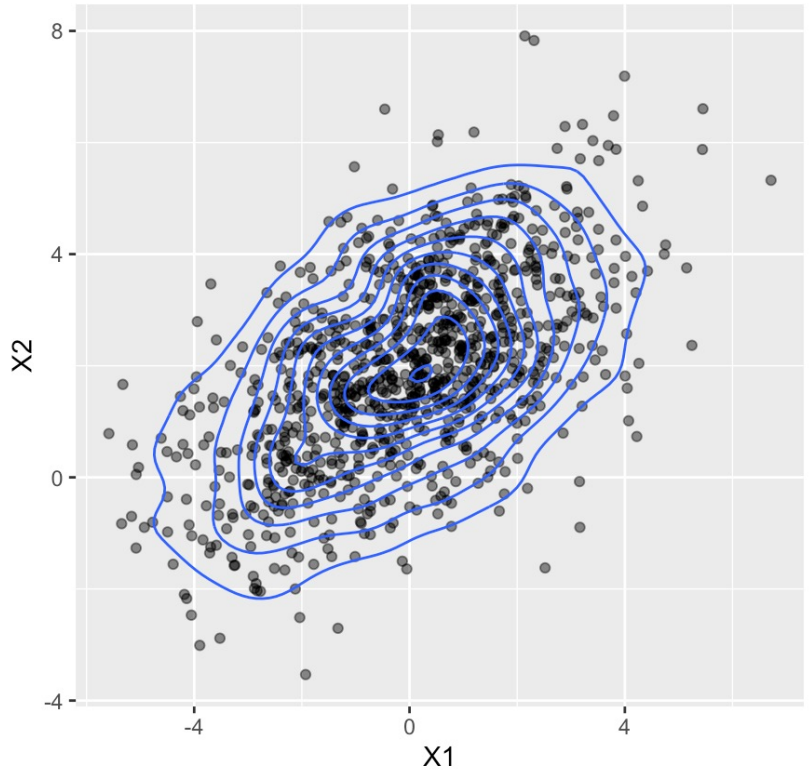
Plus, several new ones:

- Multivariate normality

Multivariate Normal Distribution







Univariate normal distribution has a bell-shaped curve set by the mean and variance

Bivariate normal distribution has most of probability mass concentrated in center with less towards the edges

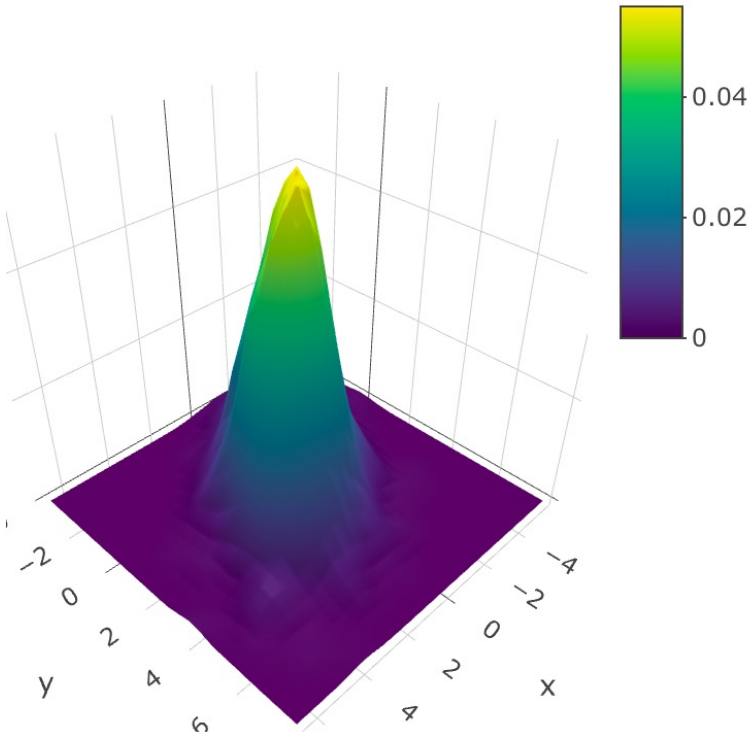
In real data, the distribution is often more elliptical in shape due to correlation between the variables

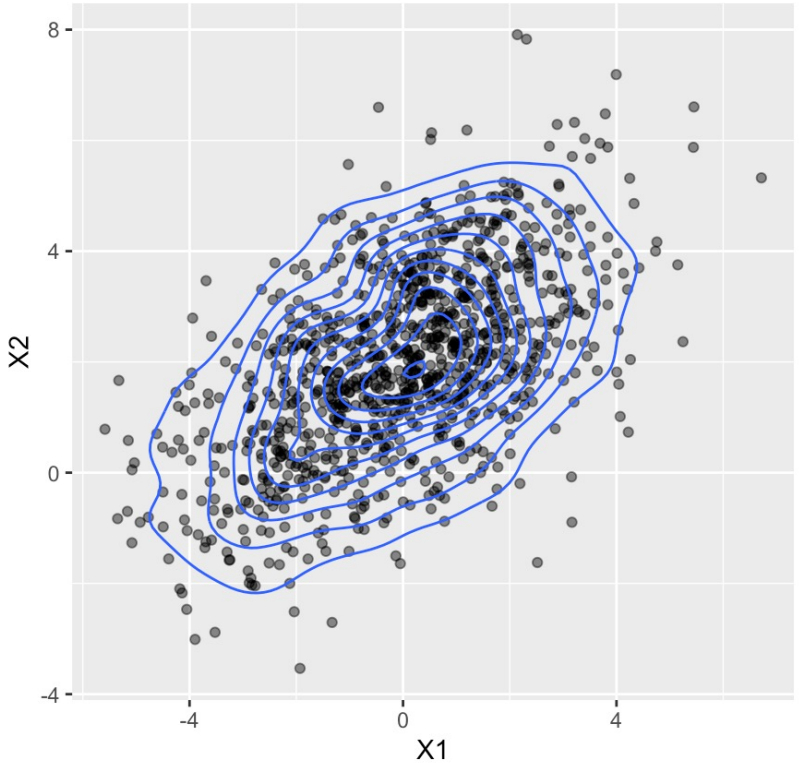
Multivariate Normal Distribution

Testing for multivariate normality in R

mshapiro.test() in package mvnormtest

You can also test each individual variable is normally distributed
BUT it is possible to have every variable be normally distributed and not have a multivariate normal distribution





Bivariate normal distribution has most of probability mass concentrated in center with less towards the edges

In real data, the distribution is often more elliptical in shape due to correlation between the variables

Assumptions of MANOVA

Some carry over from univariate ANOVA:

- Independent observations/samples
- The variance of the different groups is equal and normally distributed (homogeneity of variance)

Plus, several new ones:

- Multivariate normality
- Linearity: dependent variables have linear relationships with each level of the independent variable. If there are more than two dependent variables, the pairs of dependent variables are linearly related
- No multicollinearity between dependent variables (no very high correlations)
- No outliers in the dependent variables
- Homogeneity of covariance: covariances equal among groups
  - This is the multivariate version of homogeneity of variance

Yikes! Most biological data won't meet these assumptions

What can we do instead?

PERMANOVA!

Permutational multivariate analysis of variance

Avoids a lot of the issues associated with MANOVA by using permutation tests

Has many fewer assumptions

What can we do instead?

PERMANOVA!

Permutational multivariate analysis of variance

Avoids a lot of the issues associated with MANOVA by using permutation tests

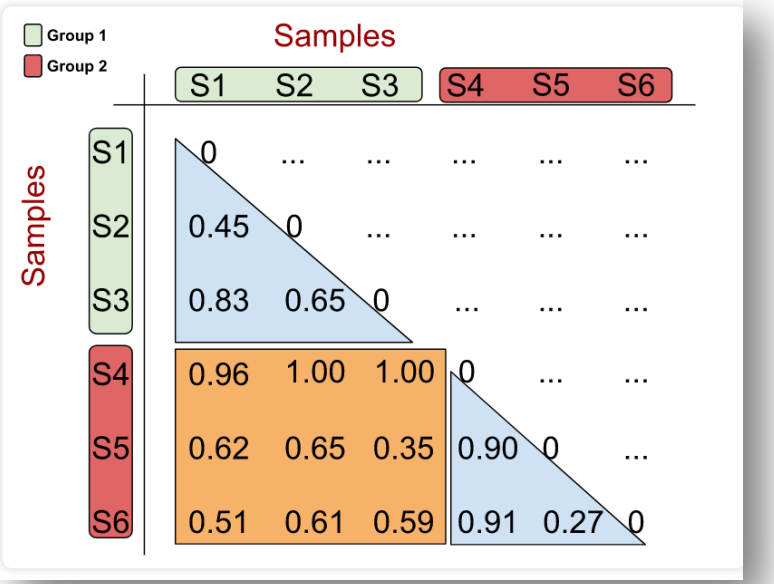
Has many fewer assumptions

It divides the variation in the distance/dissimilarity matrix into:

1. Variation within groups
2. Variation among groups

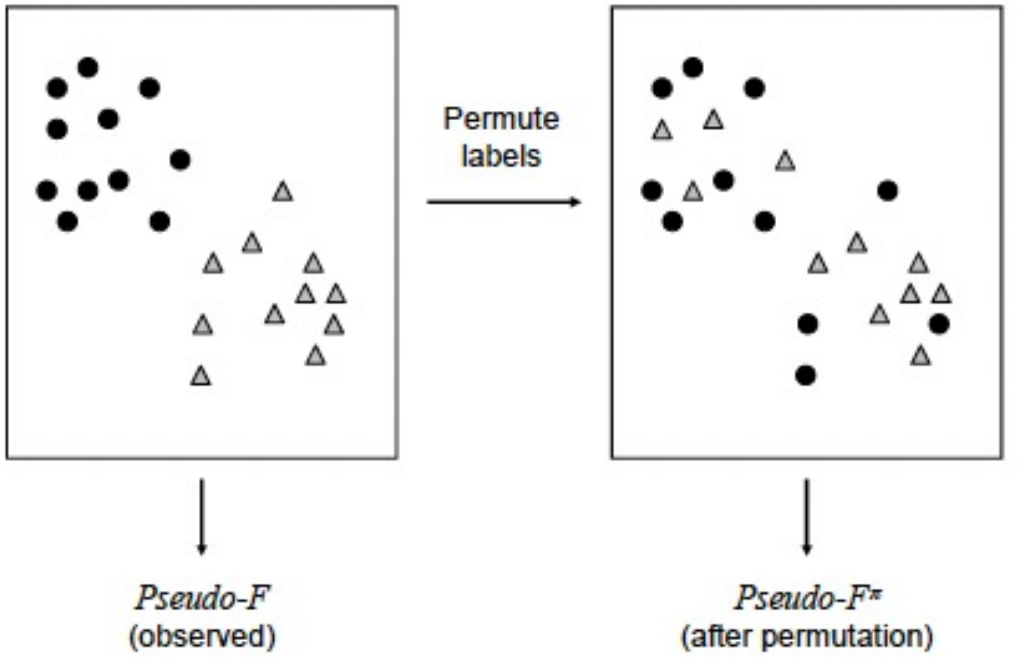
It works with any measure of distance/dissimilarity

Hypothesis testing is based on permutations



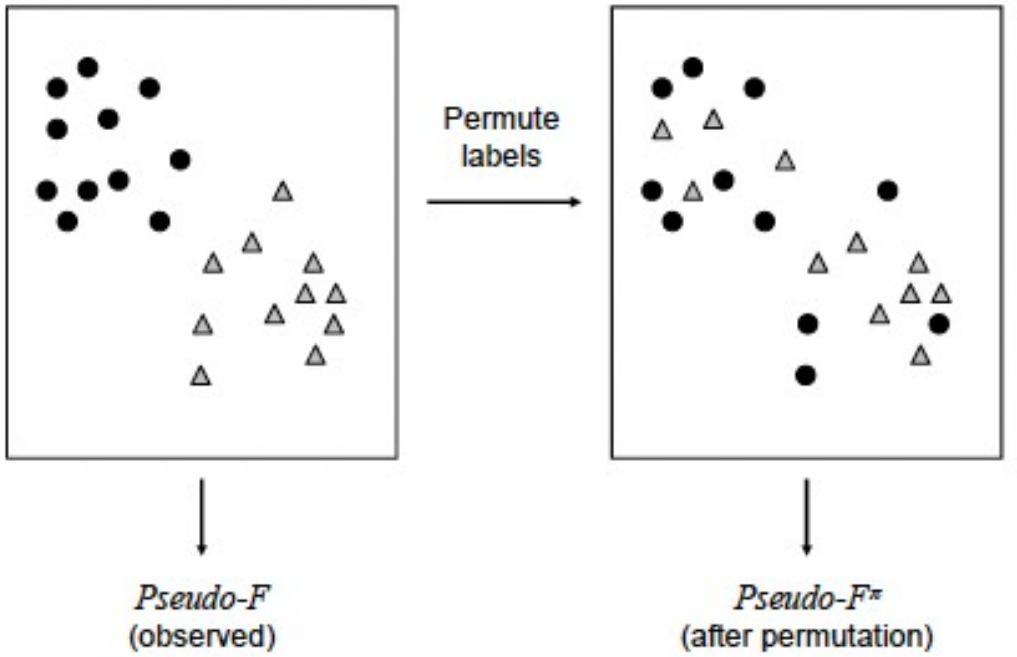
Permutations

- Samples within the raw data are permuted randomly among the groups
  - Each sample is permuted as unit
- The samples themselves have not changed position in space. They have simply been given different labels



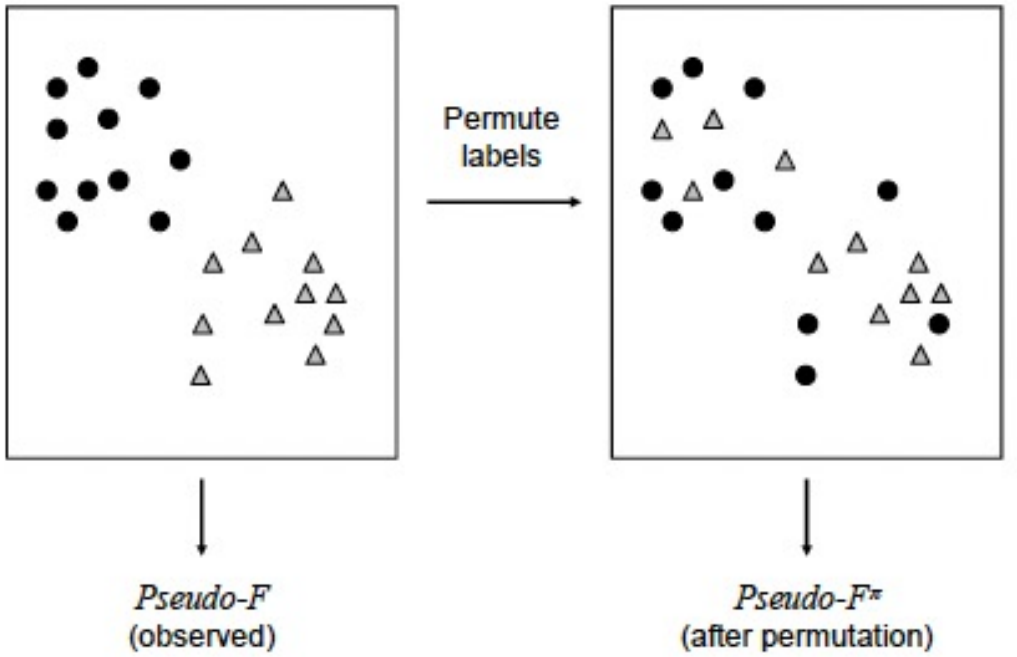
Permutations

- Samples within the raw data are permuted randomly among the groups
  - Each sample is permuted as unit
- The samples themselves have not changed position in space. They have simply been given different labels
- **If null hypothesis is true**: the pseudo-F obtained from the real ordering of data will be similar to the F values obtained using permutations
- **If the alternative hypothesis is true**: the pseudo-F obtained with the real ordering will appear large relative to distribution of values obtained from the permutations



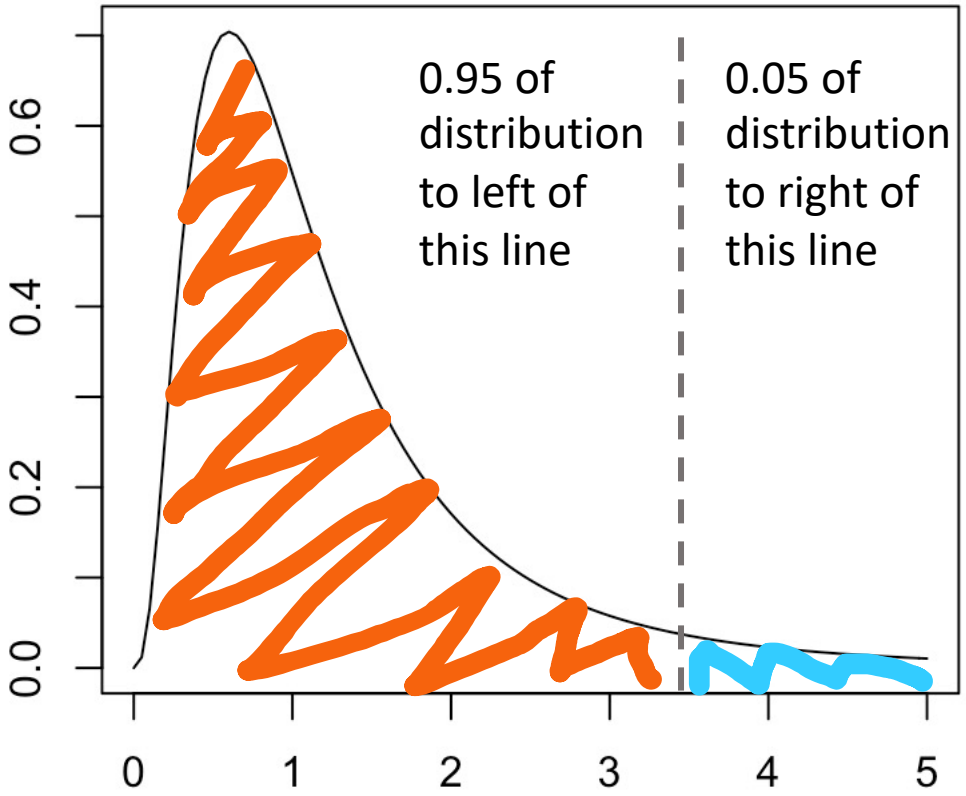
Permutations

- Samples within the raw data are permuted randomly among the groups
  - Each sample is permuted as unit
- The samples themselves have not changed position in space. They have simply been given different labels
- **If null hypothesis is true**: the pseudo-F obtained from the real ordering of data will be similar to the F values obtained using permutations
- **If the alternative hypothesis is true**: the pseudo-F obtained with the real ordering will appear large relative to distribution of values obtained from the permutations
- Because p-value is obtained using permutations, it will change slightly (usually in 3rd decimal place) when you re-run the analysis
- With 1000 permutations, smallest possible p-value is 0.001
- With 100,000 permutations, smallest possible p-value is. 0.00001



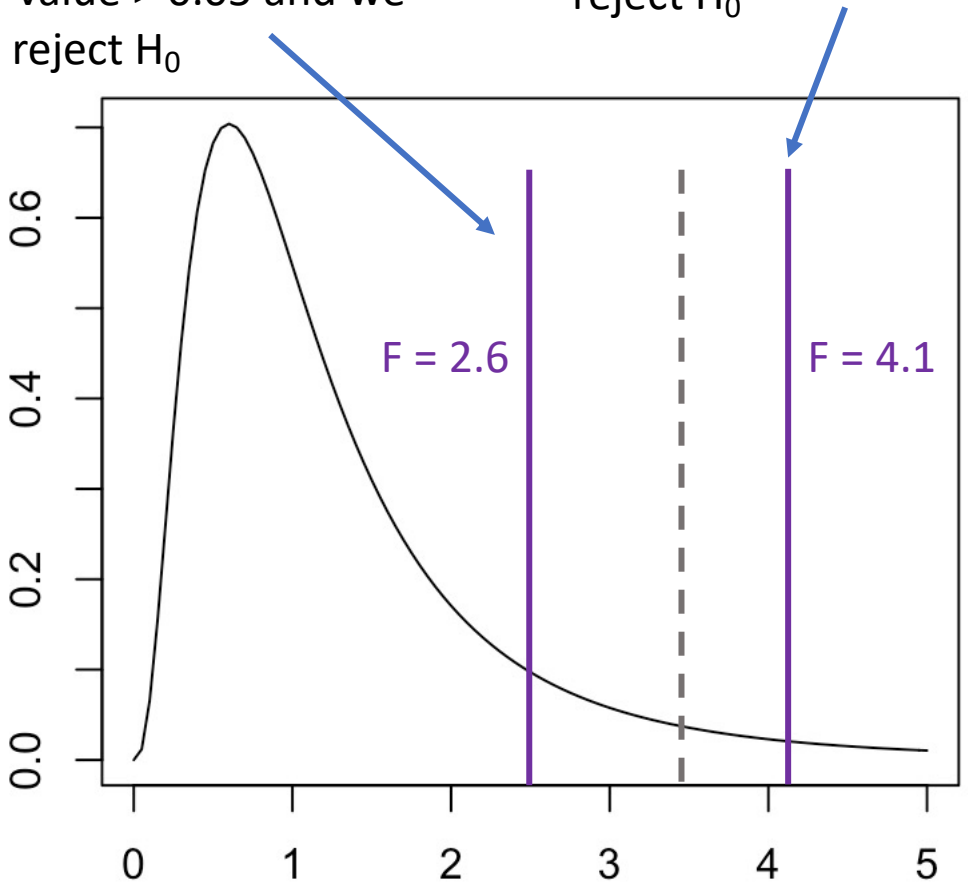
Parametric Hypothesis Testing (M)ANOVA

F-distribution (which is used for ANOVA) has a shape that is determined by the degrees of freedom associated with the data



If the F-ratio in the ANOVA falls to the left of dashed line, the p-value > 0.05 and we fail to reject H_0

If the F-ratio in the ANOVA falls to the right of dashed line, the p-value < 0.05 and we reject H_0



How do we generate the distribution for the test statistic for permutation-based statistics?

We use the data to generate the distribution!

How it works:

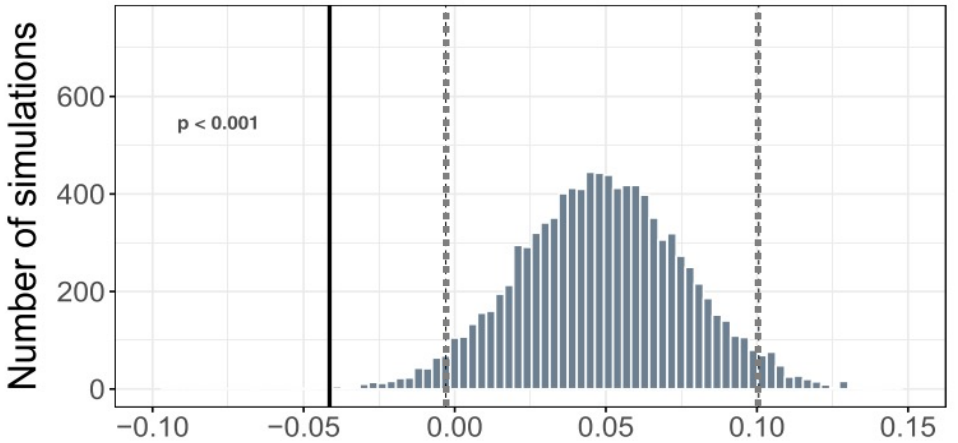
- Randomly shuffle the samples so that they are assigned to different treatment groups
- Perform the analysis to calculate the test statistic
- Repeat many times, recording the new test statistic each time

How do we generate the distribution for the test statistic for permutation-based statistics?

We use the data to generate the distribution!

How it works:

- Randomly shuffle the samples so that they are assigned to different treatment groups
- Perform the analysis to calculate the test statistic
- Repeat many times, recording the new test statistic each time
- Use the test statistics from the randomizations to create a null distribution
- Calculate the test statistic for the actual, observed data



The observed value (solid line) falls outside 95% (dashed lines), so we reject H0

If the observed was inside the dashed lines, we would fail to reject H0

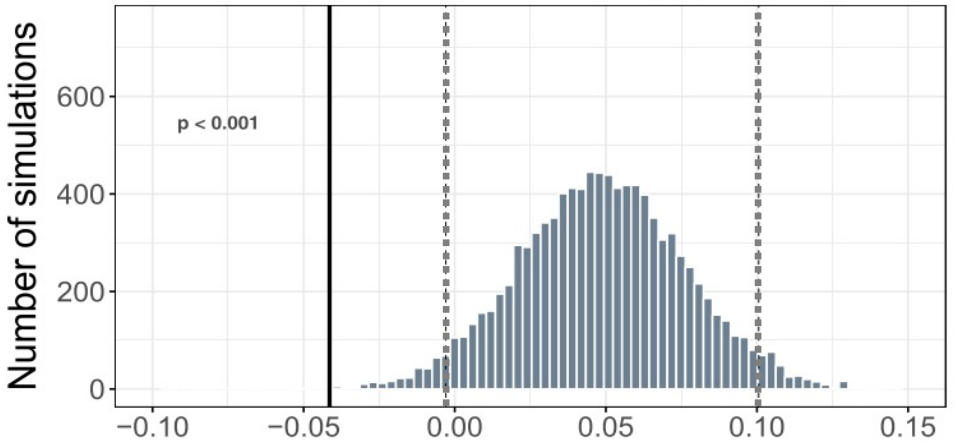
This example is a two-tailed test

How do we generate the distribution for the test statistic for permutation-based statistics?

We use the data to generate the distribution!

How it works:

- Randomly shuffle the samples so that they are assigned to different treatment groups
- Perform the analysis to calculate the test statistic
- Repeat many times, recording the new test statistic each time
- Use the test statistics from the randomizations to create a null distribution
- Calculate the test statistic for the actual, observed data
- Compare this to the null distribution generated using randomizations
- If value is extreme relative to null distribution (falls outside 0.95 of distribution) -> reject the null hypothesis



The observed value (solid line) falls outside 95% (dashed lines), so we reject H0

If the observed was inside the dashed lines, we would fail to reject H0

This example is a two-tailed test

Pseudo F-Statistic

$$\text{Pseudo F} = \frac{\text{SS}_\text{A}/(\text{a} - 1)}{\text{SS}_\text{Res}/(\text{N} - \text{a})}$$

a = number of levels within the factor A

N = number of samples

$(\text{a} - 1)$ is the degrees of freedom associated with factor A

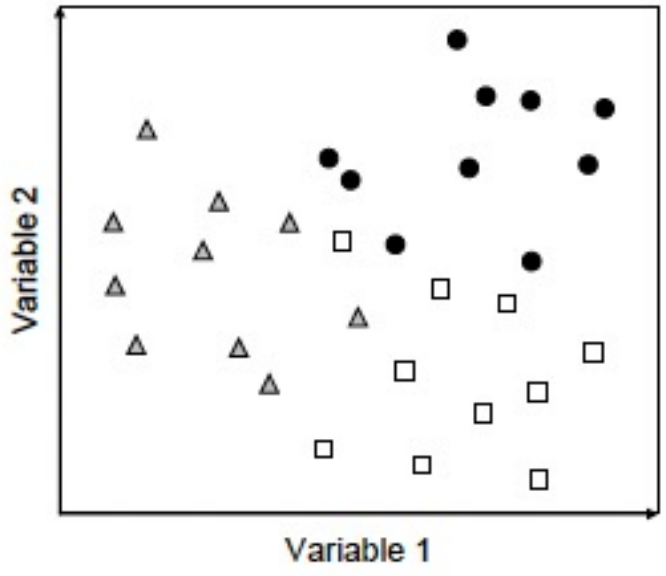
$(\text{N} - \text{a})$ is the residual degrees of freedom

As the value of Pseudo-F gets larger, the likelihood of null being true decreases

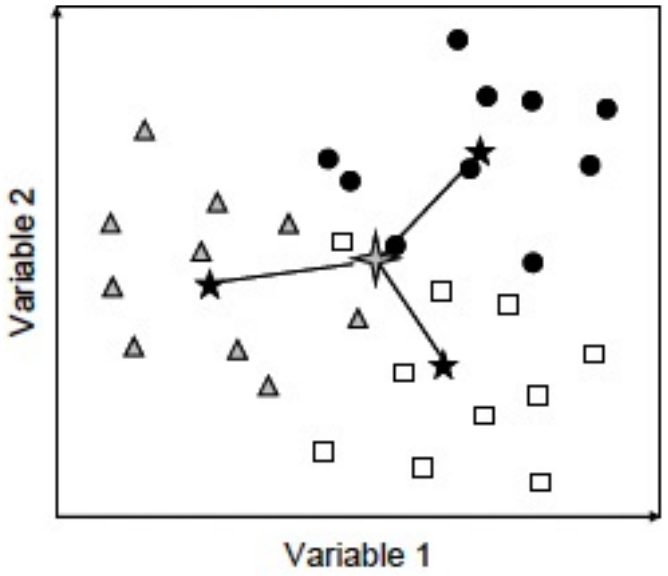
It is called “Pseudo” F because it does not have a known distribution under a null hypothesis. We use permutations of the data to generate the null distribution

Partitioning the Variance in PERMANOVA

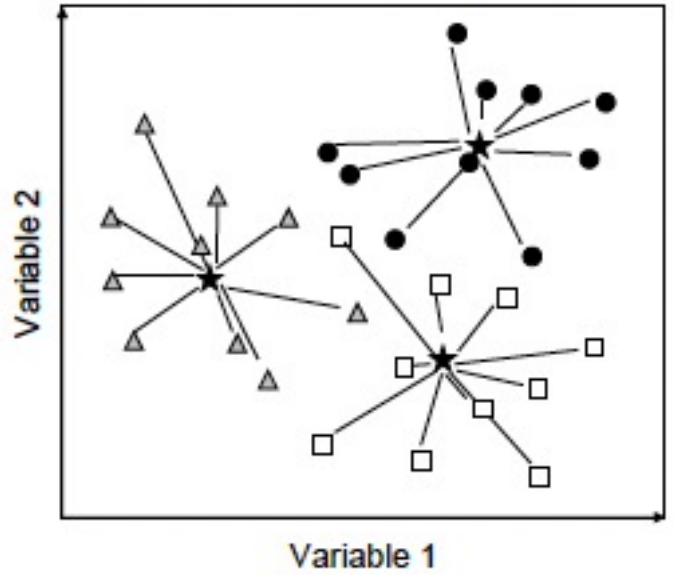
The partitioning done in multivariate space in PERMANOVA is analogous to partitioning in univariate ANOVA



Distance from group centroid to the overall centroid (SS_A)



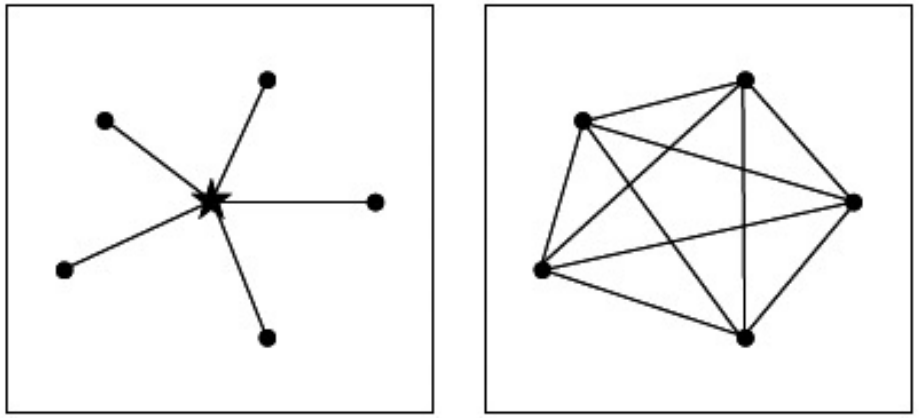
Distance from samples to group centroids (SS_{Res})



SS_A = sum of squared distances from group centroids to overall centroid
SS_{Res} = sum of squared distances from samples to their one group centroid
SS_{Total} = sum of squared distance from samples to overall centroid

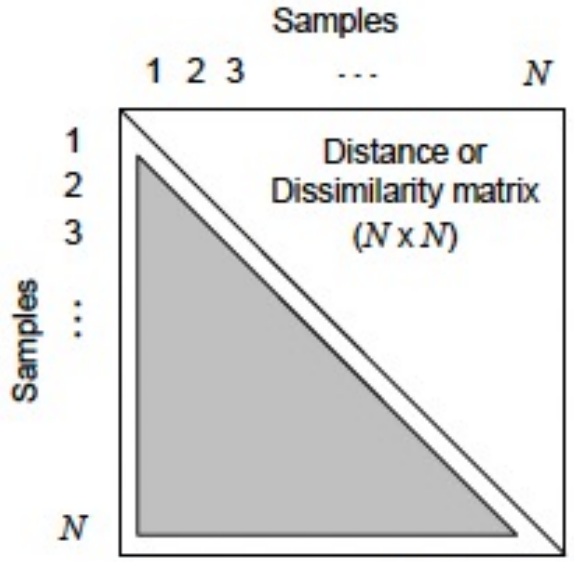
For Non-Euclidean Distance

We can't easily calculate the distance to the group centroid when using non-Euclidean distance
Instead, we use the inter-point distances to do the partitioning



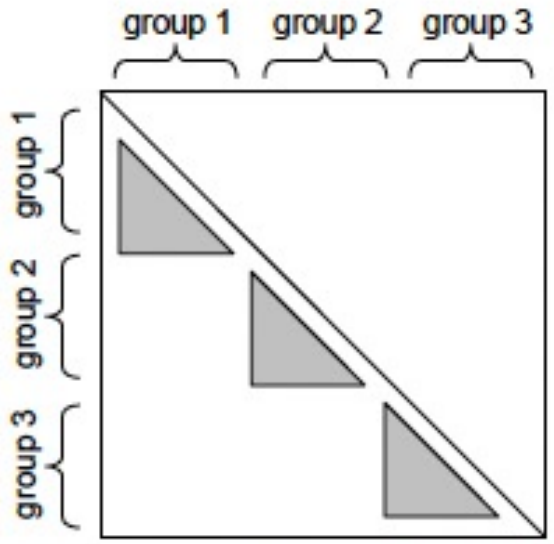
sum of squared distances from points to their group centroid = $\frac{\text{(sum of squared inter-point distances)}}{\text{(number of points)}}$

SS Total



SS_T = sum of squared dissimilarities in the sub-diagonal, divided by N.

SS Residuals



SS_{Res} = sum of squared dissimilarities within each group, divided by n.

Use the distance/dissimilarity matrix to calculate the Sum of Squares

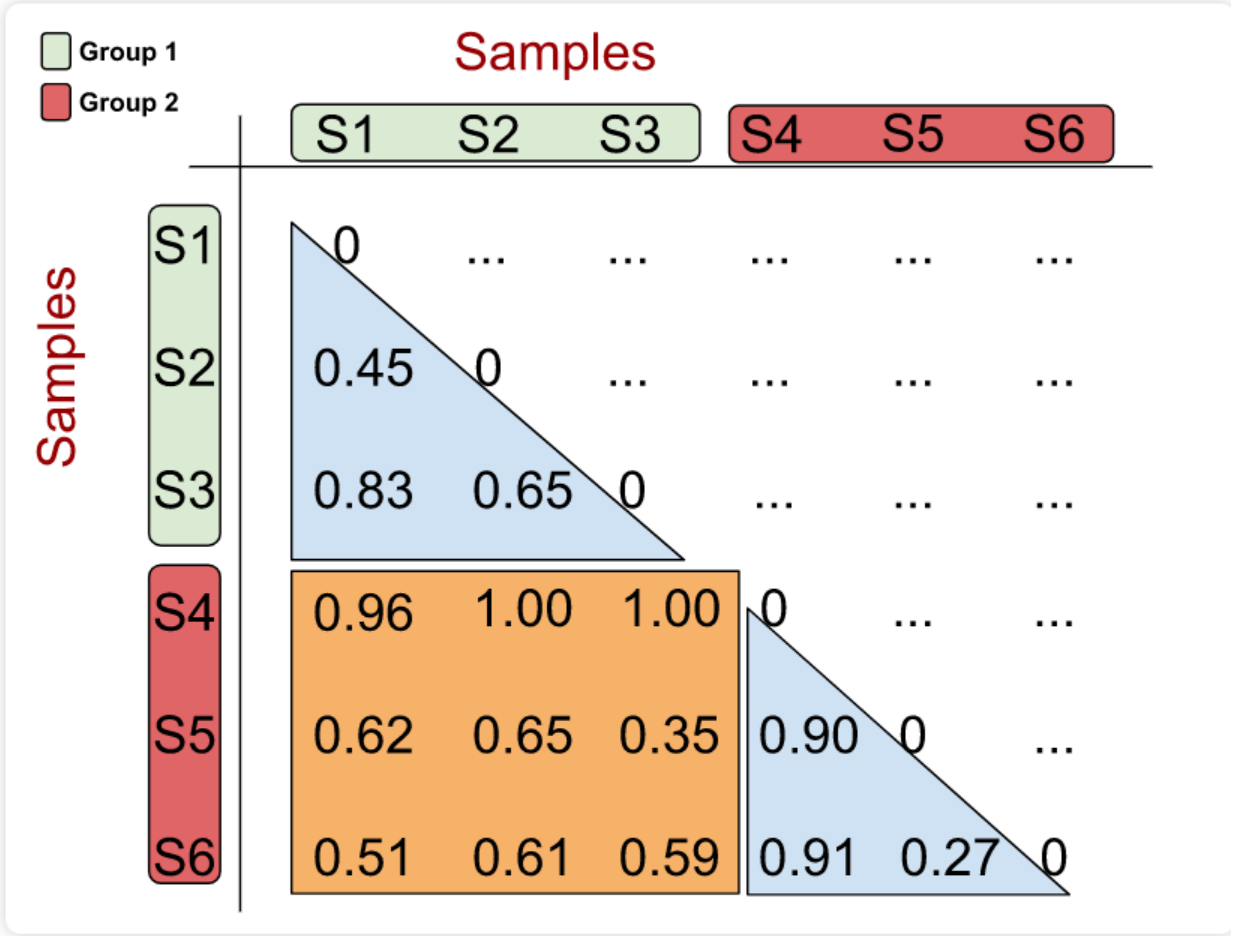


Figure 1: A grouped Bray-Curtis dissimilarity matrix. Note that the matrix is symmetrical about its diagonal. NPMANOVA will compare the within-group dissimilarities (blue triangles) to the between group dissimilarities (orange square) through a pseudo *F*-ratio (**Equation 1**)

$$F = \frac{SS_A \div (a - 1)}{SS_W \div (N - a)}$$

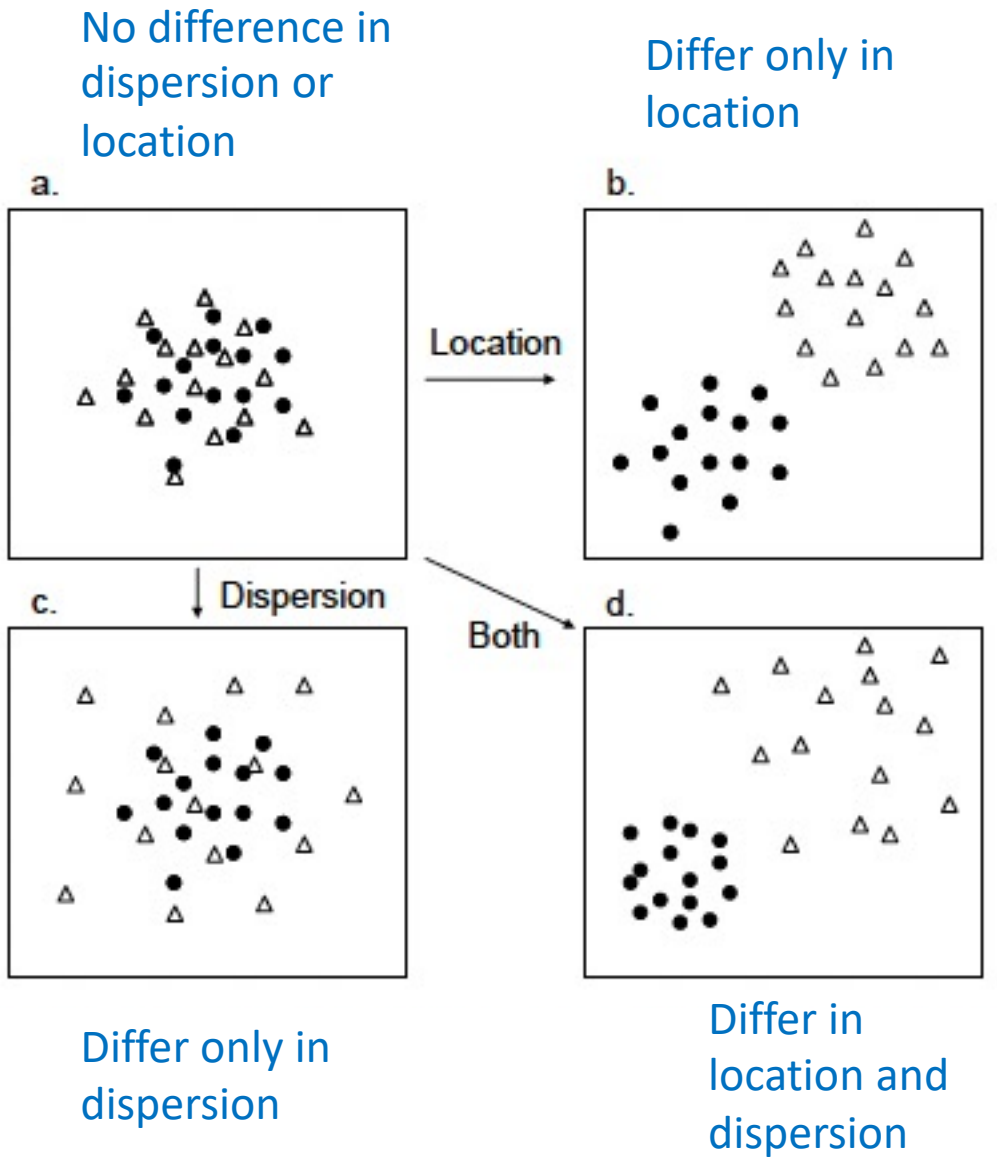
Equation 1: The *F*-ratio used in standard NPMANOVA is similar to the traditional *F*-ratio used in ANOVA, however, does not share the same distribution. SS_W is the sum of squared dissimilarities within groups, SS_A is the sum of squared dissimilarities among (between) groups, *a* is the number of groups, and *N* is the total number of objects. The terms (*a*-1) and (*N*-*a*) are the degrees of freedom associated with the explanatory factor (the grouping variable) and the residuals. See Anderson (2001) for discussion and formulae for SS_W and SS_A for simple and more complex designs.

https://archetypalecology.wordpress.com/2018/02/21/permutational-multivariate-analysis-of-variance-permanova-in-r-preliminary/

Homogeneity of Dispersion

Warning!

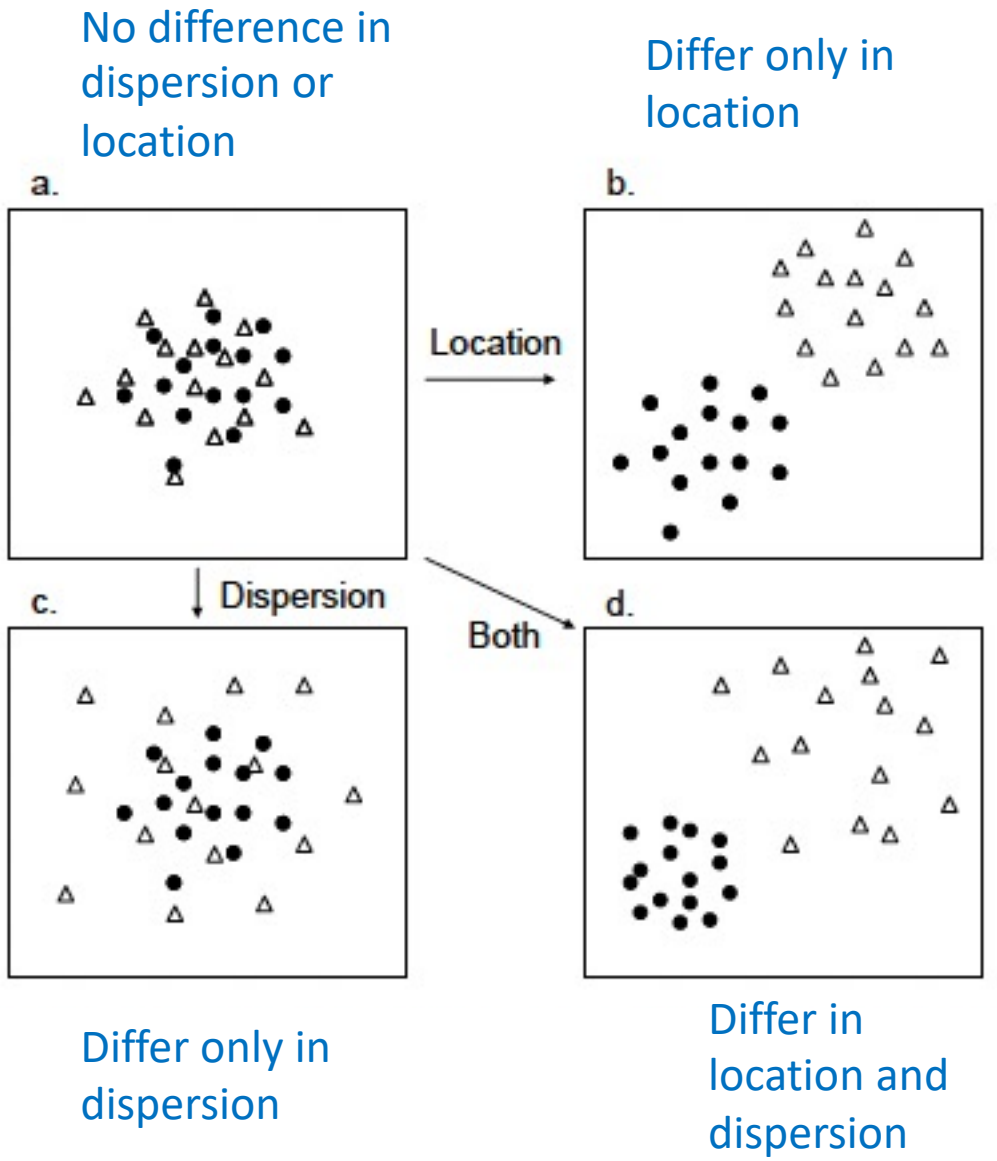
- PERMANOVA is sensitive to differences in group dispersion



Homogeneity of Dispersion

Warning!

- PERMANOVA is sensitive to differences in group dispersion
- We need to test this assumption for each factor in the model
- Use betadisper() in vegan package in R
- If the groups have different dispersions, a significant p-value in PERMANOVA could be due to differences in location, differences in dispersion, or both
  - Dispersion and location could be confounded



Using Ordination Plots and PERMANOVA Together

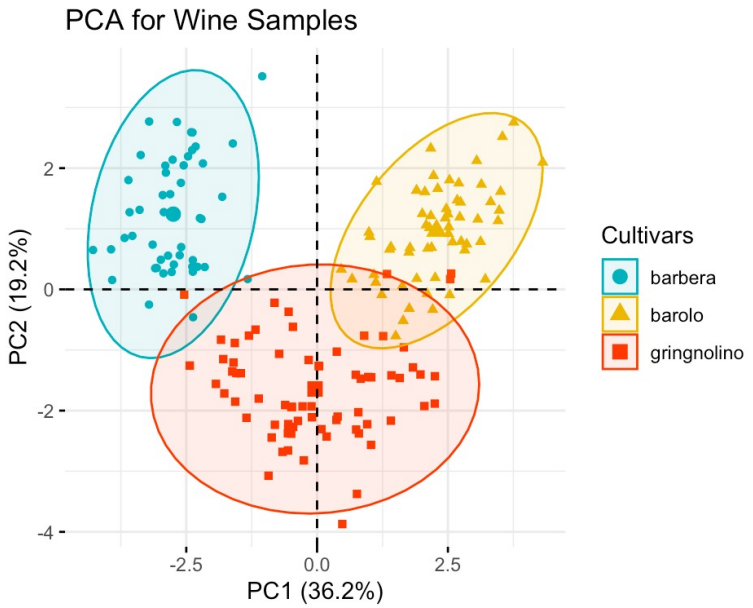
Usually, the ordination plot helps with the interpretation of the PERMANOVA and vice versa However, this is not always the case…

It is important to remember that the dimensionality of the data has been reduced to 2 or 3 dimensions when creating the ordination plot. This means some real patterns may not be obvious in the plot.

PERMANOVA works on the underlying distance/dissimilarity matrix. Its results should be trusted over any patterns (or lack of patterns) that are apparent in the ordination.

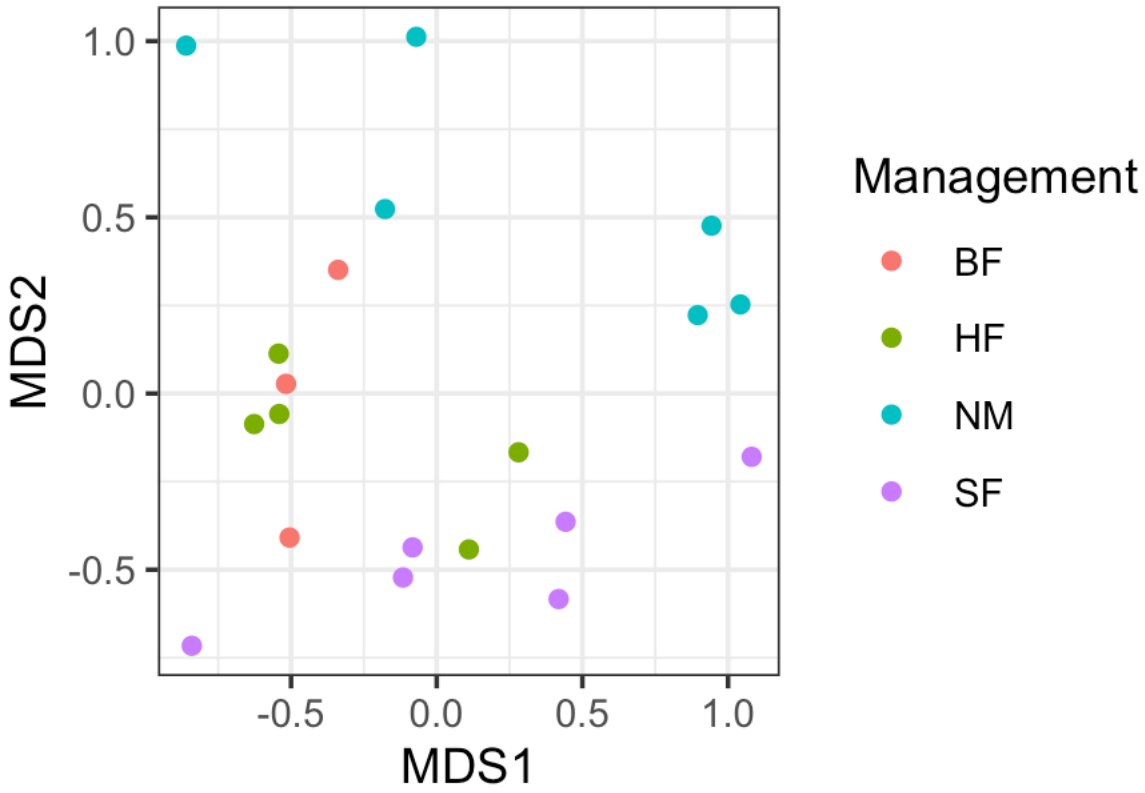
	Df	SumOfSqs	R2	F	Pr(>F)	
Cultivar	2	12359664	0.70256	206.68	0.000999	***
Residual	175	5232632	0.29744			
Total	177	17592296	1.00000			

PERMANOVA results for the Wine data set confirm that the Cultivars are different, a pattern that is very clear in the ordination plot



Today in R:

”dune” and “dune.env” dataset in *vegan* package




```
134  adonis2(dune ~ Use+Management, data = dune
     .env, permutations = 1000, method="bray",
     by="margin")
```

```
            Df SumOfSqs      R2      F   Pr(>F)
Use          2   0.3757 0.08740 1.0715 0.418581
Management   3   1.2912 0.30034 2.4547 0.007992 **
Residual    14   2.4547 0.57099
Total       19   4.2990 1.00000
```

```
143  (pairwise.management<-pairwise.adonis(dune.bc,
     dune.env[,"Management"], perm = 1000, sim.method
     = "bray"))
```

```
     pairs Df SumsOfSqs  F.Model        R2     p.value p.adjusted sig
1 SF vs BF  1 0.4016624 2.514890 0.2643110 0.058941059 0.35364635
2 SF vs HF  1 0.2828804 1.857489 0.1710790 0.120879121 0.72527473
3 SF vs NM  1 0.7575728 3.425694 0.2551595 0.002997003 0.01798202   .
4 BF vs HF  1 0.1617135 1.567531 0.2071390 0.205794206 1.00000000
5 BF vs NM  1 0.5662456 2.715242 0.2794827 0.023976024 0.14385614
6 HF vs NM  1 0.6513088 3.423068 0.2755413 0.027972028 0.16783217
```