Module 14:
Statistics comparing one dependent and one independent variable

Module 14: Learning Outcomes

- Define key terms in statistics that will help you differentiate between certain statistical test
- Differentiate between a parametric and non-parametric test, and identify the situation in which you would use one or the other
- Be able to use a t-test and Wilcoxon rank sum test to compare two means
- Be able to use ANOVA and Kruskall-wallis tests to compare multiple means
- Be able to use Pearson correlation and Kendall correlation tests to correlate two continuous variables

How many variables are we comparing?

- **Single variable** (Module 14): comparing one response to one predictor
  - Eg. antibiotic use and Shannon measures
- **Multivariable** (Module 15): comparing multiple responses and predictors
  - Eg. antibiotic use and body site to Shannon and Observed features

Defining our variable

- Independent variable ("**predictor**")= "x axis", differs by treatment

"reported.antibiotic.usage"

"Yes" "No"

- Dependent variable ("**response**") = "y axis", responds to treatment

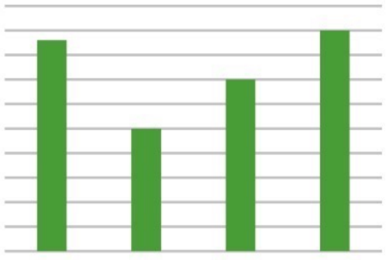
Shannon diversity measure

Species	Frequency	p_i	*ln* (p_i)	p_i * *ln* (p_i)
A	40	0.38	-0.97	-0.37
B	20	0.19	-1.66	-0.32
C	15	0.14	-1.95	-0.28
D	8	0.08	-2.57	-0.20
E	22	0.21	-1.56	-0.33
			H	**1.49**

Types of Predictors

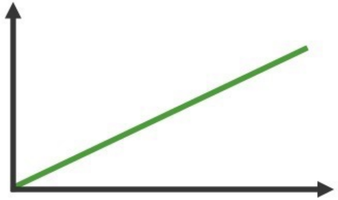
- **Categorical**: discrete variables

- **Continuous**: increasing/decreasing numerical values



Categorical
(also called "discrete")

Age group, sex, number of siblings, citizenship, race...



Continuous

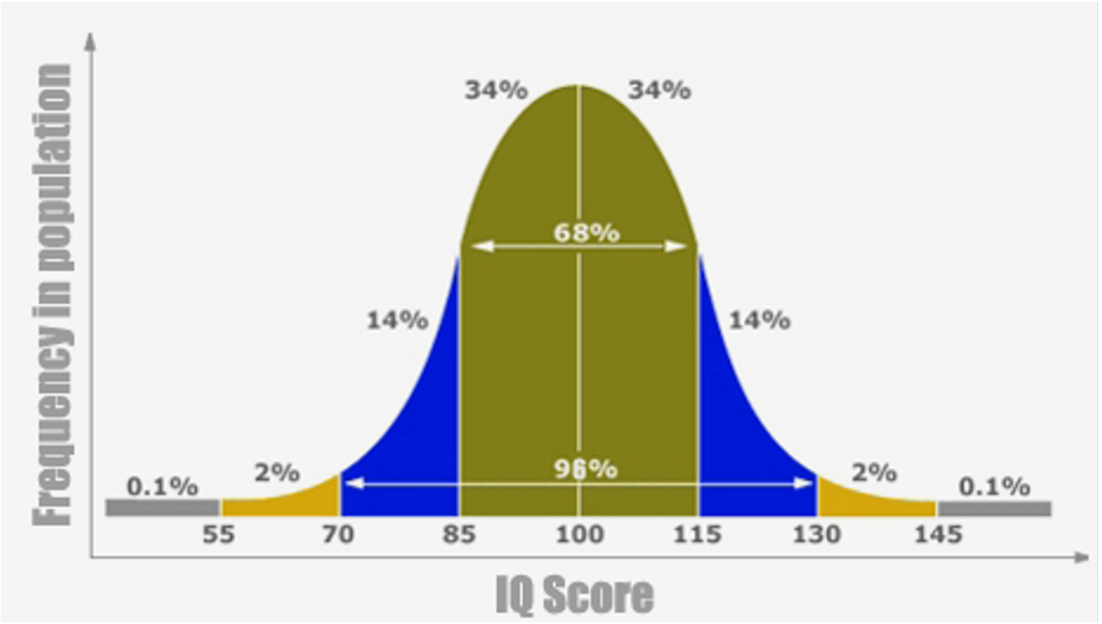
Age, height, distance, temperature...

Types of Responses

Parametric distribution: follows a normal distribution

Non-parametric distribution: does not follow a normal distribution

What is a normal distribution?

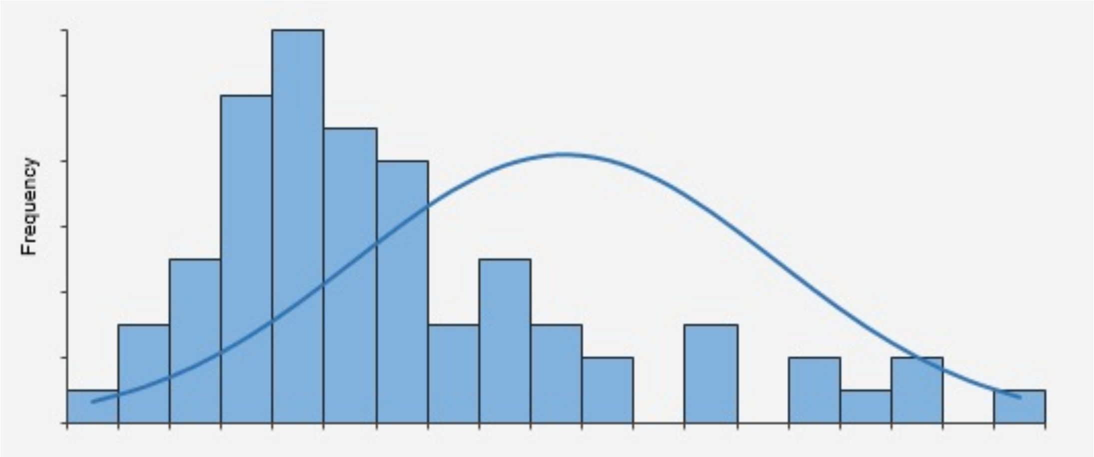


“Distribution” is the spread of data points across a set of variables

Many distributions in nature are “Normal” (Gaussian)
You can infer things from normally distributed data because there is an expected spread of values, given the **parameters**, mean and standard deviation.

However, not all data is “normally distributed”

What is a non-normal distribution?



Data can be skewed or evenly distributed

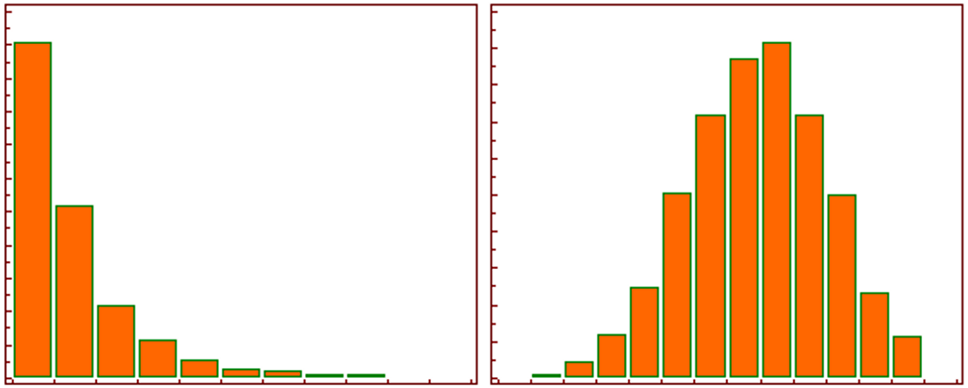
For example, microbiome data is generally right-skewed: there are lots of rare things, and very few abundant things

This has statistical implications, which we will see in our R code

To fix non-normal data you can:

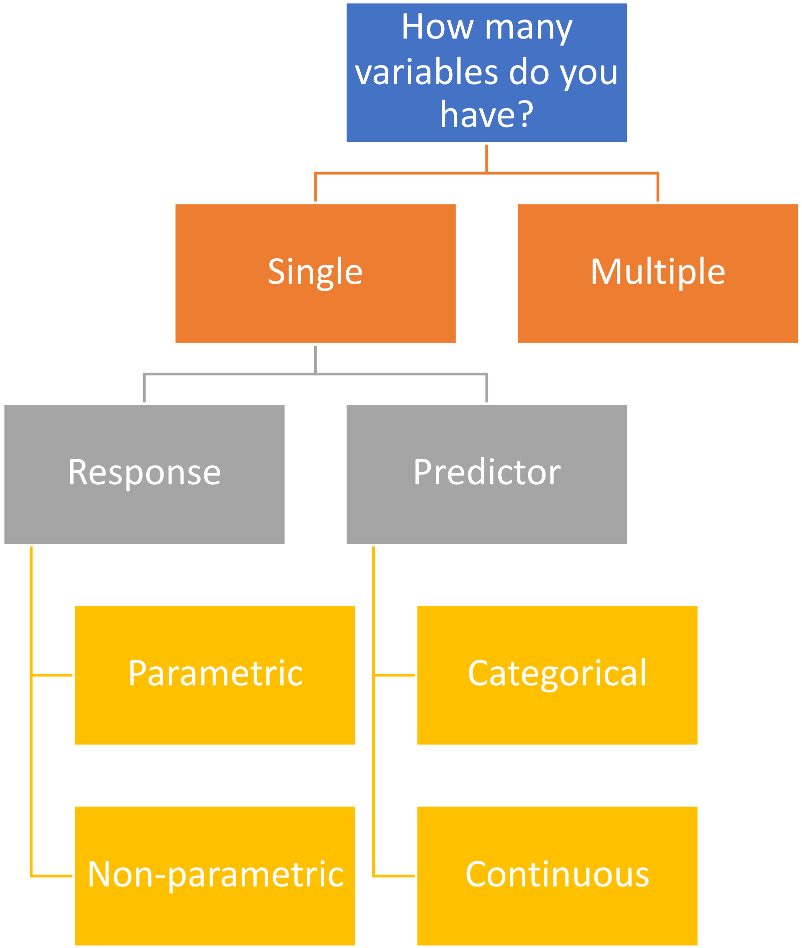
Transform data (only works with some data)

- e.g. log-transform



Use a non-parametric test

Decision tree then follows…



Statistics in R

- Plenty of basic functions to calculate statistics
- In this module, we will quickly review general statistical concepts and learn how to apply them

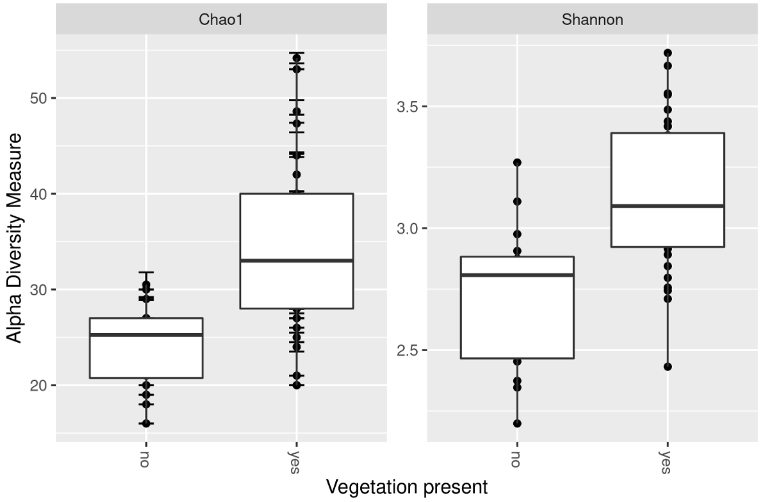
- statistical_test(x, y) or statistical_test(y ~ x)

Types of tests for one independent and one dependent variable

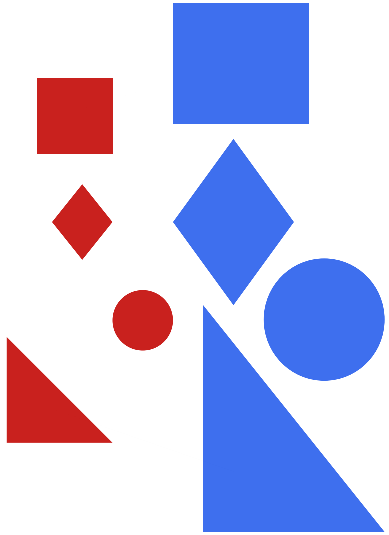
	CATEGORICAL independent variable (predictor)	CONTINUOUS independent variable (predictor)
Continuous dependent variable (response)	T-test (parametric) ANOVA (parametric, 2+ groups) Wilcoxon/Mann-Whitney test (non-parametric) Kruskall-wallis test (non-parametric, 2+groups)	Pearson's product-moment correlation (parametric) Spearman's rank correlation (non-parametric)

T-test

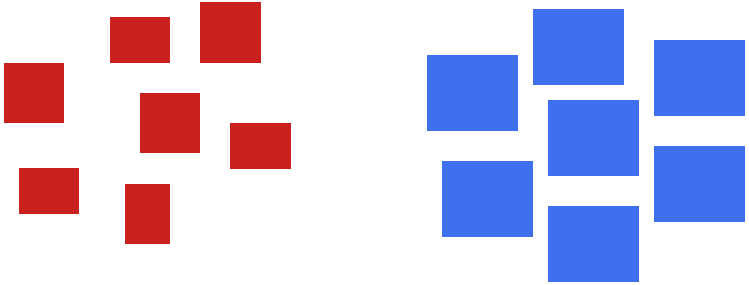
- Compares two means
- Typically visualized by boxplots
- Can be paired or unpaired



Paired



Unpaired



T-test

t.test(Y ~ X, data = dat)

t.test(dat$Y ~ dat$X)

- dat$X = c('A', 'A', 'B', 'B', 'B', 'A')
- dat$Y = c(12, 13, 40, 51, 43, 10)

- t.test(A_vec, B_vec)
    - A_vec = c(12,13,10)
    - B_vec = c(40,51,43)

Types of tests for one independent and one dependent variable

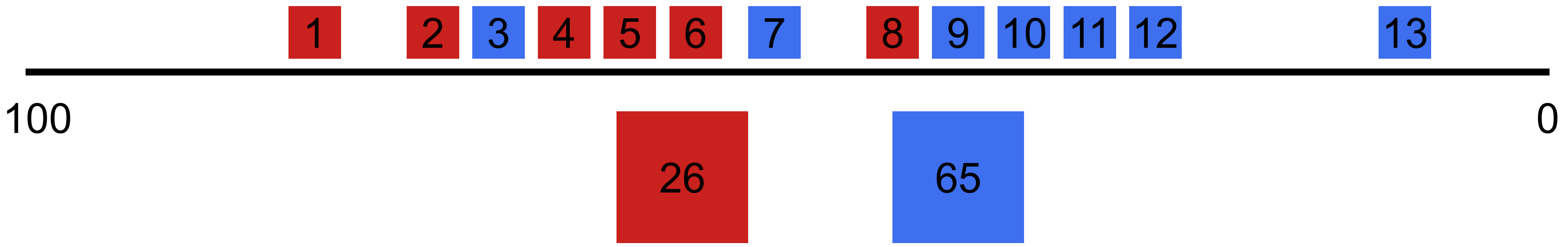
	CATEGORICAL independent variable	CONTINUOUS independent variable
Continuous dependent variable	T-test (parametric) ANOVA (parametric, 2+ groups) Wilcoxon/Mann-Whitney test (non-parametric) Kruskall-wallis test (non-parametric, 2+groups)	Pearson's product-moment correlation (parametric) Spearman's rank correlation (non-parametric)

T-test assumptions

- The data are continuous.
- The sample data have been randomly sampled from a population.
- There is homogeneity of variance (i.e., the variability of the data in each group is similar).
- **The distribution is approximately normal.**
  - **The t-test is a PARAMETRIC test**
  - **PARA = parameter-based; mean and standard deviation**

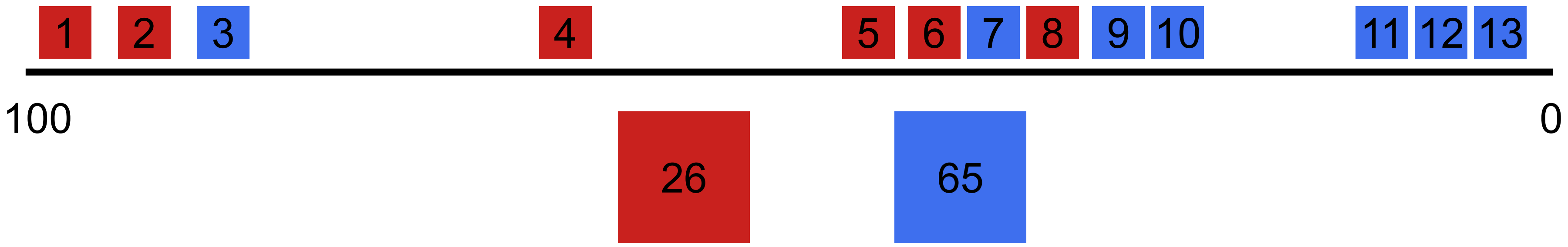
Wilcoxon rank test (or Mann-Whitney test)

- Compares two groups to see which is larger
- Alternative to t-test: uses (sums of) ranks rather than actual values to determine whether one group is “higher” than the other



Wilcoxon rank test (or Mann-Whitney test)

- Compares two groups to see which is larger
- Alternative to t-test: uses (sums of) ranks rather than actual values to determine whether one group is “higher” than the other



Wilcoxon rank test (or Mann-Whitney test)

```
wilcox.test( Y ~ X, data= dat )

wilcox.test( dat$Y ~ dat$X )


wilcox.test( A_vec, B_vec)
```

Types of tests for one independent and one dependent variable

	CATEGORICAL independent variable	CONTINUOUS independent variable
Continuous dependent variable	T-test (parametric) ANOVA (parametric, 2+ groups) Wilcoxon/Mann-Whitney test (non-parametric) Kruskall-wallis test (non-parametric, 2+groups)	Pearson's product-moment correlation (parametric) Spearman's rank correlation (non-parametric)

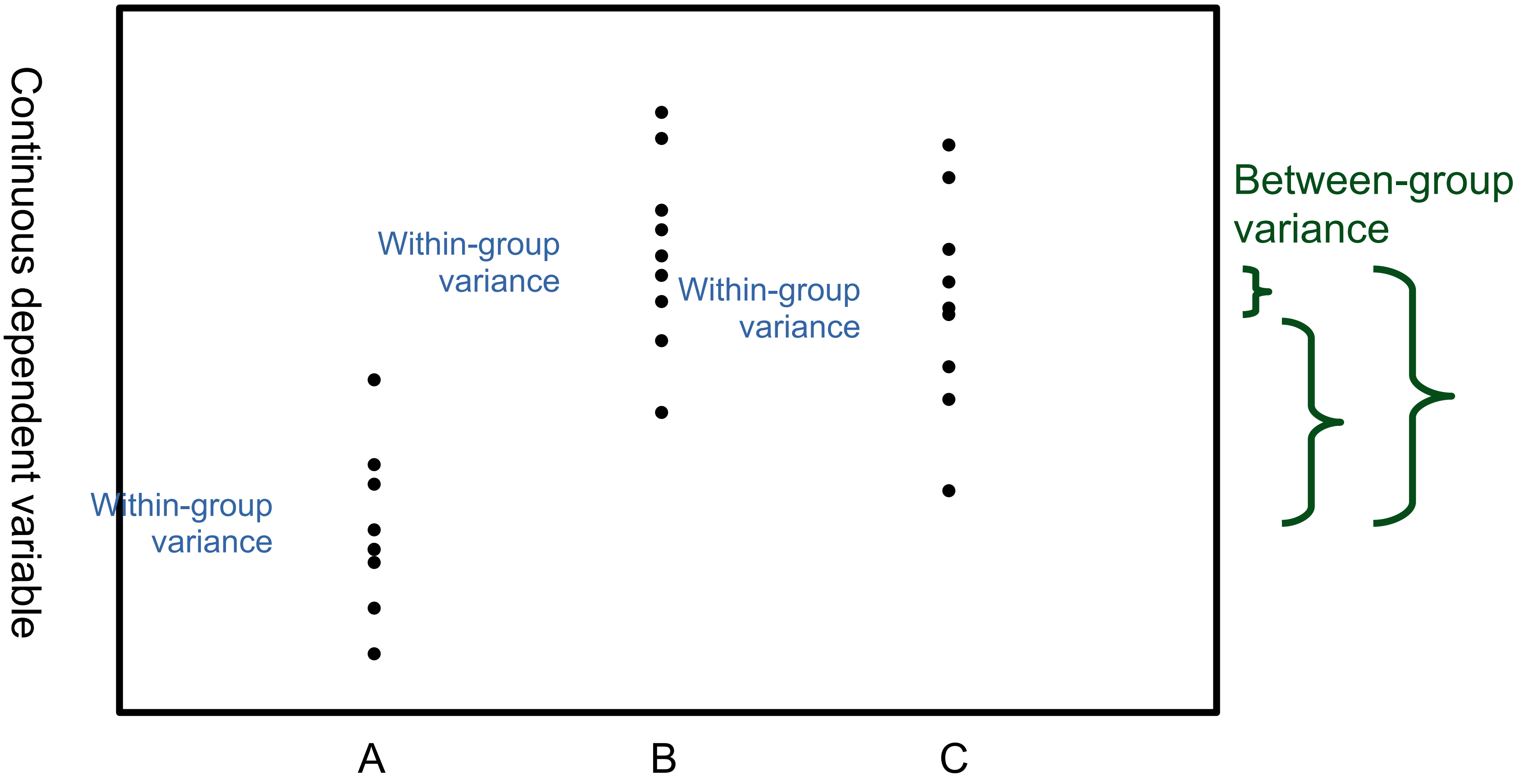
ANOVA

- Compares means across 2+ groups (in one variable)
- Mathematically identical to a t-test when there are only 2 groups

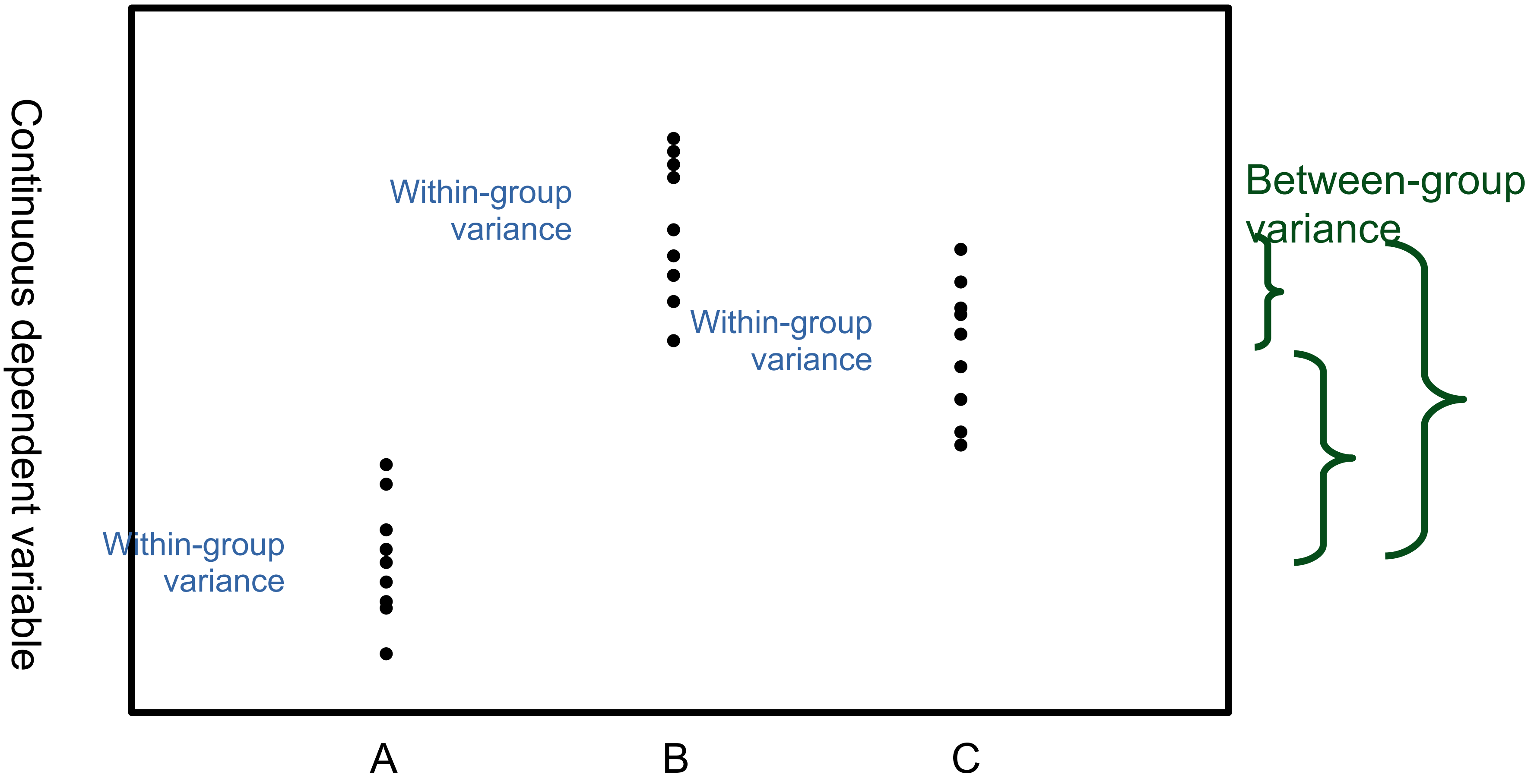
- * Tells you whether ANY group is different; then you must follow-up with Tukey post-hoc test to see WHICH group(s) are different

ANOVAs look at whether within-group variance is smaller than between group variance

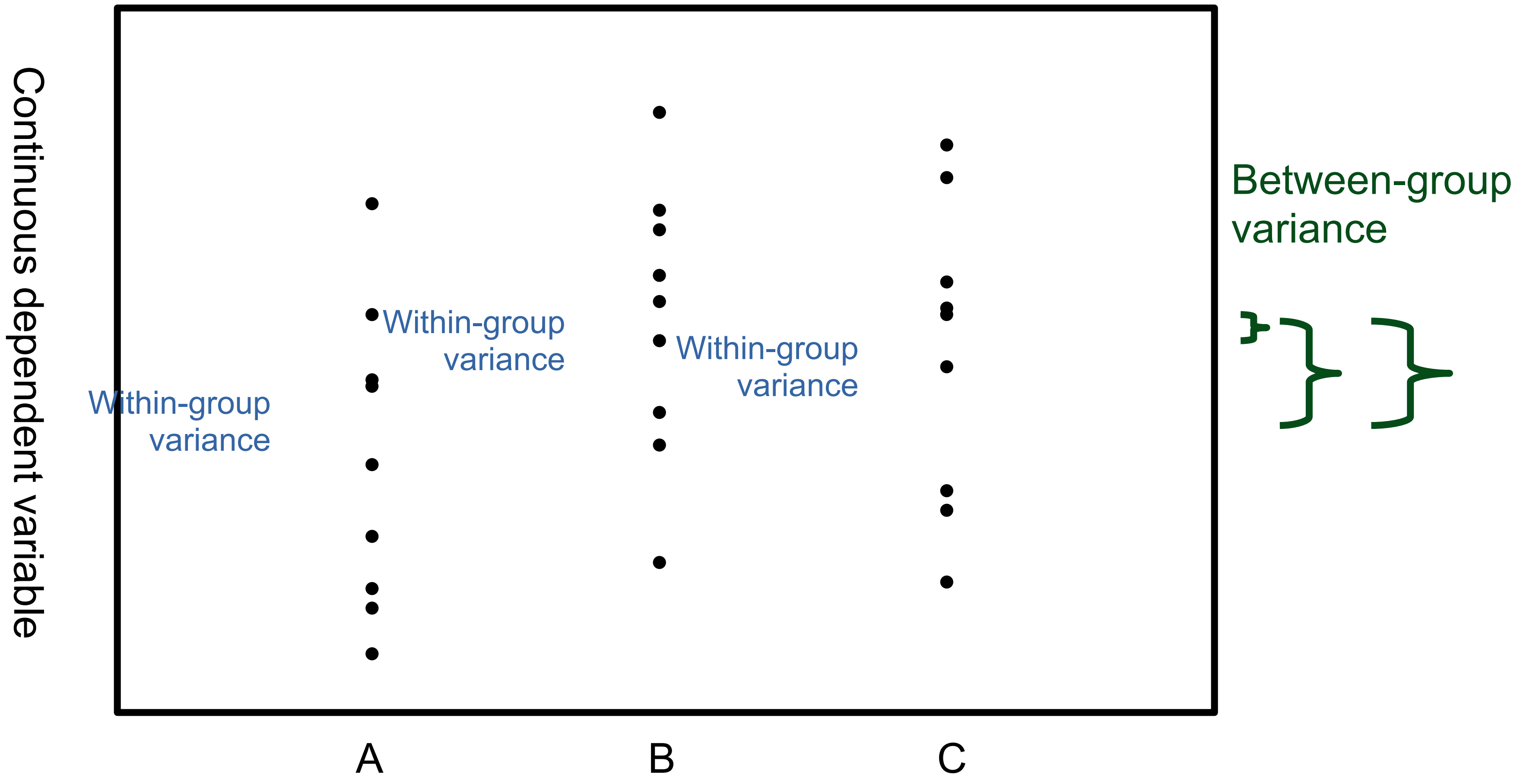
ANOVA



ANOVA


Continuous dependent variable
Within-group variance
Within-group variance
Within-group variance
Between-group variance
A
B
C

ANOVA



Continuous dependent variable
Within-group variance
Within-group variance
Within-group variance
Between-group variance
A
B
C

ANOVA

(1) Set up model

- Model <- lm(Y ~ X, data = dat)

(2) Calculate ANOVA and summarise

- model_aov <- aov(Model)
- summary(model_aov)

(3) Tukey Honest Significant Differences test to identify different groups

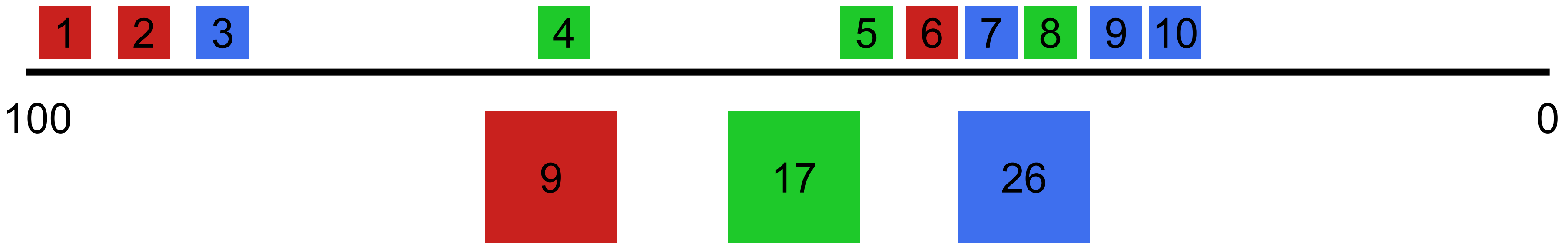
- TukeyHSD(model_aov)

Types of tests for one independent and one dependent variable

	CATEGORICAL independent variable	CONTINUOUS independent variable
Continuous dependent variable	T-test (parametric) ANOVA (parametric, 2+ groups) Wilcoxon/Mann-Whitney test (non-parametric) Kruskall-wallis test (non-parametric, 2+groups)	Pearson's product-moment correlation (parametric) Spearman's rank correlation (non-parametric)

Kruskall-Wallis test

- Compares means across 2+ groups
- Non-parametric alternative to ANOVA; similar to Wilcoxon test, where it sums ranks by group

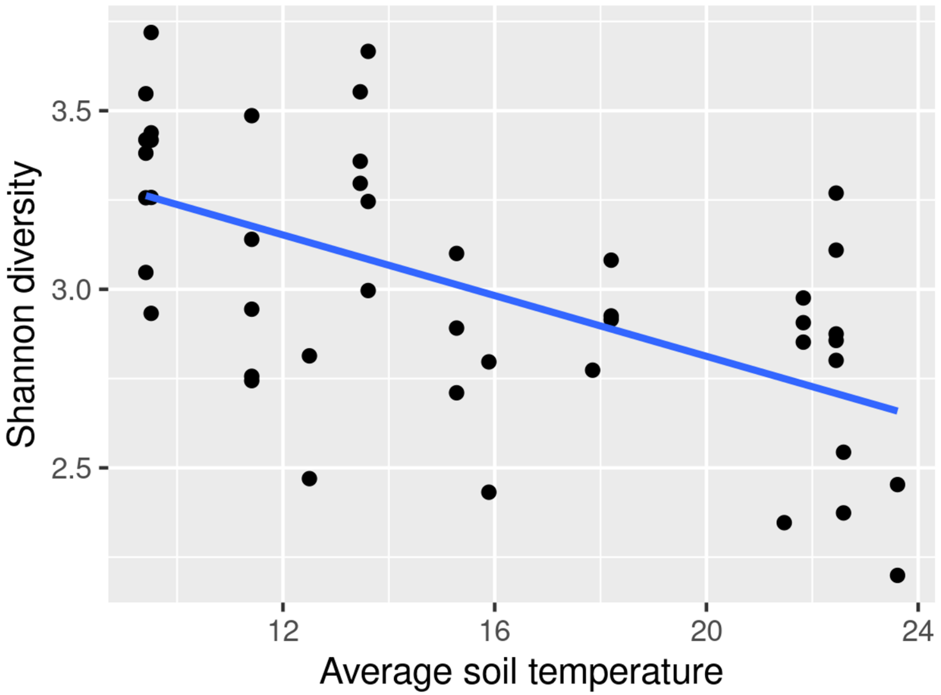


Types of tests for one independent and one dependent variable

	CATEGORICAL independent variable	CONTINUOUS independent variable
Continuous dependent variable	T-test (parametric) ANOVA (parametric, 2+ groups) Wilcoxon/Mann-Whitney test (non-parametric) Kruskall-wallis test (non-parametric, 2+groups)	Pearson's product-moment correlation (parametric) Spearman's rank correlation (non-parametric)

Pearson’s product-moment correlation

- Measures degree of correlation between two continuous variables



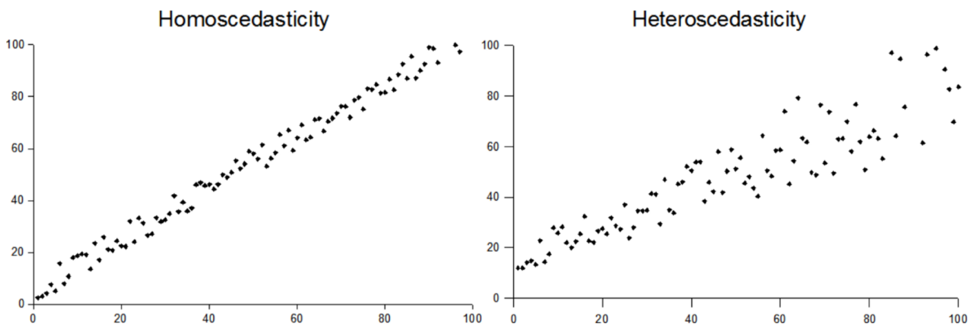
Pearson’s product-moment correlation

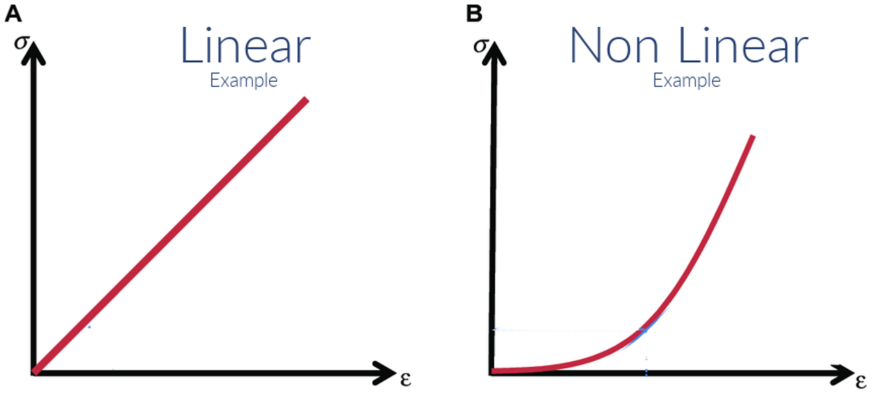
- Measures degree of correlation between two continuous variables

- cor.test(Y ~ X, data = dat)
- cor.test(Y ~ X, data = dat, method= “pearson”)

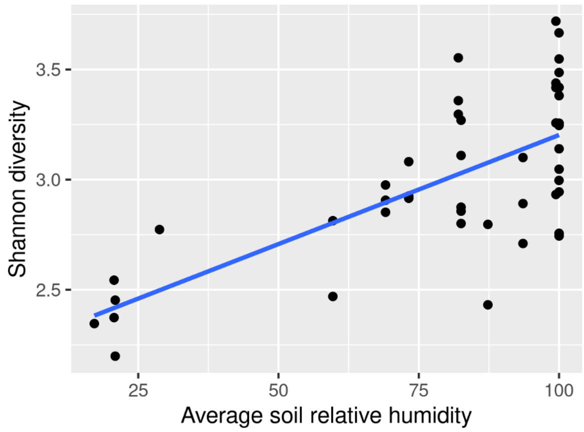
Pearson's Assumptions

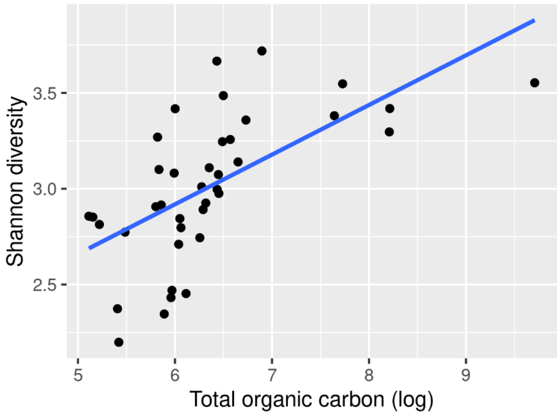
- X and Y variables are normally distributed
- Data is homoscedastic (the error is evenly distributed along the line; one side is not trumpet-shaped
- Relationship is LINEAR





Pearson’s Assumptions may not apply to our data





Spearman’s rank correlation

- Measures degree of correlation between two continuous variables
- No assumptions about normality, skedasticity, or linearity

- cor.test(Y ~ X, data = dat, method= “spearman”)

Spearman’s rank correlation

English (mark)	Maths (mark)	Rank (English)
56	66	9
75	70	3
45	40	10
71	60	4
62	65	6
64	56	5
58	59	8
80	77	1
76	67	2
61	63	7

Spearman’s rank correlation

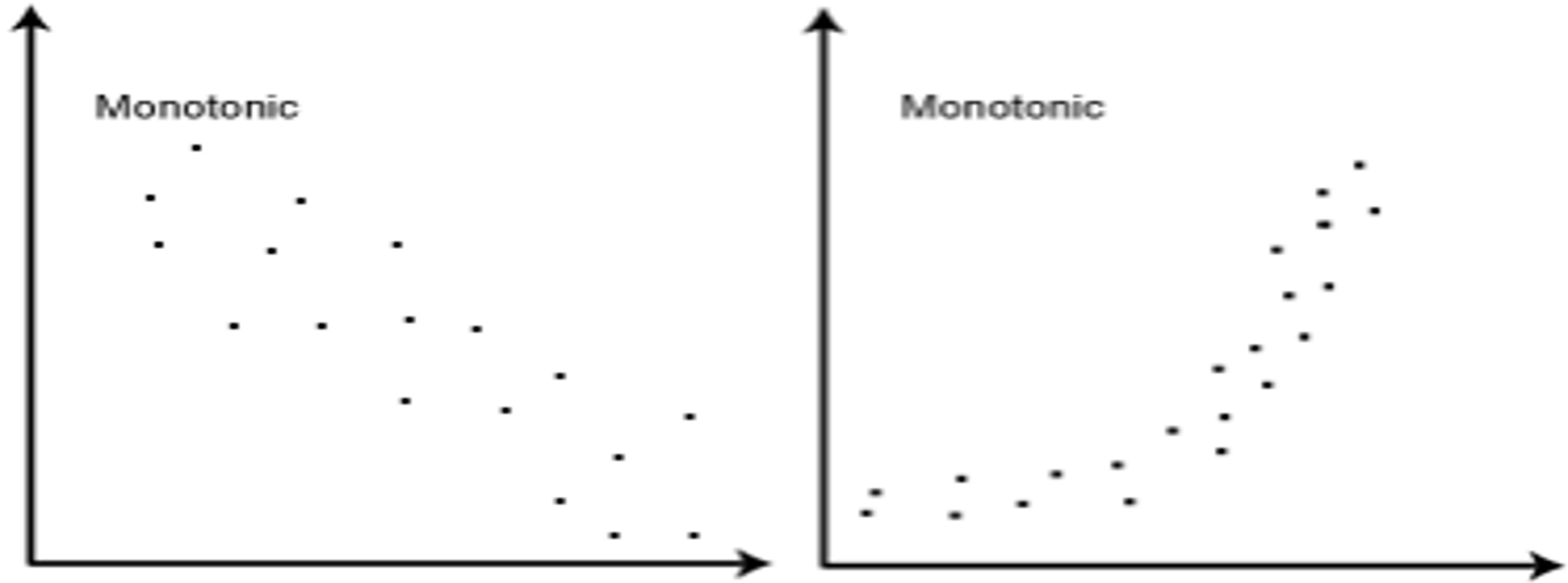
English (mark)	Maths (mark)	Rank (English)	Rank (maths)	d	d^2
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	3	9
62	65	6	5	1	1
64	56	5	9	4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	1	1
61	63	7	6	1	1

Spearman's rank correlation

English (mark)	Maths (mark)	Rank (English)	Rank (maths)	d	d^2
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	3	9
62	65	6	5	1	1
64	56	5	9	4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	1	1
61	63	7	6	1	1

Sums, then compares to a known threshold

Spearman's rank correlation can handle both linear and non-linear correlations

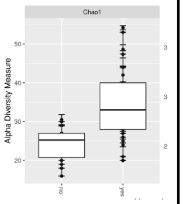
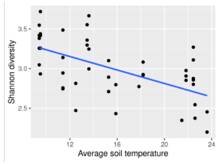


Why do we ever use parametric tests if their assumptions suck?

Why do we ever use parametric tests if their assumptions suck?

- If assumptions are met, parametric tests are more powerful (they are more likely to get significant results)
  - Because we can “infer” information from the data’s assumed distribution
- If assumptions are NOT met, non-parametric tests can be more powerful under certain conditions

SUMMARY

	CATEGORICAL independent variable 	CONTINUOUS independent variable 
Continuous dependent variable	PARAMETRIC: t.test lm + aov + summary + TukeyHSD NON-PARAMETRIC: wilcox.test kruskal.test	PARAMETRIC cor.test(, method= “pearson”) NON-PARAMETRIC cor.test(,method= “spearman”)