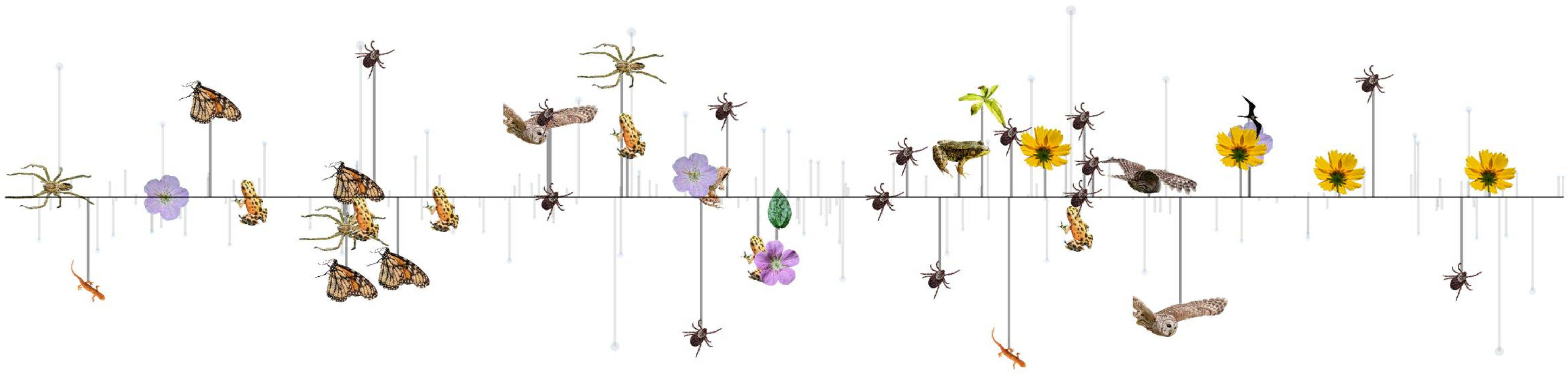
Intro to Quantitative Ecology
UMass Amherst – Michael France Nelson

Deck 8: Correlations and Associations



Announcements

- Monday is the deadline to withdraw from a course, or convert to P/F.
  - Some of you have been working with Ana and/or me to get caught up! This is great!
  - If you are very behind, and you haven't reached out, you need to consider your options wisely.

Announcements: UMass Incomplete Policy

- Students who are unable to complete course requirements within the allotted time because of severe medical or personal problems may request a grade of Incomplete from the instructor of the course and must complete an Incomplete Grade Form.
- Incomplete grades are warranted only if a student is passing the course at the time of the request and if the course requirements can be completed by the end of the following semester.

Additional info

- An incomplete cannot serve as a substitute for a non-passing grade.
- Additional complication: This is my last semester at UMass, so completing outstanding work for an incomplete won’t be simple.

Final Project Info!

Final Project Components

The final project consists of:

1. A set of take-home questions.
    - These will be made available on Moodle in finals week. They're not on the public-facing course website.
2. A take-home R guide
    - I provide a RMarkdown template file, you fill in the details!
    - The template is available now, you may take a look at any time.
    - You can (and should) adapt material from your assignments for the guide.
    - You'll fill in all material in square brackets and make all requested code adjustments.

Correlation

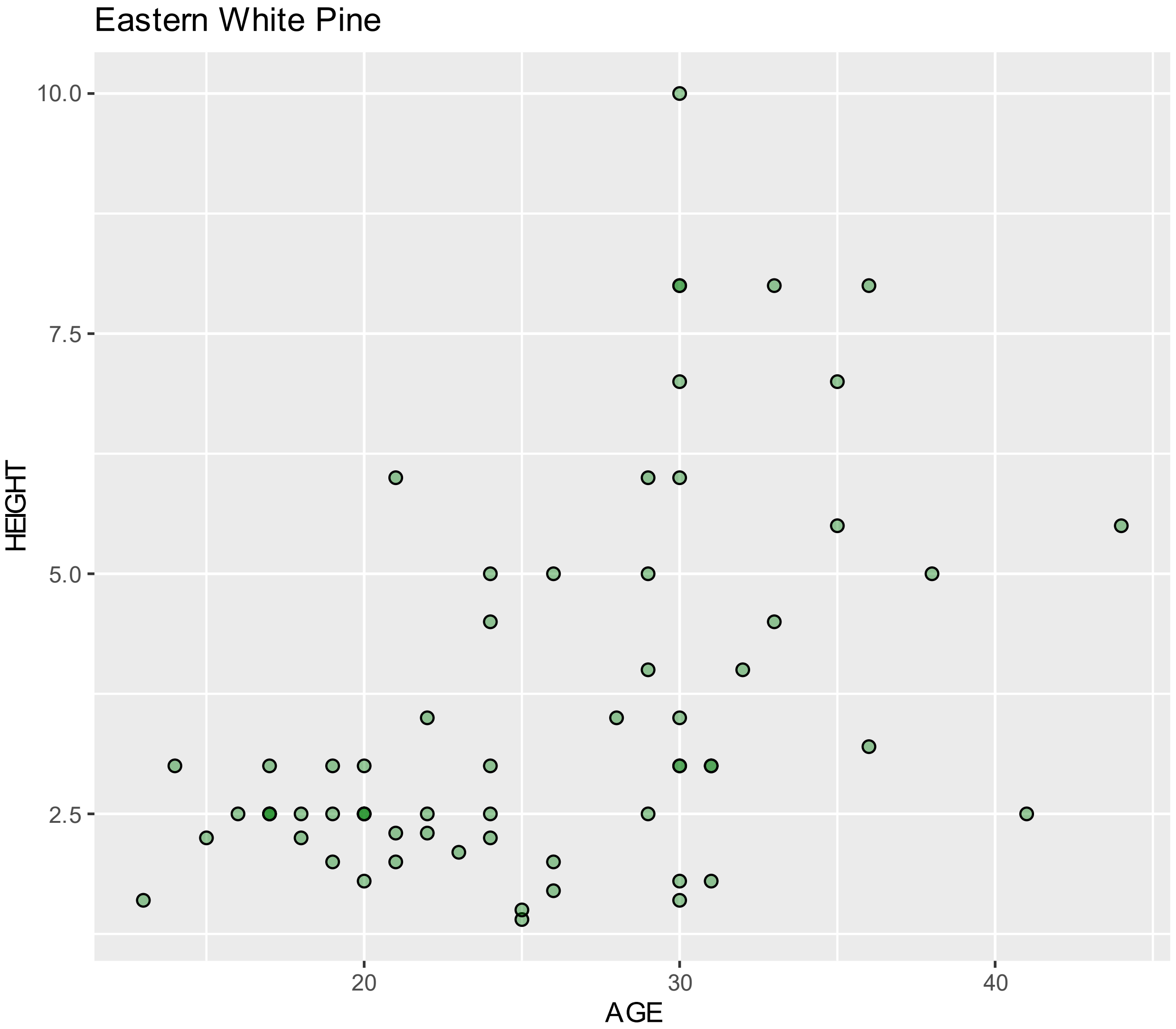
Key Terms

- Continuous/numerical variables
- Categorical variables
- Functions:
  - Linear, exponential, Logarithmic, polynomial, square root
  - Monotonic (always increasing, always decreasing, or flat; no humps)
- Variance and covariance
- Correlation: Spearman and Pearson
- Slope vs. correlation
- Contingency and two-way tables
- Chi-square test

Correlation: What is it?

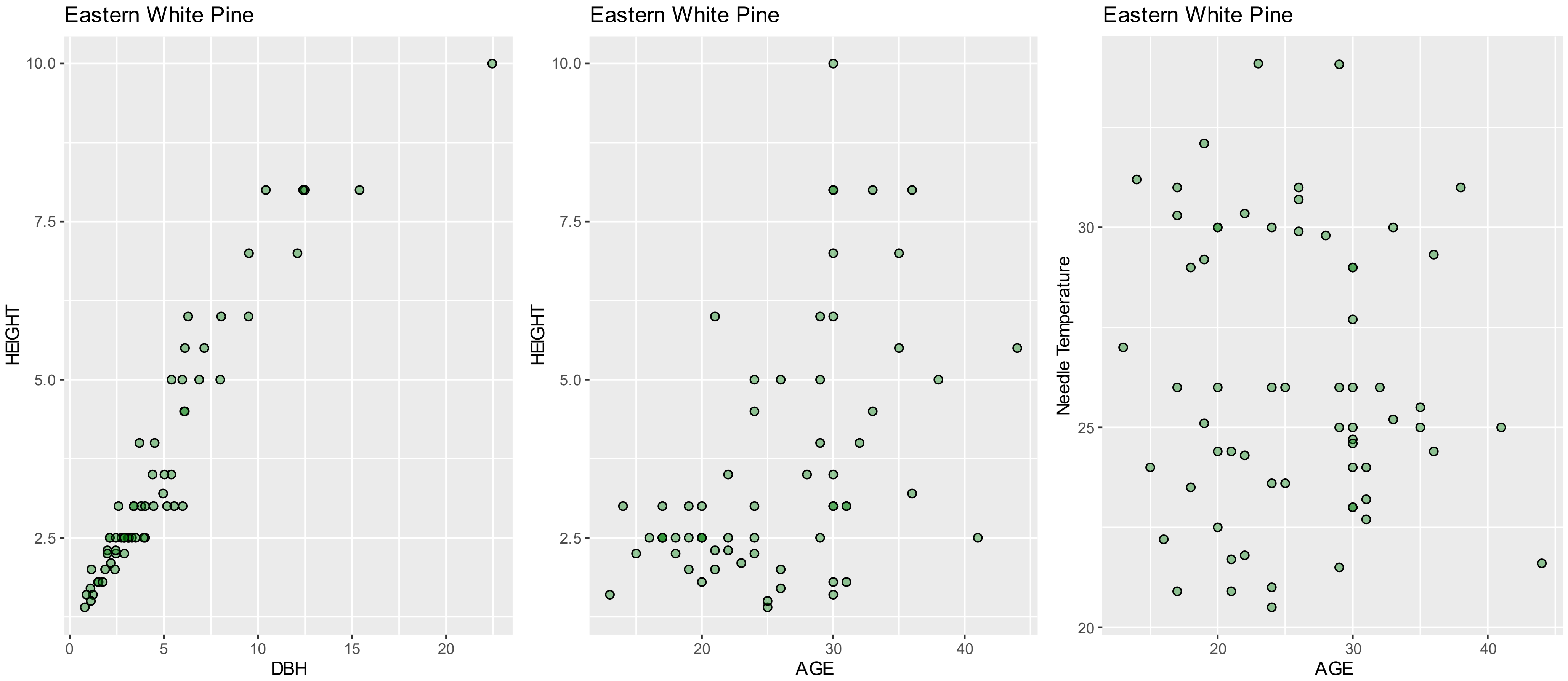
Let's step back… What is correlation?

We can ask: On a scale of -1.0 to 1.0, how highly coordinated are the variables in this scatterplot?



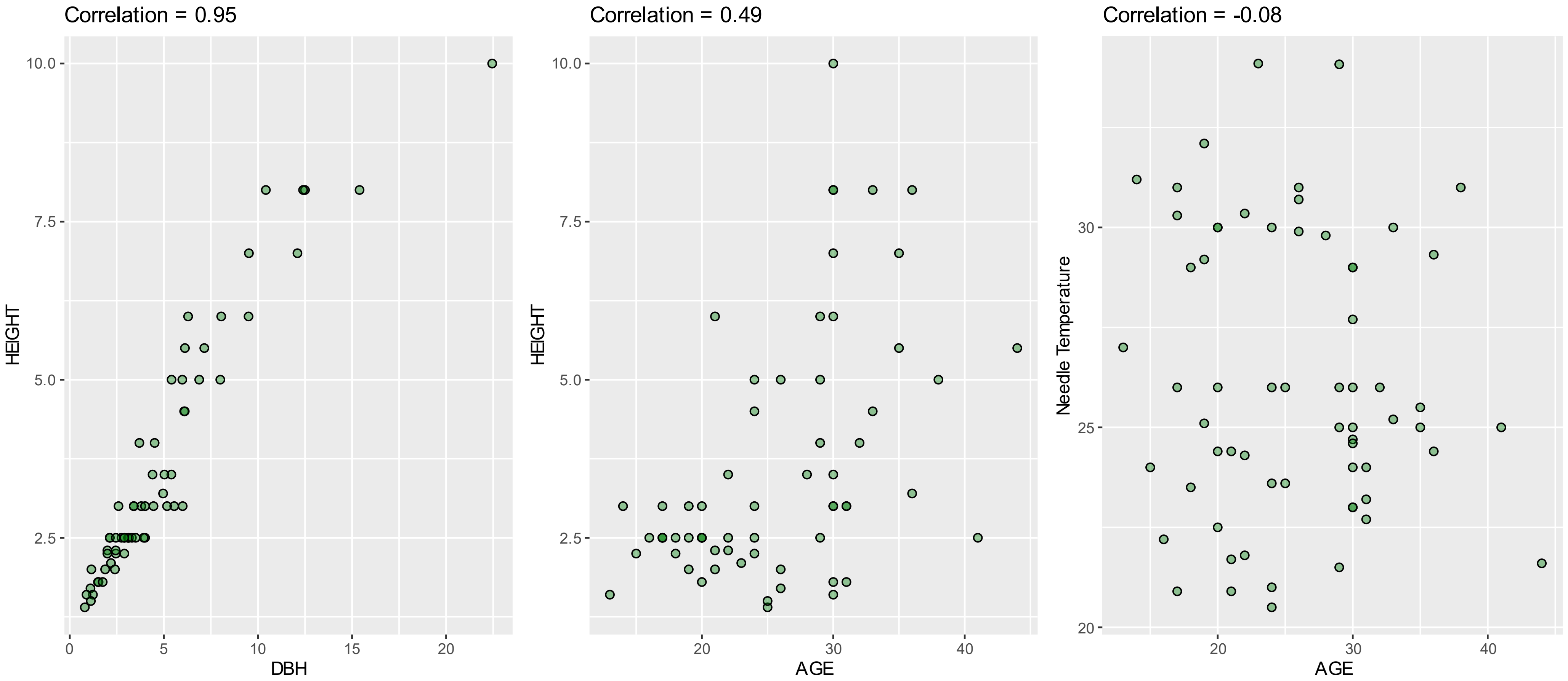
Correlation: Intuition

On a scale of -1.0 to 1.0, how highly coordinated are the variables in these plots?



Correlation: Intuition

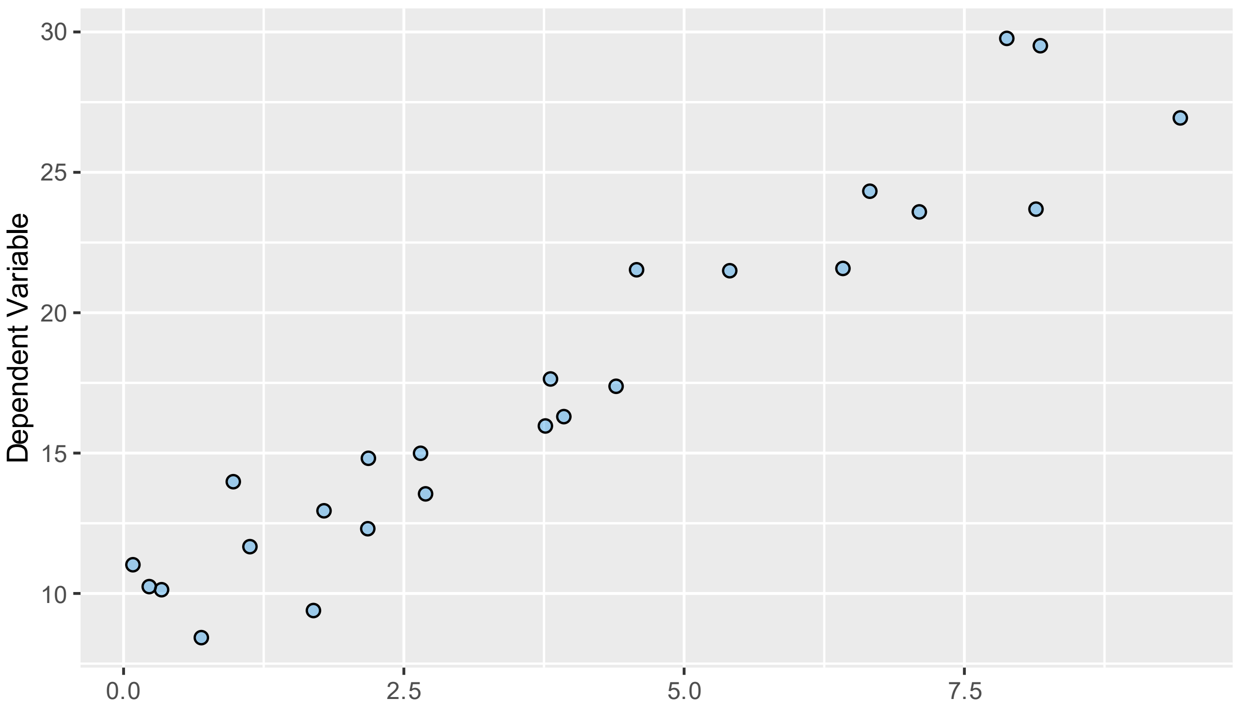
On a scale of -1.0 to 1.0, how highly coordinated are the variables in these plots?



Correlations

We're interested in the relationship between two variables:
We want to know if the variation is coordinated.

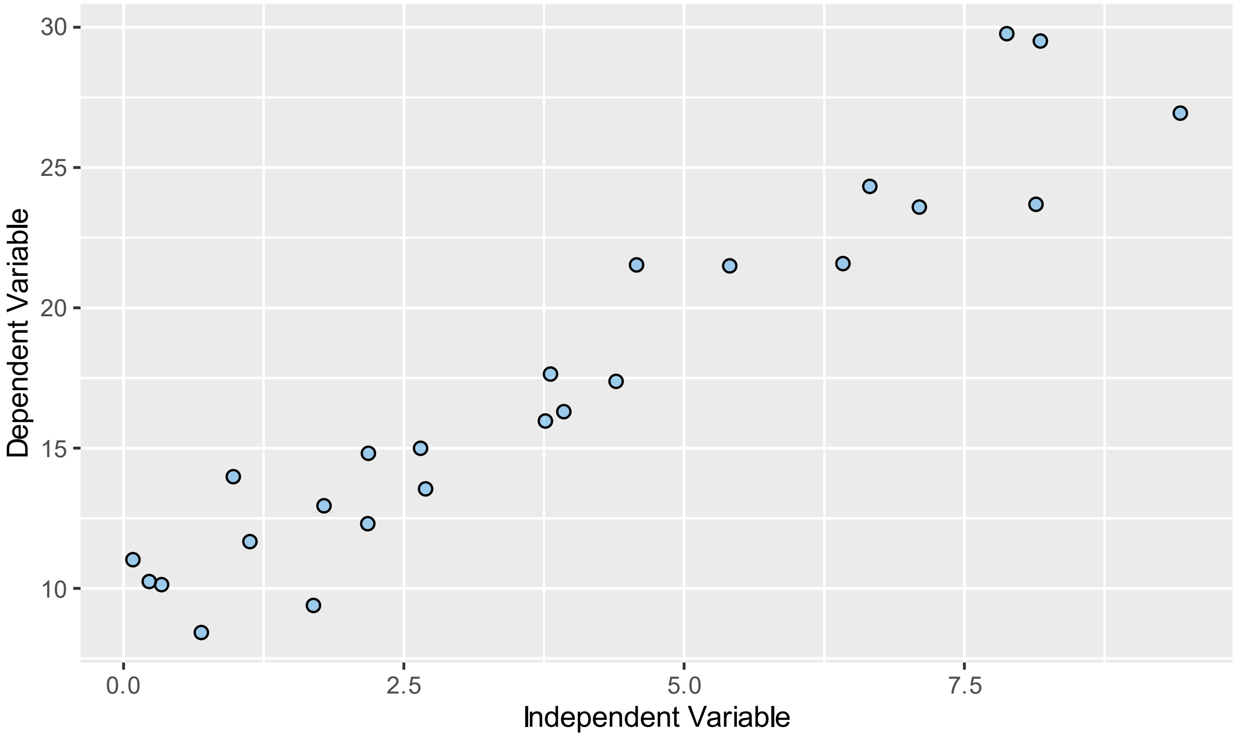
- Dependent variable:
    - data we are interested in explaining
    - Y-axis
    - *response* variable



Correlations

We're interested in the relationship between two variables:
We want to know if the variation is coordinated.

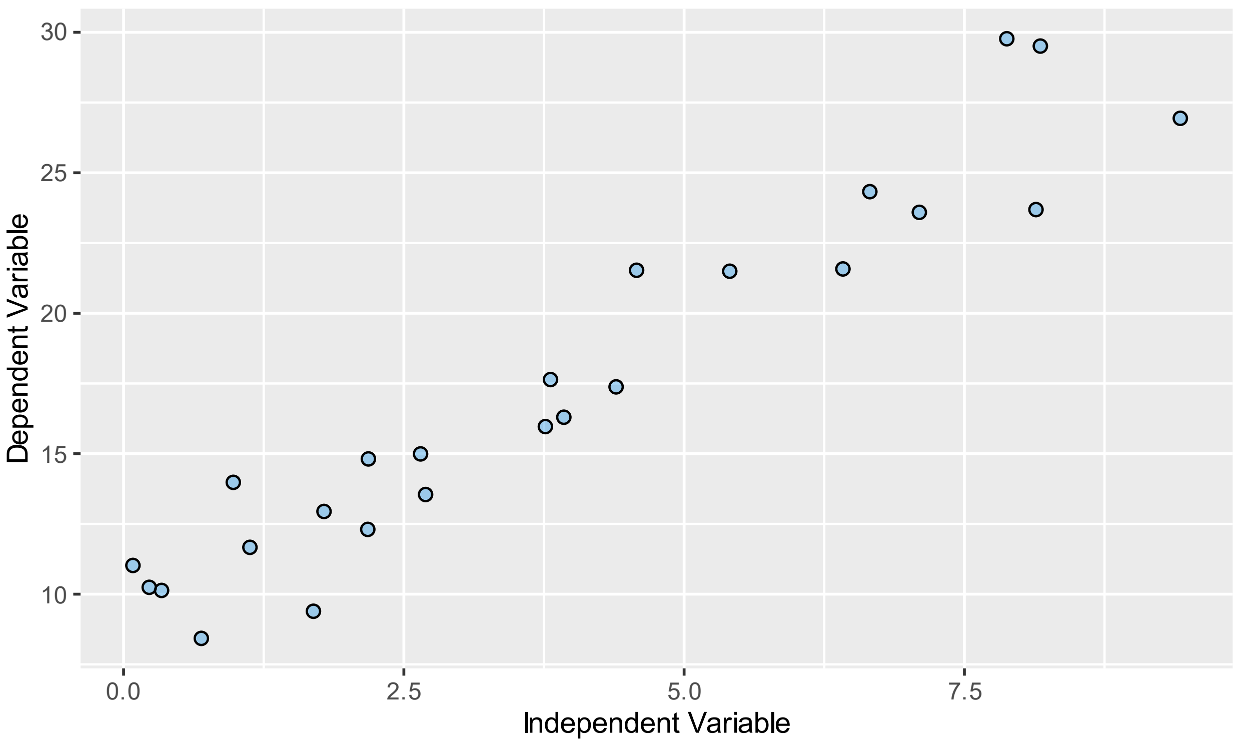
- Independent variable:
  - data used to describe variation in dependent variable
  - X-axis
  - *Explanatory* or *predictor* variable



Correlations

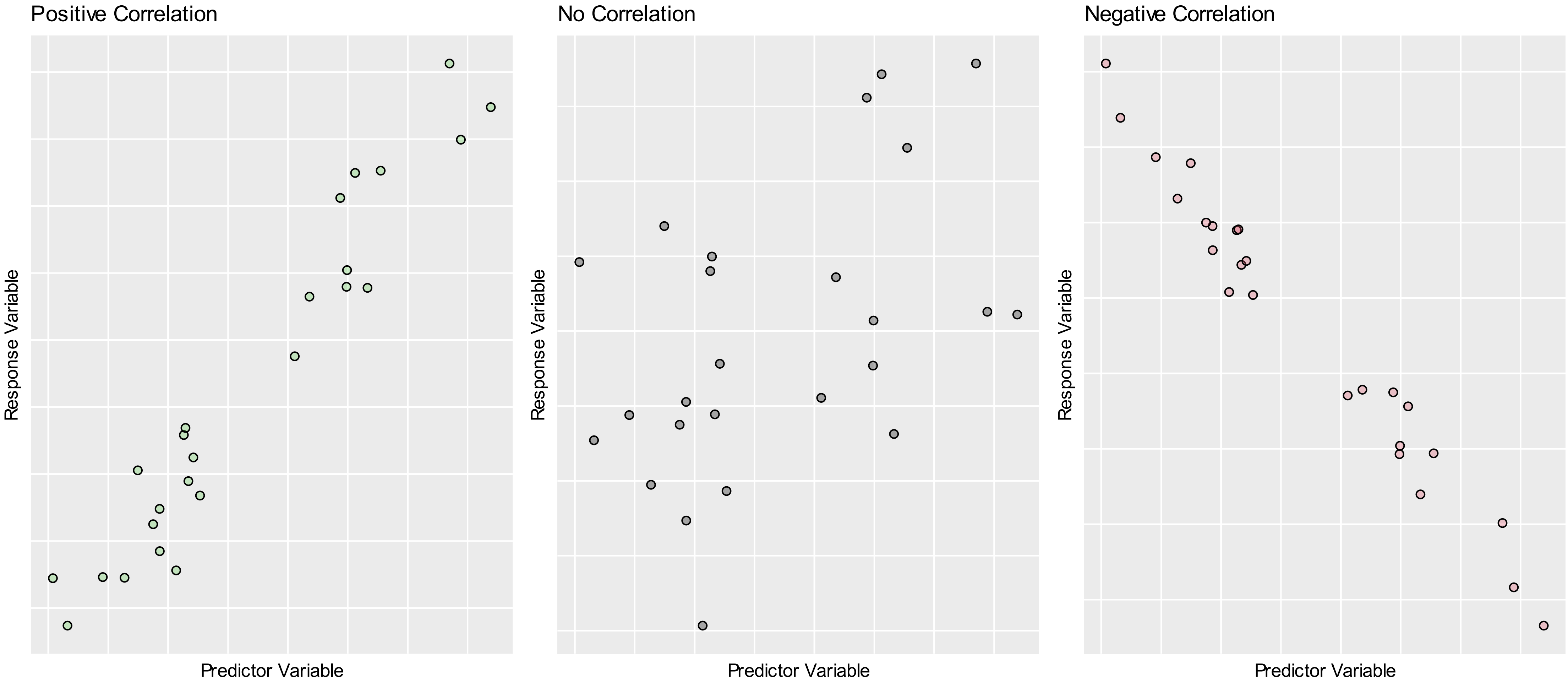
We're interested in the relationship between two variables:
We want to know if the variation is coordinated.

- **Dependent variable:**
  - data we are interested in explaining
  - Y-axis
  - *response* variable
- **Independent variable:**
  - data used to describe variation in dependent variable
  - X-axis
  - *explanatory* or *predictor* variable
- **Dealing with *pairs* of values (x, y)!**



Correlations: Sign

What is the sign of the correlation?



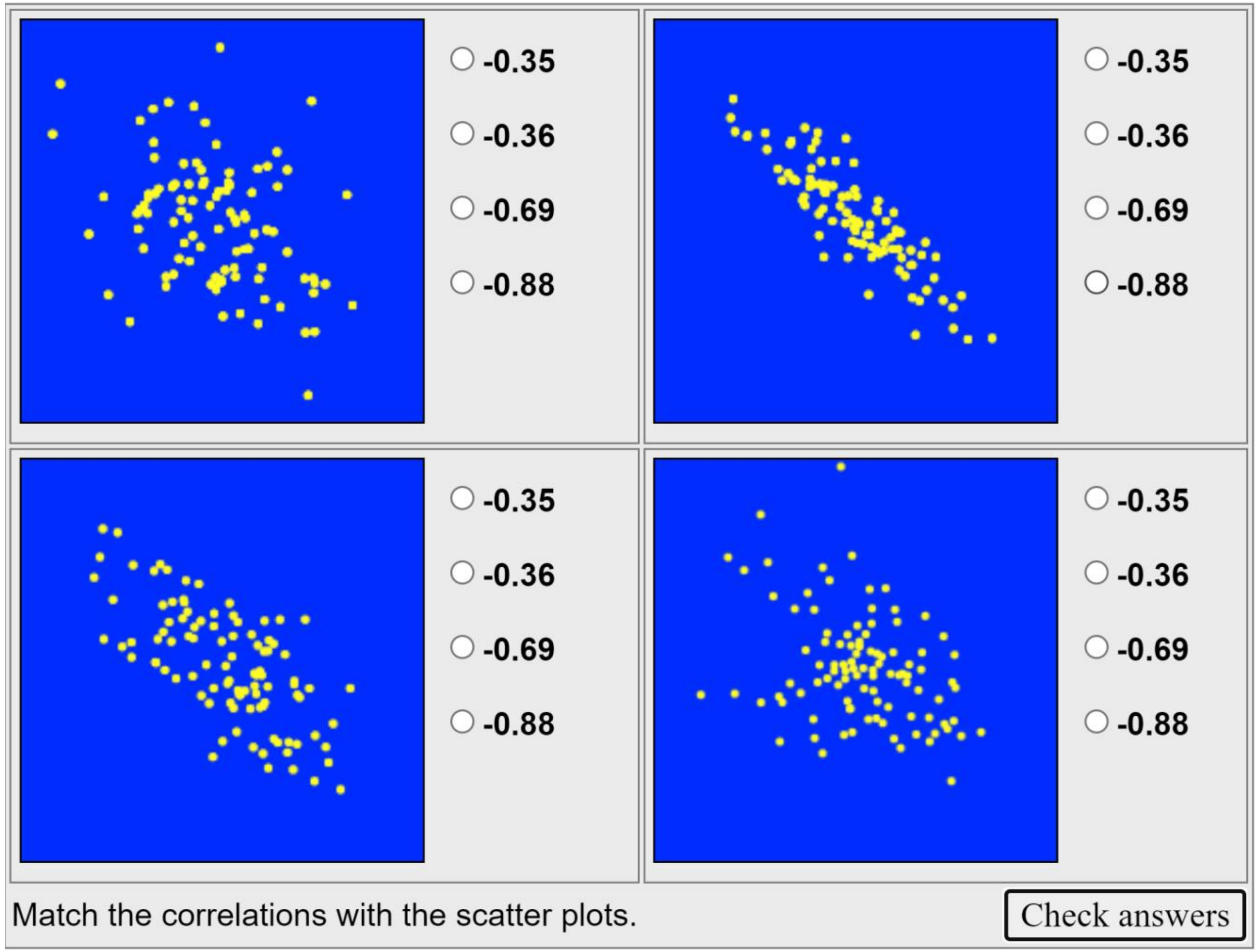
Correlations

Lets play a game!

http://www.istics.net/Correlations

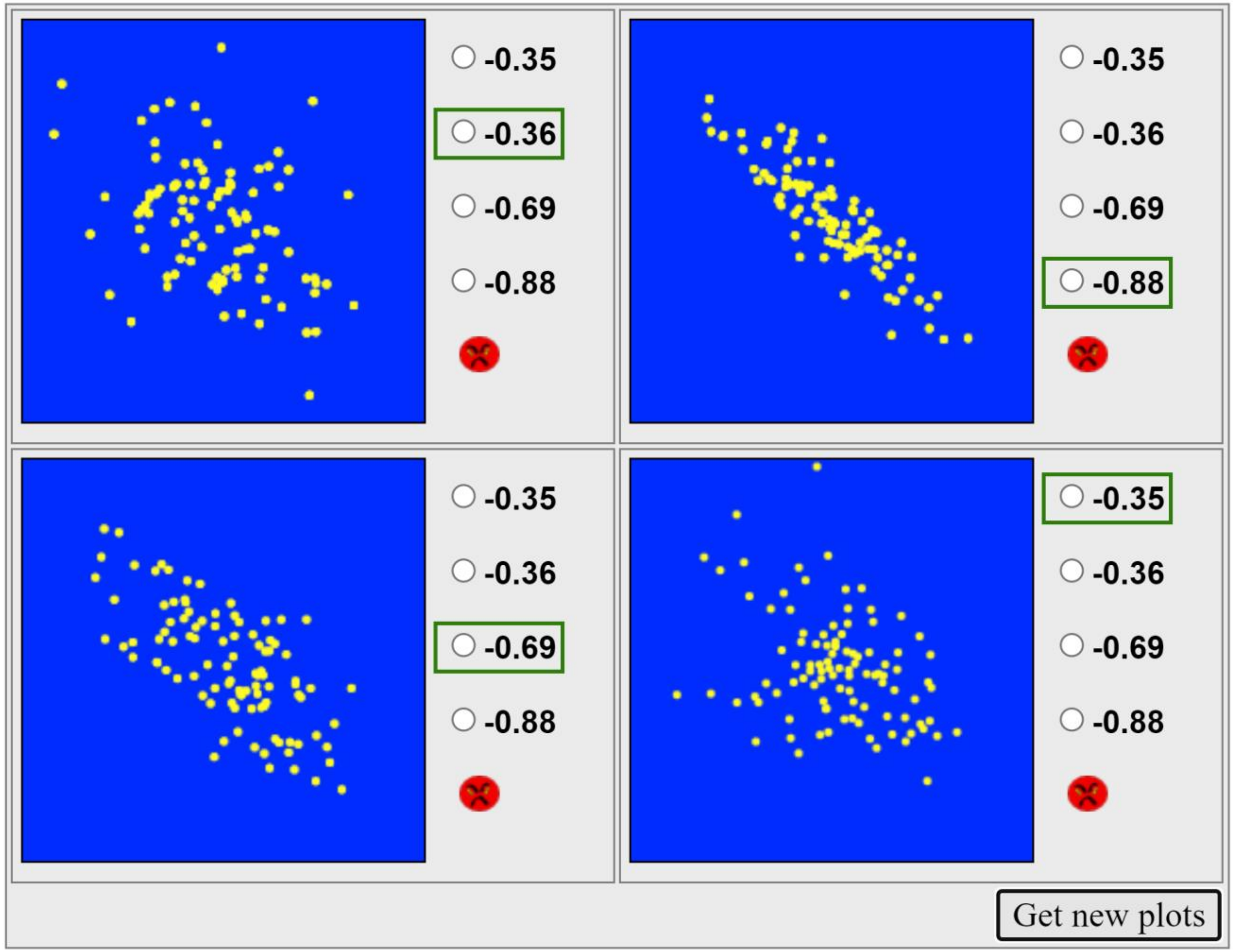
Correlations

http://www.istics.net/Correlations



Correlations

http://www.istics.net/Correlations



Variance and Covariance

Remember the standard deviation?

- It is a great way to quantify the amount of variability in a collection of numbers! It is the square root of the **variance**.

Variance

- Measures the amount of variation in a collection of numbers (just like sd).
- Normalized by the sample size: "Mean squared deviation"
- Variance calculation uses a sum of squares term:

$$Var(x) = \frac{1}{n}\sum_{i}^{n}(x_i - \bar{x})^2$$

- Variance is univariate and always positive (it's those squared terms!), but it's in weird units…

Variance and Covariance

Covariance

- Quantifies the amount of coordinated variation in **two** variables. It is a bivariate statistic.

$$\text{Cov}(x, y) = \frac{1}{n}\sum_{i}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- No squared term… can range from –infinity to infinity
- The covariance of a variable with itself is just the variance.
- Positive covariance means that as x values increase, the corresponding y value tends to increase. Vice versa for negative covariance.
- Zero covariance means there is no coordinated variation.

Variance, Covariance, and Correlation

- Covariance quantifies the amount of coordinated variation in **two** variables. It is a bivariate statistic.
- Variance quantifies the amount of variation in a single variable. It is a univariate statistic.
- Correlation is a **normalized** version of covariance. It's normalized to fall into the range -1 to 1.
    - The calculation uses the covariance of both variables, as well as the individual variances of each variable.

Correlations: Quantification and Significance

We need a way to

- quantify correlations/relationships
- assess whether correlations/relationships are *significant*

What might such a test look like?

We've got two variables: x, y

How have we compared two groups of numbers?

- Non-parametric test
- Parametric test

Correlations: Spearman Correlation

We need a *test*!

1. Spearman's rank test (r_s) - nonparametric
   - determines the strength of the link between 2 samples
   - data *do not* have to be normally distributed
   - relationship *does not* have to be linear
     - but still exhibits a *monotonic* (not u- or n-shaped) positive or negative trend
   - use the *ranks* of values
   - correlation strength ranges from -1 to 1
     - -1: perfect *negative* correlation
     - 1: perfect *positive* correlation
     - 0: no correlation

Correlations: Pearson Correlation

We need a *test*!

2.Pearson's product moment (r) - parametric

- determines the strength of the link between 2 samples
- data *must be* normally distributed
- the relationship *must be* linear
  - positive of negative trend
  - monotonic: not u- or n-shaped
- uses actual values
- correlation strength ranges from -1 to 1
  - -1: perfect *negative* correlation
  - 1: perfect *positive* correlation
  - 0: no correlation

Correlations: r_s and r_p

Spearman's rank test (rs)	Pearson's rank test (rs)
▶ determine the strength of the link between 2 samples ▶ data *do not* have to be normally distributed ▶ trend *does not* have to be linear ▶ uses the *ranks of values*	▶ determine the strength of the link between 2 samples ▶ data *must be* normally distributed ▶ trend *must be* linear ▶ uses the *actual data values*

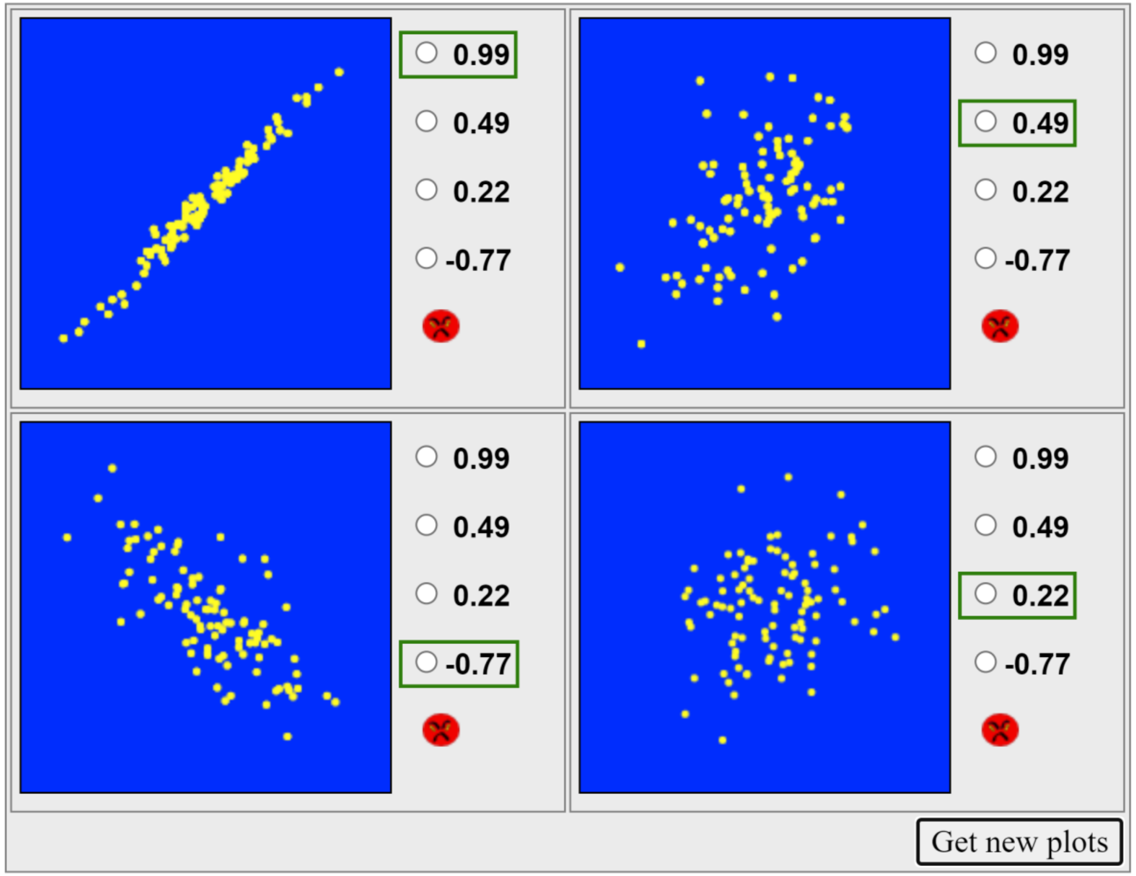
Correlations: r_s and r_p

Both the Spearman and Pearson correlation coefficients are normalized to fall in the range from -1 to 1

- -1: perfect *negative* correlation
- 1: perfect *positive* correlation
- 0: no correlation

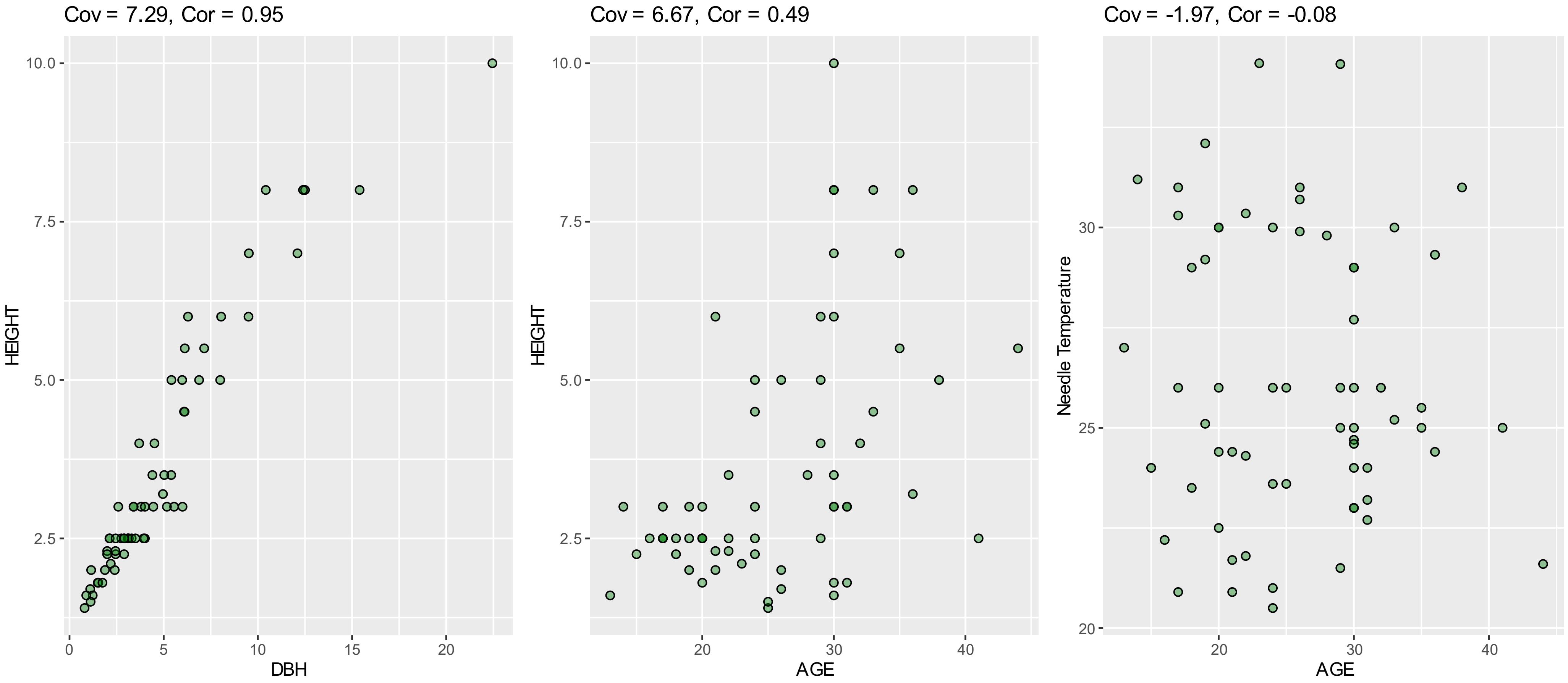
Both correlations only work for monotonic trends.

- The trend must be always increasing or always decreasing



Correlation Coefficients

We like correlation coefficients because we can interpret them easily (compared to variances, covariances, etc.).



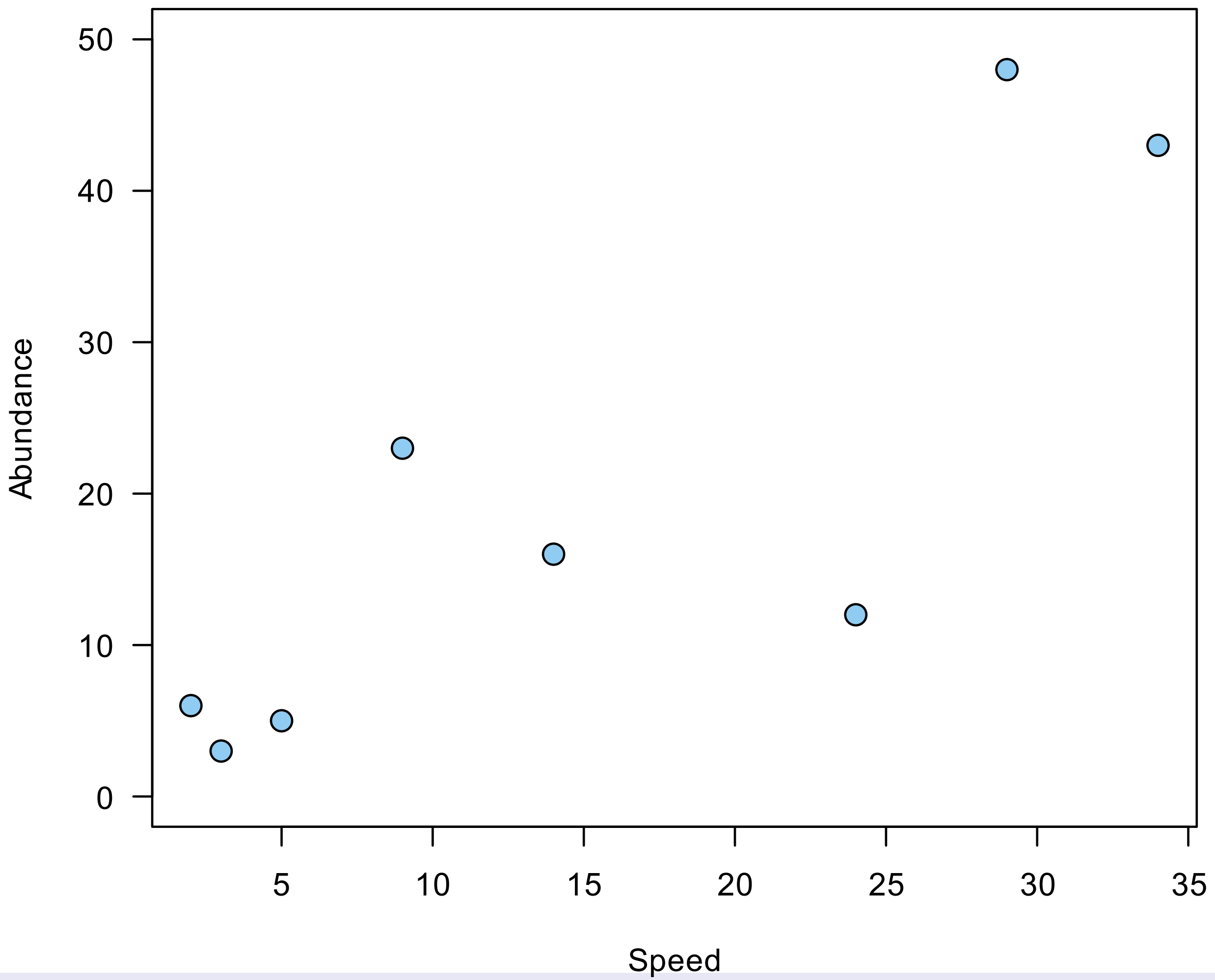
Example: the Mayfly data

Mayflies have aquatic larvae that develop in fast-moving streams.

Example: the Mayfly data

You've collected data on mayfly abundance in streams with different water flow speeds:

V1	Speed	Abundance
1	2	6
2	3	3
3	5	5
4	9	23
5	14	16
6	24	12
7	29	48
8	34	43



Spearman's rank test (r_s) - the hypothesis

Before conducting any statistical test, we need to state the hypotheses!

Spearman's rank test (r_s) - the hypothesis

Before conducting any statistical test, we need to state the hypotheses!

- H0 (the null hypothesis): There is no correlation between stream speed and mayfly abundance
- H1 (the alternative hypothesis): There is a positive correlation between stream speed and mayfly abundance

Spearman's rank test (r_s) - the statistical test

$$Rr_s = 1 - \frac{6 \times \sum D^2}{n(n^2 - 1)}$$

- Software (Excel and R) will do the math for us, BUT we should be aware of what's going on!

V1	Speed	Abundance
1	2	6
2	3	3
3	5	5
4	9	23
5	14	16
6	24	12
7	29	48
8	34	43

Spearman's rank test (r_s) - the statistical test

$$Rr_s = 1 - \frac{6 \times \sum D^2}{n(n^2 - 1)}$$

- First, calculate the *ranks* of the values: speed

V1	Speed	Abundance	Speed.rank
1	2	6	1
2	3	3	2
3	5	5	3
4	9	23	4
5	14	16	5
6	24	12	6
7	29	48	7
8	34	43	8

Spearman's rank test (r_s) - the statistical test

$$Rr_s = 1 - \frac{6 \times \sum D^2}{n(n^2 - 1)}$$

- First, calculate the *ranks* of the values: abundance

V1	Speed	Abundance	Speed.rank	Abundance.rank
1	2	6	1	3
2	3	3	2	1
3	5	5	3	2
4	9	23	4	6
5	14	16	5	5
6	24	12	6	4
7	29	48	7	8
8	34	43	8	7

Spearman's rank test (r_s) - the statistical test

$$Rr_s = 1 - \frac{6 \times \sum D^2}{n(n^2 - 1)}$$

- Then, calculate the *difference* in ranks: D

V1	Speed	Abundance	Speed.rank	Abundance.rank	Diff
1	2	6	1	3	2
2	3	3	2	1	-1
3	5	5	3	2	-1
4	9	23	4	6	2
5	14	16	5	5	0
6	24	12	6	4	-2
7	29	48	7	8	1
8	34	43	8	7	-1

Spearman's rank test (r_s) - the statistical test

$$Rr_s = 1 - \frac{6 \times \sum D^2}{n(n^2 - 1)}$$

- Then, square the *difference* in ranks: D^2
  - Why do we like to square things?
  - Do squared terms usually go with measures of center, or spread?

V1	Speed	Abundance	Speed.rank	Abundance.rank	Diff	Diff.sq
1	2	6	1	3	2	4
2	3	3	2	1	-1	1
3	5	5	3	2	-1	1
4	9	23	4	6	2	4
5	14	16	5	5	0	0
6	24	12	6	4	-2	4
7	29	48	7	8	1	1
8	34	43	8	7	-1	1

Spearman's rank test (r_s) - the statistical test

$$Rr_s = 1 - \frac{6 \times \sum D^2}{n(n^2 - 1)}$$

- Now we have all the pieces:
- n : number of observations
- D : difference in ranks
- D^2: differences squared

V1	Speed	Abundance	Speed.rank	Abundance.rank	Diff	Diff.sq
1	2	6	1	3	2	4
2	3	3	2	1	-1	1
3	5	5	3	2	-1	1
4	9	23	4	6	2	4
5	14	16	5	5	0	0
6	24	12	6	4	-2	4
7	29	48	7	8	1	1
8	34	43	8	7	-1	1

Spearman's rank test (r_s) - intuition

$$Rr_s = 1 - \frac{6 \times \sum D^2}{n(n^2 - 1)}$$

- The key part of the formula is:

$$\frac{6 \times \sum D^2}{n(n^2 - 1)}$$

- This term takes on values from 0 to 2, such that R can take on values from -1 to 1.

V1	Speed	Abundance	Speed.rank	Abundance.rank	Diff	Diff.sq
1	2	6	1	3	2	4
2	3	3	2	1	-1	1
3	5	5	3	2	-1	1
4	9	23	4	6	2	4
5	14	16	5	5	0	0
6	24	12	6	4	-2	4
7	29	48	7	8	1	1
8	34	43	8	7	-1	1

Spearman's rank test (r_s) - intuition

$$\frac{6 \times \sum D^2}{n(n^2 - 1)}$$

- The 6 and $n(n^2 - 1)$ parts of the formula are ***normalizing terms***.
- The normalizing terms ensure the value of this term is constrained to the range 0 – 2.

V1	Speed	Abundance	Speed.rank	Abundance.rank	Diff	Diff.sq
1	2	6	1	3	2	4
2	3	3	2	1	-1	1
3	5	5	3	2	-1	1
4	9	23	4	6	2	4
5	14	16	5	5	0	0
6	24	12	6	4	-2	4
7	29	48	7	8	1	1
8	34	43	8	7	-1	1

Spearman's rank test (r_s) - intuition

We've now encountered the normalization idea a few times:

- Calculating standard deviation: we divide by the sample size
- Sums of squares in ANOVA.
- Normalizing constants of probability distribution functions. We haven't worked with these directly, but they are there!

Normal Distribution PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Normalizing constant. Ensures that the indefinite integral has a value of 1.0.

Spearman's rank test (r_s) - intuition

$$\frac{6 \times \sum D^2}{n(n^2 - 1)}$$

- If ranks are all similar, the D^2 terms will all be small (or 0 for perfect for correlation).
- This overall term is then closer to zero, making R close to 1 (positive correlation)

V1	Speed	Abundance	Speed.rank	Abundance.rank	Diff	Diff.sq
1	2	6	1	3	2	4
2	3	3	2	1	-1	1
3	5	5	3	2	-1	1
4	9	23	4	6	2	4
5	14	16	5	5	0	0
6	24	12	6	4	-2	4
7	29	48	7	8	1	1
8	34	43	8	7	-1	1

Spearman's rank test (r_s) - intuition

$$\frac{6 \times \sum D^2}{n(n^2 - 1)}$$

- If ranks are all very dissimilar, all the D^2 terms will all be large.
- This overall term is then closer to 2 making R closer to -1

V1	Speed	Abundance	Speed.rank	Abundance.rank	Diff	Diff.sq
1	2	6	1	3	2	4
2	3	3	2	1	-1	1
3	5	5	3	2	-1	1
4	9	23	4	6	2	4
5	14	16	5	5	0	0
6	24	12	6	4	-2	4
7	29	48	7	8	1	1
8	34	43	8	7	-1	1

Spearman's rank test (r_s) - intuition

$$\frac{6 \times \sum D^2}{n(n^2 - 1)}$$

- If ranks are all random, the values of the D^2 terms will be a mix of large and small.
- This overall term is then closer to 1 making R closer to 0 (no correlation).

V1	Speed	Abundance	Speed.rank	Abundance.rank	Diff	Diff.sq
1	2	6	1	3	2	4
2	3	3	2	1	-1	1
3	5	5	3	2	-1	1
4	9	23	4	6	2	4
5	14	16	5	5	0	0
6	24	12	6	4	-2	4
7	29	48	7	8	1	1
8	34	43	8	7	-1	1

Spearman's rank test in R

Now that we've seen the how the formula works, we'll let R take care of the calculations:

$$r_s = 1 - \frac{6 \times \sum D^2}{n(n^2 - 1)}$$

R general syntax:

- Coefficient only:

```
cor(var1, var2, method='spearman')
```

- Coefficient *and* significance test R:

```
cor.test(var1, var2, method='spearman')
```

Spearman's rank correlation in R

R syntax with mayfly data

- Coefficient only:

```
cor(
  mayfly$Speed,
  mayfly$Abundance,
  method='spearman')
[1] 0.8095238
```

V1	Speed	Abundance
1	2	6
2	3	3
3	5	5
4	9	23
5	14	16
6	24	12
7	29	48
8	34	43

Spearman’s rank test in R

R syntax with mayfly data: Coefficient *and* significance test

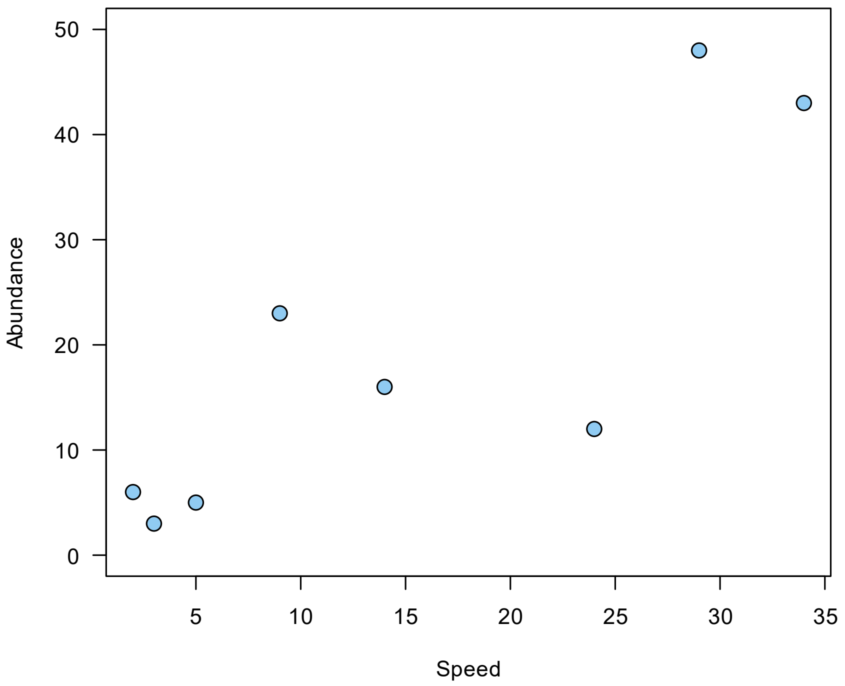
```
cor.test(
  mayfly$Speed,
  mayfly$Abundance,
  method='spearman')


    Spearman's rank correlation rho


data:  mayfly$Speed and mayfly$Abundance
S = 16, p-value = 0.02178
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8095238
```

Mayfly Data: Fitting a Linear Model

Remember that the Spearman correlation coefficient doesn't require the trend to be linear.

However… When a relationship is linear, we can use a more powerful correlation: the Pearson Correlation Coefficient.

	Speed	Abundance
1:	2	6
2:	3	3
3:	5	5
4:	9	23
5:	14	16
6:	24	12



Pearson Correlation

Pearson correlation is based on some parametric assumptions (having to do with the individual distributions of x and y).

It also assumes the relationship between x and y is linear.

Calculation uses covariance and variance terms:

$$r(x,y) = \frac{Cov(x,y)}{Var(x) \times Var(y)}$$

[You don’t need to memorize the formula!]

Correlation in R

R general syntax:

- Coefficient only:

```
cor(var1, var2, method = "pearson")
```

- Coefficient *and* significance test

```
cor.test(var1, var2, method = "pearson")
```

Pearson's product moment in R

R syntax with mayfly data

- Coefficient only:

```
cor(
  mayfly$Speed,
  mayfly$Abundance,
  method='pearson')
[1] 0.8441408
```

V1	Speed	Abundance
1	2	6
2	3	3
3	5	5
4	9	23
5	14	16
6	24	12
7	29	48
8	34	43

Pearson's product moment in R

R syntax with mayfly data: Coefficient *and* significance test

```
cor.test(mayfly$Speed, mayfly$Abundance,method='pearson')

	Pearson's product-moment correlation

data:  mayfly$Speed and mayfly$Abundance
t = 3.8568, df = 6, p-value = 0.008393
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3442317 0.9711386
sample estimates:
      cor 
0.8441408
```

Mayfly conclusions

Spearman's rank correlation coefficient:

- $r_s = 0.73$:
- the null hypothesis: no correlation
- the alternative hypothesis is?

Pearson's correlation coefficient:

- $r = 0.844$
- the null hypothesis: no correlation
- the alternative hypothesis is?

Conclusion:

- The values are slightly different, but the qualitative interpretation is same regardless for both!
- there *is* a statistically significant positive correlation between stream flow and mayfly abundance!

Correlation Summary

- Correlation tells us:
  - Strength of relationship between two continuous variables
- Correlation calculations can be:
  - Non-parametric: Spearman
  - Parametric: Pearson
- Correlation requires:
  - Monotonic function: always increasing or decreasing – no humps
- Correlation does not tell us:
  - The magnitude of the relationship. For this we need a regression model!
  - NOTE: your book presents the slope (m) and correlation coefficient (r) together in the ch. 8; they give us complementary information. We'll look at slope coefficients in detail when we talk about regression.

Linking Data: Associations

Contingency Tables and Chi-square Tests

Links: Numerical and Categorical Data

Numerical Data	Categorical Data
We've seen one way to quantify the relationship between two numeric variables: correlation coefficients: • Pearson – for linear relationships • Spearman – for monotonic relationships We'll be delving into another way to quantify these relationships: linear regression.	• Categorical data often comes in the form of counts. • For example, how many female penguins were observed on Dream Island. • We can arrange counts of categorical data observations into a table

Contingency Tables

- When observations have two categorical variables, we can arrange them in a table of counts.
- For example, we can make a table of how many penguins of each species were found on each of the islands.
- We could also make a table of how many individuals of each sex were on each island.

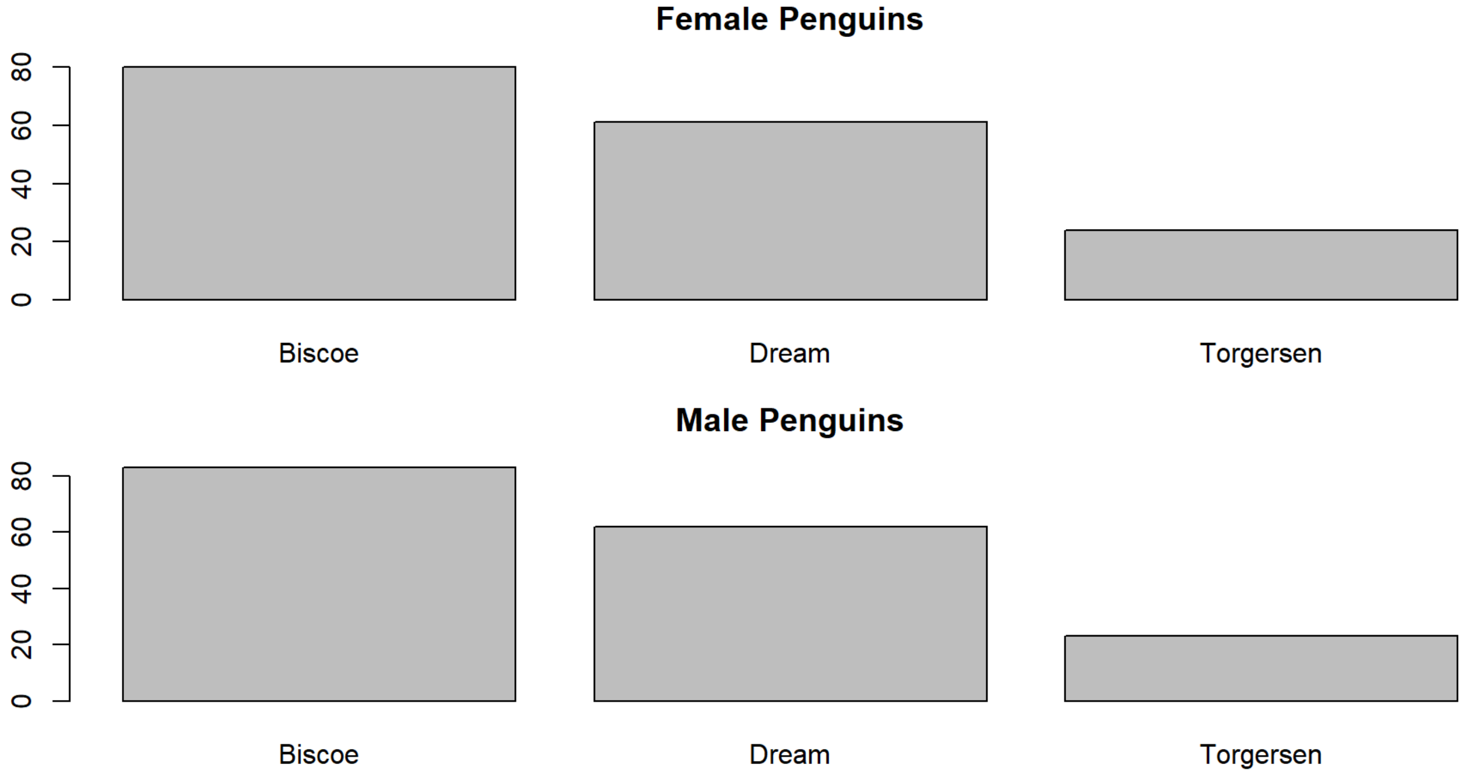
	Biscoe	Dream	Torgersen
Adelie	44	56	52
Chinstrap	0	68	0
Gentoo	124	0	0

	Biscoe	Dream	Torgersen
female	80	61	24
male	83	62	23

Contingency Tables: Visualizing Proportions

- We can visualize individual rows or columns of a contingency table using barplots.
- In this table, the proportions look similar for male and female penguins

	Biscoe	Dream	Torgersen
female	80	61	24
male	83	62	23

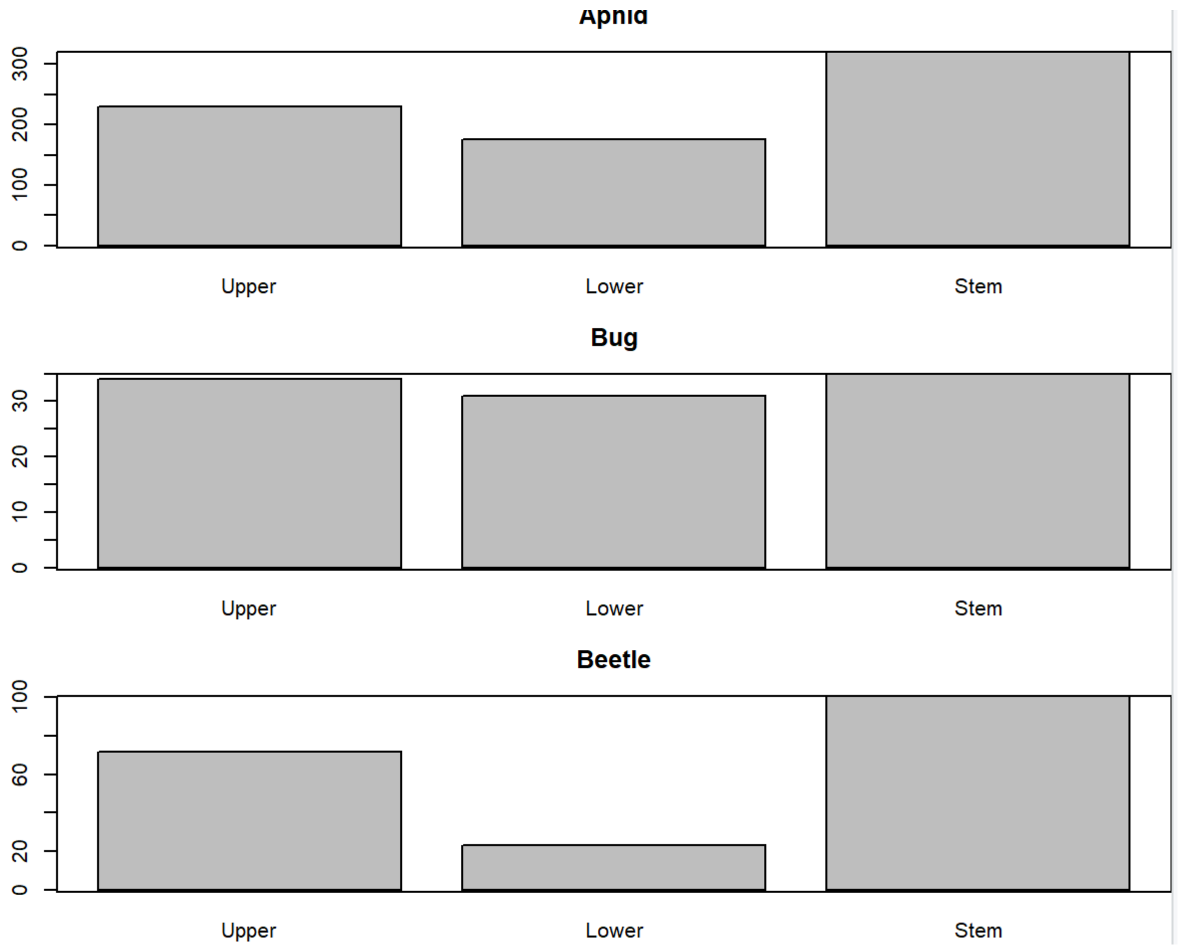


Contingency table: visualizing rows

- Here's a portion of the invertebrate table from your book:

	Upper	Lower	Stem
Aphid	230	175	321
Bug	34	31	35
Beetle	72	23	101

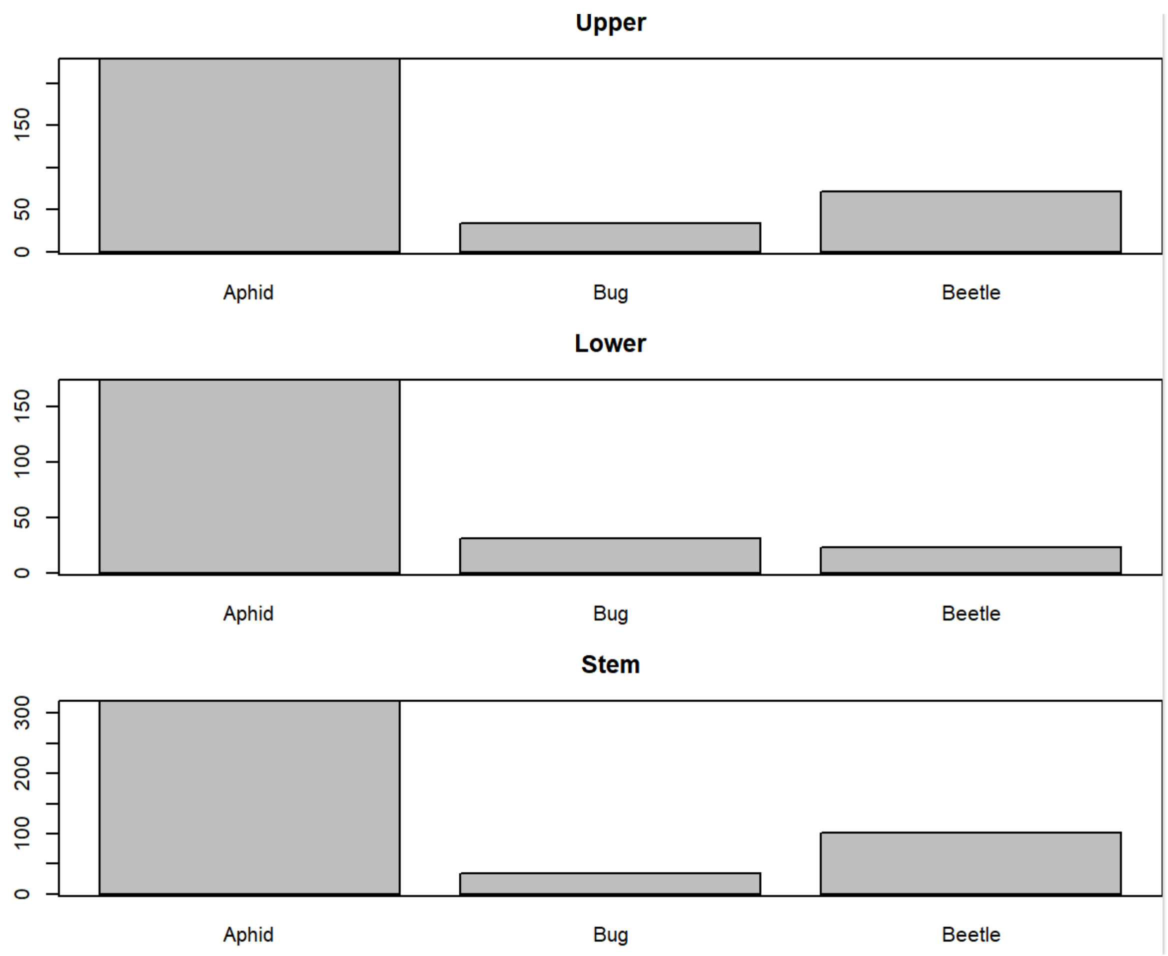
- The proportions for the rows look less similar for this table.
- Squint your eyes and look at the barplots in each row. The patterns are different!



Contingency table: visualizing columns

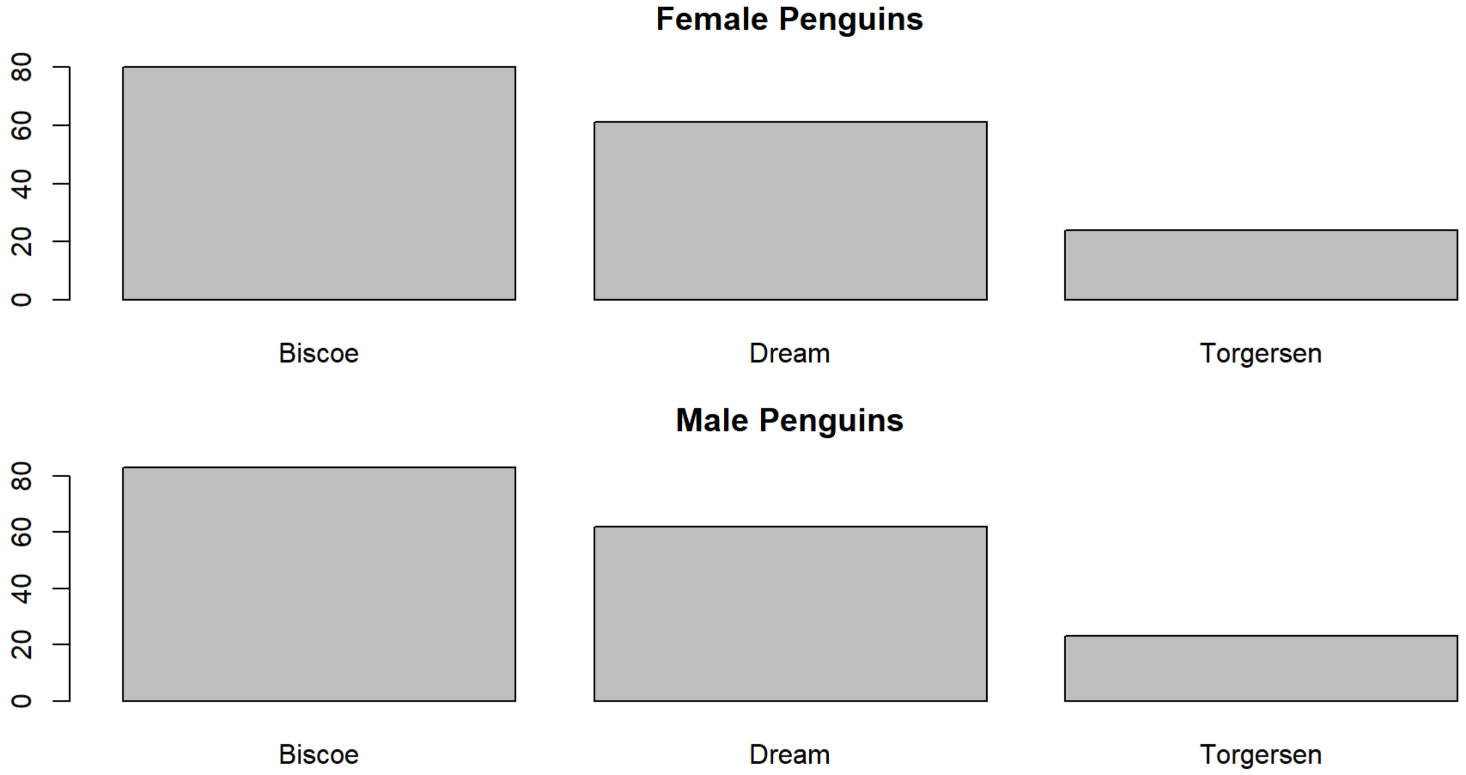
- Here's a portion of the invertebrate table from your book:
- We can also look at barplots of the columns

	Upper	Lower	Stem
Aphid	230	175	321
Bug	34	31	35
Beetle	72	23	101



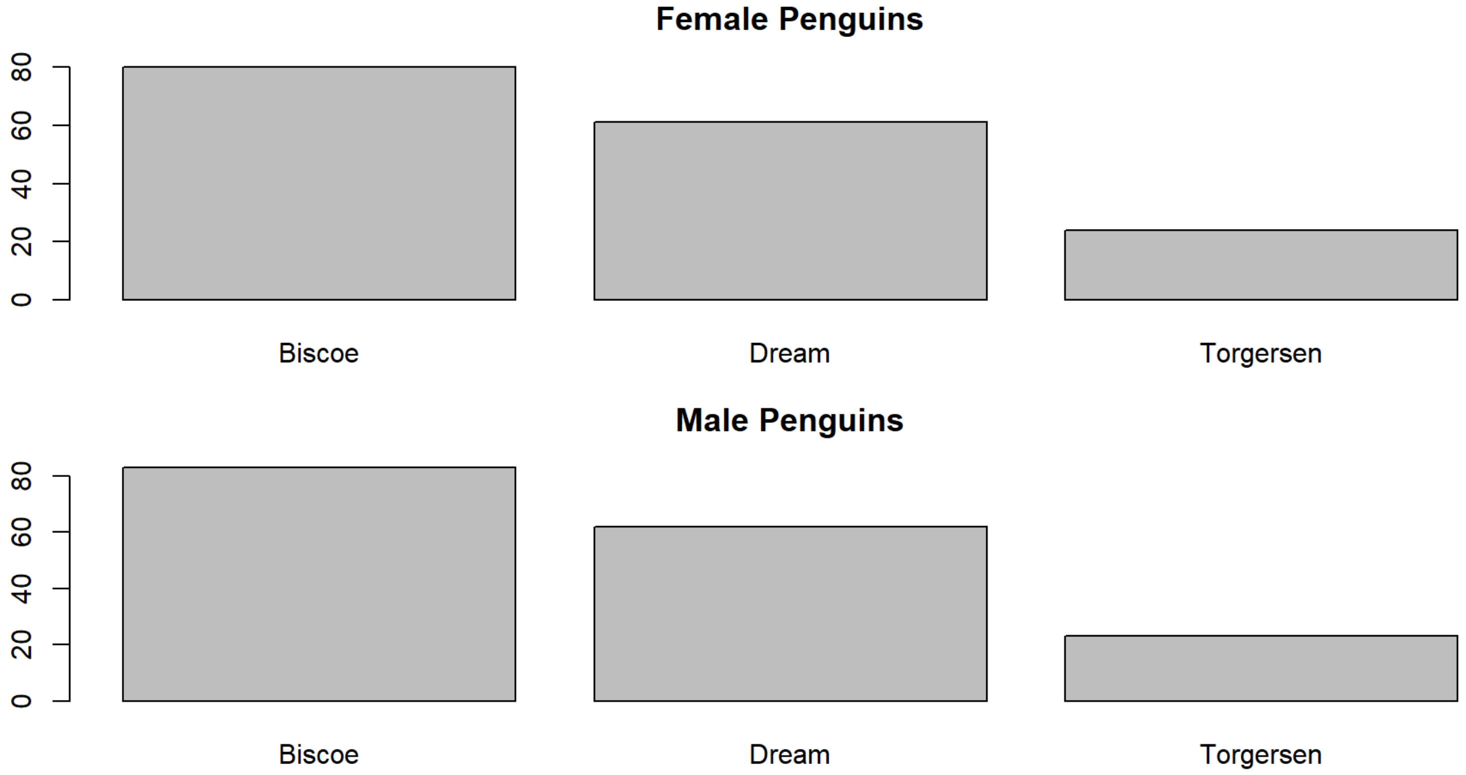
Chi-Square Tests

- Graphical exploration is a first step, but we need a test….
- The chi-square test helps us determine if there are significant differences in the proportions of different categories in rows (or columns) of a contingency table.
- Do you think the proportions are significantly different?



Chi-Square Tests

- Do you think the proportions are significantly different?
- Let's conduct a chi-square test in R to find out!



Chi-Square Tests

- Do you think the proportions are significantly different?
- Let's conduct a chi-square test in R to find out!

What is the value of the test statistic?

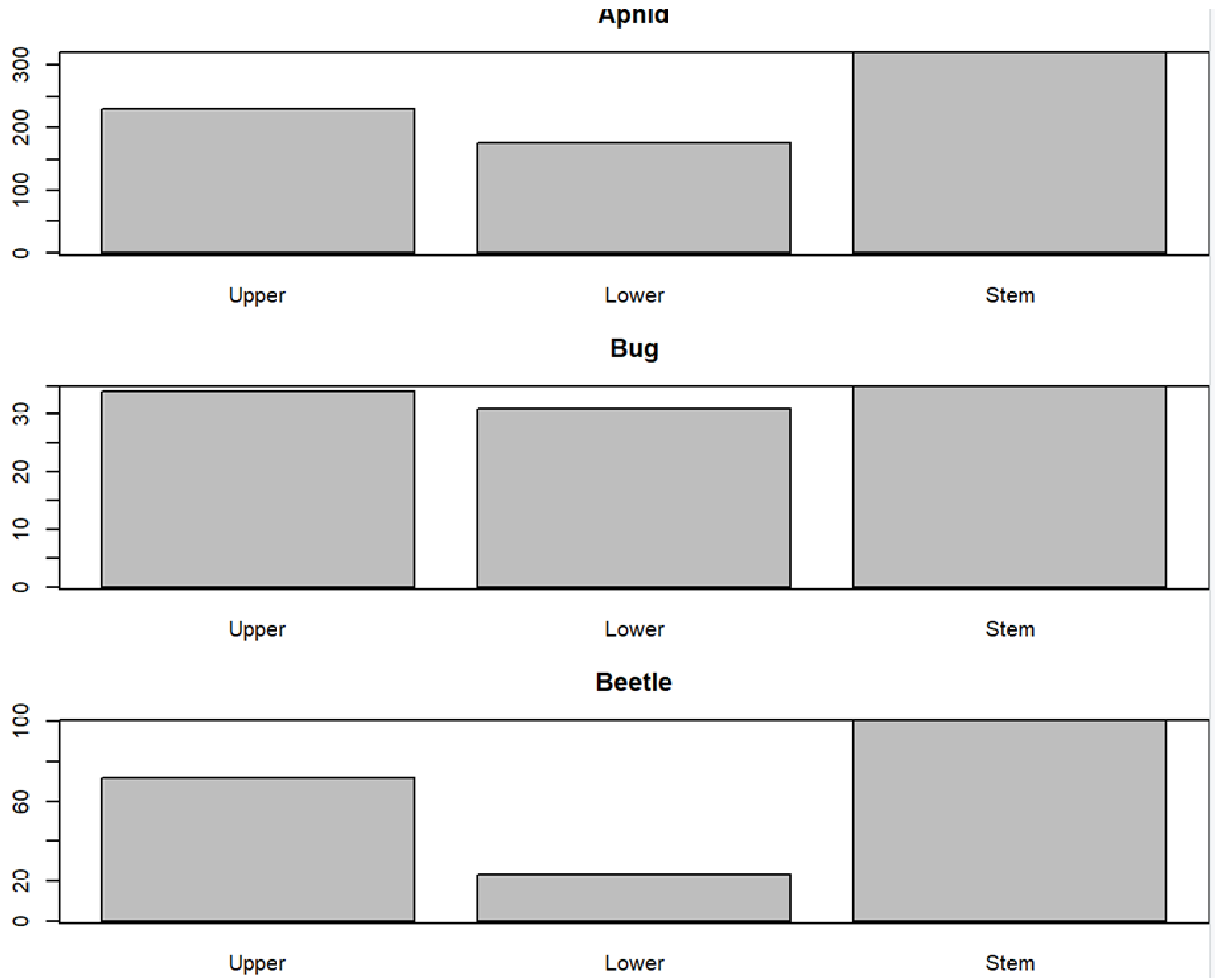
What is the p-value?

Are the proportions different?

```
        Pearson's Chi-squared test

data:  pen_table
X-squared = 0.057599, df = 2, p-value = 0.9716
```

Chi-Square Tests

- Do you think the proportions are significantly different?
- Let’s conduct a chi-square test in R to find out!



Chi-Square Tests

- Do you think the proportions are significantly different?
- Let's conduct a chi-square test in R to find out!

What is the value of the test statistic?

What is the p-value?

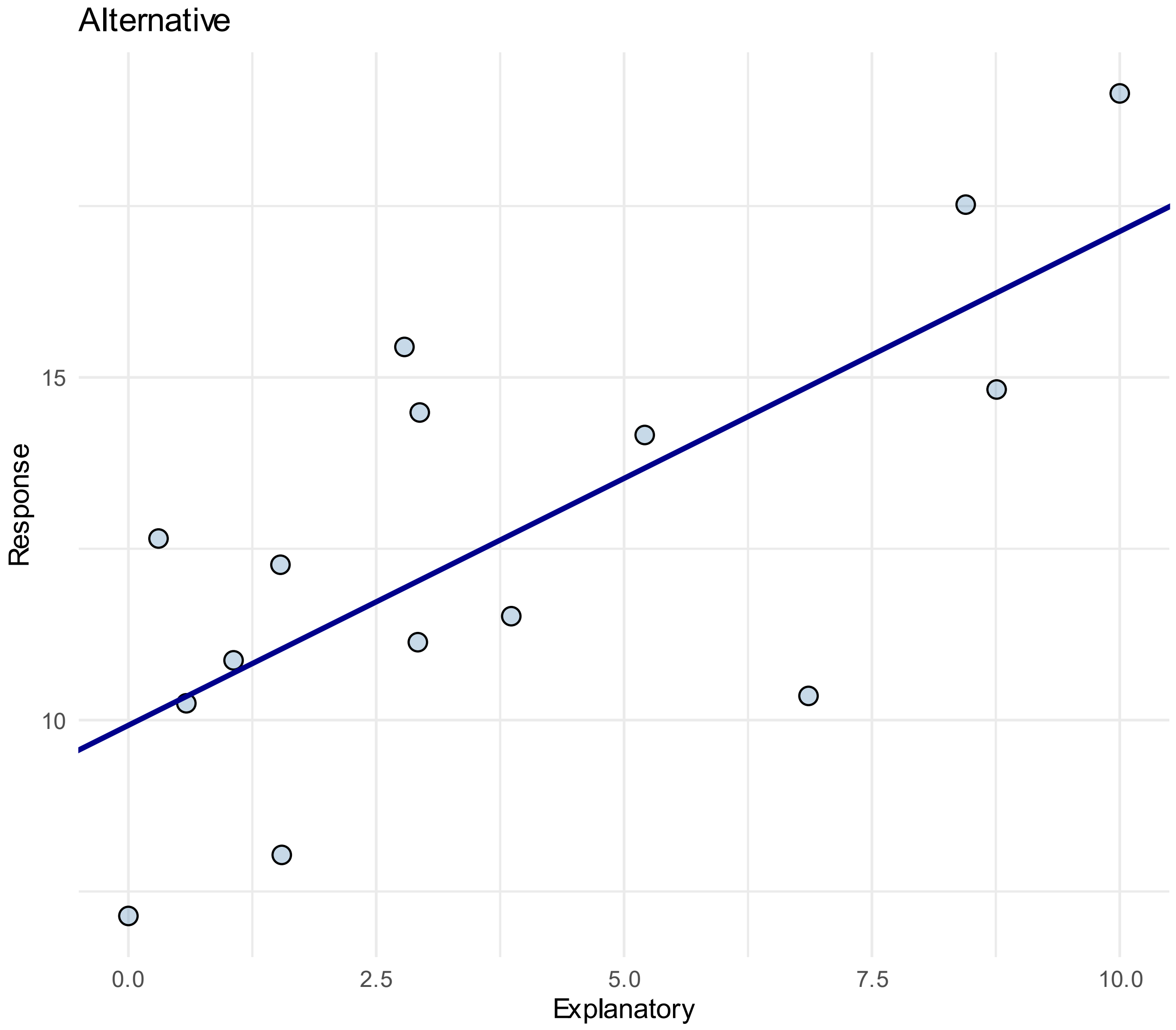
Are the proportions different?

```
        Pearson's Chi-squared test

data:  invert_table
X-squared = 19.621, df = 4, p-value = 0.0005931
```

Preview of Linear Models

Correlation and Regression

- Correlation tells us if there is a relationship, and how strong the relationship is.
  - This is not the same as the magnitude.
  - Suppose we want to know: how many units does the response variable change for a 1-unit change in the predictor variable? Correlation doesn't help with that!
- The slope of a best-fit line is the magnitude of a linear relationship



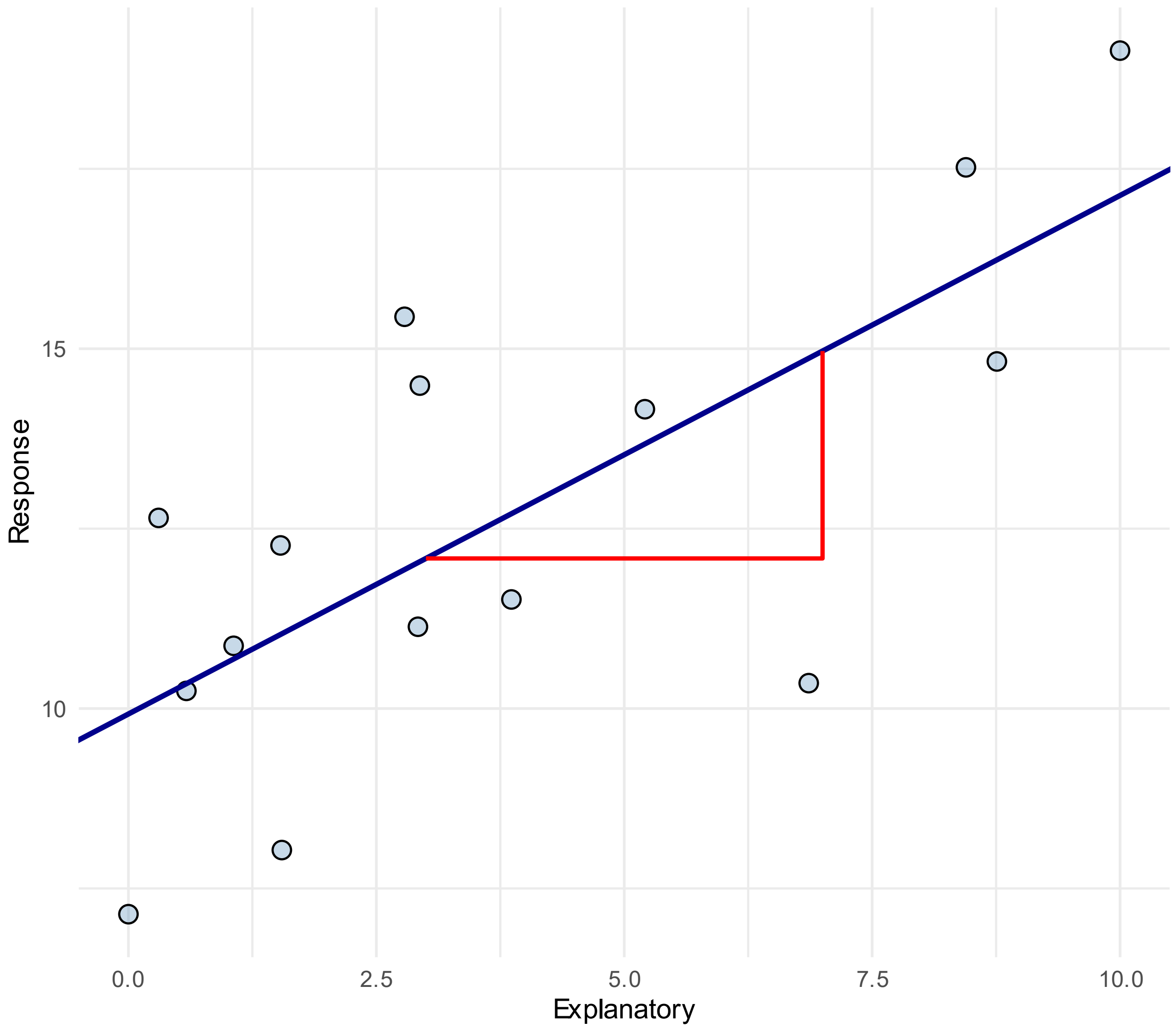
Linear equation in R

Remember the parameters of the equation of a line:

- Slope
- Intercept

To calculate the slope and the intercept of the best fit line:

- use a *linear model*
- in R use the lm() function: lm(respnse ~ explanatory)



Linear equation in R

To calculate the slope and the intercept of the best fit line:

- use a *linear model*
- in `R` use the `lm()` function: `lm(respnse ~ predictor)`

```
lm(mayfly$Abundance ~ mayfly$Speed)


Call:
lm(formula = mayfly$Abundance ~ mayfly$Speed)


Coefficients:
 (Intercept)  mayfly$Speed
       1.867         1.176
```

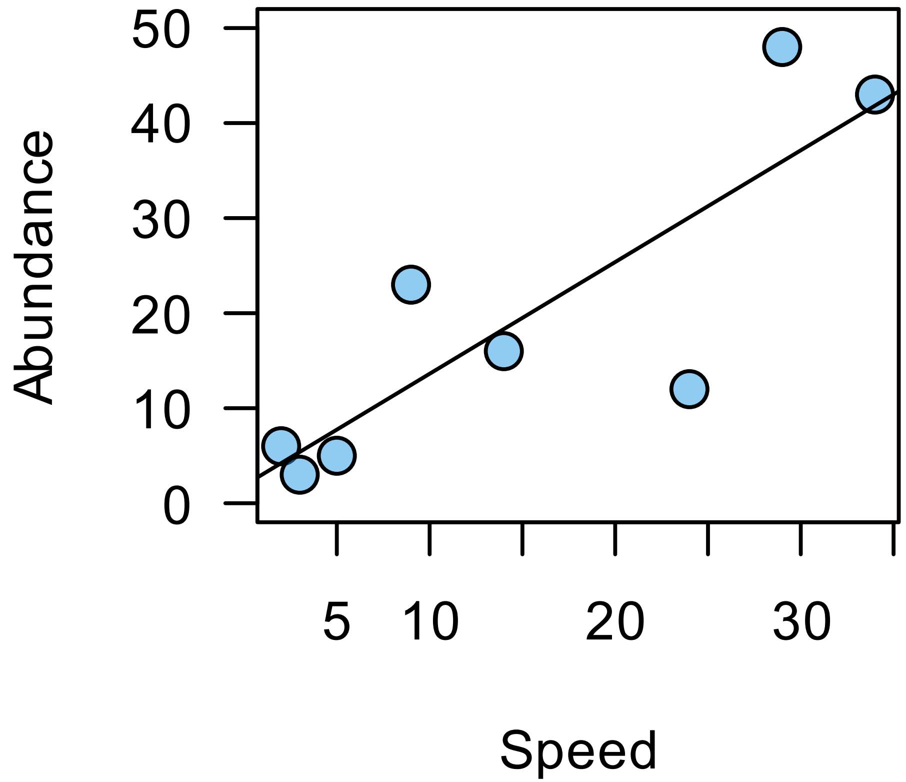
Linear model in R

To calculate the slope and the intercept of the best fit line:

- use a *linear model*
- in `R` use the `lm()` function: `lm(respnse ~ predictor)`
- Save your model object to an R variable:

```
fit_mayfly = lm(
  Abundance ~ mayfly,
  data = mayfly)
```

Note the formula notation
with data argument



Linear equation

To calculate the slope and the intercept of the best fit line:

- You can get lots of information about a fitted model object using `summary(fit_mayfly)`

```
lm(formula = mayfly$Abundance ~ mayfly$Speed)
Residuals:
    Min       1Q   Median       3Q      Max
-18.080   -2.481  -0.580    3.975   12.042
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.8667     5.7912   0.322  0.75813
mayfly$Speed   1.1756     0.3048   3.857  0.00839 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 10.05 on 6 degrees of freedom
Multiple R-squared:  0.71, Adjusted R-squared:  0.66
F-statistic: 14.87 on 1 and 6 DF,  p-value: 0.008393
```