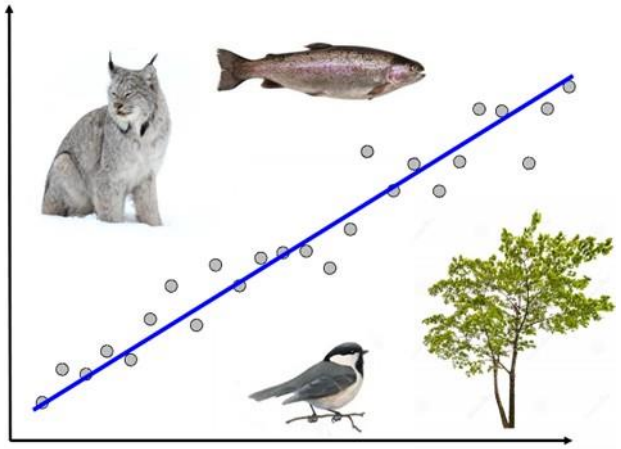
NRC 290b Introduction to Quantitative Ecology

Week 9 – Correlations and associations



Meg Graham MacLean, PhD

Department of Environmental Conservation

mgmaclean@umass.edu

2019 - Fall

This week

Monday

- Correlation
  - Pearson's product moment
- Associations
  - Chi-square

Wednesday

- Group exercise
  - Salamanders again!! Two questions:
    - Is there a correlation between SVL and Total_length?
    - Is there an association between Site and Sex?

Statistical testing

Using a t-test to test the difference between two samples, you calculate a *p*-value of 0.02. If you using a 5% significance (alpha) level – what do you conclude?

a) You reject the null hypothesis

b) You accept the null hypothesis

c) You don't reject the null hypothesis

d) You reject the alternative hypothesis

e) You don't reject the alternative hypothesis

What kind of questions do we usually ask?

When data are normally distributed…

- Differences
  - T-test, ANOVA ✓
- Correlations
  - Pearson's product moment
- Associations
  - Chi-squared

Correlations vs. Associations

What is the difference between correlations and associations?

a) Just the question you are asking and H_0
b) Correlations link continuous data and associations link categorical data
c) Associations link continuous data and correlations link categorical data
d) Nothing

Correlations

For a recent study, I looked at the effects of ash mortality due to the invasive emerald ash borer on landowner likelihood of harvesting.

What is the response variable in this study?

a) Ash mortality (due to EAB)

b) Me

c) EAB spread

d) Landowner likelihood of harvesting

Correlations

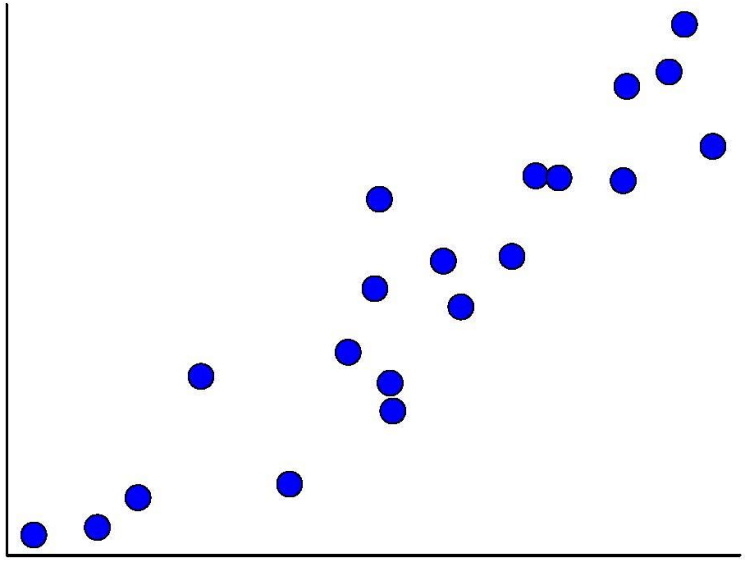
Given the graph below, what is the most likely Pearson's correlation coefficient (r)?

a) 0.90

b) 0.09

c) -0.09

d) -0.90



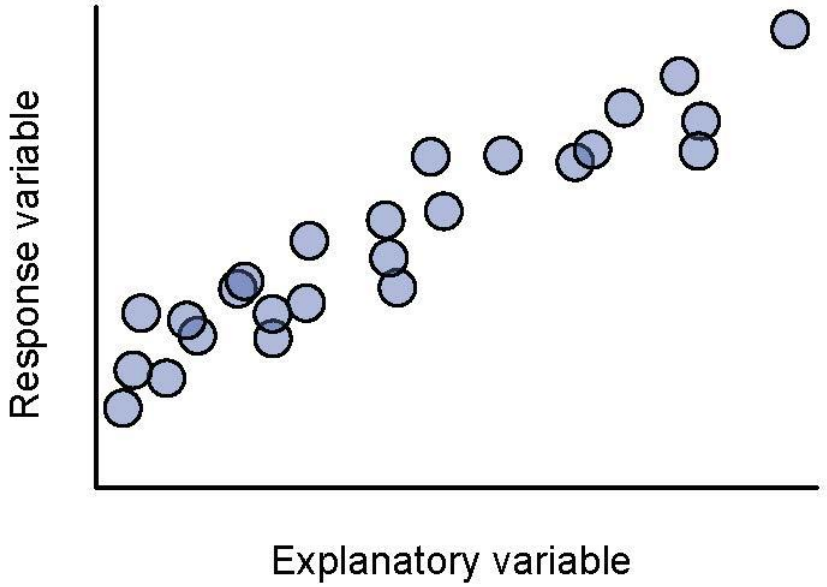
Correlations

When do we use tests for correlations?

- When we are interested in the link between two continuous samples!

What do our samples look like?

- Response variable
  - The data we are interested in explaining
  - Y-axis
- Explanatory/predictor variable
  - The data we think is influencing the response variable
  - Explains some of the variation in the response sample
  - X-axis
- We are dealing with *pairs* of values!

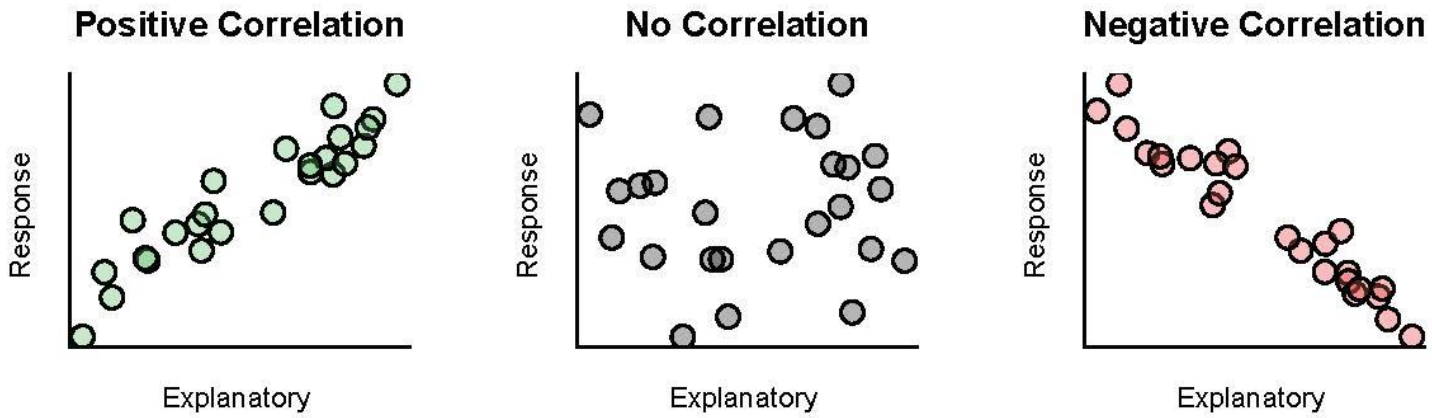


Correlation

How do we test for correlation?

For normally distributed data: Pearson's product moment (*r*)!

- H_0: there is *no significant* correlation between samples
- Assumptions:
  - Data are normally distributed
  - Relationship is linear
  - -1 = perfect negative correlation, 0 = no correlation, 1 = perfect positive correlation



Pearson's product moment

Assumes the relationship between response and explanatory variables is linear:

$$y = mx + c$$

y = response variable

x = explanatory variable

m = slope of correlation line

$$m = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

c = intercept

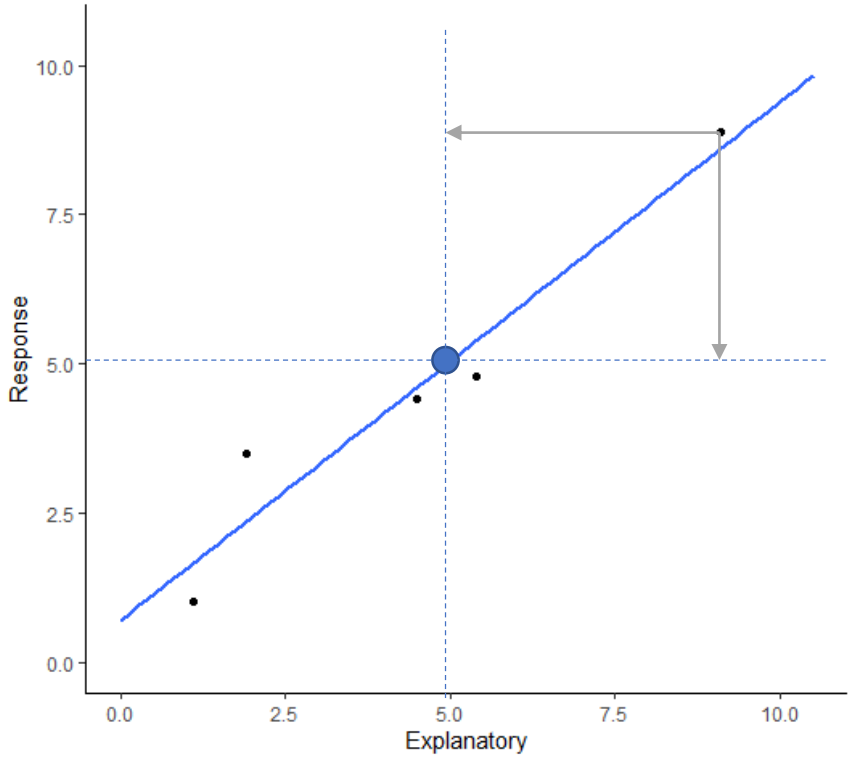
$$c = \bar{y} - m\bar{x}$$

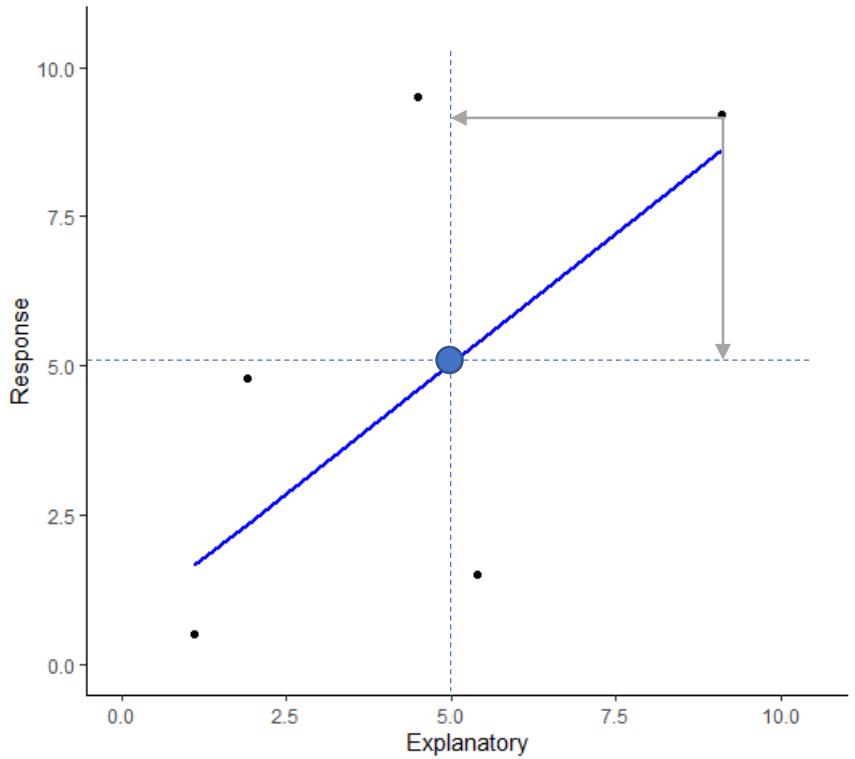
Significant correlation?

So, we can fit a straight line – but is the linear relationship significant??

Calculate Pearson's correlation coefficient:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$





Significant correlation?

So, we can fit a straight line – but is the linear relationship significant??

Calculate Pearson's correlation coefficient:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

The value of *r* indicates two things:

- Sign of the correlation
  - $-1 < r < 0$ is a negative correlation
  - $0 < r < 1$ is a positive correlation
- Strength of the correlation
  - The farther *r* is from 0, the stronger the correlation

But when can we reject H_0?

Pearson's product moment

Use a table of critical values!

- Choose a significance level (alpha)
- Determine your degrees of freedom:
  - Number of pairs-2
- Lookup your critical value

If your *r* (test statistic) is *greater* than the critical value, *reject* your null hypothesis!

Degrees of freedom	Significance	
	5%	1%
1	0.997	1
2	0.95	0.99
3	0.878	0.959
4	0.811	0.917
5	0.754	0.874
6	0.707	0.834
7	0.666	0.798
8	0.632	0.765
9	0.602	0.735
10	0.576	0.708
12	0.532	0.661
14	0.497	0.623
16	0.468	0.59
18	0.444	0.561
20	0.423	0.537
22	0.404	0.515
24	0.388	0.496
26	0.374	0.478
28	0.361	0.463

Pearson's product moment in R

OR use R!

Two tests in R for testing correlation:

- To compute Pearson's correlation coefficient (*r*):

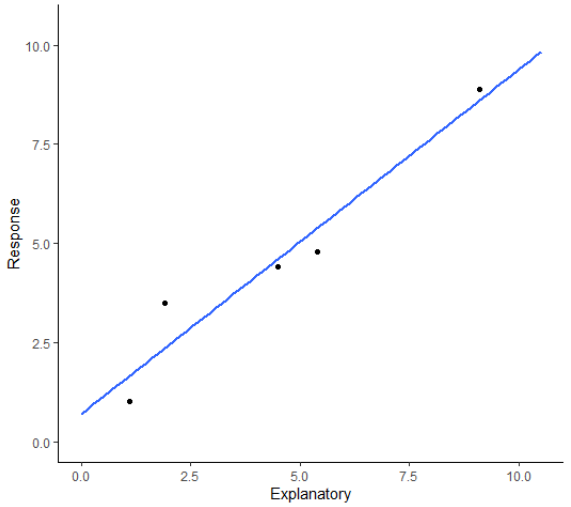
```
cor(var1, var2, method='pearson')
```

- To compute both *r* and significance test:

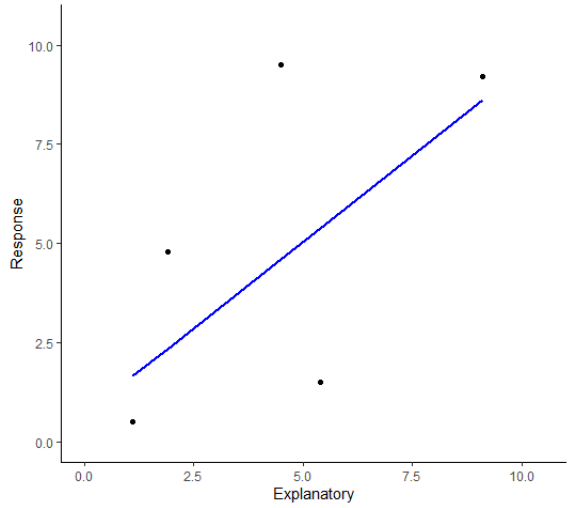
```
cor.test(var1, var2, method='pearson')
```

Pearson's product moment in R

```
close.cor <- cor.test(close$Explanatory, close$Response, method = "pearson")
> close.cor

        Pearson's product-moment correlation

data:  close$Explanatory and close$Response
t = 6.402, df = 3, p-value = 0.00772
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5596727 0.9977934
sample estimates:
      cor
0.9652952
```




```
far.cor <- cor.test(far$Explanatory, far$Response, method = "pearson")
> far.cor

        Pearson's product-moment correlation

data:  far$Explanatory and far$Response
t = 1.306, df = 3, p-value = 0.2827
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.5976823  0.9694019
sample estimates:
      cor
0.6020588
```



What happens when our data aren't normally distributed?

Spearman's rank test!

Spearman's rank test

Spearman's rank test is similar to the Pearson's product moment, but for skewed data:

- Data aren't normally distributed
  - Uses the ranks of the values rather than the values
- The relationship isn't necessarily linear!
  - But can't' be u- or n-shaped
- -1 = perfect negative correlation, 0 = no correlation, 1 = perfect positive correlation

To use the test in R:

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

Categorical data?

What if instead of continuous data – we have categorical data?

We can use a chi-square test to test for association between categorical datasets!

- H_0: there is *no significant* association between categories
- Assumptions:
  - Is non-parametric (meaning doesn't need a specific distribution of data)
  - Cells should contain frequencies or counts (not proportions) (and ≥5)
  - Levels/categories should be mutually exclusive

	Maple	Oak	Beech
Leaves off (<50%)	38	8	9
Leaves on (≥50%)	5	20	8

Chi-square

Calculate chi-square (χ^2)

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- O = observed

O	Maple	Oak	Beech	Total
Leaves off (<50%)	38	8	9	55
Leaves on (≥50%)	5	20	8	33
Total	43	28	17	**88**

- E = expected

$$E = \frac{\sum row \sum col}{\sum grand}$$

E	Maple	Oak	Beech
Leaves off (<50%)	(55*43)/88	(55*28)/88	(55*17)/88
Leaves on (≥50%)	(33*43)/88	(33*28)/88	(33*17)/88

Chi-square

Calculate chi-square (χ^2)

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- O = observed

O	Maple	Oak	Beech	Total
Leaves off (<50%)	38	8	9	55
Leaves on (≥50%)	5	20	8	33
Total	43	28	17	**88**

- E = expected

$$E = \frac{\sum row \sum col}{\sum grand}$$

E	Maple	Oak	Beech
Leaves off (<50%)	27	18	11
Leaves on (≥50%)	16	11	6

Chi-square

Calculate chi-square (χ^2)

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- $O - E$

O-E	Maple	Oak	Beech	Total
Leaves off (<50%)	11	-10	-2	0
Leaves on (≥50%)	-11	10	2	0
Total	0	0	0	**0**

- $(O - E)^2$

$(O-E)^2$	Maple	Oak	Beech	Total
Leaves off (<50%)	121	100	4	
Leaves on (≥50%)	121	100	4	
Total				**450**

Chi-square

Calculate chi-square (χ^2)

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$(O-E)^2$	Maple	Oak	Beech	Total
Leaves off (<50%)	121	100	4	
Leaves on (≥50%)	121	100	4	
Total				**450**

E	Maple	Oak	Beech
Leaves off (<50%)	27	18	11
Leaves on (≥50%)	16	11	6

- $\chi^2 = 26.7$

χ^2	Maple	Oak	Beech	Total
Leaves off (<50%)	4.6	5.2	0.2	
Leaves on (≥50%)	7.7	8.6	0.4	
Total				**26.7**

Chi-square

χ^2	Maple	Oak	Beech	Total
Leaves off (<50%)	4.6	5.2	0.2	
Leaves on (≥50%)	7.7	8.6	0.4	
Total				**26.7**

$$df = (\#\ columns\ - 1) * (\#\ rows\ - 1)$$

$\chi^2 = 26.7$
$df = 2$

$26.7 > 5.99$

We can *reject* H_0!

df	0.05
1	3.84
2	5.99
3	7.81
4	9.49
5	11.07
6	12.59
7	14.07
8	15.51
9	16.92
10	18.31

Chi-square goodness of fit

What if you already know what to expect of your data??

- You can still use chi-square with your known expected values (instead of calculating them)!
  - H_0: the observed values are *not significantly* different from the expected values

	O	E	$(O-E)^2/E$
Maple, leaf off	38	37	0.03
Maple, leaf on	5	6	1
Oak, leaf off	8	5	9
Oak, leaf on	20	23	0.4
Beech, leaf off	9	9	0
Beech, leaf on	8	8	0
Total	88	88	10.4

df	0.05
1	3.84
2	5.99
3	7.81
4	9.49
5	11.07
6	12.59
7	14.07
8	15.51
9	16.92
10	18.31

EAB

Let's say I compute an $r = 0.96$, and a $p = 0.008$ for my H_0: there is no significant correlation between EAB caused tree mortality and likelihood of salvage logging.

What does this result mean:

a) EAB tree mortality is causing the likelihood of salvage logging to go up
b) EAB tree mortality is positively associated with likelihood of salvage logging
c) EAB tree mortality is positively correlated with likelihood of salvage logging
d) EAB tree mortality is not linked to likelihood of salvage logging

Week 9 – Correlations and associations

Part II - Wednesday

Today's Exercise

Back to the redbacked salamander (*Plethodon cinereus)* data! We're trying to answer the questions:

1. Is there a correlation between SVL and Total_length?
2. Is there an association between Site and Sex?

To do so – split yourselves into 3 teams:

1. Correlation team
    1. Figure out which test you must do (Pearson/Spearman)
    2. Create a new R script to complete the test
    3. Help write your section of the report
2. Association team
    1. Set up the data so you can perform the chi-square (you will have to summarize the data!)
    2. Perform the chi-square
    3. Help write your section of the report
3. Summarization team
    1. Create a report doc and organize how it will be submitted (you don't need to write it all)
    2. Take notes today
    3. Combine the codes and make sure it runs as one
    4. Finalize report and turn it in

Today's Exercise

Back to the redbacked salamander (*Plethodon cinereus)* data! We're trying to answer the questions:

1. Is there a correlation between SVL and Total_length?
2. Is there an association between Site and Sex?

The report and code are due by TUESDAY, November 5th at midnight

The PDF report should include:

1. Which test you used for the correlation and why
2. Is there a correlation? How do you know? What does the correlation mean (in lay-person terms)? What is the slope and intercept? What does it look like (graph)?
3. Is there an association? How do you know? What does the association mean (in lay-person terms)?

The .R script should include:

1. All of your code (as one file), that I can run as-is (no extraneous typing that isn't actually run-able code!)
2. Comments about what you are doing in each part and with names and date

CATME is due by Wednesday, November 6th at midnight

For Monday:

1) Read Ch 11 in Gardner (2017) – Tests for linking several factors – *it's all regression!*
2) Answer the individual evaluation questions on moodle

All before 11:55pm on Sunday