R Stats Bootcamp

2.9 - Correlations

Megan Lewis

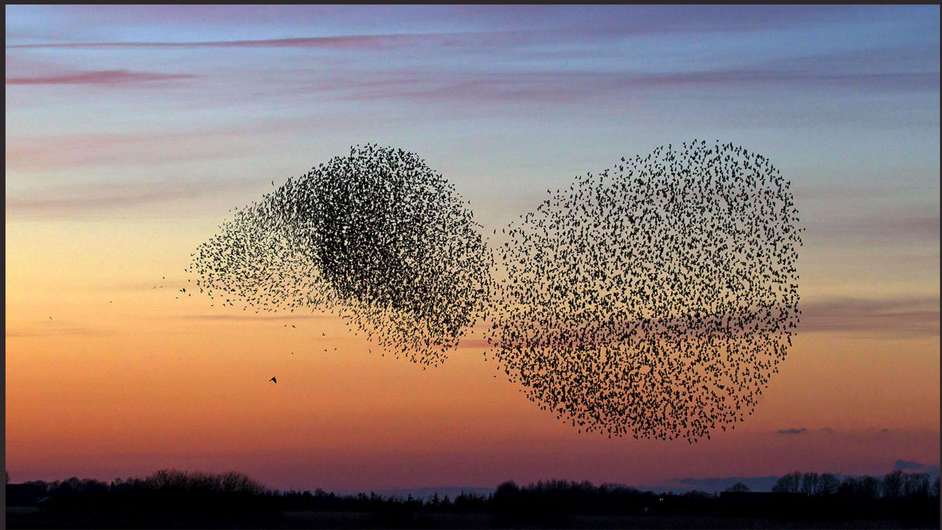
2025-02-20

R stats bootcamp - Module 2

Schedule:

- ~~Session 7: Explore data~~
- ~~Session 8: Distributions~~
- **Session 9: Correlation**
- Session 10: Regression
- Session 11: T-test
- Session 12: ANOVA

R Stats Bootcamp



Ice cream sales and forest fires are correlated because both occur more often in the summer heat. But, **correlation does not imply causation**. - Nate Silver

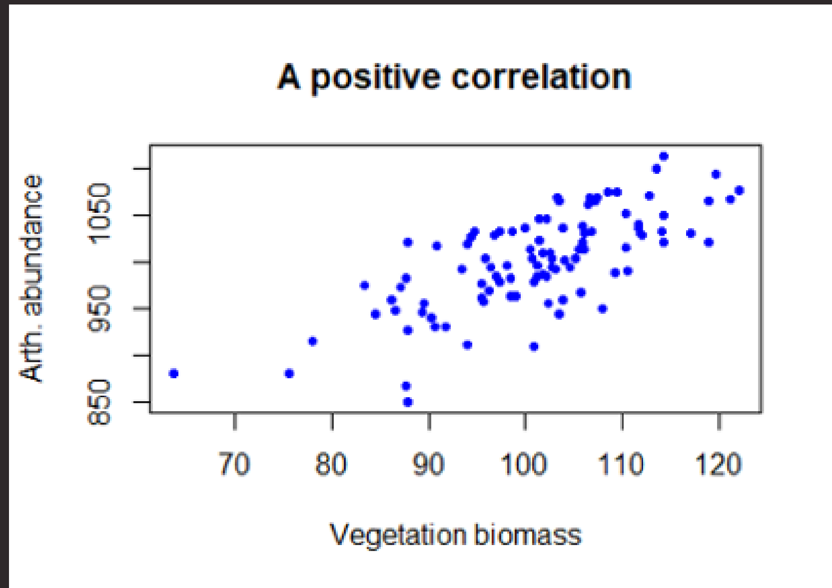
The data were formless like a cloud of tiny birds in the sky…

Session 9 objectives:

- The question of correlation
- Data and assumptions
- Graphing
- Tests and alternatives
- Practice exercises

The question of correlation

- Is there a demonstrable association between two numerical variables?
- Do they “co-vary”?
- Positive vs negative
- Strong vs weak



Flash challenge: Off to R!

Data and assumptions

- Pearson's Correlation
  - Important assumptions
    - Linear relationship between variables
    - Numeric values are Gaussian
  - Technically:
    - Covariance of two variables divided by the product of the standard deviations

Pearson's Correlation coefficient

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$n = sample\ size$$

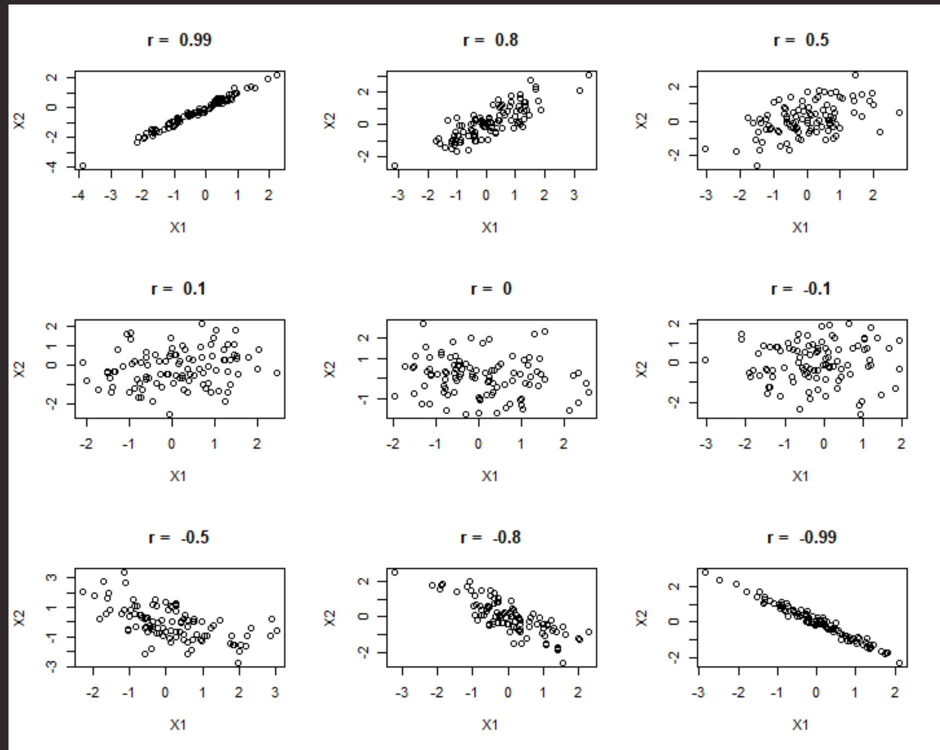
$$x_i, y_i = values\ of\ x\ and\ y\ for\ row\ i$$

Pearson's Correlation coefficient

But we can calculate this in R using the `cor()` function...

Graphing

- Useful tools to assess correlations visually
- Lots of variables
- Correlation matrices in R



Tests and alternatives

- Testing correlation coefficients
  - Null hypothesis testing; `cor.test()`
- Pearson’s assumption
  - Linear relationship
  - Bivariate Gaussian Distribution
  - Homoscedasticity (similar variance)
  - Independent observations
  - No outliers

What if my data doesn't meet assumptions

- Alternative options available
- E.g., Spearman's rank correlation, Kendall-tau etc.
- Statistical test of correlations, a process…

Results and reporting

- Think about your audience
  - Yourself
  - Others

Results and reporting - for yourself

- Comment on R script
  - Reproducible format
  - Think tidy and organised
- Can also be beneficial for colleagues, supervisors and collaborators

Results and reporting - for others

“You should NEVER PRESENT RAW COPIED AND PASTED STATISTICAL RESULTS (O.M.G!)”

- Ed Harris (Always)

Results and reporting - for others

- Format output and figures for ease of consumption
- Potential formatting options:
  - R Markdown/Quarto
  - Word processing document

Statistical summary

- Null hypothesis statistical tests
  - Test statistic (varies between tests)
  - Sample size or degrees of freedom
  - The p-value
- e.g., We found a significant correlation between petal width and length (Pearson's $r = 0.96$, $df = 148$, $P < 0.0001$).

Statistical summary

- e.g., We found a significant correlation between petal width and length (Pearson's $r = 0.96$, $df = 148$, $P < 0.0001$).
- NB:
  - Rounding of decimal accuracy
    - Usually 2, but be consistent!
  - P-value format
    - If smaller than 0.0001, then $P < 0.0001$, don't use scientific notation (no one likes that)

Flash Challenge

- Validate - histograms

Correlation alternatives to Pearson's

- Spearman's rank correlation
  - Data are ranked or otherwise ordered
  - Data rows are independent

Spearman's Rank example

- Off to R!

Practice Exercises

Practice exercise 1

- Load the `waders` data and read the help page.
- Use the pairs function on the data and make a statement about the overall degree of intercorrelation between variables based on the graphical output.

Practice exercise 2

- Think about the variables and data themselves in waders.
- Do you expect the data to be Gaussian?
- Formulate hypothesis statements for correlations amongst the first 3 columns of bird species in the dataset.
- Show the code to make three good graphs (i.e., one for each pairwise comparison for the first three columns), and perform the three correlation tests.

Practice exercise 3

- Validate the test performed in question 2.
- Which form of correlation was performed, and why.
- Show the code for any diagnostic tests performed, and any adjustment to the analysis required.
- Formally report the results of your validated results.

Practice exercise 4

- Load the `2.3-cfseal.xlsx` data and examine the information in the data dictionary.
- Analyse the correlations among the weight, heart, and lung variables, utilizing the 1 question, 2 graph, 3 test and 4 validate workflow.
- Show your code and briefly report the results.

Practice exercise 5

- Comment on the expectation of Gaussian for the age variable in the `cfseal` data.
- Would expect this variable to be Gaussian?
- Briefly explain you answer and analyse the correlation between weight and age, using our four-step workflow and briefly report your results.

Practice exercise 6

Write a plausible practice question involving any aspect of the use of correlation, and our workflow. Make use of the data from either the `waders` data, or else the `cfseal` data.