R Stats Bootcamp

2.7 - Exploring data

Megan Lewis

2024-11-28

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

Session 7 objectives:

- Question formulation and hypothesis testing
- Summarize: Weighing the Pig
- Variables and graphing
- “Analysis” versus “EDA”
- Statistical Analysis Plan: the concept
- Practice exercises

Question, explore, analyze

Weighing the pig

A dataset often comes to the Data Scientist in an imperfect state, possibly incomplete, containing errors, and with a minimal description. Likewise, it may contain wonderful knowledge, there to discover. Either way, your first task is to weigh the pig.

Understanding the data

- The first task for data analysis
- Typically involves examining the variables
- Are they what we expect? Do we need to adjust the variable types?

Exploratory data analysis

- Part practical, part philosophical
- Require skills and experience
- Subjective
- Considered important, can take large amount of time, but only reported briefly if at all

Order of operation

- Question
- Explore
- Analyze

You choose your data analysis prior to collecting the first data point.

Order of operation

- Focus on the question
- Choose an analysis that can resolve the question, given the data
- Data collection should be DESIGNED to fit the question and chosen analysis prior to collection
- Explore data to examine assumptions
- Make use of graphing, diagnostic and summary statistics

Question formulation and hypothesis testing

- Null Hypothesis Testing Framework

Question formulation and hypothesis testing

- “population of interest”
  - Cannot be directly measured
  - Too large
  - Inconvenient or impossible to measure

Question formulation and hypothesis testing

- “population of interest”
- samples and sampling
  - drawn randomly from population
  - may be subject to experimental conditions

Question formulation and hypothesis testing

- test statistics
  - Magnitude of observed differences or measured associations
  - How likely such an observed difference or association would be to observe in the absence of a hypothesized effect
  - Smaller the test statistic generally means no effect

Question formulation and hypothesis testing

- Null hypothesis
  - Consistent with no effect or difference
  - Evaluate whether to reject the null hypothesis using the P-value

Question formulation and hypothesis testing

Question formulation and hypothesis testing

Question formulation and hypothesis testing

- See bootcamp page for some links to further reading

Summarize: Weighing the pig

The best way to gain skill in handling data is to practice.

- Weighing the pig
  - Creating a summary-at-a-glance of a dataset
  - Graphics
  - Statistical summary
  - Description of how much data we have
  - Specification of variables

Chick Data Demo

Downloading and exploring chick data

Chick data

Hypothesis

The hypothesis voices “how you think the world works” or what you predict to be true.

Hypothesis

- Minimum amount of info usually interested in:
- How much data is there?
- What is the central tendency (e.g., mean, variance, etc.)?
- Are there rare values?

Hypothesis

- Typically start graphing the data
- Questions and hypotheses should inform initial peek at the data
- i.e., for chick weight data, the question is related to chick weight for each individual feed type, not overall tendency for chick weight

Hypothesis

- Shouldn’t approach as a check box ticking exercise
- Looking for details that give us insight into what the data is like
- i.e. are the mean and median close to each other? Are the standard deviation, variance or standard error of a numeric variable relative to different levels of a factor?

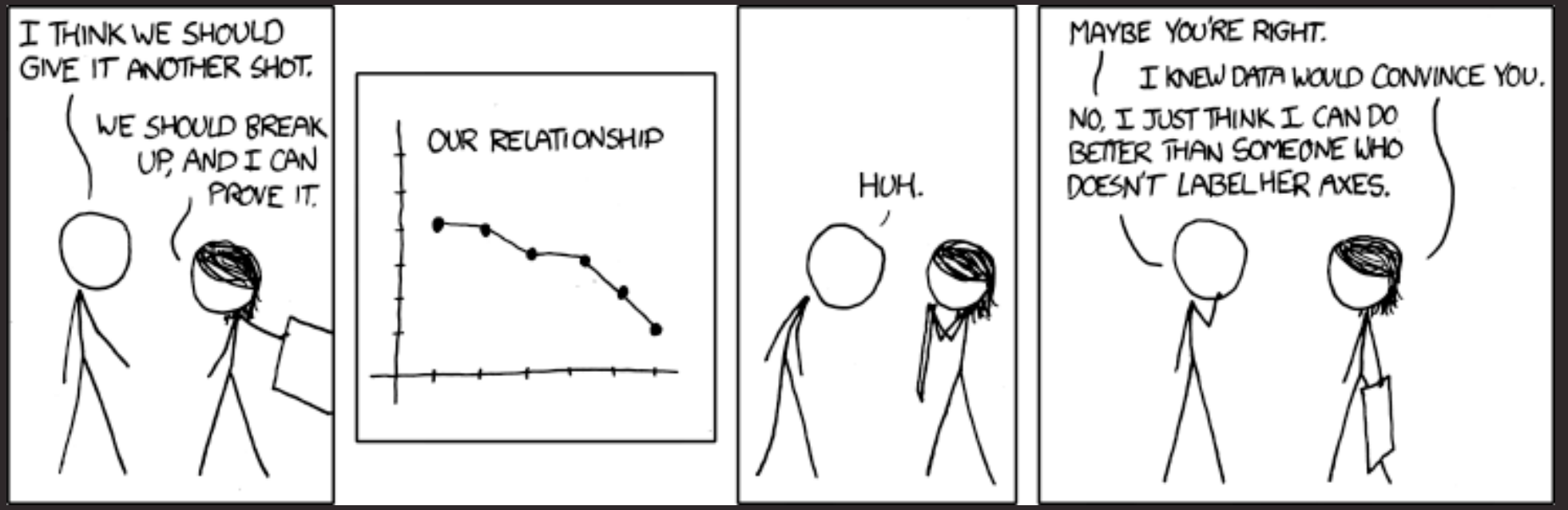
Chick Data Demo

HARUG R Stats Bootcamp by Ed Harris

Variables and graphing

Graphing

A good graph usually tells the whole story, but a bad graph is worse than no graph at all.



Scientific graphs

- Must convey relevant information
- Should be consistent in aesthetics
- Must be self contained (meaning 100% within figure and legend)
- Should reflect a hypothesis or statistical concept
- Should be appropriate to the data

Layering information

- Graphs are made up of a series of layers
  - Colours, lines, text, legends etc.
- Different approaches in R:
  - Base graphics
  - Packages like `ggplot`

Layering information

- Build features on a graph using arguments:
  - A main title with argument `main`
  - x axis title with argument `xlab`
  - Adding lines with functions `abline()` or `lines`

Types of graphs

- Choose a graph
  - Fits the data
  - Conveys the information you want to examine or showcase

Types of graphs

- For a single numeric variable:
  - Distribution with a histogram `hist()`
  - Central tendency relative to a factor with a boxplot `boxplot()`

Chick Data Demo

HARUG R Stats Bootcamp by Ed Harris

‘Analysis’ verses ‘EDA’

- Exploratory Data Analysis
  - Important part of the complete data analysis process
  - Can make a distinction between the “Analysis” part that generates evidence

Analysis

- A data analysis is:
  - Designed to fit specific question or hypothesis
  - Part of a workflow
  - Designed and formatted to present to others e.g., in a report or manuscript
  - Contains only bare essentials as relating to hypothesis
  - Strictly reproducible via script and archived data
  - Done in conjunction with EDA

EDA

- Informal, may be haphazard
- Designed to explore or gain understanding of data
- Assumptions testing
- Usually not designed to document or show to others
- Occurs primarily before (every) analysis
- May or may not be documented to be reproducible
- Done before the final evidence-generating analysis

Statistical Analysis Plan: The concept

- A formal document used to design data analysis
- Makes a formal connection between the hypothesis, the data collected and the method of analysis used
- Should occur before data collection

Statistical Analysis Plan: Components

- Hypothesis stated in plain language
- Each hypothesis translated into specific statistical model
- Specification of data and data collection methods
- Specification of effect size
- Justification of sample size (power analysis or other means)

The scientific method - classic version

(Ye Olde) Scientific Process model

The scientific process model we teach to children

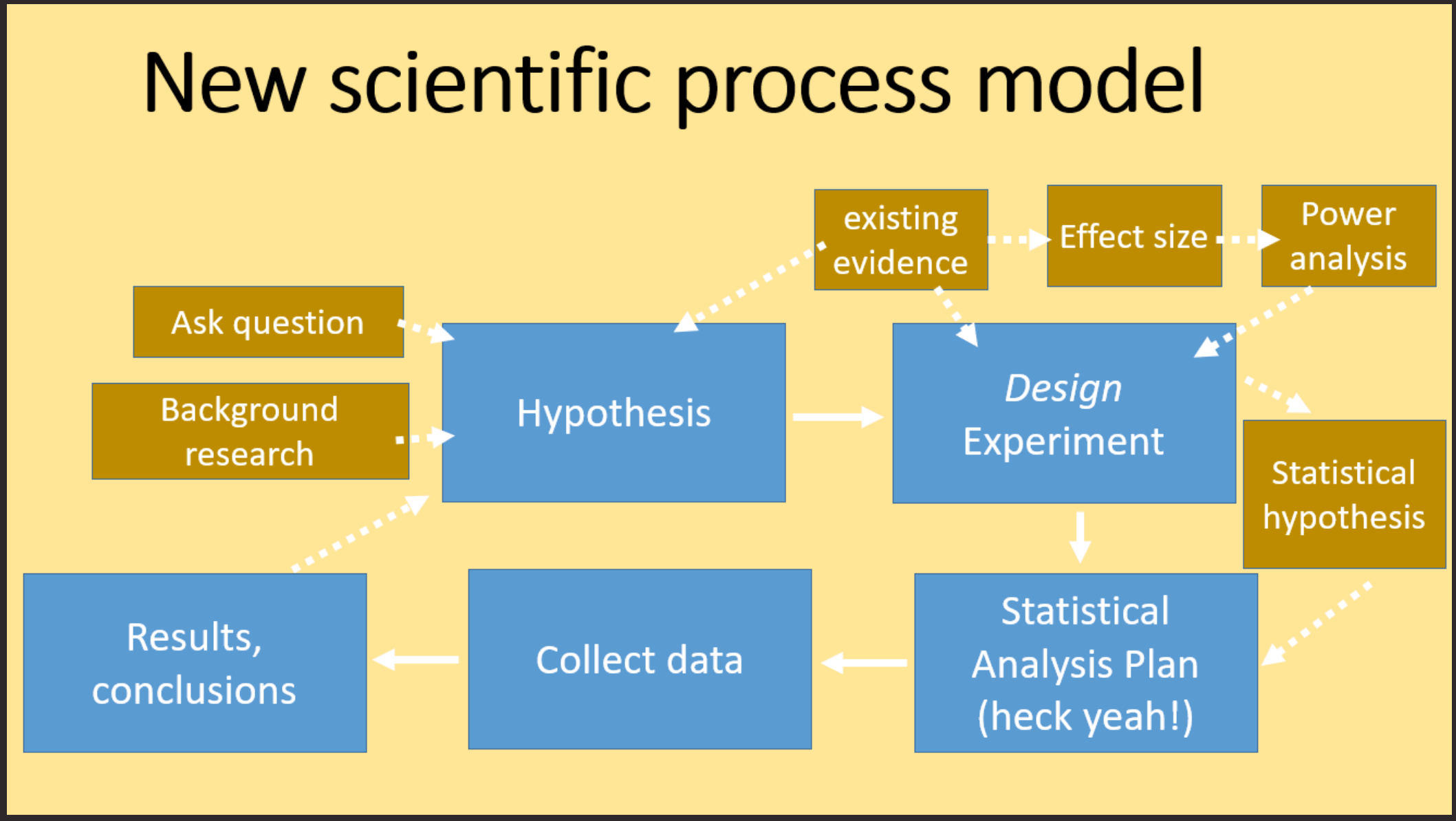
The scientific method - classic version

- Implies we plan analysis only after we conduct experiment and collect data
- Widely used across science, but considered **poor practice**!

The scientific method - classic version

- Why poor practice?
  - Expected difference or relationship (i.e., effect size) should be part of the hypothesis and should be quantified BEFORE collecting data
  - Statistical test should be chosen prior to data collection to ensure evidence matches the expectation
  - Sample size should be justified
    - Collecting too little data may lead to failing to detect a difference
    - Collecting too much data: waste of resources & ethical considerations

Best practice scientific method



Practice exercises

Practice exercise 01

Show code to set up an R analysis file with a header, table of contents, and a setup section that sets your working directory, loads any required libraries and reads in the data. Call the `data.frame` object you create `seed`.

Practice exercise 02

- `pct`, `wet` and `dry` should be numeric
- `block` and `trial` should be factors
- `treatment` should be a factor with the level “Control” set as the reference.

Show the code to do this.

Practice exercise 03

Use `aggregate()` to calculate the mean, standard deviation, standard error, and the count (e.g. `length()`) of `pct` for each level of treatment. Show the code.

Practice exercise 04

Make a fully labelled boxplot of the `pct` variable as a function of treatment.

Add a horizontal line (red and dashed) for the overall mean of `pct`, and two horizontal lines (gray, dotted) for the overall mean of pct +/- 1 standard deviation.

Practice exercise 05

(hard: may require tinkering and problem solving)

Experiment making a boxplot showing `pct ~ treatment` separated for each trial

Practice exercise 06

Write a plausible practice question involving `aggregate()` and `boxplot()` in-built R dataset `iris`.