AP Research Handbook

2019-05-22

Contents

Research Philosophy & Ethics

- No Plagiarism
- Institutional Review Board
- Data privacy standards

This is a AP Research handbook written in **Markdown**.

The **bookdown** package can be installed from CRAN or Github:

```r
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

Chapter 1

Research Question

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 1. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter 6.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

Chapter 2

Literature Review

List resources for conducting literature review. Show example of literature review with inline citations. Show ways to keep track of sources for bibliography.

- How to Write a Literature Review
  - contains example literature reviews from political science, philosophy, and chemistry.

Consider using a reference management system like Mendeley to organize your sources as you conduct your literature review. In fact, Mendeley has a Literature Search function, so you can manage sources and conduct literature reviews at the same time. See the Bibliography Management Section for more information on managing sources.

- Databases for Literature Reviews
  - Directory of Open Access Journals
    - Browse by subjects in the humanities and sciences. This can be your starting point if you have not developed a research topic.
  - arXiv
    - Open-access journal articles in fields such as mathematics, statistics, economics, physics, quantitative biology, quantiative finance, and electrical engineering
    - arXiv to BibTex: Outputs automated citations in BibTeX and other formats by typing the arXiv number of the article. For instance, just type in 1905.03758 into the search engine if the article is labeled arXiv: 1905.03758.
    - Alternatively, use Mendeley Web Importer to import article into Mendeley Desktop for automated citation outputs.
  - Mendeley Literature Search
    - Download Mendeley Desktop and register for a free account. Mendeley Desktop syncs with your online Mendeley account, but the literature search is currently only available in the desktop version.
    - Mendeley is primarily a reference managements software, so you can organize your citations as you conduct your literature review.
  - CORE
    - Search engine with the world's largest collectin of open-access research papers.
    - For batch searches of metadata and full texts, you may consider requesting a free API key to use the Core API.
  - ScienceOpen
    - Search for content, authors, collections, and journals in the advanced search, where you have the option to search by discipline or key word.
  - Dimensions
    - Search for articles in clincial sciences, biochemistry, public health, physical chemistry, and materials engineering.

  - EBSCO Open Access
    - Search open-access journals and dissertations. Note that dissertations can vary in quality, since they have not gone through peer review.
    - AP Research students should have access to a free EBSCO account from the AP Capstone program.
  - SSRN
    - Many of the social science articles are free access.
  - ERIC: Institute of Educadtion Sciences
    - Search for articles related to to education research.
    - The search engine includes the open to search for full-text articles.
  - dblp: Computer Science Bibliography
    - Index of major computer science publications.
    - Option to search for open-access articles.
  - EconBiz
    - Search for journal articles, working papers, and conference papers in economics and business.
    - Option to search for open-access articles.
  - MyJSTOR
    - You can sign up for a free MyJSTOR account to access up to six articles a month for free.
    - This may be helpful for accessing articles that are not open access.

- Tips for Accessing Paywalled Articles
  - Search for the author's website. Many researchers have draft manuscripts on their websites or research profiles on sites such as ResearchGate.
  - Consult your school's research librarian for other ways to access the article.
  - Send the author an e-mail to request for a digital copy of the article. You should provide context in the e-mail request by including a brief description of your AP Research project and its relevance and connection to the author's article.

Chapter 3

Bibliography Management

While the bibliography is placed at the end of research papers, reference management begins as soon as you begin your literature review.

- Managing Citations in LaTeX
- Mendeley Desktop Download
  - The download will prompt you to create a free account. Mendeley Desktop is synced with your online account.
- Mendeley Web Importer
  - As you search for research articles online, you can use the Mendeley Web Importer to import citations into your Mendeley Desktop. If the importer can recognize the online article's metadata, it will automatically populate the citation entries. If not, you can still enter the citation entries manually and import into the Mendeley Desktop to keep track of your sources.
- Mendeley Tutorials
- Exporting .bib files from Mendeley Desktop
- Install MS Word plugin
- Import Mendelay sources into LyX
- Import .RIS Files into Mendeley

Chapter 4

Paper Guidelines

- AP Research Proposal Guidelines
- AP Research Paper Guidelines (LaTeX)
  - draft in progress

Chapter 5

File Organization

Chapter 6

Qualitative Research Methods

- Qualitative Research Methods Field Guide

6.1 Case Study

6.2 Narrative

6.3 Phenomenological

6.4 Ethnography

6.5 Grounded Theory

Chapter 7

Quantitative Methods

Causal Inference

These notes are based on Professor Masten's online course on Causal Inference at the Social Science Research Institute at Duke.

7.0.1 Introduction

- Causal effect is often easy to detect with simple actions for which the effect immediately follows (e.g., you caused the alarm clock to stop ringing by pressing the snooze button)
- With multiple causes and delayed effects, causality is much harder to detect.
- Measurement:
  - Unit of analysis: countries, city blocks, people, firms, etc.
  - Outcome variable: the characteristic of the unit of analysis that we want to affect
  - Policy/treatment variable: the characteristic that we use to change the outcome varialbe
- A lot of characteristics cannot readily be quantified, so we often use proxy variables. For example, GDP could be a proxy for economic development.
- Causality: how an intervention in the policy variable affects the outcome variable
- Data:
  - The value of the policy variable has to vary in the dataset. Without this variation, you can't analyze how changes in the policy variable might affect the outcome variable.
  - Larger standard deviation = larger variation
- Correlation vs. Causation
  - If the policy and outcome variables are correlated, this does not necessarily imply a casual relationship.
  - Selection Problem: when units get to choose their policy variable, correlations between policy and outcome variables are unlikely to be causal.
    - ∗ Example: Neighborhoods with a lot of trees tend to have less crime.
    - ∗ If this were a casual relationship, then we could plant more trees in a neighborhood and expect crime to go down. However, this is unlikely. More likely, people who tend to commit less crimes chose to live in neighborhoods with tree-lined streets.
- Average Treatment Effect
  - Causal effects vary among people, so there is a distribution of causal effects in the population.

  - Theoretical ideal: you would know the unit level of causal effect for each person and thus make individualized treatment decisions. This is impossible in practice. You can't know the effect of receiving and not receiving treatment for an individual.
  - Unit-level causal effect: difference in outcome between treatment & control, holding all other variables fixed
  - Avg. treatment effect (ATE): avg. of all values for unit-level causal effects in a population
  - Avg. outcome under the policy: avg. outcome when everyone is affected by the policy (i.e., receives treatment)
  - Avg. outcome without the policy: avg. outcome when everyone is not affected by policy (i.e., does not receive treatment)
  - ATE = Avg. outcome under policy - Avg. outcome w/o policy

7.0.2 Experiments

- Controlled Experiments
  - Control group does not receive treatment
  - Experimental group receives treatment
  - All possible factors that could affect the outcome are identical for both groups, except for the treatment
  - Difference in the outcome between the two groups is the treatment effect
  - Typically used in hard sciences, but difficult to achieve in social sciences given the myriad of factors, many of which are difficult to measure and control
- Randomized experiments
  - Split units randomly into two large groups: treatment or control
  - Right after randomization and before the experiment, both groups should be similar (i.e., avg. values of factors should be about the same), because the split was done randomly and the groups are very large
  - Since the two groups are similar in all factors except treatment, changes in the *average* outcomes are due to treatement

7.0.3 Regression as Causality

7.0.4 Instrumental Variables

7.1 Statistical Tests

- Choosing a Statistical Test

- Hypothesis Testing Roadmap

- Chosing the Correct Statistical Test in SAS, Stata, SPSS, and R

- Uses & Misuses of Statistics

- 1 group

  - interval variables
    - 1-sample t test for the mean
    - chi-squared test for variance
  - categorical variables
    - z test for proportions (2 categories)
    - chi-squared goodness-of-fit
  - ordinal or interval
    - one-sample median test

- 2 groups (independent groups)

  - interval variables

    - ∗ 2 independent sample t-test (equal variances)
    - ∗ 2 independent sample t-test (unequal variances)
    - ∗ F test for difference between 2 variances
  - categorical variables
    - ∗ z test for difference between 2 proportions
    - ∗ chi-squared test for difference between 2 proportions
    - ∗ Fisher's exact test

- 2 groups (dependent or paired groups)
  - paired t-test (interval variables)
  - McNemar's test (categorical variables)
  - Wilcoxon signed ranks test (oridinal or interval variables)
- more than 2 groups (independent groups)
  - one-way ANOVA (for interval variables)
  - Kruskal Wallis (for ordinal or interval variables)
  - chi-squared test (for categorical variables)
- more than 2 groups (dependent groups)
  - one-way repeated measures ANOVA (for interval variables)
  - repeated measures logistic regression (for categorical variables)
  - Friedman test (for ordinal or interval)

7.1.1 1-sample t-test

- Assumptions:
  - data is a simple random sample from population
  - data follows normal distribution
  - by Central Limit Theorm, with sample size $n >= 30$, the sample mean is normally distributed regardless of the population distribution
- Two-tailed Hypothesis:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

- Test Statistic:

$$T = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}$$

  - $\overline{X}$ = sample mean
  - μ_0 = hypothesized population mean
  - S = sample standard deviation
  - $t_{(n-1)}$ = t distribution with $n-1$ degrees of freedom

7.1.2 chi-squared test for variance

7.1.3 z test for proportions

- Assumptions:
  - sample proportion $p = \frac{X}{n}$ comes from random sample in population, where X is number of events of interest in sample size n.
  - p follows a binomial distribution, but we can assume normality when X and $n - X$ are each at least 5 (old standards) or at least 15 (current standards)

- Two-tailed Hypothesis:

$$H_0 : \pi = \pi_0$$
$$H_1 : \pi \neq \pi_0$$

- Test Statistic:

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim \mathcal{N}(0,1)$$

  - π_0 = hypothesized proportion
  - p = sample proportion

7.1.4 t-test for 2 independent samples

- Assumptions:
  - two independent samples are randomly selected from two populations with the same variance
  - if you cannot use the assumption of same variance, use the Welch two-sample t-test
    * test statistic is the same as below, but degrees of freedom are adjusted
    *
  - if populations are not normally distributed, the sample sizes n_1 and n_2 from the two populations needs to be at least 30 to ensure that the distribution of the sample means are normal by the Central Limit Theorem

- Two-tailed Hypothesis:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

  - μ_1 = population mean of 1st sample
  - μ_2 = population mean of 2nd sample

- Test Statistic:

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

  - S_p = pooled variance
  - $\overline{X}_1$ = mean of 1st sample
  - $\overline{X}_2$ = mean of 2nd sample
  - S_1^2 = variance of 1st sample
  - S_2^2 = variance of 2nd sample

- More Info

- R Example
  - includes examples under both assumptions of equal and unequal variances
  - Andrew Heiss provides a brief tutorial with frequentist, simulation-based, and Bayesian approaches to comparing means between two groups. Also see Matti Vuorre's tutorial for more details.

7.1.5 paired t-test

- Assumptions:

- More Info with R Example

7.1.6 chi-squared test for proportions

- The chi-squared test for 2 x 2 frequency tables is equivalent to the square of the z-test for two proportions. See this link for detailed explanation.

7.1.7 chi-squared test for independence

- Explain connection between chi-squared test for independence and log-linear models, which are Poisson models for categorical data.

7.1.8 ANOVA

7.2 Numerical Methods

In AP Calculus, you mostly encountered problems that can be solved analytically. However, in research, many differential equation models do not have analytical forms and must be solved numerically. Matlab is often used in applied math, engineering, and physical sciences for such cases as well as other modeling applications. Octave is an open-source alternative to Matlab. While R not the first language that comes to mind for numerical methods, many numerical R packages have been developed as well as integration with Matlab, Octave, and Julia.

- Numerical Computing with Matlab
  - This site has PDF versions of Cleve Moler's textbook on numerical computing alongside a video series with lectures on differential equations and linear algebra by Prof. Gilbert Strang and computational video tutorials by Moler.
- Numerically Solving Differential Euqations with R

7.2.1 Root-Finding Algorithms

- Newton-Raphson Method Using R
- Bisection Method Using R
- Secant Method Using R

7.2.2 Numerical Solutions to Differential Equations

- Euler Method Using Matlab
- Runge-Kutta Methods

Resources by Discipline

Biology & Biostatistics

- Handbook of Biological Statistics
- An R Companion for the Handbook of Biological Statistics

Economics & Econometrics

- Introduction to Econometrics with R
- Principles of Econometrics with R
- Introduction to Data Science
- Using R for Introductory Econometrics
- Examples:
  - Annotated Sample Econometrics Paper
  - Microeconomic example of utility maximization constrained by budget lines

Psychology

- Psychology Research Methods

Public Health & Epidemiology

- Examples:
  - SIR Model Using R

Social Sciences

- Social Science Methods Modules
- Applied Causal Analysis

Chapter 8

Data

8.1 Data Sources by Discipline

8.1.1 Demography and Official Statistics

- U.S. Census Data
  - American Fact Finder
  - IPUMS
    - U.S. census microdata with social, economic, and health variables.
    - Create custom data sets or use online tool.
- UK Office for National Statistics
- Statistics Canada

8.1.2 Economics

- Panel Study of Income Dynamics
- University of Michigan Surveys of Consumers

8.1.3 Education

- Institute of Education Sciences: Data Files
- National Assessment of Educational Progress Data Explorer

8.1.4 Law

- Caselaw Access Project
  - Digital access to U.S. state and federal cases from the 1600s to present.

8.1.5 Social Sciences

- ICPSR

8.2 Data Documentation

Cite the source of your data. Provide links to the original data source and accompanying codebook, if any. Your data documentation will document your data analysis from the download of the raw data to the final steps of data analysis.

- Create a Codebook
  - List of codebook creation tools with guides and download links.
- Guide to Writing a Codebook
- How to Use R Codebook Package

Chapter 9

Analysis

- Logical Fallacies

  Read about the common fallacies in social research. Summary below:

  - fallacies of authority
  - fallacies of logic
  - fallacies of emotion

- Statistical Biases

  - Sampling bias
    - e.g., 1948 U.S. presidential election (see this case study)
    - even very large samples could have sampling biases if sampling methods are poor and unrepresentative of the population (e.g., 1936 Literary Digest Poll)
  - Omitted variable bias
  - Nonresponse bias
  - Selection bias
  - Survivorship bias
    - e.g., when bankrupt companies are removed from a stock index and replaced with profitable companies, the index would experience an upward bias. Business failures would not be accounted for in time series data.

Data Programming

- R
  - To download R, choose a CRAN mirror closest to your geographic location.
  - In order to build R packages, you should also download the latest recommended version of Rtools. Currently, the latest recommended version is `Rtoools35.exe`.
  - During the installation of Rtools, you may need to add in `"C:\Rtools\mingw_64\bin;"` to the path.
- R Studio
  - R Studio is an integrated development environment (IDE) for R. After downloading R Studio, you should be able to type the following command at the console to download some common R packages for data analysis and visualization.

```r
install.packages(c("dplyr", "tidyr", "ggplot2", "esquisse", "stats", "xtable"))
```

9.1 Cleaning and Reshaping Data

```r
library(reshape2)
library(tidyr)
library(xtable)
library(stringr)
library(knitr)
options(kableExtra.latex.load_packages = FALSE)
library(kableExtra)
library(pander)


#original data is organized by id/trial (two locations per entry)
game <- data.frame(id = c(rep("X",3), rep("Y",3), rep("Z",3)),
           trial = rep(c(1,2,3), 3),
           location_A = round(rnorm(9, mean = 0, sd = 1), 1),
           location_B = round(rnorm(9, mean = 0, sd = 1), 1))

# reshape data from wide to long (each entry is unique by id/trial/location)
game_long <- melt(game, id = c("id","trial"), value.name = "score")
game_long$variable <- str_sub(game_long$variable,-1,-1)
colnames(game_long)[3] <- "location"

# reshape data back to wide (same as original data)
game_wide <- dcast(game_long, id + trial ~ location, value.var = "score")
# reshape data into even wider form (one entry per id with 6 value columns: 2 locations X 3 trials)
```

```
game_wider <- dcast(game_long, id ~ location + trial, value.var = "score")

# using tidyr and dplyr to reshape data
game_long2 <- game %>% gather(label, score, location_A, location_B) %>%
    separate(label, c("label_p1","location"), sep = "_") %>%
    dplyr::select(-label_p1)

game_wide2 <- game_long2 %>% spread(location, value = score)

#unite() function creates the location X trial combinations first in long format # then apply the sprea
#just like in game_wide, each entry in game_wide2 is unique by id
game_wider2 <- game_long2 %>% unite(location_trial, location, trial) %>%
    spread(location_trial, value = score)

#xtable method
#print(xtable(game, caption = "Wide Data Listed by Person/Trial (Scores by Location)"), type="html")

#kable method
#kable(game, caption = "Wide Data Listed by Person/Trial (Scores by Location)", booktabs = TRUE) %>%
#    kable_styling(latex_options = c("hold_position"))

#pander method (most flexible)
pandoc.table(game, caption = "(\\#tab:wide) Wide Data Listed by Person/Trial (Scores by Location)")
```

Table 9.1: Wide Data Listed by Person/Trial (Scores by Location)

id	trial	location_A	location_B
X	1	-0.8	-0.3
X	2	0.4	1
X	3	1.1	-0.3
Y	1	-0.4	0.4
Y	2	-0.2	2.6
Y	3	-0.3	0.8
Z	1	-0.1	0.3
Z	2	0.9	-0.4
Z	3	0.2	0.5

```
pandoc.table(game_wider, caption = "(\\#tab:wider) Wider Data Listed by ID (Scores by Location/Trial)")
```

Table 9.2: Wider Data Listed by ID (Scores by Location/Trial)

id	A_1	A_2	A_3	B_1	B_2	B_3
X	-0.8	0.4	1.1	-0.3	1	-0.3
Y	-0.4	-0.2	-0.3	0.4	2.6	0.8
Z	-0.1	0.9	0.2	0.3	-0.4	0.5

```
pandoc.table(game_long, caption = "(\\#tab:long) Long Data")
```

Table 9.3: Long Data

id	trial	location	score
X	1	A	-0.8
X	2	A	0.4
X	3	A	1.1
Y	1	A	-0.4
Y	2	A	-0.2
Y	3	A	-0.3
Z	1	A	-0.1
Z	2	A	0.9
Z	3	A	0.2
X	1	B	-0.3
X	2	B	1
X	3	B	-0.3
Y	1	B	0.4
Y	2	B	2.6
Y	3	B	0.8
Z	1	B	0.3
Z	2	B	-0.4
Z	3	B	0.5

- Data Wrangling with dplyr and tidyr

9.2 Regular Expressions

- Regular Expressions in R
- Basic Regular Expressions in R Cheat Sheet

Literate Programming

9.3 LaTeX

- MiKTeX
  - First, download MiKTeX. Choose the version corresponding to your operating system (Windows, Mac, or Linux). Skip this step if you decide to use ShareLaTeX, which is an online LaTeX editor and does not require your computer to have underlying LaTeX packages via MiKTeX.
  - Recommended, download the basic installer, which will download other uninstalled packages on the fly on an as-needed basis. If you want to download all packages, you can choose the Net Installer, but this may take up a lot of space.
- Review of LaTeX Editors
  - Overleaf/ShareLaTeX
  - TeXstudio
  - LyX
- LaTeX Guides
- LaTeX Cheat Sheet
- Q and A:
  - Reference File in Parent Folder

9.4 Beamer

Beamer is a LaTeX class for presentations.

9.5 knitr (R + LaTeX)

- Using knitr in LyX
- Configure Texstudio to use knitr
- Create LaTeX Tables with kable
  - To avoid a incompatibility warning about the LaTeX `xcolor` package, place `options(kableExtra.latex.load_pac = FALSE)` in your R chunk before `library(kableExtra)`. See Hao Zhu's explanation in page 4 of the link above.
- kableExtra Vignettes
  - vignettes for using outputting tables from R into HTML, LaTeX, and Word
- xtable and stargazer Examples

- pander Tutorial

9.6 R Markdown

- Markdown Reference
- R Markdown Cheat Sheet
- Writing a Reproducible Paper in R Markdown

9.7 R Bookdown

- Authoring Books with R Bookdown
- R Markdown: The Definitive Guide
- Writing Thesis with Bookdown
  - Section on outputting into Microsoft Word using `bookdown::preview_chapter()`

Version Control

- Git
- Github
- Best Practices Using Github in RStudio
- Tutorial on Git for Behavioral Sciences