STAT107 Data Science Discovery

Lab: Hypothesis Testing

Man Fung (Heman) Leung

Fall, 2021

University of Illinois at Urbana-Champaign

Collaboration

- Please work in a group of 2–4 students
  - collaboration is important in data science!
  - meet new friends and discuss :)
  - let us know if you have any questions

Comment: Random Variable

- Common/potential mistakes
    - each reflection worth 2 points
    - 2.1: note that some implementations are more efficient (we do not grade computational efficiency but it is good to know)
    - 2.2: I let you go if you forgot to use function in 2.1 but your simulation was correct
    - 2.4: wrong possible outcome, e.g., 1 water because there are only 3 non-water pokemons
    - 3.1: some obviously wrong guess, e.g., $\hat{p}_2 > \hat{p}_3$, get 1 point deducted
    - 3.3: wrong formula, e.g., forgot to take sqrt
    - 3.4: 30 observations is a rule-of-thumb in CLT but definitely NOT sufficient to obtain accurate estimate. To achieve desired accuracy, we should perform fixed width analysis
- Running the test cases successfully do not imply full score

Project: You and Data Science

- Sample essay I wrote for another course
  - Note that I used R and my course is NOT about data science
  - Some topics are discussed on the next slide
- Recommended structure
  1. Introduction/Motivation
     - why do you want to work on this problem?
     - who has worked on this problem before?
  2. Data and Methodology
     - where/how do you get the data?
     - what are your assumptions/models/goals?
  3. Result
     - what are your findings? Are they different from previous work?
     - which model do you used/prefer?
     - how do you choose the model parameters?
  4. Conclusion/Discussion
     - what do you want to investigate in the future?

Project: You and Data Science

- Exploratory data analysis
  - visualize your data
  - handle missing data
- General-to-specific modeling
  - I have not figured out its difference from backward stepwise regression; see this post
  - however, this is quite popular in econometrics (at least I know this during my internship in the central bank of HK)
- Advanced modeling issues
  - supervised vs unsupervised
  - correlation vs causation
  - inference vs prediction

Today's lab: Hypothesis Testing

- Main page
- Retrieve the lab using git
- Complete the notebook
- Submit your work. Feel free to:
    - ask us questions
    - leave whenever you finish the lab

Checking completion

Default total number of cells: 67

- 0.0 in cell 4
- 0.1 in cell 6 (reflection)
- 1.0 in cell 9
- 1.1 in cell 12 (reflection)
- 1.2a in cell 14
- 1.2b in cell 17
- 1.2c in cell 20
- 1.3a in cell 23
- 1.3b in cell 26 (reflection)
- 2.0a in cell 28 (reflection)
- 2.0b in cell 30
- 2.1 in cell 33 (reflection)
- 2.2a in cell 35
- 2.2b in cell 38
- 2.2c in cell 41
- 2.3a in cell 44
- 2.3b in cell 47 (reflection)
- 3.0 in cell 49
- 3.1 in cell 52 (reflection)
- 3.2a in cell 54
- 3.2b in cell 57
- 3.2c in cell 60
- 3.3a in cell 63
- 3.3b in cell 66 (reflection)