STAT107 Data Science Discovery

Lab: Hypothesis Testing

Man Fung (Heman) Leung

Spring, 2022

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
- Attendance form
    - you can come up if you do not want to use this form
    - submit before you leave the lab

Comment: lab_random_variable

- Check email for score decomposition
- 2.4: -0.5 to -1 if you write down wrong possible outcomes (they should be 2, 3, 4 and 5 waters)
- 3.2: -0.5 if your formula is wrong or there is no formula for `W_mean`. Putting wrong numbers as compared with 3.1 also counts as wrong. You should avoid using raw numbers
  - no penalty if you swap `W_mean` and `W_mean_est`. However, the correct answer is to use the expected value formula but not `df["waters"].mean()` for `W_mean`
- 3.3: -0.5 if your formula is wrong or there is no formula for `W_sd`. -0.5 if you compute `W_mean_est`. I take off formula points once only in 3.2 and 3.3
  - no penalty if you write increasing the number of replications will lead to a more “normal” distribution. However, the reason behind this is complicated

Project: You and Data Science

- Due at 23:59 on 4th of May (Wed)
- Requirements
    - a non-trivial dataset (at least 200 data points)
    - some Python code
    - a pdf report with at least 1 page
        - at least 1/2 page must be text
        - line spacing of up to 1.15
        - font size up to 12
- Some modeling topics that you should think in advance
    - supervised vs unsupervised
    - correlation vs causation
    - inference vs prediction

Project: You and Data Science

- Sample essay I wrote for another course
  - Note that I used R and my course was NOT about data science
- 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?

Today's lab: Hypothesis Testing

- Main page
- Hints
  - 1.1/2.1: remember to type the hypotheses
  - 1.2b: for critical value, use `scipy.stats.norm.ppf()`. For *p*-value, use `2*(1-scipy.stats.norm.cdf())`
  - 2.2b: same as 1.2b except that *t*-distribution should be used
- Submit your work. Feel free to:
  - ask us questions
  - leave whenever you finish the lab