STAT107 Data Science Discovery

Lab: Central Limit Theorem

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

Random fact of the day

Biased sampling is very common in real life and can easily affect your analysis if you do not put in enough thoughts. Let's look at an example on US air force during WW2.

Practical experience of the day

When $n \to \infty$, the conclusions of LLN and CLT are similar. However, CLT provides variability assessment, which tells us more information. This is useful if we want to compare different estimators.

Comment: Project Mosaic

- Rubric
  - section 1.2: 1 pt
  - sections 2.4, 2.5: 2 pts each
  - sections 3, 4, 7: 5 pts each
  - sections 6, 8: 10 pts each
  - +2EC if you use your own tiles
    - I also do not know this until I see the rubric...
  - +5EC if you try something new in section 9
- Common/potential mistakes
  - undefined/redefined variables in loop (I let you go)
  - did not add up color in section 4. You still pass the test case but you will notice that the mosaic is obviously wrong (-2 pts)
  - did not upload the mosaic (-2 pts)
- Running the test cases successfully do not imply full score

Today's lab: Central Limit Theorem

- Main page
- Retrieve the lab using git
- Complete the notebook
  - 2.1: use `.sample()` and `.mean()`. Check previous labs if you forget what do they do. Remember to subset the correct column first
  - 3.1: groupby appropriate column name and aggregate with sum
- Submit your work. Feel free to:
  - ask us questions
  - leave whenever you finish the lab

Checking completion

Default total number of cells: 40

- 1.1 in cell 7
- 1.2 in cell 9
- 1.3 in cell 12
- 2.1 in cell 16
- 2.2 in cell 18–22
- 2.3 in cell 25 (reflection)
- 3.1 in cell 29–32
- 3.2 in cell 35
- 3.3 in cell 38 (reflection)