STAT107 Data Science Discovery

Lab: Justice

Man Fung (Heman) Leung

Spring, 2022

University of Illinois Urbana-Champaign

Collaboration

ILLINOIS

- 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
- Lab next week
  - optional attendance if you complete project_mosaic and notify me in advance
  - required otherwise as extra office hour

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 (survivorship bias).

Comment: lab_birthday

- Check email for score decomposition
- For similar mistakes, I take off points once only (e.g., 1.4/2.3)
- I notice some reflections are similar. I am fine with group work but make sure you write in your own words and review your answers carefully (e.g., "less likely" vs "likely" in 2.3)
- 1.2.1: no points taken off if you forget to try `birthday.myBirthday(n)`
- **1.1/2.1**: -0.5 if you use the functions in `birthday` before the question introduce them
- 1.2/1.3/2.1/2.2: -0.5 if you pass in the wrong `n` into the function
- **1.4/2.3**: -0.5 if you do not point out which of the two events has a higher probability, nor provide a numerical analysis (originally you must do both in 2.3)

Comment: lab_birthday

- 2.3: the correct answer is at least one shared birthday between anyone has a higher probability, which can be seen from the graphs
- 3.2: you should check <= 2000
- **4.4**: the question asks you to find an estimate using the histogram. -0.5 if you only compute the sample median without graphical justification, or if you make the guess in 4.5 only
- **4.5**: the question asks you to use the estimate from 4.3 (i.e., the number you guessed from the graph in 4.4). -0.5 if you fail to do so. You should also check >= but I do not take off point from this mistake (while other TAs may because they consider the situation instead of the statistical meaning here)

- Overview
  - similar to a k-nearest neighbors algorithm
  - use your own tiles (2 extra points last semester)
- Ideas for extra credit in Section 9
  - vectorization using `np.mean` (sections 4 and 6)
  - subsampling (section 7)
  - other distance measure, e.g., sum of absolute difference (section 7)
  - filter the best tile, e.g., brighter/darker (section 7)
  - increase k in k-nearest neighbors (section 7)

Today's lab: Justice

- Main page
- Hints
  - pre-lab and post-lab survey for 2 extra points each
  - if you do not pass a test case, try the simulation once more
  - 1.2: for each iteration, generate a random integer (1–100). If it is ≤ 8, increase count by 1. Repeat 100 times and return count
  - 2.3: check here (or previous lab) if you forget how to use `df.sample`
  - 3: to rigorously justify our claim, we should conduct a hypothesis test (you can ignore this comment when you do the lab. Just trying to tell you something more)
- Submit your work. Feel free to:
  - ask us questions
  - leave whenever you finish the lab