PHP2510

ASSIGNMENT #1 WEEKS 1-4

PURPOSE

Probability is the language we use to quantify and model the uncertainty of the world around us. The purpose of this assignment is to provide more practice working with random variables to calculate probabilistic outcomes. This assignment will also help you achieve more mastery of common statistical notation and showcase foundational coding skills in R.

Topics Covered: Basic Probability; Random Variables, R Coding

INSTRUCTIONS

Collaboration between students is allowed, but students should submit their own, individual work. Answers will be a mix of code and short-form responses; please include all relevant code. A templated R Markdown file is available for use (optionally).

Grading: Each question (or subquestion) below is worth 3 points, awarded as follows: 1 point for the correct answer or final results; 1 point for displaying appropriate reasoning; 1 point for showing work/code and/or clear communication.

Logistics: Per the syllabus, the problem set will be accepted up to 5 days late, with a 5% penalty for each late day. If you have extenuating circumstances that prevent you from submitting the assignment on time, please contact me **before** the original due date to arrange an exception, which will be determined on a case-by-case basis.

Feedback: Homework grades, along with any feedback, will be available within one week of the due date for all students that submitted the problem set on time. Late submissions will be graded shortly thereafter.

PROBLEM SET

In this homework assignment, we will be working with an open source dataset on births, which can be found here. Please download the dataset and familiarize yourself with the variables.

1. [9 points] Determine whether you would model each variable below using a discrete or continuous random variable. What distribution would be most reasonable? Justify your choice.
    a. premie
    b. visits
    c. mage
2. [9 points] Calculate the mean and the standard deviation of the "gained" variable. Compare the actual distribution of "gained" against a reasonably-parametrized normal distribution. How similar or different are they?
3. [9 points] Derive the variance of a bernoulli(p) random variable by hand. Calculate the variance of the "sex" variable (coded to 1/0 for male/female) using your result and using R's 'var' function. Explain your findings.
4. [6 points] What percent of babies were born more than 2 standard deviations away from the mean in terms of weeks? Write this in probabilistic notation and then calculate it in R.

5. [3 points] Generate code, using the tidyverse library and pipes, that converts the full dataset to a subset that only contains premature female babies with fathers under 40.
6. [3 points] Let's say I randomly select n observations (with replacement) from this dataset and calculate the oldest mother's age. Use simulations to determine the smallest value of n such that I find the true maximum over 50% of the time.
7. [9 points] A colleague suggests that the number of doctor visits is predictive of whether the baby will be born premature. They hypothesize that if there are 10 or more visits, there is less of a chance of the baby being premature.
    a. Does the data support that statement?
    b. What is the sensitivity and specificity of this criteria?
    c. What do you say to your colleague about their hypothesis and its strength/usefulness?
8. [3 points] Why do we model real-world data using a theoretical distribution? What does this modeling accomplish, and what are the limitations of fitting a tidy, mathematical model to messy, human data? Contextualize your answer by drawing on an example from the birth dataset.
9. [3 points] In your own words, what is the core purpose of using simulation in this homework? When would you choose to simulate a process rather than just calculating a single, definitive answer, and what unique insights does the simulation approach provide? Contextualize your answer by drawing on an example from the birth dataset.