Chris Piech
CS 109

Section #9
Dec 1, 2021

Section 9

Problem 1, 4 by Oishi Banerjee. Problem 2 by Chris Piech. Problem 3 by David Varodayan

1. **Debugging Null Hypotheses Code**

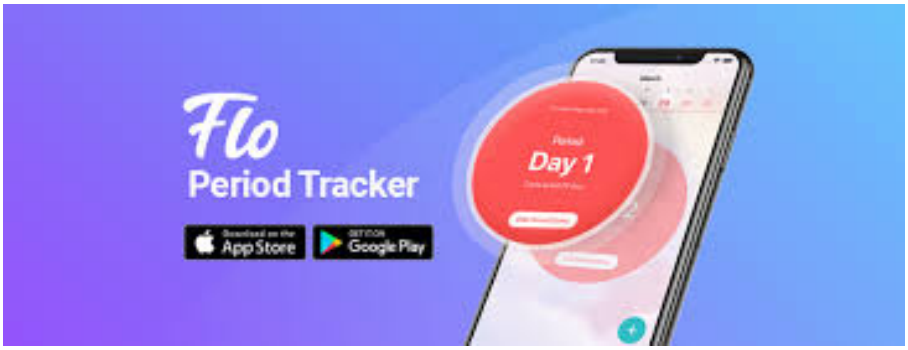
   While testing the efficacy of a new drug, Skylar Pharmaceuticals has collected 1000 data samples. Most of the samples came from patients who were treated with the drug, but the rest came from patients who received a placebo. Skylar observed that the sample mean blood pressure in the treated group was 80, while the sample mean blood pressure in the placebo group was 86.

   To demonstrate the difference is statistically significant, Skylar implemented the following to produce a p-value. The code showed that the result was not statistically significant. However! Their code is not right. Point out the errors and the corresponding fixes:

```
import numpy as np
# list_treat has blood pressures of each patient who was treated
# list_placebo has blood pressures of each patient who received a placebo
# both are ordinary 1-d numpy array
def pvalue(list_treat, list_placebo):
    whole = np.concatenate([list_treat,list_placebo])
    threshold = np.mean(list_treat) - np.mean(list_placebo)
    counter, num_trials = 0, 100000
    for trial in range(num_trials):
        sample_treat = resample(list_treat, 500)
        sample_placebo = resample(list_placebo, 500)
        mean_treat = np.mean(sample_treat)
        mean_placebo = np.mean(sample_placebo)
        new_diff = np.abs(mean_treat - mean_placebo)
        if new_diff == threshold: counter += 1
    return counter/num_trials

def resample(whole,num_samples):
    return np.random.choice(whole, num_samples, replace=True)
```

2. **Flo. Tracking Menstrual Cycles**



Let X represent the length of a menstrual cycle: the number of days, as a continuous value, between the first moment of one period to the first moment of the next, for a given person. X is parameterized by α and β with probability density function:

$$f(X = x) = \beta \cdot (x - \alpha)^{\beta-1} \cdot e^{-(x-\alpha)^2}$$

a. For a particular person, $\alpha = 27$ and $\beta = 2$. Write a simplified version of the PDF of X.

b. For a particular person, $\alpha = 27$ and $\beta = 2$. Write an expression for the probability that they have their period on day 29. In other words, what is the $P(29.0 < X < 30.0)$?

c. For a particular person, $\alpha = 27$ and $\beta = 2$. How many times more likely is their cycle to last **exactly** 28.0 days than exactly 29.0 days? You do not need to give a numeric answer. Simplify your expression.

d. A person has recorded their cycle length for 12 cycles stored in a list: $m = [29.0, 28.5, \ldots, 30.1]$ where m_i is the recorded cycle length for cycle i. Use MLE to estimate the parameter values α and β. Assume that cycle lengths are IID.

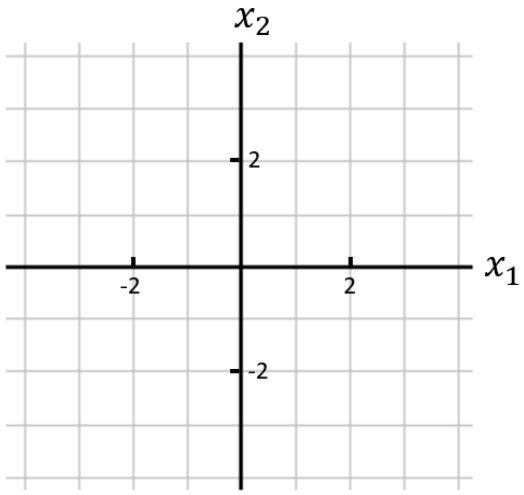
   You don't need a closed form solution. Derive any necessary partial derivatives and write up to three sentences describing how a program can use the derivatives in order to chose the most likely parameter values.

Note: Flo is a real "AI based" app that helps people track their period lengths. The real world distribution of periods is thought to be a mixture distribution between a normal and a weibell distribution [1]. This problem only has you estimate parameters for a simplified Weibull [2].

[1]: Modeling menstrual cycle length using a mixture distribution.
`https://academic.oup.com/biostatistics/article/7/1/100/243078`

[2]: Weibull Distribution.
`https://en.wikipedia.org/wiki/Weibull_distribution`

3. **Logistic regression**



Suppose you have trained a logistic regression classifier that accepts as input a data point (x_1, x_2) and predicts a class label $\hat{Y}$. The parameters of the model are $(\theta_0, \theta_1, \theta_2) = (2, 2, -1)$. On the axes, draw the decision boundary $\theta^T \mathbf{x} = 0$ and clearly mark which side of the boundary predicts $\hat{Y} = 0$ and which side predicts $\hat{Y} = 1$.

4. **Beta Exam Distributions**

Suppose hundreds of thousands (that is, a sufficiently large number) of student scores on a 150-question exam are distributed according to the following random variable:

$$R = \sum_{i=1}^{50} M_i + 0.5 \sum_{j=1}^{100} W_j \tag{1}$$

Each of the M_i are independent and identically distributed (IID) Beta random variables, $M_i \sim \text{Beta}(a_M = 10, b_M = 2)$. The W_j are separate IID Beta random variables $W_j \sim \text{Beta}(a_W = 8, b_W = 4)$, where all W_j are independent of all M_i. If we sample 100 student scores $R_1, \ldots, R_n$ IID according to the distribution of R above, what is the distribution of the sample mean $\overline{R}$?