Practical Homework 2

ECE399 - Probability Models and Inference for Engineering

April 22, 2019

In this homework, you'll play with and learn about discrete random variables as well as conditional probabilities. Throughout this homework, you're only allowed to use the random generators random.uniform and random.randint, along with the functions written in the Practical Homework 1 wherever needed. Every other random generator requires to be implemented by yourselves.

1 Part 1: Discrete Random Variables

In this section, we're going to play with some discrete random variables and will learn about some characteristics of these variables.

1.1 Coupon Collector's Problem

Remember section 2.3 part E from the Practical Homework 1. In that question you were asked to keep rolling the die until all eight numbers appear at least once, and then record the number of total rollings attempted. This problem in general is called *Coupon collector's problem*. The name is because the problem can be equivalently stated as: Given n coupons, how many coupons do you expect you need to draw with replacement before having drawn each coupon at least once? [1]

Therefore let's define the *Random Variable* X as the number of rolling attempts it takes to observe all the sides of the die at least once.

A) Similar to the previous homework, run the experiment for many times, and generate samples of the random variable X. (In the previous assignment, we asked for 1000 samples. However, here we recommend a way larger number, at least 100000).

B) For the results of part A, for each value x observed for the random variable X, count the number of times that $X = x$ appeared and divide it by the total number of experiments. For example, if you have repeated the experiment for 100000 times and observed the value 10 for 2000 times, you need to store the number $2000/100000 = 0.02$ corresponding to $X = 10$. These values will give you an estimate of the *probability* $P(X = x)$ for each x, or equivalently an estimate of the X's *probability mass function* (PMF). Store the values for all different x values observed .

C) Plot, and report the PMF for X, using matplotlib.pyplot.bar function. What is the x with the largest probability?

D) Calculate the average of all - say 100000- attempts, and report the numerical value. Compare it to $8\sum_{i=1}^{8}\frac{1}{i}$ as well as $8(\log 8 + \gamma) + 1/2$ where $\gamma \approx 0.5772156649$ is the *Euler-Mascheroni constant*.

[1] Read more:
https://en.wikipedia.org/wiki/Coupon_collector%27s_problem
http://mat.uab.cat/matmat/PDFv2014/v2014n02.pdf

1.2 Monty Hall Problem

The Monty Hall problem is based on the American television game show "Let's Make a Deal" and named after its original host, "Monty Hall".

Problem Description: Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice? [2]

A) What do you think? Is it better to switch to the third door, or remain with the current selection? Will it make any difference at all? Why do you think so? (This question won't be graded based on the correctness of the answer. Just speak your own mind!)

B) Now let's write a function in python named MontyHallProblem emulating the Monty Hall problem. Follow the steps below:
1) Emulate the doors: Create three doors, and assign a big prize to one of them **randomly**.
2) Emulate the contestant: Pick one of the doors **randomly**.
3) Reveal what's behind one of the two remaining doors: If one of the doors is the prize, remove the non-prize one. If both are non-prize, remove one of them **randomly**.
4) Moment of the final choice: The contestant keeps the current selection with a probability α and switches to the alternative door with a probability $1-\alpha$. The parameter α is given to the function as an input.
5) Return the results: "WIN" or "LOSE".

C) For each $\alpha = 0, 0.1, 0.2, \ldots, 1.0$ run the experiment for many times (such as 10000), and estimate the probability of winning as a function of α. Plot the results. What is the probability of winning if you never switch? What's the probability of winnig if you always switch? Is it in concordance with your initial guess in part A?

1.3 Birthday Paradox Problem

Suppose there are a group of n people. We're interested in examining the event in which at least two people share the same birthday and its probability as a function of n. Let's define the random variable X as follows: If in a group of n people there are people (two or more) that share the same day as their birthday, X is equal to 1, otherwise $X = 0$. Assume a year is always 365 days, hence no leap year's hassle.

A) What is the value of n for which $P(X = "1") = 0$? What is the value of n for which $P(X = "1") = 1$?

B) What is the theoretical value of $P(X = "1")$ for $n = 2$?

C) Write a function named BirthdayParadox that emulates X. This function should take the number of people n as the input, assign a day of the year from 1 to 365 to each person randomly, and then scan the results. If there are people with the same birthday it returns "1", otherwise it returns "0".

D) For each population of people $n = 1, 2, \ldots, 366$, run the function BrithdayParadox for 100000 times. (Brace yourselves, this can take a bit long!). For each n, estimate $P(X = "1")$ from simulations. Store the results.

E) Plot the $P(X = "1")$ values from the part D versus n. What trend do you see? Plot the approximate function $1 - \exp(-\frac{n(n-1)}{730})$ in the same plot. Compare the two curves.

2 Part 2: Discrete Conditional Probabilities

In this part, we will examine the conditional events and their probabilities through some simple examples.

[2] Read more: `https://www.montyhallproblem.com/`

2.1 Coin Toss - Die Rolling

Consider the following random experiment:
First we toss a coin (biased in general with parameter α), if Heads appears, we roll a four-sided die. If Tails appears, we roll an eight-sided die.

A) Write a function named CoinTossDieRolling that emulates this experiment. The function should take the parameter α as input, and needs to output a tuple containing the results of coin toss and die rolling (For example: $("H", 3)$ or $("T", 7)$).

B) For $\alpha = 0.5$ (i.e. suppose the coin toss is fair), generate 10000 samples of the experiment above.

C) Estimate the conditional probability $P("1"|"H")$: Of the 10000 experiments above, count the number of times that Heads appeared for the coin. Let's denote it by $n("H")$. Next, count the number of times that Heads appeared for the coin and "1" appeared for the die. Let's denote it by $n("H", "1")$. Calculate the fraction $n("H", "1")/n("H")$. Compare to the theoretical value for $P("1"|"H")$.

D) Repeat part C for $P("5"|"H")$, $P("1"|"T")$, and $P("5"|"T")$. Report the results.

E) Repeat part C for $P("H"|"5")$, $P("T"|"5")$, $P("T"|"1")$, and $P("H"|"1")$. Report the results.

F) Repeat parts B,C,D,E for $\alpha = 0.25, 0.75$. How do the results change compared to $\alpha = 0.5$?

2.2 Pregnancy Test

A medical lab has developed a new pregnancy test. If the test is applied to a pregnant (denoted by "P") case, it will return positive results (denoted by +) with probability α. If the test is applied to a non-pregnant case (denoted by "N") it will return negative results (denoted by -) with probability β. Suppose that of all the study population, a fraction γ are pregnant and the rest are non-pregnant.

A) Find the theoretical expressions of $P(+|"P")$, $P(-|"P")$, $P(+|"N")$, and $P(-|"N")$ in terms of α, β, and γ.

B) Find the theoretical expressions of $P("P"|+)$, $P("P"|-)$, $P("N"|-)$, and $P("N"|-)$ in terms of α, β, and γ.

C) Download and read the file "PregnancyData.csv", containing the pregnancy test data for a population of size 10000. The first column demonstrates the index of each case, the second column shows the Pregnancy/Non-pregnancy status, and the third column shows the result of the test on each subject.

D) Estimate the parameters α, β and γ from the data. Explain how you do it.

E) Substitute your estimates of part D in the formulas of parts A and B, and calculate the numerical values of each of the eight conditional probabilities.

F) Now directly estimate the conditional probabilities of parts A and B from the data (Hint: Follow the logic in section 2.1 part C). Compare the results to the numerical results of part E.