Homework1

Dr. Purna Gamage

8/18/2020

Explain your work, give concise reasoning, and . Attach R code with comments if applicable. Using Markdown is the best way to do this.

Problem 1 (5 points)

Let a be the 11th digit to the right of the decimal point of $\sin 1.23$. Let $b = \sqrt{a^4 + \pi a^3}$. Let c be the number of digits to the left of the decimal point of $e^{(\ln b)^3}$. Let $d = \sum_{j=1}^{c} j^{7/2}$. Let $e = \lfloor d \rfloor \mod 103$ where $\lfloor \rfloor$ is the floor function. Compute a, b, c, d, e.

Problem 2 (6 points)

Problem 6 in ch. 1 of **Chihara / Hesterberg**. A typical Gallup poll surveys about n = 1000 adults. Suppose the sampling frame contains 100 million adults (including you). Now, select a random sample of 1000 adults.

(a) What is the probability that you will be in this sample?

(b) Now suppose that 2000 such samples are selected, each independent of the others. What is the probability that you will not be in any of the samples?

(c) How many samples must be selected for you to have a 0.5 probability of being in at least one sample?

*Read the chapter and review basic probability. Use **R** to calculate all probabilities. For part b, first compute the probability that you will not be in any single sample. For part c, It is sufficient to give the answer as a multiple of 10,000.*

Problem 3 (10 points)

Consider the baby names data for 2014. Write a function that computes the conditional probability P(gender = F |name = XXX) for a given character string XXX and use it to compute these conditional probabilities for the 10 most common female baby names of that year. For which female baby name of the top 10 is this conditional porbability maximal? What does this mean?

Problem 4 (5 points)

The number of adult internet users who take photos or videos they have found online and post them on sites designed for sharing images with many people continues to increase. Looking only at adult internet users, aged 18 and over, about 15% are 18 to 29 years old, another 26% are 30 to 49 years old, and remaining 59% are 50 and over. The Pew Internet and American Life Project finds that 68% of Internet users aged 18 to 29 have posted photos or videos they have found online, along with 54% of those aged 30 to 49 and 26% of those 50 or older.

What percent of all adult internet users post photos or videos that they have found online?

Problem 5 (10 points)

Suppose X and Y are independent random variables that both have uniform $U(0,1)$ distributions. Consider the events

$$A = \{X \leq \frac{1}{3}\}, \quad B = \{Y < \sin \pi X\}\,.$$

Then $P(A) = \frac{1}{3}$. Use **R** and simulations to estimate $P(B)$, $P(A|B), P(B|A)$.

R *will automatically generate independent random variables in a simulation.Use "runif(n)" to simulate 100,000 Uniform(0,1) distributed random values.*

make a data frame for X and Y uniform random varibales, compute two additional columns for the events A and B as given in the problem, and estimate the relevant probabilities by subsetting.

Problem 6 (4 points)

In the data science community, there is an ongoing debate on the comparison of **R** and Python. Much of this debate is happening on the Internet. Look up some arguments for and against both Python and **R** and summarize them in no more than half a page.