Will Monroe
CS 109

Problem Set #4
July 22, 2017

Problem Set #4
Due: 12:30pm on Monday, July 31st

With problems by Mehran Sahami and Chris Piech

For each problem, briefly explain/justify how you obtained your answer. Brief explanations of your answer are necessary to get full credit for a problem even if you have the correct numerical answer. The explanations help us determine your understanding of the problem whether or not you got the correct answer. Moreover, in the event of an incorrect answer, we can still try to give you partial credit based on the explanation you provide. It is fine for your answers to include summations, products, factorials, exponentials, or combinations; you don't need to calculate those all out to get a single numeric answer.

Unless otherwise stated, you may also use functions in a library like Python's `scipy.stats` to compute values of PMFs and CDFs; if you use these, provide your code that calls these functions and explain how you arrived at each parameter to a function or constructor.

1. On average, 7.5 users sign-up for an on-line social networking site each minute. What is the probability that:
   a. More than 10 users will sign-up for the social networking site in the next minute?
   b. More than 15 users will sign-up for the social networking site in the next 2 minutes?
   c. More than 20 users will sign-up for the social networking site in the next 3 minutes?

2. Let X be a continuous random variable with probability density function:

$$f_X(x) = \begin{cases} c(2 - 2x^2) & -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

   a. What is the value of c in order for $f_X(x)$ to be a valid probability density function?
   b. What is the cumulative distribution function (CDF) of X?
   c. What is $E[X]$?

3. Say we have a cable of length n. We select a point (chosen uniformly randomly) along the cable, at which we cut the cable into two pieces. What is the probability that the shorter of the two pieces of the cable is less than 1/3 the size of the longer of the two pieces? Explain formally how you derived your answer.

4. Let X be a normal (Gaussian) random variable with $\mu = 6$. If $P(X > 9) = 0.3$, what is the approximate value of $\text{Var}(X)$?

5. The **median** of a continuous random variable having cumulative distribution function F is the value m such that $F(m) = 0.5$. That is, a random variable is just as likely to be larger than its median as it is to be smaller. Find the median of X (in terms of the respective distribution parameters) in each case below.

   a. $X \sim \text{Uni}(a, b)$
   b. $X \sim \text{N}(\mu, \sigma^2)$
   c. $X \sim \text{Exp}(\lambda)$

6. Let X_i = the number of weekly visitors to a web site in week i, where $X_i \sim \text{N}(2200, 52900)$ for all i. Assume that all X_i are independent of each other.

   a. What is the probability that the total number of visitors to the web site in the next two weeks exceeds 5000?
   b. What is the probability that the weekly number of visitors exceeds 2000 in at least 2 of the next 3 weeks?

7. The joint probability density function of continuous random variables X and Y is given by:

$$f_{X,Y}(x, y) = c\frac{y}{x} \qquad \text{where } 0 < y < x < 1$$

   a. What is the value of c in order for $f_{X,Y}(x, y)$ to be a valid probability density function?
   b. Are X and Y independent? Explain why or why not.
   c. What is the marginal density function of X?
   d. What is the marginal density function of Y?
   e. What is $E[X]$?
   f. What is $E[Y]$?

   (Hint: At some point, integration by parts may be your friend on this problem.)

8. Let X, Y, and Z be independent random variables, where $X \sim \text{N}(\mu_1, {\sigma_1}^2)$, $Y \sim \text{N}(\mu_2, {\sigma_2}^2)$, and $Z \sim \text{N}(\mu_3, {\sigma_3}^2)$.

   a. Let $A = X + Y$. What is the distribution (along with parameter values) for A?
   b. Let $B = 4X + 3$. What is the distribution (along with parameter values) for B?
   c. Let $C = aX - b^2Y + cZ$, where a, b, and c are real-valued constants. What is the distribution (along with parameter values) for C? Show how you derived your answer.

9. Say we have a coin with unknown probability X of coming up heads when flipped. However, we believe (subjectively) that the prior probability (before seeing the results of any flips of the coin) of X is a Beta distribution, where $E[X] = 0.5$ and $\text{Var}(X) = 1/36 \approx 0.0278$.

   a. What are the values of the parameters a and b (where $a, b > 1$) of the prior Beta distribution for X?
   b. Now say we flip the coin 12 times, obtaining 8 heads and 4 tails. What is the form (and parameters) of the posterior distribution of $(X \mid 12 \text{ flips resulting in 8 heads and 4 tails})$?
   c. What is $E[X \mid 12 \text{ flips resulting in 8 heads and 4 tails}]$?
   d. What is $\text{Var}(X \mid 12 \text{ flips resulting in 8 heads and 4 tails})$?

10. Say we have an array of n doubles, `arr[n]` (indexed from 0 to $n - 1$), which contains independent and identically distributed *non-negative* real values (where each value in the array is unique). What is the expected number of times that "max update" (as noted by the comment in the code) is executed in the function below (assuming the function is passed the array `arr` and its size n)? Give an expression (not a big-Oh running time) for the expectation, and explain how you derived your answer.

```
double max(double arr[], int n) {
  double max = -1;  // note: all elements in arr[] are > -1.
  for(int i = 0; i < n; i++) {
    if (arr[i] > max) {
      max = arr[i];  // max update: (max = arr[i])
    }
  }
  return max;
}
```

11. **[Coding]**[1] Below are two sequences of 300 "coin flips" (H for heads, T for tails). One of these is a true sequence of 300 independent flips of a fair coin. The other was generated by a person typing out H's and T's and trying to *seem* random. Which sequence is the true coin flips? Make an argument that is justified with probabilities calculated on the sequences.

    We'll save you a bit of time by telling you that both sequences have 148 heads, two less than the expected number for a 0.5 probability of heads. It won't be as simple as finding out which one is closer to half heads!

    Sequence 1:

```
TTHHTHTTHTTTHTTTHTTTHTTHTHHTHHTHTHHTTTHHTHTHTTHTHH
TTHTHHTHTTTHHTTHHTTHHHTHHTHTTHTHTTHHTHHHTTHTHTTTHH
TTHTHTHTHTHTTHTHTHHHTTHTHTHHTHHHTHTHTTHTTHHTHTHTHT
THHTTHTHTTHHHTHTHTHTTHTTHHTTHTHHTHHHTTHHTHTTHTHTHT
HTHTHTHHHTHTHTHTHHTHHTHTHTTHTTTHHTHTTTHTHHTHHHHTTT
HHTHTHTHTHHHTTHHTHTTTHTHHTHTHTHHTHTTHTTHTHHTHTHTTT
```

    Sequence 2:

```
HTHHHTHTTHHTTTTTTTTHHHTTTHHTTTTHHTTHHHTTHTHTTTTTTH
THTTTTHHHHTHTHTTHTTTHTTHTTTTHTHHTHHHHTTTTTHHHHTHHH
TTTTHTHTTHHHHTHHHHHHHHTTHHTHHTHHHHHHHHTHTHTTTHHTTT
THTHHTTHTTHTHTHTTHHHHHTTHTTTHTHTHHTTTTHTTTTTHHTHTH
HHHTTTTHTHHHTHHTHTHTHTHHHTHTTHHHTHHHHHHTHHHTHTTTHH
HTTTHHTHTTHHTHHHTHTTHTTHTTTHHTHTHTTTTHTHTHTTHTHTHT
```

[1] Problem 11 is solvable without code, but it would require some tedious counting. You're encouraged to copy and paste the sequences from the PDF and put your computer to good use on this one.