Math 372: Homework 06

Due Wednesday, March 4

Instructions: Do not write your solutions on this sheet—you don't have enough space to show your work. So use separate sheets of paper. For all problems you need to show your work.

Problem 1 (Variance shortcut formula)**.** Let X be a random variable, and let $\mu = \mathbb{E}[X]$. By definition of variance, we know that $\mathrm{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right]$. Show that $\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mu^2$.

Problem 2 (Geometric random variable)**.** Recall that a random variable X is said to be a ***geometric random variable*** if

$$\mathbb{P}[X=k] = \begin{cases} (1-p)^{k-1}p & : \quad k = 1, 2, \dots \\ 0 & : \quad \text{otherwise} \end{cases}$$

for some $p \in (0,1)$. The parameter p is called the "success probability".

(a) Compute $\mathbb{P}[X \le k]$ for $k = 1, 2, \dots$. *Hint:* first compute $\mathbb{P}[X > t]$.

(b) Plot the cdf of X for $p = 1/2$. *Hint:* it should be a step function.

(c) Suppose that X is a geometric random variable with success parameter $p \in (0,1)$. Show that

$$\mathbb{P}[X = m + k \mid X > m] = \mathbb{P}[X = k] \quad \text{for integers } m > 0 \text{ and } k > 1.$$

Explain why this is called the *memoryless property* of the geometric distribution.

Problem 3. The human genome consists of 3.2 billion nucleotides (A, G, C, T). For this problem we can think of the human genome as a sequence of nucleotides

`CAGGAAGCATTTATATAGAGGTAATAATTAAATTTAACTTTTTATTATTCTCTATGCCCAAGAGAATGTG...`

and the sequence consists of 3,200,000,000 nucleotide letters.

The DNA mutation rate in humans is approximately $2 \times 10^{-8} = 0.00000002$ mutations per nucleotide per generation. Assume that nucleotides mutate *independently*, that is, whether one nucleotide undergoes mutates is independent from any other nucleotides.

Let X be the number of mutations in the genome of a randomly selected newborn baby.

(a) Explain why X is a binomial random variable.

(b) What is $\mathbb{E}[X]$? (Interpretation: this is the average number of mutations in a human baby's genome).

(c) Attempt to compute some probabilities, such as $\mathbb{P}[X = 1]$ and $\mathbb{P}[X = 60]$, and $\mathbb{P}[X \le 100]$. I do not expect you to be successful with this. What makes this task difficult?

Problem 4. Let K be the number of Prussian cavalry officers killed by horsekick in a random year. Because of careful data collection by the Prussian military[1], we know that K has the following distribution:

$$p(0) = .51, \quad p(1) = .33, \quad p(2) = .11, \quad p(3) = .04, \quad p(4) = .01, \quad \text{and} \quad \mathbb{P}[K \ge 5] = 0.$$

(a) What is the expectation of the random variable K?

(b) What is the variance of K?

(c) Let $\lambda = \mathbb{E}[K]$, rounded to 1 decimal place, which you computed in part (a). Using this value, let Y be a Poisson random variable with rate parameter λ, and let $q(y) := \mathbb{P}[Y = y]$ be the pmf of Y. Compute $q(0), q(1), q(2), q(3), q(4)$, and $\mathbb{P}[Y \ge 5]$, and compare the probability mass functions p and q.

Problem 5. Let Y be a Poisson random variable with parameter λ. Find $\mathbb{E}\left[\frac{1}{Y+1}\right]$.

[1] This problem actually comes from real data collected by the Prussian army, and was one of the first known applications of the Poisson distribution, by Russian mathematician L. Bortkewitsch in 1898; see `https://www.jstor.org/stable/2348169?seq=3`

Problem 6 (Extra credit)**.** Suppose you draw a pair of cards from a standard deck of cards. Let B denote the event that both cards are aces. Let A be the event that you have at least one ace. Let S be the event that you have the ace of spaces. In Homework 4, we saw that

$$\mathbb{P}[B \mid A] = \frac{1}{33} \approx 0.03$$

and that

$$\mathbb{P}[B \mid S] = \frac{1}{17} \approx 0.06$$

This is such a strange result! Why is the second one twice as big as the first? Could we have made a mistake?

Write program using any language that estimates the conditional probability $\mathbb{P}[B \mid A]$. Also write a program that estimates conditional probability $\mathbb{P}[B \mid S]$. What estimates do you get?