DS 100/200: Principles and Techniques of Data Science **Date: Fall 2019**

Extra Probability Problems Solutions

Name:

1. (a) Let p denote the probability that a particular item A appears in a simple random sample (SRS). Suppose we collect 5 independent simple random samples, i.e., each SRS is obtained by drawing from the entire population. Let X denote the random variable for the total number of times that A appears in these 5 samples. What is the expected value of X, i.e., $\mathbb{E}[X]$? Your answer should be in terms of p.

   Solution: Let X_i denote a Bernoulli random variable equal to 1 if A appears in the i^{th} sample and 0 otherwise, $i = 1, \ldots, 5$. Then, the X_i are independent and identically distributed Bernoulli(p) random variables and $\mathbb{E}[X] = \mathbb{E}[X_1 + X_2 + X_3 + X_4 + X_5] = 5\mathbb{E}[X_1] = 5p$.

   (b) What is $Var(X)$? Again, your answer should be in terms of p.

   Solution: By the linearity property for the variance of sums of independent random variables, $Var(X) = Var(X_1 + X_2 + X_3 + X_4 + X_5) = 5Var(X_1) = 5p(1 - p)$.

2. Show that if two random variables X and Y are independent, then $Var(X - Y) = Var(X) + Var(Y)$. You may not use the fact that $Var(X + Y) = Var(X) + Var(Y)$ if X and Y are independent. Instead, use linearity of expectations and the definition of variance. *Hint*: If two random variables are independent, then their covariance is 0 and $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

   Solution:

$$
\begin{aligned}
Var(X - Y) &= \mathbb{E}[(X - Y)^2] - \mathbb{E}[X - Y]^2 \\
&= \mathbb{E}[X^2 - 2XY + Y^2] - (\mathbb{E}[X] - \mathbb{E}[Y])^2 \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - \mathbb{E}[X]^2 + 2\mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]^2 \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2 + \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 + 2\mathbb{E}[X]\mathbb{E}[Y] - 2\mathbb{E}[XY] \\
&= Var(X) + Var(Y) + 2\mathbb{E}[X]\mathbb{E}[Y] - 2\mathbb{E}[X]\mathbb{E}[Y] \\
&= Var(X) + Var(Y).
\end{aligned}
$$

3. Consider rolling (independently) one fair six-sided die and one loaded six-sided die.

   Let X_1 and X_2 denote, respectively, the number of spots from one roll of the fair die and one roll of the loaded die. Suppose the distribution for the loaded die is

$$\begin{aligned}
\Pr(X_2 = 1) = \Pr(X_2 = 2) &= \frac{1}{16} \\
\Pr(X_2 = 3) = \Pr(X_2 = 4) &= \frac{3}{16} \\
\Pr(X_2 = 5) = \Pr(X_2 = 6) &= \frac{4}{16}.
\end{aligned}$$

   Let $Y = X_1X_2$ denote the product of the two numbers of spots.

   (a) What is the expected value of Y.

   Solution:

$$\begin{aligned}
\mathbb{E}[X_1X_2] &= \mathbb{E}[X_1]\mathbb{E}[X_2] \\
&= \frac{1+2+3+4+5+6}{6} \cdot \left(\frac{1}{16} \cdot (1+2) + \frac{3}{16} \cdot (3+4) + \frac{4}{16} \cdot (5+6)\right) \\
&= \frac{7}{2} \cdot \frac{17}{4} \\
&= \frac{119}{8} = 14.875
\end{aligned}$$

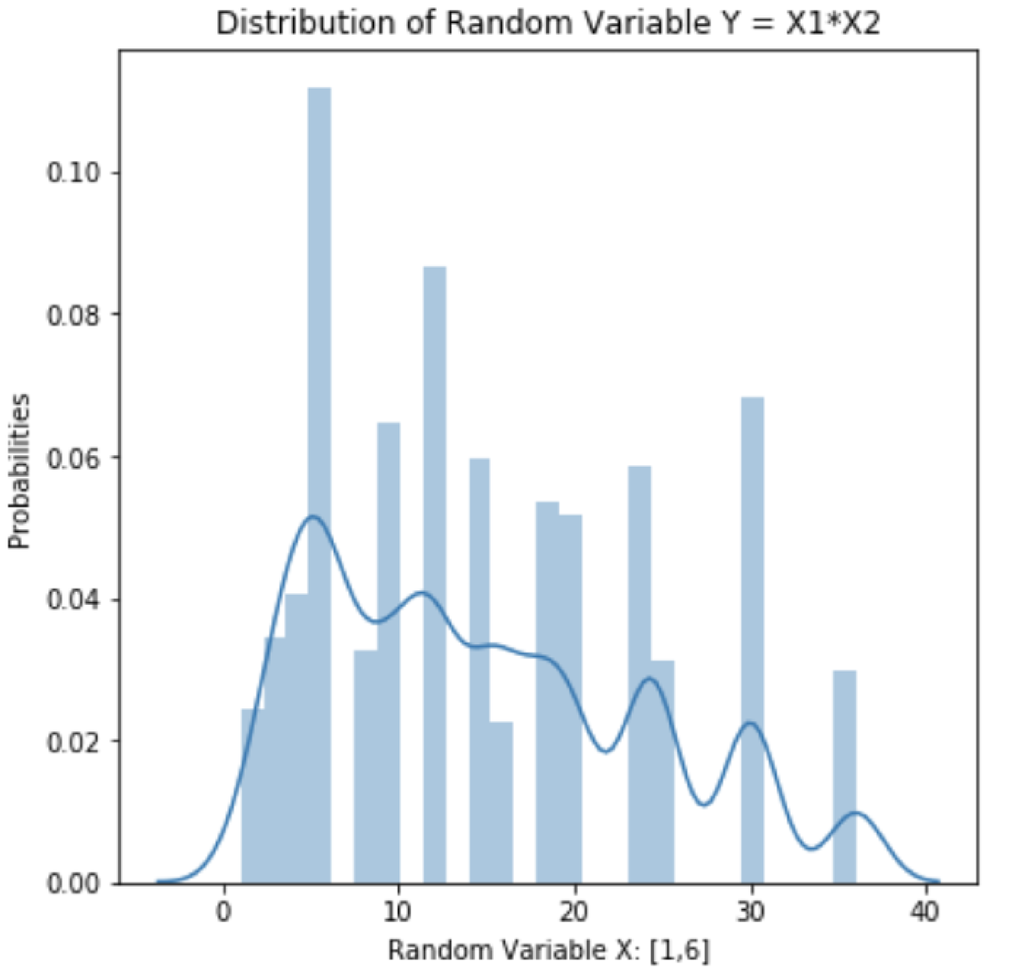
   (b) What is the variance of Y.

   Solution:

$$\begin{aligned}
Var(X_1) &= \frac{1+4+9+16+25+36}{6} - \frac{7}{2}^2 = \frac{35}{12} \\
Var(X_2) &= \frac{1}{16} \cdot (1+4) + \frac{3}{16} \cdot (9+16) + \frac{4}{16} \cdot (25+36) - \frac{17}{4}^2 = \frac{35}{16} \\
Var(X_1X_2) &= Var(X_1)Var(X_2) + Var(X_1)\mathbb{E}[X_2]^2 + Var(X_2)\mathbb{E}[X_1]^2 \\
&= \frac{35}{12} \cdot \frac{35}{16} + \frac{35}{12} \cdot \frac{17}{4}^2 + \frac{35}{16} \cdot \frac{7}{2}^2 \\
&= \frac{5495}{64} = 85.859375
\end{aligned}$$

   (c) Estimate the sampling distribution of Y by simulating 10,000 rolls of the pair of dice. Provide a graphical display of the distribution. Compare the mean and variance from this estimate to the values you computed above.

Solution: The plot should look something like:



The mean and variance from the sampling distribution should be similar to the values calculated in the previous two parts.