DS 100/200: Principles and Techniques of Data Science **Date: March 8, 2019**

Discussion #6 Solutions

Name:

Probability

1. Suppose John visits your store to buy some items. He buys toothpaste for \$2.00 with probability 0.5. He buys a toothbrush for \$1.00 with probability 0.1. Let the random variable X be the total amount John spends. Find $\mathbb{E}[X]$.

Solution: Let $X_{\text{toothpaste}}$ be the amount John spends on toothpaste, and $X_{\text{toothbrush}}$ be the amount John spends on a toothbrush.

From the linearity of expectation, we have:

$$\mathbf{E}[X] = \mathbf{E}[X_{\text{toothpaste}} + X_{\text{toothbrush}}] = \mathbf{E}[X_{\text{toothpaste}}] + \mathbf{E}[X_{\text{toothbrush}}]$$

We know that $\mathbf{E}[X_{\text{toothpaste}}] = (0.5)(0) + (0.5)(2) = 1$, and $\mathbf{E}[X_{\text{toothbrush}}] = (0.9)(0) + (0.1)(1) = 0.1$. Thus, $\mathbf{E}[X] = 1.1$.

2. Suppose we have a coin that lands heads 80% of the time. Let the random variable Y be the *proportion* of times the coin lands heads out of 100 flips. What is Var$[Y]$?

Solution: Let X_i be the outcome of the i^{th} flip. If the i^{th} flip lands heads then we say $X_i = 1$ and otherwise $X_i = 0$. Let random variable Y be the *proportion of times* X_i lands heads:

$$Y = \frac{1}{100} \sum_{i=1}^{100} X_i$$

We can compute the variance of Y using the following identities:

$$\begin{aligned}
\textbf{Var}\left[Y\right] &= \textbf{Var}\left[\frac{1}{100}\sum_{i=1}^{100} X_i\right] && (1)\\
&= \frac{1}{100^2}\textbf{Var}\left[\sum_{i=1}^{100} X_i\right] && \text{(Squared variance of constant multiple.)}\\
&= \frac{1}{100^2}\sum_{i=1}^{100} \textbf{Var}\left[X_i\right] && \text{(Ind. Variables implies linearity of var.)}\\
&= \frac{1}{100^2}\sum_{i=1}^{100} p(1-p) = \frac{p(1-p)}{100}\\
&= \frac{.8(1-.8)}{100} = \frac{.16}{100} = .0016
\end{aligned}$$

3. Let X be a random variable with mean $\mu = \mathbb{E}[X]$. Using the definition $\text{Var}(X) = \mathbb{E}[(X-\mu)^2]$, show that for any constant c,

$$\mathbb{E}[(X-c)^2] = (\mu - c)^2 + \text{Var}(X).$$

Solution: One way to show this is to write $X - c = X - \mu + \mu - c$. Squaring both sides,

$$\mathbb{E}[(X-c)^2] = \mathbb{E}[(X-\mu)^2 + (\mu-c)^2 + 2(X-\mu)(\mu-c)]$$

Now using linearity of expectation and pulling out the constants,

$$\begin{aligned}
\mathbb{E}[(X-c)^2] &= \mathbb{E}[(X-\mu)^2] + (\mu-c)^2 + 2\underbrace{\mathbb{E}[X-\mu]}_{=0}(\mu-c)\\
&= \text{Var}(X) + (\mu-c)^2.
\end{aligned}$$

4. Use the above result to prove that

   - $\text{Var}(X) \leq \mathbb{E}[(X-c)^2]$ for any c
   - $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

Solution: The first bullet follows from using $(\mu - c)^2 \geq 0$, and the second bullet follows from plugging in $c = 0$ (as the equation holds for all c).

Alternatively, we can also show the first bullet by finding the value for c that minimizes the RHS.

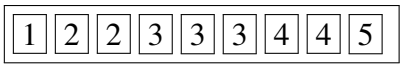
$\mathbb{E}[(X - c)^2] = \mathbb{E}[X^2 - 2cX + c^2] = \mathbb{E}[X^2] - 2c\mathbb{E}[X] + c^2$

Taking a derivative wrt c and setting the expression to 0, we get:

$-2\mathbb{E}[X] + 2c = 0$, so $\hat{c} = \mathbb{E}[X]$.

$\hat{c}$ minimizes the RHS, and if we plug it into the RHS, we get the definition of variance. Therefore Var(X) is a lower bound for the RHS.

5. We roll a die 9 times and record the value of each roll on slips of paper. Then, we place all 9 slips of paper in a box:



The numbers in the box have the following summary statistics:

Statistic	Sum	Sum of Squares	Mean	Median
Value	27	93	3	3

For each of the following, answer the following questions: Is this value calculable from the information given? If so, either calculate it by hand or describe how you would calculate this value. If not, then suggest an estimate for the quantity. All draws are with replacement.

(a) The expected value of a single draw from the box

Solution: $\mathbb{E}[\text{Single draw}] = \text{Average of the Box} = \dfrac{\text{Sum of Tickets}}{\text{Number of Tickets}} = \dfrac{27}{9} = 3$

(b) The expected value of the average of nine draws from this box

Solution: $\mathbb{E}[\text{Average of nine draws}] = \text{Average of the Box} = 3$
If X_i is the value of the ith draw, $\mathbb{E}[\frac{1}{9}\sum_{i=1}^{9} X_i] = \frac{1}{9}\sum_{i=1}^{9}\mathbb{E}[X_i] = (\frac{1}{9})9\mathbb{E}[X_1] = 3$.

(c) The exact variance of the tickets in the box

Solution: Using $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, we can use the sum of squares above to compute $\mathbb{E}[X^2] = \frac{93}{9}$. Therefore, $\text{Var}[X] = \frac{93}{9} - 3^2 = \frac{4}{3}$.

(d) The exact variance of the average of nine draws from the box

Solution: $\text{Var}[Y] = (\frac{1}{9^2})\text{Var}[\sum_{i=1}^{9} X_i] = (\frac{1}{81})9\text{Var}[X_1] = \frac{\frac{4}{3}}{9} = \frac{4}{27}$

Modeling

6. We wish to model exam grades for DS100 students. We collect various information about student habits, such as how many hours they studied, how many hours they slept before the exam, and how many lectures they attended and observe how well they did on the exam. Propose a model to predict exam grades and a loss function to measure the performance of your model.

Solution: Example solution: Let x_1, x_2, x_3 correspond to hours studied, hours slept and number of lectures attended respectively. Let y be their score on the exam.

$$f(x) = \theta_1 * x_1 + \theta_2 * x_2 + \theta_3 * x_3$$
$$L(\theta, x, y) = (f(x) - y)^2$$

7. Suppose we collected even more information about each student, such as their eye color, height, and favorite food. Do you think adding these variables as features would improve our model?

Solution: These features are most likely not going to improve our model. This problem is meant to emphasize overparameterization/overfitting using too many features that do not contribute to the performance of the model. Overfitting, bias variance and feature engineering will be discussed later on in the semester.