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

Discussion #6

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 $\mathbb{E}[X]$.

   B: \$ brush, P: \$ paste
   $X = P + B$
   $E[X] = E[P] + E[B]$
   $= 1 + 0.1 = \boxed{1.1}$
   $E[P] = 2 \cdot 0.5 + 0 \cdot 0.5 = 1$
   $E[B] = 1 \cdot 0.1 + 0 \cdot 0.9 = 0.1$

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 $\text{Var}[Y]$?

   at bottom

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,

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

   at bottom

4. Use the above result to prove that

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

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:

   $\boxed{1}\,\boxed{2}\,\boxed{2}\,\boxed{3}\,\boxed{3}\,\boxed{3}\,\boxed{4}\,\boxed{4}\,\boxed{5}$

   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

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

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

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

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.

f : information $\longrightarrow$ prediction

$f(x) = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$ $\theta_i \in \mathbb{R}$

x_1 : hours studied

x_2 : hours slept

x_3 : lectures attended

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?

2.

$$X_i = \begin{cases} 1 & \text{if flip i is heads} \\ 0 & \text{'' '' tails} \end{cases} \qquad p = 0.8$$

"indicator r.v."

$$Y = \frac{1}{100} \sum_{i=1}^{100} X_i \qquad X = \sum_{i=1}^{100} X_i \Rightarrow X \sim Bin(100, 0.8)$$

$$\begin{aligned} var(Y) &= var\left(\frac{1}{100} \sum_{i=1}^{100} X_i\right) \\ &= \left(\frac{1}{100}\right)^2 var\left(\sum_{i=1}^{100} X_i\right) \qquad var(\alpha X) = \alpha^2 var(X) \\ &= \left(\frac{1}{100}\right)^2 \sum_{i=1}^{100} var(X_i) \qquad \text{each } X_i \text{ independent} \\ &= \left(\frac{1}{\cancel{100}}\right)^2_{100} \cdot \cancel{100} \cdot var(X_1) \\ &= \frac{1}{100} \cdot var(X) \\ &= \frac{1}{100} \cdot p \cdot (1-p) \\ &= \frac{0.8 \cdot 0.2}{100} = \frac{0.16}{100} = \boxed{0.0016} \end{aligned}$$

Using binomial

$$X = \sum_{i=1}^{100} X_i \qquad X \sim Bin(100, 0.8)$$

$$Y = \frac{1}{100} X$$

$$var(Y) = var\left(\frac{1}{100} X\right) = \left(\frac{1}{100}\right)^2 var(X) = \frac{\cancel{100} \cdot 0.8 \cdot 0.2}{\cancel{100^2}\ 100} = \frac{0.16}{100}$$

3.

RTP $E[(X-c)^2] = (\mu-c)^2 + var(X)$

$X-c = (X-\mu) + (\mu-c)$

$(a+b)^2 = a^2+2ab+b^2$

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

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

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

$$= var(X) + 2(\mu-c)E[X-\mu] + (\mu-c)^2$$

$$= var(X) + 2(\mu-c)(\underbrace{E[X]-\mu}_{0}) + (\mu-c)^2$$

$$= var(X) + (\mu-c)^2$$ as required.

4.

RTP $var(X) \leq E[(X-c)^2] \quad \forall c \in \mathbb{R}.$

1) $E[(X-c)^2] = var(X) + (\mu-c)^2$

$\geq var(X)$ as required.

a) $f(c) = E[(X-c)^2]$

$= E[X^2 - 2cX + c^2]$

$= E[X^2] - 2c\,E[X] + c^2$

$\frac{d f(c)}{dc} = 0 - 2E[X] + 2c = 0$

$\Rightarrow \quad c^* = E[X] = \mu$

since $\frac{d^2 f(c)}{dc^2} > 0$, min.

$\Rightarrow \quad f(c^*) = E[(X-\mu)^2] = var(X)$

$\therefore \quad \underline{E[(X-c)^2] \geq var(X)}$

b) RTP $\quad var(X) = E(X^2) - E(X)^2$

$E((X-c)^2) = var(X) + (\mu - c)^2$

setting $c = 0$:

$E(X^2) = var(X) + \mu^2$

$\Rightarrow \quad var(X) = E(X^2) - \mu^2$

$= E(X^2) - E(X)^2$

different!