# STA 247 — Solutions to Assignment #2, Part I

1. Ten records have to be accessed for a request of the first type, and each has probability 0.3 of having to be read from disk, independently for each record. The distribution of the number of records read is therefore binomial, with $n = 10$ and $p = 0.3$. From standard properties of binomial distributions, we know that the mean of this distribution is $np = 3$ and the variance is $np(1-p) = 2.1$, so the standard deviation is $\sqrt{2.1} = 1.449$. The probability that more than five records will have to be read is one minus the probability that five or fewer records will have to be read. We can find this from the binomial cumulative distribution function, using either Table 2 in the back of the book, or the R expression `pbinom(5,10,0.3)`. The answer is $1 - 0.9527 = 0.0473$.

2. Let $X$ be the number of records read from disk to satisfy a request, and let $Y$ be the type of this request (either 1 or 2). In Question 1, we found the mean and variance of $X$ given that $Y$ is 1. The result was $E(X|Y=1) = 3$ and $\mathrm{Var}(X|Y=1) = 2.1$. We can similarly find the mean and variance of $X$ given that $Y$ is 2, with the result being $E(X|Y=2) = 20 \times 0.3 = 6$ and $\mathrm{Var}(X|Y=2) = 20 \times 0.3 \times 0.7 = 4.2$. From this, and the fact that $P(Y=1) = 0.8$ and $P(Y=2) = 0.2$, we can find the mean and variance of $X$ not conditional on anything, using formulas from Section 2.9 in the text:

$$\begin{aligned} E(X) &= E(E(X|Y)) \\ &= P(Y=1)E(X|Y=1) + P(Y=2)E(X|Y=2) \\ &= 0.8 \times 3 + 0.2 \times 6 \\ &= 3.6 \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(X) &= E(\mathrm{Var}(X|Y)) \ + \ \mathrm{Var}(E(X|Y)) \\ &= P(Y=1)\mathrm{Var}(X|Y=1) + P(Y=2)\mathrm{Var}(X|Y=2) \ + \ E[(E(X|Y) - E(E(X|Y)))^2] \\ &= P(Y=1)\mathrm{Var}(X|Y=1) + P(Y=2)\mathrm{Var}(X|Y=2) \ + \ E[(E(X|Y) - E(X))^2] \\ &= P(Y=1)\mathrm{Var}(X|Y=1) + P(Y=2)\mathrm{Var}(X|Y=2) \\ &\quad + P(Y=1)(E(X|Y=1) - E(X))^2 + P(Y=2)(E(X|Y=2) - E(X))^2 \\ &= 0.8 \times 2.1 + 0.2 \times 4.2 \ + \ 0.8(3-3.6)^2 + 0.2(6-3.6)^2 \\ &= 3.96 \end{aligned}$$

   The standard deviation of $X$ is therefore $\sqrt{3.96} = 1.99$.

   We can use Chebychev's inequality to give an upper bound for $P(X > 7)$ as follows:

$$\begin{aligned} P(X > 7) &= P(X - 3.6 > 3.4) \\ &\leq P(|X - 3.6| > 3.4) \\ &= P(|X - 3.6| > 1.71 \times 1.99) \\ &= P(|X - \mu| > 1.71\sigma) \\ &\leq 1/1.71^2 = 0.342 \end{aligned}$$

However, we get a stronger upper bound if we remember that $X$ is an integer, and use the stronger form of Chebychev's inequality (with $\geq$ rather than $>$) mentioned in class:

$$\begin{aligned} P(X > 7) &= P(X \geq 8) \\ &= P(|X - 3.6| \geq 4.4) \quad \text{(since } X \text{ can't be negative)} \\ &= P(|X - 3.6| \geq 2.21 \times 1.99) \\ &= P(|X - \mu| \geq 2.21\sigma) \\ &\leq 1/2.21^2 = 0.205 \end{aligned}$$

The exact probability can be found as follows:

$$\begin{aligned} P(X > 7) &= P(Y=1)P(X > 7 \,|\, Y=1) \;+\; P(Y=2)P(X > 7 \,|\, Y=2) \\ &= P(Y=1)(1 - P(X \leq 7 \,|\, Y=1)) \;+\; P(Y=2)(1 - P(X \geq 7 \,|\, Y=2) \\ &= 0.8 \times (1 - 0.9984) \;+\; 0.2 \times (1 - 0.7723) = 0.04682 \end{aligned}$$

Here, $P(X \leq 7 \,|\, Y=1)$ and $P(X \leq 7 \,|\, Y=2)$ can be found from Table 2 in the book or using `pbinom`, since we know that the distribution of $X$ conditional on $Y=1$ is binomial with $n = 10$ and $p = 0.3$, and the distribution of $X$ conditional on $Y=2$ is binomial with $n = 20$ and $p = 0.3$.

3. The mean for 50 requests is just 50 times the mean for one request, which is $50 \times 3.6 = 180$. Since we are assuming independence from one request to another, the variance for 50 requests is just 50 times the variance for one request, which is $50 \times 3.96 = 198$. The standard deviation is therefore $\sqrt{198} = 14.1$.

4. The independence assumption may not be reasonable. The cache is designed to store the commonly-accessed records. When one uncommon record is accessed, other uncommon records may also be accessed soon thereafter. For instance, if a request is made for the students in a class that is usually taken by part-time students, it may be unlikely that most of these students' records will be in the cache, since they aren't taking many other classes, and therefore aren't the subject of many other requests. However, in this problem, such dependencies may be small enough that assuming independence nevertheless produces a good enough model.

   If the independence assumption does not hold, the means computed above will not be affected (provided the other assumptions do hold). However, the standard deviations might be larger than computed above, since events such as a whole class not being in the cache (or all of them being in the cache) will increase the variability of the number of accesses needed.