CS 70 Discrete Mathematics and Probability Theory
Spring 2017 Rao

# DIS 10a

## 1 Deriving Chebyshev's Inequality

Recall Markov's Inequality, which applies for non-negative $X$ and $\alpha > 0$:

$$\Pr[X \geq \alpha] \leq \frac{\mathbf{E}[X]}{\alpha}$$

Use an appropriate substitution for $X$ and $\alpha$ to derive Chebyshev's Inequality:

$$\Pr[|Y - \mu| \geq k] \leq \frac{\text{Var}(Y)}{k^2}$$

**Solution:**

Let $X = (Y-\mu)^2$. Note that this satisfies the criterion that $X$ is non-negative. Let $\alpha = k^2$ for $k > 0$. Again, this satisfies the criterion that $\alpha > 0$. Note also that the event $|Y - \mu| \geq k$ is equivalent to the event $(Y-\mu)^2 \geq k^2$. Then

$$\begin{aligned}
\Pr(|Y-\mu| \geq k) &= \Pr((Y-\mu)^2 \geq k^2) \\
&= \Pr(X \geq \alpha) \\
&\leq \frac{\mathbf{E}(X)}{\alpha} \\
&= \frac{\mathbf{E}((Y-\mu)^2)}{k^2} \\
&= \frac{\text{Var}(Y)}{k^2}.
\end{aligned}$$

This is equivalent to Chebyshev's Inequality.

## 2 Working with the Law of Large Numbers

(a) A fair coin is tossed and you win a prize if there are more than 60% heads. Which is better: 10 tosses or 100 tosses? Explain.

(b) A fair coin is tossed and you win a prize if there are more than 40% heads. Which is better: 10 tosses or 100 tosses? Explain.

(c) A coin is tossed and you win a prize if there are between 40% and 60% heads. Which is better: 10 tosses or 100 tosses? Explain.

(d) A coin is tossed and you win a prize if there are exactly 50% heads. Which is better: 10 tosses or 100 tosses? Explain.

**Solution:**

(a) 10 tosses. By LLN, the sample mean should have higher probability to be close to the population mean as n increases. Therefore the average proportion of coins that are heads should be closer to 0.50, and has a lower chance of being greater than 0.60 if there are 100 tosses compared with 10 tosses.

(b) 100 tosses. Based on the first part, consider the inverse of the event "more than 60% heads" and the symmetry of heads and tails.

(c) 100 tosses. Based on the first part, consider the union of the events "more than 60% heads" and "more than 60% tails" ("less than 40% heads").

(d) 10 tosses. Compare the probability of getting equal number of heads and tails between $2n$ and $2n+2$ tosses.

$$\begin{aligned}
\Pr[n \text{ heads in } 2n \text{ tosses}] &= \binom{2n}{n}/2^{2n} \\
\Pr[n+1 \text{ heads in } 2n+2 \text{ tosses}] &= \binom{2n+2}{n+1}/2^{2n+2} \\
&= \frac{(2n+2)!}{(n+1)!(n+1)!}\cdot\frac{1}{2^{2n+2}} \\
&= \frac{(2n+2)(2n+1)2n!}{(n+1)(n+1)n!n!}\cdot\frac{1}{2^{2n+2}} \\
&= \frac{2n+2}{n+1}\cdot\frac{2n+1}{n+1}\binom{2n}{n}\cdot\frac{1}{2^{2n+2}} \\
&< \left(\frac{2n+2}{n+1}\right)^2\binom{2n}{n}\cdot\frac{1}{2^{2n+2}} \\
&= 4\binom{2n}{n}\cdot\frac{1}{2^{2n+2}} = \binom{2n}{n}/2^{2n} = \Pr[n \text{ heads in } 2n \text{ tosses}]
\end{aligned}$$

The larger $n$ is, the less probability we'll get 50% heads. □

# 3 Easy A's

A friend tells you about a course called "Laziness in Modern Society" that requires almost no work. You hope to take this course next semester to give yourself a well-deserved break after mastering CS 70. At the first lecture, the professor announces that grades will depend only a midterm and a final. The midterm will consist of three questions, each worth 10 points, and the final will consist of four questions, also each worth 10 points. He will give an A to any student who gets at least 60 of the possible 70 points.

However, speaking with the professor in office hours you hear some very disturbing news. He tells you that, in the spirit of the class, the GSIs are very lazy, and to save time the grading will be done as follows. For each student's midterm, the GSIs will choose a real number randomly from a distribution with mean $\mu = 5$ and variance $\sigma^2 = 1$. They'll mark each of the three questions with that score. To grade the final, they'll again choose a random number from the same distribution, independent of the first number, and will mark all four questions with that score.

If you take the class, what will the mean and variance of your total class score be? Use Chebyshev's inequality to conclude that you have less than a 5% chance of getting an A.

**Solution:**

Let $X$ be the total number of points you receive in the class. Then $X = X_m + X_f$ where $X_m$ are the points you receive on the midterm and $X_f$ are the points you receive on the final. Your midterm score is generated as $X_m = 3Y_m$, where the r.v. $Y_m$ represents the real number that the GSI chose when grading your midterm. Similarly, $X_f = 4Y_f$. The problem statement tells us that $Y_m$ has mean 5 and variance 1 and $Y_f$ has mean 5 and variance 1, so $\mathbf{E}[Y_m] = \mathbf{E}[Y_f] = 5$ and $\text{Var}(Y_m) = \text{Var}(Y_f) = 1$. Thus,

$$\mathbf{E}[X] = \mathbf{E}[X_m] + \mathbf{E}[X_f] = 3\mathbf{E}[Y_m] + 4\mathbf{E}[Y_f] = 35,$$
$$\text{Var}(X) = \text{Var}(X_m) + \text{Var}(X_f) = 9\text{Var}(Y_m) + 16\text{Var}(Y_f) = 25.$$

Using Chebyshev's Inequality, we get

$$\Pr[X \geq 60] \leq \Pr[|X - 35| \geq 25] \leq \frac{\text{Var}(X)}{25^2} = \frac{1}{25}.$$

Unfortunately, you have at most a 4% chance of getting an A. So, the answer is: your mean score will be 35, the variance will be 25, and yes, you can conclude that you have less than a 5% chance of getting an A.

Note that although we calculated a bound for $\Pr[|X - 35| \geq 25]$, which is the probability that you will get 60 or above or 10 or below, we cannot divide by 2 to refine our bound unless the distribution is symmetric about its mean. In this case, the distribution is not symmetric.

# 4 Playing Pollster

As an expert in probability, the staff members at the Daily Californian have recruited you to help them conduct a poll to determine the percentage $p$ of Berkeley undergraduates that plan to participate in the student sit-in. They've specified that they want your estimate $\hat{p}$ to have an error of at most $\varepsilon$ with confidence $1 - \delta$. That is,

$$\Pr(|\hat{p} - p| \leq \varepsilon) \geq 1 - \delta.$$

Assume that you've been given the bound

$$\Pr(|\hat{p} - p| \geq \varepsilon) \leq \frac{1}{4n\varepsilon^2},$$

where $n$ is the number of students in your poll.

(a) Using the formula above, what is the smallest number of students $n$ that you need to poll so that your poll has an error of at most $\varepsilon$ with confidence $1-\delta$?

(b) At Berkeley, there are about 26,000 undergraduates and about 10,000 graduate students. Suppose you only want to understand the frequency of sitting-in for the undergraduates. If you want to obtain an estimate with error of at most 5% with 98% confidence, how many undergraduate students would you need to poll? Does your answer change if you instead only want to understand the frequency of sitting-in for the graduate students?

(c) It turns out you just don't have as much time for extracurricular activities as you thought you would this semester. The writers at the Daily Californian insist that your poll results are reported with at least 95% confidence, but you only have enough time to poll 500 students. Based on the bound above, what is the worst-case error with which you can report your results?

**Solution:**

(a) We know we need to have

$$\Pr(|\hat{p}-p| \leq \varepsilon) \geq 1-\delta.$$

Subtracting both sides from 1, it follows that we must have

$$\Pr(|\hat{p}-p| > \varepsilon) \leq \delta.$$

Therefore if we choose $n$ such that

$$\frac{1}{4n\varepsilon^2} \leq \delta,$$

we will have

$$\Pr(|\hat{p}-p| \geq \varepsilon) \leq \delta,$$

and since $\Pr(|\hat{p}-p| > \varepsilon) \leq \Pr(|\hat{p}-p| \geq \varepsilon)$, this will meet the requirement that

$$\Pr(|\hat{p}-p| > \varepsilon) \leq \delta.$$

Thus we must have that

$$\begin{aligned} \frac{1}{4n\varepsilon^2} &\leq \delta \\ \frac{1}{n} &\leq 4\varepsilon^2\delta \\ n &\geq \frac{1}{4\varepsilon^2\delta}. \end{aligned}$$

(b) Plugging in $\varepsilon = 0.05$ (our maximum error) and $\delta = 0.02$ (probability of being off by at least this error) to the bound you found above, you get that $n \geq 5000$. The answer is the same for graduate students; the size of the population does not affect the number of samples you need.

(c) If you only have time to poll 500 people and want to report your results with 95% confidence, you must report that the error in your estimate is at most 10%. You can find this by plugging in $1/(4 \cdot 500 \cdot \varepsilon^2) = .05$ and solving for $\varepsilon$.