1 Tightness of Inequalities

(a) Show by example that Markov's inequality is tight; that is, show that given $k > 0$, there exists a discrete non-negative random variable X such that $\mathbb{P}(X \geq k) = \mathbb{E}[X]/k$.

(b) Show by example that Chebyshev's inequality is tight; that is, show that given $k \geq 1$, there exists a random variable X such that $\mathbb{P}(|X - \mathbb{E}[X]| \geq k\sigma) = 1/k^2$, where $\sigma^2 = \operatorname{Var} X$.

Solution:

(a) In the proof of Markov's Inequality ($\mathbb{P}[X \geq \alpha] \leq \frac{\mathbb{E}[X]}{\alpha}$), the first time we lose equality is at this step:

$$\mathbb{E}[X] = \sum_a (a \cdot \mathbb{P}[X = a]) \geq \sum_{a \geq \alpha} (a \cdot \mathbb{P}[X = a])$$

We get an inequality because we drop all $a \cdot \mathbb{P}[X = a]$ terms where $a < \alpha$. Thus, we can only maintain equality if all of these dropped terms were actually 0. This would mean either $a = 0$ or $\mathbb{P}[X = a] = 0$ for an $a > 0$, which means X can put probability on 0, but should put no probability on any other value $< \alpha$.

The next time we lose equality in the proof is the step following the one above:

$$\sum_{a \geq \alpha} (a \cdot \mathbb{P}[X = a]) \geq \alpha \cdot \sum_{a \geq \alpha} \mathbb{P}[X = a]$$

We get an inequality because we treat all $a \geq \alpha$ in the summation as just α, so we can pull out the α term. The only way for us to maintain equality is if we never have to substitute α for some larger a. This tells us that X should not put probability on any value $> \alpha$.

Both of these facts drive the intuition behind our example: that X can only take values 0 and α.

Let X be the random variable which is 0 with probability $1 - p$ and k with probability p, where $k > 0$. Then, $\mathbb{E}[X] = kp$, and Markov's inequality says

$$\mathbb{P}(X \geq k) \leq \frac{\mathbb{E}[X]}{k} = \frac{kp}{k} = p,$$

which is tight.

(b) The proof of Chebyshev's Inequality ($\mathbb{P}[|X - \mathbb{E}[X]| \geq \alpha] \leq \frac{\operatorname{Var}(X)}{\alpha^2}$) comes from an application of Markov's Inequality to the variable $Y = (X - \mathbb{E}[X])^2$ being $\geq \alpha^2$. The only ways we can lose equality in the proof of Chebyshev's is if we lose equality in the application of Markov!

Therefore, we need the variable Y to satisfy the conditions from Part (a) that ensure the application of Markov will be tight. To recap, we would need Y to only take values 0 and α^2. Thus, $(X - \mathbb{E}[X])$ can take on the values $\{-\alpha, 0, \alpha\}$.

Let

$$X = \begin{cases} -a & \text{with probability } k^{-2}/2 \\ a & \text{with probability } k^{-2}/2 \\ 0 & \text{with probability } 1-k^{-2} \end{cases}$$

for $a > 0$. Note that $\operatorname{Var} X = a^2 k^{-2}$, so $k\sigma = a$, so Chebyshev's inequality gives

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq k\sigma) = \mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{1}{k^2},$$

which is tight.

2 Just One Tail, Please

Let X be some random variable with finite mean and variance which is not necessarily non-negative. The *extended* version of Markov's Inequality states that for a non-negative function $\phi(x)$ which is monotonically increasing for $x > 0$ and some constant $\alpha > 0$,

$$\mathbb{P}(X \geq \alpha) \leq \frac{\mathbb{E}[\phi(X)]}{\phi(\alpha)}$$

Suppose $\mathbb{E}[X] = 0$, $\operatorname{Var}(X) = \sigma^2 < \infty$, and $\alpha > 0$.

(a) Use the extended version of Markov's Inequality stated above with $\phi(x) = (x+c)^2$, where c is some positive constant, to show that:

$$\mathbb{P}(X \geq \alpha) \leq \frac{\sigma^2 + c^2}{(\alpha + c)^2}$$

(b) Note that the above bound applies for all positive c, so we can choose a value of c to minimize the expression, yielding the best possible bound. Find the value for c which will minimize the RHS expression (you may assume that the expression has a unique minimum). Plug in the minimizing value of c to prove the following bound:

$$\mathbb{P}(X \geq \alpha) \leq \frac{\sigma^2}{\alpha^2 + \sigma^2}.$$

(c) Recall that Chebyshev's inequality provides a two-sided bound. That is, it provides a bound on $\mathbb{P}(|X - \mathbb{E}[X]| \geq \alpha) = \mathbb{P}(X \geq \mathbb{E}[X] + \alpha) + \mathbb{P}(X \leq \mathbb{E}[X] - \alpha)$. If we only wanted to bound the probability of one of the tails, e.g. if we wanted to bound $\mathbb{P}(X \geq \mathbb{E}[X] + \alpha)$, it is tempting to just divide the bound we get from Chebyshev's by two. Why is this not always correct in general? Provide an example of a random variable X (does not have to be zero-mean) and a constant α such that using this method (dividing by two to bound one tail) is not correct, that is, $\mathbb{P}(X \geq \mathbb{E}[X]+\alpha) > \frac{\mathrm{Var}(X)}{2\alpha^2}$ or $\mathbb{P}(X \leq \mathbb{E}[X]-\alpha) > \frac{\mathrm{Var}(X)}{2\alpha^2}$.

Now we see the use of the bound proven in part (b) - it allows us to bound just one tail while still taking variance into account, and does not require us to assume any property of the random variable. Note that the bound is also always guaranteed to be less than 1 (and therefore at least somewhat useful), unlike Markov's and Chebyshev's inequality!

(d) Let's try out our new bound on a simple example. Suppose X is a positively-valued random variable with $\mathbb{E}[X] = 3$ and $\mathrm{Var}(X) = 2$. What bound would Markov's inequality give for $\mathbb{P}[X \geq 5]$? What bound would Chebyshev's inequality give for $\mathbb{P}[X \geq 5]$? What about for the bound we proved in part (b)? (*Note*: Recall that the bound from part (b) only applies for zero-mean random variables.)

Solution:

(a) Note that $\sigma^2 = \mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2]$. Using the inequality presented in the problem, we have:

$$\mathbb{P}(X \geq \alpha) \leq \frac{\mathbb{E}[(X+c)^2]}{(\alpha+c)^2} = \frac{\mathbb{E}[X^2+2cX+c^2]}{(\alpha+c)^2} = \frac{\mathbb{E}[X^2]+2c\,\mathbb{E}[X]+c^2}{(\alpha+c)^2} = \frac{\sigma^2+c^2}{(\alpha+c)^2}$$

(b) We set the derivative with respect to c of the above expression equal to 0, and solve for c.

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}c}\frac{\sigma^2+c^2}{(\alpha+c)^2} &= 0 \\ \frac{2c(\sigma+c)^2 - 2(\alpha+c)(\sigma^2+c^2)}{(\alpha+c)^4} &= 0 \\ 2c(\sigma+c)^2 - 2(\alpha+c)(\sigma^2+c^2) &= 0 \\ \alpha c^2 + (\alpha^2-\sigma^2)c - \sigma^2\alpha &= 0 \\ c &= \frac{\sigma^2}{\alpha} \end{aligned}$$

To get the last step we use the quadratic equation and take the positive solution. Plugging in this value for c yields us the desired inequality.

This bound is also known as Cantelli's inequality.

(c) It is possible for one of the tails to contain more probability than the other. One example of a random variable which demonstrates this is X, where $\mathbb{P}(X=0) = 0.75$ and $\mathbb{P}(X=10) = 0.25$, with $\alpha = 7$. Here, $\mathbb{E}[X] = 2.5$ and $\mathrm{Var}(X) = 100 \cdot 0.25 \cdot 0.75$, so we have:

$$\mathbb{P}(X \geq \mathbb{E}[X]+7) = 0.25 > \frac{\mathrm{Var}(X)}{2\cdot 7^2} \approx 0.19$$

(d) Using Markov's: $\mathbb{P}(X \geq 5) \leq \frac{\mathbb{E}[X]}{5} = \frac{3}{5}$

Using Chebyshev's: $\mathbb{P}(X \geq 5) \leq \mathbb{P}(|X - \mathbb{E}[X]| \geq 2) \leq \frac{\text{Var}(X)}{2^2} = \frac{1}{2}$

Using bound shown above (Cantelli's):

Since we have the condition that this bound applies to zero-mean random variables, let us define $Y = X - \mathbb{E}[X] = X - 3$. Note that $\text{Var}(Y) = \text{Var}(X)$.

Then we get: $\mathbb{P}(X \geq 5) = \mathbb{P}(Y \geq 2) \leq \frac{\text{Var}(Y)}{2^2 + \text{Var}(Y)} = \frac{1}{3}$.

We see that Cantelli's inequality (the bound from part (b)) does better than Chebyshev's, which does better than Markov's (note that having a smaller upper bound is better)! This is a good demonstration on how we might derive better bounds using Markov's inequality, if we know further information about the random variable like its variance.

3 Erasures, Bounds, and Probabilities

Alice is sending 1000 bits to Bob. The probability that a bit gets erased is p, and the erasure of each bit is independent of the others.

Alice is using a scheme that can tolerate up to one-fifth of the bits being erased. That is, as long as Bob receives at least 801 of the 1000 bits correctly, he can decode Alice's message.

In other words, Bob becomes unable to decode Alice's message only if 200 or more bits are erased. We call this a "communication breakdown", and we want the probability of a communication breakdown to be at most 10^{-6}.

1. Use Markov's inequality to upper bound p such that the probability of a communications breakdown is at most 10^{-6}.
2. Use Chebyshev's inequality to upper bound p such that the probability of a communications breakdown is at most 10^{-6}.
3. As the CLT would suggest, approximate the fraction of erasures by a Gaussian random variable (with suitable mean and variance). Use this to find an approximate bound for p such that the probability of a communications breakdown is at most 10^{-6}.

Solution:

1. Let the indicator random variable for the i^{th} bit's erasure be X_i. That is, $X_i = 1$ if the i^{th} bit is erased, and 0 otherwise. Pr $(X_i = 1) = p$, and the X_is are all independent of one another.

   Let X be the total number of erasures. Then $X = X_1 + X_2 + \ \ldots \ + X_{1000}$. A communications breakdown happens if and only if $X \geq 200$. So we would like to have $\Pr(X \geq 200) \leq 10^{-6}$.

   Further, let μ_X denote the mean of X and σ_X^2 denote the variance of X. From previous discussion sections, $\mu_X = 1000p$ and $\sigma_X^2 = 1000p(1-p)$.

   Applying Markov's inequality to X (which is non-negative), we have:

$$\Pr(X \geq 200) \leq \frac{E[X]}{200} = \frac{1000p}{200} = 5p$$

So, if p is such that $5p \leq 10^{-6}$, then our objective of $\Pr(X \geq 200) \leq 10^{-6}$ is met. So the upper bound for p is $10^{-6}/5$, or 2×10^{-7}.

2. Chebyshev's inequality states the following:

$$\Pr(\|X - \mu_X\| \geq k\sigma_X) \leq \frac{1}{k^2}$$

So we need to choose a k given by:

$$k = \frac{200 - 1000p}{\sqrt{1000p(1-p)}}$$

Note: The above is valid only for $200 - 1000p > 0$, or $p < 0.2$, since k has to be positive. But as we will see below, our upper bound for p will be below 0.2, so there is no problem.

Proceeding with the above value of k, and substituting for μ_X and σ_X, we obtain:

$$\Pr(\|X - 1000p\| \geq 200 - 1000p) \leq \frac{1}{\left(\frac{(200-1000p)^2}{1000p(1-p)}\right)}$$

Simplifying, we get:

$$\Pr(\|X - 1000p\| \geq 200 - 1000p) \leq \frac{p(1-p)}{40(1-5p)^2}$$

Now we know the following:

$$\begin{aligned}\Pr(X \geq 200) &= \Pr(X - 1000p \geq 200 - 1000p)\\ &\leq \Pr(\|X - 1000p\| \geq 200 - 1000p)\\ &\leq \frac{p(1-p)}{40(1-5p)^2}\end{aligned}$$

As before, to meet our objective, we just have to ensure that

$$\frac{p(1-p)}{40(1-5p)^2} \leq 10^{-6} ,$$

which yields an upper bound of about 3.998×10^{-5} for p.

3. Let Y be equal to the fraction of erasures, i.e. $\frac{X}{1000}$. Using properties of expectation and variance, we can see that

$$\mathbb{E}[Y] = p$$
$$\mathrm{Var}(Y) = \mathrm{Var}(X) \cdot \frac{1}{1000^2} = \frac{p(1-p)}{1000}$$

Therefore, by Central Limit Theorem, we can say that Y is roughly a normal distribution with that mean and variance. Since we are interested in the event that $Y \geq 0.2$, let's figure out how many standard deviations above the mean 0.2 is:

$$\frac{0.2-p}{\sqrt{\frac{p(1-p)}{1000}}} = \frac{(0.2-p)\sqrt{1000}}{\sqrt{p(1-p)}}.$$

Therefore, the probability that we get a failure should be approximately (by CLT),

$$1 - \Phi\left(\frac{(0.2-p)\sqrt{1000}}{\sqrt{p(1-p)}}\right)$$

where Φ is the CDF of a standard normal variable. Setting this to be at most 10^{-6} gives us

$$\Phi\left(\frac{(0.2-p)\sqrt{1000}}{\sqrt{p(1-p)}}\right) \geq 1 - 10^{-6}$$

And, since $\Phi^{-1}(1-10^{-6}) \approx 4.753$, we solve the inequality

$$\frac{(0.2-p)\sqrt{1000}}{\sqrt{p(1-p)}} \geq 4.753$$

This yields that we need $p \leq 0.1468$.

Note that this gives quite a different value from the previous parts. This is because the Central Limit Theorem gives a much tighter approximation for tail events than Markov's and Chebyshev's. Therefore, we do not need p to be so low to achieve a communication breakdown probability of 10^{-6}. The other bounds required us to need a probability of on the order of 10^{-5}, but here we realize that we only need it to be less than 0.1468. Quite drastic!

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 ε with confidence $1-\delta$. That is,

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

Assume that you've been given the bound

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

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 ε 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

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

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

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

Therefore if we choose n such that

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

we will have

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

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

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

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 ε.

5 Oski's Markov Chain

When Oski Bear is studying for CS70, he splits up his time between reading notes and working on practice problems. To do this, every so often he will make a decision about what kind of work to do next.

When Oski is already reading the notes, with probability a he will decide to switch gears and work on a practice problem, and otherwise, he will decide to keep reading more notes. Conversely, when Oski is already working on a practice problem, with probability b he will think of a topic he needs to review, and will decide to switch back over to the notes; otherwise, he will keep working on practice problems.

Assume that (unlike real life, we hope!) Oski never runs out of work to do.

(a) Draw a 2-state Markov chain to model this situation.

(b) In the remainder of this problem, we will learn to work with the definitions of some important terms relating to Markov Chains. These definitions are as follows:

  (a) (Irreducibility) A Markov chain is irreducible if, starting from any state i, the chain can transition to any other state j, possibly in multiple steps.
  (b) (Periodicity) $d(i) := \gcd\{n > 0 \mid P^n(i,i) = \mathbb{P}[X_n = i \mid X_0 = i] > 0\}$, $i \in \mathscr{X}$. If $d(i) = 1\ \forall i \in \mathscr{X}$, then the Markov chain is aperiodic; otherwise it is periodic.
  (c) (Matrix Representation) Define the transition probability matrix P by filling entry (i, j) with probability $P(i, j)$.
  (d) (Invariance) A distribution π is invariant for the transition probability matrix P if it satisfies the following balance equations: $\pi = \pi P$.

  For what values of a and b is the Markov chain irreducible?

(c) For $a = 1$, $b = 1$, prove that the Markov chain is periodic.

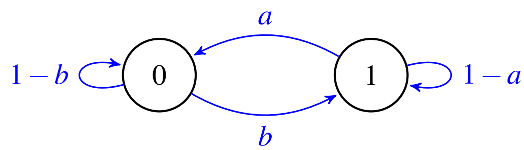
(d) For $0 < a < 1$, $0 < b < 1$, prove that the Markov chain is aperiodic.

(e) Construct a transition probability matrix using the Markov chain.

(f) Write down the balance equations for the Markov chain and solve them. Assume that the Markov chain is irreducible.

Solution:

(a) Let state 0 represent Oski working on a practice problem, and let state 1 represent reading notes.



(b) The Markov chain is irreducible if both a and b are non-zero. It is reducible if at least one is 0.

(c) We compute $d(0)$ to find that:

$$d(0) = \gcd\{2,4,6,\ldots\} = 2.$$

Thus, the chain is periodic.

(d) We compute $d(0)$ to find that:

$$d(0) = \gcd\{1,2,3,\ldots\} = 1.$$

Thus, the chain is aperiodic.

(e)

$$\begin{bmatrix} 1-b & b \\ a & 1-a \end{bmatrix}$$

(f)

$$\begin{aligned} \pi(0) &= (1-b)\pi(0) + a\pi(1), \\ \pi(1) &= b\pi(0) + (1-a)\pi(1). \end{aligned}$$

One of the equations is redundant. We throw out the second equation and replace it with $\pi(0)+\pi(1) = 1$. This gives the solution

$$\pi = \frac{1}{a+b}\begin{bmatrix} a & b \end{bmatrix}.$$

6 Markov Chains: Prove/Disprove

Prove or disprove the following statements, using the definitions from the previous question.

(a) There exists an irreducible, finite Markov chain for which there exist initial distributions that converge to different distributions.

(b) There exists an irreducible, aperiodic, finite Markov chain for which $\mathbb{P}(X_{n+1} = j \mid X_n = i) = 1$ or 0 for all i, j.

(c) There exists an irreducible, non-aperiodic Markov chain for which $\mathbb{P}(X_{n+1} = j \mid X_n = i) \neq 1$ for all i, j.

(d) For an irreducible, non-aperiodic Markov chain, any initial distribution not equal to the invariant distribution does not converge to any distribution.

Solution:

(a) False. Every finite irreducible Markov chain has a unique stationary distribution. If it's possible for the Markov chain to converge to two different distributions given different starting distributions, it implies there are two stationary distributions. To elaborate further, we know in the long run the fraction of time spent in each state converges to the stationary distribution. So if the distribution converges, the long-run fraction of time will be whatever distribution it converges to, which we see must be the stationary distribution.

(b) True, you can have one state pointing to itself. However for number of states > 1 it is false. Consider the initial distribution of having a probability of 1 of being in an arbitrary state. After a transition, the resulting distribution must be a probability 1 of being in a different state (if it were the same state, this would immediately imply that the Markov chain is reducible). Further transitions have the same effect. Therefore this initial distribution does not converge. Therefore this Markov chain cannot be aperiodic and irreducible (since it would converge in that case).

(c) True. Consider the states $\{0,1,2,3\}$. Set $P(i,j) = 1/2$ if $i \equiv j \pm 1 \mod 4$ and 0 otherwise. In other words, the Markov chain is a square with each side replaced with two links pointing in opposite directions with probabilities of $1/2$. Consider the period of state 0. Any path from 0 back to itself, such as $0-1-2-1-0$, alternates in parity of each consecutive state since each state only points to the state above or below it mod 4. Therefore state 0 has period 2. Therefore this Markov chain is not aperiodic (and all states have period 2).

(d) False. Take the initial distribution $\begin{bmatrix} 0.25 & 0.30 & 0.25 & 0.20 \end{bmatrix}$ for the above Markov chain. After one transition it goes to the invariant distribution, $\begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$.

7 Balls, meet Bins

Alice and Bob are tasked with throwing balls into bins (to set up a probability problem for later). They decide to make a game out of it: Alice and Bob will each take a ball, and once per minute, they will both simultaneously (and independently) attempt to throw their balls into a bin.

Once Alice or Bob successfully lands a ball in a bin, that person stops while the other person continues to try until they also land a throw. When this happens, the game is over.

Suppose that on every try, the probability of successfully landing the throw in a bin is p. What is the expected number of minutes until the game is over? Solve this using a Markov chain with three states. Then, state how your solution can be interpreted in terms of two geometric random variables.

Solution:

The three states are $\{N, B, BB\}$. N represents neither person having landed a throw, B represents one person having a landed throw, and BB is when both have a landed a throw. We can write hitting time equations:

$$\begin{aligned}
\mathbb{E}[N] &= 1 + p^2\,\mathbb{E}[BB] + (1-p)^2\,\mathbb{E}[N] + 2p(1-p)\,\mathbb{E}[B] \\
\mathbb{E}[B] &= 1 + (1-p)\,\mathbb{E}[B] + p\,\mathbb{E}[BB] \\
\mathbb{E}[BB] &= 0
\end{aligned}$$

$p^2, (1-p)^2, 2p(1-p)$ respectively are the probabilites of both Alice and Bob landing a throw at the same time, neither landing a throw, and one of them landing a throw. In the state D, we're only concerned with the remaning person who hasn't yet landed a ball, so the probabilities simplify to p and $1-p$. $\mathbb{E}[BB] = 0$ because it's the success state. Solving the system we get

$$\mathbb{E}[N] = \frac{2(1-p)+1}{1-(1-p)^2} = \frac{2}{p} - \frac{1}{1-(1-p)^2}.$$

Notice this is the max of two geometric random variables. We could solve this problem by finding $\mathbb{P}[\max > k]$ and using the tail sum to find the expectation, which would give the right expression (which is equal to the left but looks different).

8 Homework Process and Study Group

You must describe your homework process and study group in order to receive credit for this question.