Solutions to CS 70 Challenge Problems: Expectation, Variance, and Bounds

Other worksheets and solutions at https://alextseng.net/teaching/cs70/

Alex Tseng

1 Expectation

(a) Short tandem repeats (or STRs) occur in some segments of DNA. They consist of very short sequences that are repeated usually around 20-30 times. Consider a strand of DNA with n base pairs, where each base could be **A**, **C**, **G**, or **T**, randomly and independently. We are interested in the STR **AG**, so we are looking for segments of DNA that look like **AGAGAG**. This segment would have 3 repeats. What is the probability that you have at least r repeats starting at position p? Assume that all $2r$ bases would fit in the n base pairs starting at p.

$(\frac{1}{4})^{2r}$

If there is a sequence of at least r repeats starting at p, then the first base at p must be **A**, and then the base at $p+1$ must be **G**, and so on. The $2r$ bases starting at p must all be specific bases, each with an *independent* chance of $\frac{1}{4}$. Thus, the probability is $(\frac{1}{4})^{2r}$. Note that we don't care if the repetition continues past $2r$ base pairs, so don't need to factor those into our probability. Also note that it does not depend on p.

(b) Continuing from the above part, what is the expected number of positions where a sequence of at least r repeats start?

$(n-2r+1)(\frac{1}{4})^{2r}$

Let X_i be the indicator random variable that reflects whether or not the STR repeats at least r times starting at position i. For any X_i, $X_i = 1$ if there is a sequence of r repetitions starting at i. From the part above, this happens with probability $(\frac{1}{4})^{2r}$. With the remainder probability $1-(\frac{1}{4})^{2r}$, $X_i = 0$.

Let X be the random variable that is the number of positions where a sequence of at least r repeats start. Then $X = \sum_{i=1}^{n-2r+1} X_i$. Note that the X_i's are not all independent! Clearly, if $X_i = 1$, then X_{i+1} must be 0, since the base at position $i+1$ would be **G**. However, our expression for X is still valid. The number of positions where such a sequence starts is the sum of the X_i's. This is true regardless of independence/dependence relations.

What we want is $\mathrm{E}[X] = \mathrm{E}[\sum_{i=1}^{n-2r+1} X_i]$. By the linearity of expectations, this is $\sum_{i=1}^{n-2r+1} \mathrm{E}[X_i]$. The expectation of an indicator variable is just the probability that the indicator variable is 1 (a fact that should be memorized). Then $\mathrm{E}[X] = \sum_{i=1}^{n-2r+1} \mathrm{E}[X_i] = \sum_{i=1}^{n-2r+1} (\frac{1}{4})^{2r} = (n-2r+1)(\frac{1}{4})^{2r}$

(c) You draw a random number from the set $\{1, ..., 100\}$, and then another number from the set $\{1, ..., 50\}$. What is the expectation of the sum of the numbers? What is the expectation of the product?

Sum: 75, Product: 1287.75

Let X be the number we draw from the first set, and Y be the number we draw from the second set.

$\mathrm{E}[X] = \sum_{i=1}^{100} \frac{1}{100} i = \frac{1}{100} \sum_{i=1}^{100} i = \frac{1}{100} \frac{100(101)}{2} = 50.5$

$\mathrm{E}[Y] = \sum_{j=1}^{50} \frac{1}{50} j = \frac{1}{50} \sum_{j=1}^{50} = \frac{1}{50} \frac{50(51)}{2} = 25.5$

The expectation of the sum is $\mathrm{E}[X+Y]$, which by the linearity of expectations, is $\mathrm{E}[X] + \mathrm{E}[Y] = 50.5 + 25.5 = 76$.

The expectation of the product is $\mathrm{E}[XY]$. Normally, we would have to appeal back to the summations and solve from scratch, but since X and Y are independent ($P(X \cap Y) = P(X)P(Y)$), we can simplify $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y] = 50.5 \times 25.5 = 1287.75$.

(d) *Challenge* Suppose you have chips numbered 1 to k in a bag, and you draw 2 from the bag randomly, *with replacement*. Let A be the first number you drew, and B be the second number you drew. What

is $\mathrm{E}[\max\{A,B\}]$? What is $\mathrm{E}[\min\{A,B\}]$. Show that $\mathrm{E}[\max\{A,B\}] + \mathrm{E}[\min\{A,B\}] = \mathrm{E}[A] + \mathrm{E}[B]$ *without appealing to the linearity of expectations.*
Hint: Start with the definition of expected value. You will need to be a little clever with the algebra.
We will start with the maximum.
Recall the definition of expected value is simply the sum of all possible outcomes, weighted by the probabilities they occur.
$\mathrm{E}[\max\{A,B\}] = \sum_{a=1}^{k}\sum_{b=1}^{k}\frac{1}{k^2}\max\{a,b\}$. This follows from the definition of expected value. The possible outcomes range over all combinations of A and B, each with a probability of $\frac{1}{k^2}$ (A and B are independent), and each outcome has a value of the maximum.
$\sum_{a=1}^{k}\sum_{b=1}^{k}\frac{1}{k^2}\max\{a,b\} = \frac{1}{k^2}\sum_{a=1}^{k}\sum_{b=1}^{k}\max\{a,b\}$
Now we do something a little tricky to rewrite the max in terms of simple summations. Notice that we are essentially iterating through all values of A, and within a value of A we iterate through all values of B, and add the maximum. To take away the maximum, we rewrite the summation to iterate through all values of A, and within a value of A we iterate through the values of B in two steps. If the value of B is less than the value of A, we take B. Otherwise, we take A:
$\frac{1}{k^2}\sum_{a=1}^{k}\sum_{b=1}^{k}\max\{a,b\} = \frac{1}{k^2}\sum_{a=1}^{k}(\sum_{b=1}^{a}a + \sum_{b=a+1}^{k}b)$
It turns out this is as far as we can (or need to) go for simplifying.

The case for the minimum is completely symmetric:
$\mathrm{E}[\min\{A,B\}] = \frac{1}{k^2}\sum_{a=1}^{k}\sum_{b=1}^{k}\min\{a,b\} = \frac{1}{k^2}\sum_{a=1}^{k}(\sum_{b=1}^{a}b + \sum_{b=a+1}^{k}a)$
Notice the swapped a and b in the inner summations for the minimum; now if $B \le A$, we take B, and A otherwise.

Now we want to show that $\mathrm{E}[\max\{A,B\}] + \mathrm{E}[\min\{A,B\}] = \mathrm{E}[A] + \mathrm{E}[B]$.
$\mathrm{E}[\max\{A,B\}] + \mathrm{E}[\min\{A,B\}] = \frac{1}{k^2}\sum_{a=1}^{k}(\sum_{b=1}^{a}a + \sum_{b=a+1}^{k}b) + \frac{1}{k^2}\sum_{a=1}^{k}(\sum_{b=1}^{a}b + \sum_{b=a+1}^{k}a)$
$= \frac{1}{k^2}\sum_{a=1}^{k}(\sum_{b=1}^{a}a + \sum_{b=a+1}^{k}b + \sum_{b=1}^{a}b + \sum_{b=a+1}^{k}a) = \frac{1}{k^2}\sum_{a=1}^{k}((\sum_{b=1}^{a}a + \sum_{b=a+1}^{k}a) + (\sum_{b=1}^{a}b + \sum_{b=a+1}^{k}b))$
$= \frac{1}{k^2}\sum_{a=1}^{k}(\sum_{b=1}^{k}a + \sum_{b=1}^{k}b) = \frac{1}{k^2}\sum_{a=1}^{k}(ka + \sum_{b=1}^{k}b) = (\frac{1}{k^2}\sum_{a=1}^{k}ka) + (\frac{1}{k^2}\sum_{a=1}^{k}\sum_{b=1}^{k}b) = (\sum_{a=1}^{k}\frac{1}{k}a) + (\frac{1}{k^2}k\sum_{b=1}^{k}b)$
$= (\sum_{a=1}^{k}\frac{1}{k}a) + (\sum_{b=1}^{k}\frac{1}{k}b) = \mathrm{E}[A] + \mathrm{E}[B]$
The calculation above uses many properties of summations. Make sure you understand each step.

If we could appeal to the linearity of expectations, this problem would be extremely easy:
$\mathrm{E}[\max\{A,B\}] + \mathrm{E}[\min\{A,B\}] = \mathrm{E}[\max\{A,B\} + \min\{A,B\}]$ by the linearity of expectations
$\mathrm{E}[\max\{A,B\} + \min\{A,B\}] = \mathrm{E}[A+B]$ because one of them is the max and the other is the min
$\mathrm{E}[A+B] = \mathrm{E}[A] + \mathrm{E}[B]$ by the linearity of expectations

2 Variance

(a) Imagine we roll a standard 6-sided die 100 times, and add up the resulting values. What is the variance of this distribution?
$\frac{875}{3}$
Let X_i be the result of the ith die roll. Notice that all the X_i's are i.i.d. (independent and identically distributed).
We start with the variance of a single X_i. We know $\mathrm{Var}[X_i] = \mathrm{E}[X_i^2] - \mathrm{E}[X_i]^2$.

Using the definitions of expected value, we now calculate these two values: $\mathrm{E}[X_i] = \sum_{i=1}^{6} \frac{1}{6} i = \frac{7}{2}$, so $\mathrm{E}[X_i]^2 = \frac{7}{2}$

$\mathrm{E}[X_i^2] = \sum_{i=1}^{6} \frac{1}{6} i^2 = \frac{91}{6}$. In this last step, you could have added $(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2)$ together, or used the formula $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$.

Then $\mathrm{Var}[X_i] = \frac{91}{6} - (\frac{7}{2})^2 = \frac{35}{12}$

Note that since the X_i's are i.i.d., the variances are all the same, and the variance of their sum is the sum of their variances. We are interested in $X = \sum_{i=1}^{100} X_i$, or the sum of the dice.

$\mathrm{Var}[X] = \mathrm{Var}[\sum_{i=1}^{100} X_i] = \sum_{i=1}^{100} \mathrm{Var}[X_i] = \sum_{i=1}^{100} \frac{35}{12} = \frac{3500}{12} = \frac{875}{3}$

(b) Prove or give a counterexample: if X and Y are independent, then $\mathrm{Var}[X - Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$.

This is true. $\mathrm{Var}[X - Y] = \mathrm{E}[(X - Y)^2] - \mathrm{E}[X - Y]^2 = \mathrm{E}[X^2 - 2XY + Y^2] - (\mathrm{E}[X] - \mathrm{E}[Y])^2$

$= \mathrm{E}[X^2] - 2\mathrm{E}[XY] + \mathrm{E}[Y^2] - \mathrm{E}[X]^2 + 2\mathrm{E}[X]\mathrm{E}[Y] - \mathrm{E}[Y]^2$

We know that in general, if A and B are independent, then $\mathrm{E}[AB] = \mathrm{E}[A]\mathrm{E}[B]$, so two of the terms cancel out, leaving us with:

$\mathrm{E}[X^2] - \mathrm{E}[X]^2 + \mathrm{E}[Y^2] - \mathrm{E}[Y]^2 = \mathrm{Var}[X] + \mathrm{Var}[Y]$

In order to not leave any loose threads dangling, we will now prove that if A and B are independent, then $\mathrm{E}[AB] = \mathrm{E}[A]\mathrm{E}[B]$:

$\mathrm{E}[AB] = \sum_{a,b} P(a,b)ab = \sum_{a,b} P(a)P(b)ab = \sum_{a} \sum_{b} P(a)P(b)ab = \sum_{a} P(a)a \sum_{b} P(b)b = \mathrm{E}[A]\mathrm{E}[B]$

3 Bounds

(a) If the national IQ average is 100, and $\sigma = 10$ (recall σ is standard deviation), what is the probability of finding someone with an IQ of 300 or more? Give the tightest bound you can.

$\frac{1}{400}$

Let I be the IQ of a random person we choose.

Note that here, $\mathrm{E}[I] = \mu = 100$, and the $\mathrm{Var}[X] = \sigma^2 = 100$.

We have 2 tools at our disposal: Markov's bound and Chebyshev's bound.

Markov's bound states that for a *non-negative* random variable X, $P(X \geq a) \leq \frac{\mathrm{E}[X]}{a}$, assuming $a \neq 0$

Chebyshev's bound states that $P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$

Let us first try Markov's bound. As always, we ensure that our random variable is non-negative. In case of I, which is IQ, it is never negative. Thus, our use of Markov's bound is justified: $P(I \geq 300) \leq \frac{\mu}{300} = \frac{100}{300} = \frac{1}{3}$. Thus, Markov's bound gives us an upper bound of $\frac{1}{3}$.

Let us now try Chebyshev's bound. Our value of interest is an IQ of 300, which is 200 away from the mean of 100. Normally, we need to keep in mind that Chebyshev's bound is a 2-way estimate, since it deals with probability in the distribution that is both 200 above and 200 below the mean. In this case, however, 200 IQ points below μ would be -100, and negative IQ's do not exist (at least, not yet). Thus, we are guaranteed that all the estimated probability is on the higher side above 300: $P(I \geq 300) = P(|I - 100| \geq 200)$. Given that $\sigma = 10$, we know that $k = 20$ in this query. Chebyshev's bound tells us that $P(|I - 100| \geq 200) \leq \frac{1}{20^2} = \frac{1}{400}$, which is a much tighter bound than the one calculated from Markov's bound.

(b) Consider a distribution X with $\mu = 13, \sigma = 5$. We know that X never exceeds 17. Is it possible that $P(X \leq 1) > \frac{1}{4}$?
No.
Think of X as a collection of probabilities on the number line. We can assign each number on the number line a certain weight. However, we have the stipulation that the weighted sum of numbers must add to 13. This is the definition of expected value.
The intuition here is that if more than $\frac{1}{4}$ of the probability lies at $X = 1$ or below, then there is not enough probability left to pull the mean back up to 13, because the highest number X can be is 17.

Imagine that we have p probability, where $p > \frac{1}{4}$, that is assigned to values of X that are 1 and below. In order to maximize the expected value, we assign it to the highest value in that range: 1. The rest of the probability $1 - p$ is assigned to the highest possible value for X in order to maximize the expected value, so we assign the rest to 17. To say it another way, if we assume that $P(X \leq 1) = p > \frac{1}{4}$, then to maximize the expected value, we push the probability as far up as possible, so $P(X = 1) = p$, and $P(X = 17) = 1 - p$. Then the expected value is $\mathrm{E}[X] = p + 17(1 - p) = 17 - 16p$. Since $p > \frac{1}{4}$, then $16p > 4$, so $\mathrm{E}[X] = 17 - 16p < 13$. Thus, if we assign more than $\frac{1}{4}$ probability to $X \leq 1$, even when maximizing the expected value, we will always get an expected value below 13. Thus, it is impossible that $P(X \leq 1) > \frac{1}{4}$.

Another way to see this is to use Markov's inequality. Since X is not guaranteed to be non-negative, we cannot use Markov's inequality on X. However, we can define another random variable Y where $Y = 17 - X$. Since $X \leq 17$, then $Y \geq 0$. Thus, we can use Markov's inequality on Y:
We are interested in $P(X \leq 1)$. Since $X = 17 - Y$, $P(X \leq 1) = P(17 - Y \leq 1) = P(Y \geq 16)$.
Using Markov's inequality, $P(Y \geq 16) \leq \frac{\mathrm{E}[Y]}{16} = \frac{\mathrm{E}[17-X]}{16} = \frac{\mathrm{E}[17]-\mathrm{E}[X]}{16} = \frac{17-13}{16} = \frac{4}{16} = \frac{1}{4}$.
This precisely proves that $P(X \leq 1) = P(Y \geq 16) \leq \frac{1}{4}$