Math 416

Worksheet 22. Random variables and selection

Random variables and expectation We have talked a bit about probabilities related to dice rolls and coin-flips. How do we express something like 'the average roll on a die'?

Definition. Fix a probability space (Ω, P). A **(discrete) random variable** is a function $X : \Omega \to \mathbb{R}$.

We often write e.g. $\{X = 4\}$ for the event $\{\omega \in \Omega : X(\omega) = 4\}$. A **Bernoulli** or **indicator** random variable is one that only takes the value 0 or 1. For such an X, there is some $p \in [0, 1]$ for which the following holds.

$$P(X = 1) = p$$
$$P(X = 0) = 1 - p$$

Definition. Two random variables X and Y (on the same probability space (Ω, P)) are **independent** iff for all real numbers x and y the events $\{X = x\}$ and $\{Y = y\}$ are independent, i.e.,

$$P(X = x \text{ and } Y = y) = P(X = x)P(Y = y).$$

Definition. The **expectation** (or **average** or **mean**) of a random variable X is defined to be

$$E[X] = \sum_x xP(X = x),$$

where the sum is over all (real numbers) x in the range of X.

Problem 1. Roll two fair 6-sided dice and let X be the sum of the results. From the definition, find $E[X]$.

Problem 2. Suppose that $X \sim \text{Bernoulli}(p)$, meaning that X is a Bernoulli random variable satisfying $P(X = 1) = p$. Find $E[X]$.

Linearity of expectation Computing expectations from the definition is a pain, as even a sum of two dice shows. Luckily, we don't have to.

Proposition (Linearity of expectation)**.** If $X_1, \ldots, X_n$ are random variables, then ($X_1 + X_2 + \cdots + X_n$ is also a random variable and)

$$E[X_1 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n].$$

Also $E[cX] = cE[X]$.

Problem 3. Compute the expectation of the sum of two fair 6-sided dice now!

Problem 4. We perform n independent Bernoulli trials, each with the same probability p of success. Let

$$X_k = \begin{cases} 1 & \text{if the } k^{\text{th}} \text{ trial succeeds} \\ 0 & \text{if the } k^{\text{th}} \text{ trial fails} \end{cases}$$

(a) What is the expected number Y of successes?
(b) The number of successes is called a binomial(n, p) random variable. Explain this name by computing $P(Y = k)$ directly.

Problem 5. Suppose that we perform independent Bernoulli trials, each with probability p of success, until we see a successful trial. (E.g. roll a die until a 3 appears.) Let Y be the total number of trials (after we have seen the first success). We say that Y follows a geometric(p) distribution.

(a) Explain why $P(Y = n) = (1 - p)^{n-1}p$.
(b) Prove that $E[Y] = \frac{1}{p}$. (This is a calculus exercise.)

Problem 6. Suppose that n hats are taken from n people and returned randomly. Let X be the number of people who get their own hat back. Find $E[X]$.

(*Hint:* Find simpler random variables $I_1, \ldots, I_n$ so that $X = I_1 + I_2 + \cdots + I_n$.)

Order statistics The problem: given an (unsorted) array $x[1 \cdots n]$ of numbers and $k \leq n$, find the k^{th}-smallest element of the array x. (E.g. for $k = 1$ this is the min; for $k = n$ this is the max; for $k = \lfloor n/2 \rfloor$ this is the median.)

Problem 7. Briefly discuss a deterministic algorithm to find the k^{th} order statistic. What is its running time?

For simplicity, let's suppose that all entries of x are distinct. The idea is to choose a *splitter* or *pivot* x_i (e.g. we could always take $x[1]$); split the array into the elements x^- that are below x_i and the elements x^+ that are above x_i; and then determine which of x^-, x^+ has the k^{th} order statistic and run the algorithm recursively on that part.

Problem 8. Fill in the blanks in the algorithm to make it work. Then briefly discuss the proof of correctness and why the algorithm always terminates!

Algorithm 1: the `Select` subroutine

```
1  Select(x[1 ··· n], k):
2  Choose a splitter x_i ;                                   // How?  TBD
3  Form new arrays x^- = ∅ and x^+ = ∅ ;
4  foreach j = 1 to n do
5  |  if x_j < x_i then
6  |  |  append [          ] ;
7  |  if x_j > x_i then
8  |  |  append [          ] ;
9  if |x^-| = k − 1    // splitter is correct
10 then
11 |  return x_i ;
12 else if |x^-| ≥ k    // k^th-smallest is in x^-
13 then
14 |  Select(x^-, k) ;
15 else if |x^-| = l < k − 1    // k^th-smallest is in x^+
16 then
17 |  Select(x^+, [          ]) ;
```

But we have not yet determined how to choose the splitter x_i.

Problem 9. If somehow we managed to choose x_i to be the median every time, then what recurrence relation would the worst-case running time $T(n)$ of this algorithm satisfy? Verify that the solution to your recurrence relation is $O(n)$.

But that's not likely to happen, since the median is one of the order-statistics we might be looking for!

Problem 10. Suppose that the splitter always turns out to be the smallest element (not the one listed first) of the array. (Again, unlikely to happen.) Show that the running time of `Select` in this case is $\Theta(n^2)$.

Problem 11. Suppose that we somehow have a method of producing a splitter (in linear time) x_i such that there are at least $\frac{1}{100}n$ elements of the array smaller than x_i and at least $\frac{1}{100}n$ elements of the array larger than x_i. What recurrence would $T(n)$ satisfy then, and what is its solution?

What happens if we choose the splitter uniformly at random? Let X be the total running time. Say that a splitter x_i is **central** if $\geq 25\%$ of x is above x_i and $\geq 25\%$ of x is below x_i.

Problem 12. What is the probability that we choose a central splitter?

Problem 13. Say that the algorithm is in **phase** j if the number of remaining elements under consideration lies in the interval $(n(\frac{3}{4})^{j+1}, n(\frac{3}{4})^j]$.

(a) What is the expected number of iterations before a central splitter is found?
(b) If the algorithm is in phase j and then chooses a central splitter, then it will move to phase $j+1$. Explain. So what is the expected number of iterations that the algorithm will spend in phase j?

Let X_j be the number of steps taken by the algorithm during phase j.

(c) How is X, the total running time, related to X_1, X_2, etc.?
(d) Explain why the number of steps required for one iteration in phase j is at most $cn(\frac{3}{4})^j$ for some constant c.
(e) Use linearity of expectation to prove that $E[X] \leq 8cn$, and conclude that the expected running time is linear.

Extra practice

Problem 14 (Coupon-collector)**.** Suppose that each kid's meal at your favorite restaurant comes with one of n toys. How many meals do you expect to buy in order to collect all n toys?

Let X be the number of meals purchased until at least one of each toy has been obtained. For each $k \in \{0, \ldots, n-1\}$, let X_k be the number of meals purchased after k toys have been gotten, until the $(k+1)^{\text{st}}$ toy. (So e.g. $X_0 = 1$.)

(a) How is X related to $X_0, \ldots, X_{n-1}$?
(b) How is X_k distributed? What is its expected value?
(c) Show that $E[X] = nH(n)$, where $H(n) = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$ is the n^{th} **harmonic number**.
(d) Use the "estimating sums by integrals" trick from much earlier in the semester (end of Worksheet 4) to show that

$$\ln n \leq H(n) \leq \ln n + 1.$$

Conclude that $E[X]$ is $n \ln n + \Theta(n)$.

15. Find an example of a random variable X for which $P(X = E[X]) = 0$.
16. Let X be a random variable that takes on the value 2^k with probability $\frac{1}{2^k}$ for $k = 1, 2, \ldots$. Convince yourself that such a random variable exists, and show that its expected value diverges (i.e., is infinite).
17. Prove that expectation is linear.
18. In the hats example, see whether you can show that $E[X^2] = 2$.
19. Let X be a geometric(p) random variable. Prove the following *memoryless* property of X:

$$P(X = n + k \mid X > k) = P(X = n).$$