Notes 24

Review sessions in 4-270 on 5/18, 6-8pm and 5/21, 7-9pm.

Slowly straightening out grades. Over finals week will do one last check. How many would be willing to mail in what they think their grades are?

1 Conditional Expectation

We can define a conditional version of expectation, conditioned on a particular event:

Definition 1.1 The conditional expectation of the random variable R given an event A is defined as

$$E[R \mid A] = \sum_{s \in A} R(s) \Pr(s \mid A)$$

.

Theorem 1.2

$$E[R \mid A] = \sum_{r} r \cdot Pr(R = r \mid A).$$

Proof. Analogous to the unconditional version of this statement, as proved in Theorem (1.2) in lectures notes 23. ∎

Example 1 Let R be the number of spots we see in a typical die roll. Then

$$E[R \mid R \leq 4] = 2.5$$

An important use of conditional expectation is in doing a "case analysis" to compute an expectation.
The following theorem provides a way to breaks down the expectation into conditional expectations:

Theorem 1.3 Let $\{A_i\}$ be a set of disjoint events that exhaustively partition the sample space. Then for any random variable R,

$$E[R] = \sum_{i} E[R \mid A_i] \Pr(A_i)$$

The proof of this theorem depends on the following basic conditional probability lemma:

Lemma 1.4 If $\{A_i\}$ partition the sample space, then

$$\Pr(B) = \sum_i \Pr(B|A_i)\Pr(A_i)$$

.

Proof. RHS is just $\sum_i \Pr(B \cap A)$, which adds up to $\Pr(B)$ by definition of a probability measure. ■

We are now ready to prove the theorem (1.3):

Proof.

$$\begin{aligned} E[R] &= \sum_r r \cdot \Pr(R = r) \\ &= \sum_r r \cdot \left(\sum_i \Pr(R = r \mid A_i)\Pr(A_i) \right) \\ &= \sum_i \left(\sum_r r \cdot \Pr(R = r \mid A_i) \right) \Pr(A_i) \\ &= \sum_i E[R \mid A_i]\Pr(A_i) \end{aligned}$$

The first equation follows from lemma (1.4). ■

Example 2 (Assembly plants)

Suppose there are two computer assembly plants F and S.

Plant F is fast but sloppy:
They make 70% of the computers, but a computer from their plant has a 2% probability of breaking down on any given day of use.

Plant S is slow and careful:
They make 30% of the computers, but the breakdown probability is only 1%.

These computers get randomly mixed, so I buy one at random from the store.
What is the expected number of days my computer works?

Let D be the random variable giving the number of days my computer works.
Let I_S be the event that I bought the computer from S, and let I_F be that I bought it from F.

We decompose this problem using conditional expectations.
That is, we calculate separately the condition expectation of the number of working days for S and for F,
then combine using the conditional expectation theorem.

The expected number of working days for a computer from S is obtained from an MTF calculation, that is, a negative binomial distribution with failure probability $p = .01$.
Recall this is just

$$E[D \mid I_S] = \frac{1-p}{p} = 99$$

Similarly, the expected number of working days for a computer from F is

$$E[D \mid I_F] = \frac{1-.02}{.02} = 49$$

.
Now to calculate the expected number of working days for the computer when we don't know which factory it came from, we use the conditional expectation theorem:

$$E[D] = E[D|I_S]\Pr(I_S) + E[D|I_F]\Pr(I_F) = 99(.30) + 49(.70) = 64$$

2 Expectations of Products

Last time, we saw expectation of sums.

Now consider the expectation of the *product* of two random variables.

We might hope for the identity:
$E[R_1 R_2] = E[R_1] \cdot E[R_2]$

But this does not hold in general.
However, it does hold in the important special case when R_1 and R_2 are independent:

Theorem 2.1 Let R_1 and R_2 be independent random variables.
Then $E[R_1 R_2] = E[R_1] \cdot E[R_2]$.

Proof. We have

$$\begin{aligned}
E[R_1R_2] &= \sum_r r \cdot Pr(R_1R_2 = r) \\
&= \sum_r r \sum_{r_1,r_2 | r_1r_2=r} Pr(R_1 = r_1 \text{ and } R_2 = r_2) \\
&= \sum_r \sum_{r_1,r_2 | r_1r_2=r} rPr(R_1 = r_1 \text{ and } R_2 = r_2) \\
&= \sum_r \sum_{r_1,r_2 | r_1r_2=r} r_1r_2Pr(R_1 = r_1 \text{ and } R_2 = r_2) \\
&= \sum_{r_1,r_2} r_1r_2Pr(R_1 = r_1 \text{ and } R_2 = r_2) \\
&= \sum_{r_1,r_2} r_1r_2Pr(R_1 = r_1) \cdot Pr(R_2 = r_2).
\end{aligned}$$

The first equality arises by decomposing the probability of an event into the sum of the probabilities of disjoint events whose union is the original event.
The next 3 steps are just calculations.
The last equality follows from the independence of R_1 and R_2.

Now, we can factor this expression as

$$\left(\sum_{r_1} r_1 \cdot Pr(R_1 = r_1)\right)\left(\sum_{r_2} r_2 \cdot Pr(R_2 = r_2)\right).$$

which is equal to $E[R_1]E[R_2]$.
Therefore,

$$E[R_1R_2] = E[R_1] \cdot E[R_2].$$

■

Example 3 Rolling 2 fair, 6-sided dice independently.
Let R_1, R_2 be values of the first and second dice, respectively.
Since these are independent, we have that

$$E[R_1R_2] = E[R_1] \cdot E[R_2] = \frac{7}{2} \cdot \frac{7}{2} = \frac{49}{4}.$$

Now, what if the two dice are rigged so that they always give the same value as one another? In this case, R_1 and R_2 always assume the same values, and we have

$$E[R_1R_2] = \sum_{i=1}^{6} i^2 \cdot \frac{1}{6} = \frac{91}{6} \neq E[R_1]^2 = (\frac{7}{2})^2.$$

So this is a case of two dependent random variables for which the expectation of their product is not the product of their expectations.

Finally, by induction on n, it is easy to prove the following corollary of Theorem 2.1

Theorem 2.2 If $R_1, \ldots, R_n$ are mutually independent, then

$$E[R_1 R_2 \cdots R_n] = (E[R_1])(E[R_2]) \cdots (E[R_n]).$$

3 Wald's Theorem

The next topic is a theorem that allows another kind of combination of experiments – one in which a probability distribution determines how many experiments are performed.
Let's start with an example:

Example 4 Suppose I flip 3 standard coins and then roll a die for each coin that comes up heads.
How many total spots do I expect to see?
Well, if H is a random variable denoting the number of heads I see, and $D_1, D_2, \ldots$ are random variables for die rolls, then the random variable describing the number of spots can be written as

$$N = \sum_{i=1}^{H} D_i.$$

This notation is a bit strange, because the upper summation index is a random variable.
It makes sense, though – for each element of the sample space, I calculate the value of the random variable N by summing up as indicated, based on the values of the other random variables at the same point.

How can I compute the expectation? One way is by conditional expectations, conditioning on the number of coins that come up heads:

$$\begin{aligned}
E[N] &= \sum_{k=0}^{3} E[N \mid H = k] \Pr(H = k) \\
&= \sum_{k=0}^{3} E[\sum_{i=1}^{k} D_i] \Pr(H = k) \\
&= \frac{1}{8}0 + \frac{3}{8}(E[D_1]) + \frac{3}{8}(E[D_1] + E[D_2]) \\
&\quad + \frac{1}{8}(E[D_1] + E[D_2] + E[D_3]) \\
&= \frac{3}{8}3.5 + \frac{3}{8}7 + \frac{1}{8}10.5 \\
&= \frac{21}{4}
\end{aligned}$$

The first equality uses the conditional expectation theorem.
The second specializes the sum to the case where the upper bound is k.

Is there an easier way to figure this out?
Notice that $E[H] = \frac{7}{2}$, and $E[D] = \frac{3}{2}$, so the expectation turns out to be $E[H] \cdot E[D]$.
But why?

Doesn't follow from linearity of expectation (rule for expectation sums), because we don't know how many copies of $E[D]$ to add up.

Doesn't follow from the product rule.
That would apply if I multiplied the number of heads by the outcome of rolling one die.

But it is true and arises from a new result called *Wald's Theorem*:

Theorem 3.1 Let Q and R be independent random variables, where Q takes on natural number values.
Let T be the sum of Q samples of R, that is:

$$T = \sum_{i=1}^{Q} R_i.$$

Then $E[T] = E[Q] \cdot E[R]$.

Proof.

Consider the following sample space: generate an infinite sequence $R_1, R_2, \ldots$ of samples of R, and generate a random Q, and then consider $\sum_{j=1}^{Q} R_j$. "Clearly," this is the same as our original experiment, since we are just ignoring the extra R_i (actually this takes some proving, but we'll ignore this technical problem). Define indicator variables

$$I_j = \begin{cases} 1 & \text{if } Q > R_j, \\ 0 & \text{otherwise.} \end{cases}$$

Then it is easy to see that

$$T = \sum I_j R_j.$$

It follows that

$$E[T] = \sum E[I_j R_j].$$

Now we use the fact that I_j and R_j are independent to write this as

$$\begin{aligned}
\sum E[I_j]E[R_j] &= \sum E[I_j]E[R] \\
&= (\sum E[I_j])E[R] \\
&= E[\sum I_j]E[R] \\
&= E[Q]E[R].
\end{aligned}$$

∎

Wait, why are I_j and R_j independent? Well, because Q and R_j are:

$$\begin{aligned}
\Pr[I_j = 1 \wedge R_j = r] &= \Pr[(Q \geq j) \wedge (R_j = r)] \\
&= \Pr[(\vee_{k \geq j}(Q = k)) \wedge R_j = r] \\
&= \Pr[\vee_{k \geq j}((Q = k) \wedge R_j = r)] \\
&= \sum_{k \geq j} \Pr[Q = k \wedge R_j = r] \\
&= \sum_{k \geq j} \Pr[Q = k \cdot] \Pr[R_j = r] \\
&= (\sum_{k \geq j} \Pr[Q = k) \Pr[R_j = r] \\
&= \Pr[\vee_{k \geq j} Q = k] \Pr[R_j = r)] \\
&= \Pr[I_j = 1] \Pr[R_j = r]
\end{aligned}$$

We can obtain alternative proof by using conditional expectations as in the example:

Proof. Picture the random experiment as follows.
First we generate an infinite number of samples of R, call them $R_1, R_2, \ldots$.; then we use Q to decide how many of these copies to add up in computing T.
This experiment works on a different sample space from the original, but gives the same distribution.

Then:

$$
\begin{aligned}
E[T] &= \sum_k E[T \mid Q = k] \cdot \Pr(Q = k) \\
&= \sum_k E[\sum_{i=1}^{k} R_i] \cdot \Pr(Q = k) \qquad (*) \\
&= \sum_k \left(\sum_{i=1}^{k} E[R_i] \right) \cdot \Pr(Q = k) \\
&= \sum_k \left(\sum_{i=1}^{k} E[R] \right) \cdot \Pr(Q = k) \\
&= \sum_k (kE[R]) \cdot \Pr(Q = k) \\
&= E[R] \sum_k k \cdot \Pr(Q = k) \\
&= E[R]E[Q]
\end{aligned}
$$

■

One of the lines in this proof – needs a little closer attention.
Namely, look at the line labeled (*), where we say that

$$E[\sum_{i=1}^{Q} R_i \mid Q = k] = E[\sum_{i=1}^{k} R_i].$$

In fact, this is the place where we use the independence of the random variables Q and R.

A bit more carefully:
The LHS can be rewritten as

$$E[\sum_{i=1}^{k} R_i \mid Q = k]$$

We want to show this is equal to the RHS.
Now, independence of R and Q can be used to show that, for any r:

$$\Pr(\sum_{i=1}^{k} R_i = r | Q = k) = \Pr(\sum_{i=1}^{k} R_i = r)$$

This can be used to show that the conditional expectation is the same as the unconditional expectation:

$$E[\sum_{i=1}^{k} R_i \mid Q = k] = E[\sum_{i=1}^{k} R_i]$$

as needed.

If Q and R are dependent, the theorem might not hold.

3.1 Branching Processes

We close this section with an application of Wald's Theorem to *branching processes*. Branching processes are used to model population growth, nuclear chain reactions, mutations, and contagious diseases.

Example 5 (Contagious disease)

Suppose someone is infected with a new, incurable disease.
Every day, he encounters n healthy people and infects each of them with probability p.
What is the expected number of infected people after k days?

Let S_k be a random variable giving the number infected people after k days.
We can write a recurrence for S_{k+1} in terms of S_k.

Let the sick people after day k be numbered $1, \ldots, S_k$.
Let H_i be a random variable giving the number of healthy people infected by sick person i in one day.
Then H_i has a binomial distribution with parameters n and p.
Then:

$$S_{k+1} = S_k + \sum_{i=1}^{S_k} H_i$$

It follows from linearity of expectation that

$$E[S_{k+1}] = E[S_k] + E[\sum_{i=1}^{S_k} H_i].$$

Then by Wald's theorem,

$$E[S_{k+1}] = E[S_k] + E[S_k] \cdot E[H]$$

where H is a variable with the binomial distribution with parameters n and p.
Thus, $E[H] = np$.

So

$$E[S_{k+1}] = E[S_k] + E[S_k] \cdot np = (1 + np)E[S_k]$$

Also, $E[S_0] = 1$.
So we have a recurrence involving these expectations.
Writing e_i for $E[S_i]$, we have:

$$e_{k+1} = (1 + np)e_k$$

and

$$e_0 = 1$$

It follows that $e_k = (1 + np)^k$.

Eventually, we will have more sick people than people. What erroneous assumption is in our model? That each sick person meets a different collection of 10 healthy people. Once there are enough sick people, it will become less likely for a sick person to meet so many healthy people. Also, many sick people will encounter the same healthy person. But our "approximation" will be valid as long as the sick population is not too big.

Now suppose any sick person recovers from the disease each day with probability r. Let B_i be the indicator variable for whether the i^{th} sick person gets better. Then $E[B_i] = r$. Then

$$S_{k+1} = S_k + \sum_{i=1}^{S_k}(H_i - B_i)$$

It follows that

$$E[S_{k+1}] = E[S_k] + E[S_k] \cdot E[H] - E[S_k]E[B_i]$$

and so

$$\begin{aligned} e_{k+1} &= e_k(1 + np - r) \\ &= (1 + np - r)^k \end{aligned}$$

Therefore, the progress of the disease is determined entirely by the parameter $(1+np-r)$. If this exceeds 1, it looks like we have an epidemic on our hands (but expectation isn't telling the whole story: there's a chance the first person gets better without making anyone sick). If less than 1, it looks pretty innocuous. In fact, next time we will see that if less than 1, the disease almost certainly dies out after a certain number of days.

4 Markov's Theorem

We spent a lot of time talking about the expected values of random variables. One thing we saw is that it is possible for a random variable to have expected value 1 even though it is never actually anywhere near 1 (e.g., if I flip two coins that are tied together, what is the expected number of heads?). This week, we are going to study *deviations from expectation.* We will study the likelihood that a random variable has a value distant from its expectation and give methods for proving that probability is small. Intuitively, the expected value of a random variable is the value that we expect it to obtain.

For example, if we flip 100 fair coins, the expected number of heads is 50. In fact, the probability that we get exactly 50 heads is low, but there is a pretty good chance that we get close to 50 heads. In particular, the probability that we deviate substantially from the expected value in this case is extremely small: The probability of getting more than 75 or less than 25 heads is less than one in a billion. This is because the distribution of the number of heads, which is a binomial distribution, is very tightly centered around its mean. This property holds for any binomial distribution, even biased ones. Now, for arbitrary random variables, it is not always the case that the distribution is tightly centered around the mean. For example, if we glue 100 coins together, the probability of getting close to the mean of 50 vanishes, as the random variable always takes on either the value 0 or 100. Even so, it turns out that we can usually say something about the probability of deviating from the mean. For example, in the latter example, the random variable never takes on values that are more than 2 times the mean.

Markov's Theorem gives the most general bound on the probability of deviating from the mean. Markov's Theorem says that if a random variable takes on only non-negative values, then it is unlikely to greatly exceed its mean;

Theorem 4.1 If $R : S \to \mathbb{R}^+$ is a random variable, then for all $x \geq 0$,

$$Pr(R \geq x) \leq \frac{E[R]}{x}.$$

Proof. (of theorem)

$$\begin{aligned} E[R] &= \sum_z z \Pr(R = z) \\ &\geq \sum_{z \geq x} z \Pr(R = z) \\ &\geq \sum_{z \geq x} x \Pr(R = z) \\ &= x \sum_{z \geq x} \Pr(R = z) \\ &= x \Pr(R \geq x) \end{aligned}$$

■

In other words, if $E[R]$ is small, then R is not likely to be large. Markov's Theorem is often expressed in the following alternative form, which can be stated as a corollary:

Corollary 4.2 If $R : S \to \mathbb{R}^+$, then for all positive real numbers c,

$$\Pr(R \geq cE[R]) \leq \frac{1}{c}.$$

Proof. (of corollary) Just write x=cE[R] in the theorem. Then

$$\Pr(R \geq cE[R]) \leq \frac{E[R]}{cE[R]} = 1/c$$

■

Example 6 Let R be the weight of a randomly selected person. Suppose we know that $E[R] = 100$, but we don't know what the distribution of R looks like. Then the probability that R is greater than or equal to 200 is, by Markov's Theorem, $\Pr(R \geq 200) = \Pr(R \geq 2 \cdot E[R]) \leq 1/2$. Similarly, $\Pr(R \geq 300) = \Pr(R \geq 3 \cdot E[R]) \leq 1/3$. Why can't everyone have weight ≥ 200 in this example? The answer is that, if this were the case, then the expected weight of a randomly selected person would be greater than 200. Not everyone can "beat" the mean!

Example 7 If I give each person a random hat back at a hat check, what is the probability that at least k people get their own hat back? Let R be the number of people who get their own hat back. Markov's Theorem doesn't give us an exact answer, but, since $E[R] = 1$, it tells us that $\Pr(R \geq k) \leq 1/k$.

Note that Markov's theorem doesn't hold if R can be negative. For example, consider a variable R that is ± 1000 with probabilities 1/2 each. It's expectation is 0! So Markov's theorem is clearly violated.

So where in the proof of Markov's Theorem did we use the non-negativity of R? If we go back and check, we see that non-negativity is required for $E[R] = \sum_x x \Pr(R = x) \geq \sum_{x \geq K} x \Pr(R = x)$ to hold.

There is a symmetric result that gives an upper bound on the probability that a random variable takes on a small value.

Theorem 4.3 If $R \leq L$ for some $L \in \mathbb{R}$, then, for all $x \leq L$:

$$\Pr[R \leq x] \leq \frac{L - E[R]}{L - x}$$

Proof. Consider the event $L - R$. Since $R \leq L$, we have that $L - R \geq 0$, and hence the conditions of Markov's Theorem are satisfied. Hence we have that $\Pr(R \leq x) = \Pr(L - R \geq L - x) \leq \frac{E[L-R]}{L-x} = \frac{L-E[R]}{L-x}$ ■

Example 8 Let R be the score of a randomly selected student on a test, and suppose the average score on a test is $E[R] = 75$, and the max score was 100. Then we can compute $\Pr(R \leq 50)$.

$$\Pr[R \leq 50] \leq \frac{100 - 75}{100 - 50} = 1/2$$

One problem with Markov's Theorem and its corollaries is that it does not give an exact probability, just an upper bound. Sometimes the upper bound it gives is exactly the right answer, but sometimes it can also be pretty far off from the actual value.

Example 9 Suppose we flip 100 fair coins, and let R be the number of heads. Then $E[R] = 50$, and so from Markov's Theorem, $\Pr(R \geq 75) \leq 2/3$. However, $\Pr(R \geq 75)$ is in fact less than one in a billion, so Markov's Theorem is way off the mark!