ORIE 5530: Modeling Under Uncertainty

Lecture 4 (Random Variables 1)

Professor Mark S. Squillante,
Adapted from Professor Itai Gurvich's original notes

We move to random variables. We start with the world of discrete variables where things are kind of more intuitive but nevertheless most of the stuff can be covered. We will later move to continuous random variables. A random variable (RV) is *discrete* if it takes on a finite or countable number of possible values.

We are sometimes interested in a function of the uncertainty rather than in the uncertainty itself. For example, with the roll of two die, instead of being interested in the outcome pair, I might be interested in the sum of the two die. Then, I could define a random variable X for this sum and I will be able to compute the *probability mass function* (pmf) of X: $\mathbb{P}(X = k)$ for different values of k that the RV X can take on by aggregating the corresponding outcomes. So, for example,

$$\mathbb{P}(X = 5) = \mathbb{P}(\{(1, 4), (4, 1), (2, 3), (3, 2)\}) = \frac{4}{36}.$$

We can also compute the *distribution* of X which is $F(k) = \mathbb{P}(X \leq k)$ for each k. For example

$$\mathbb{P}(X \leq 5) = \sum_{k=2}^{5} \mathbb{P}(X = k).$$

This is often referred to as the *cumulative distribution function* (CDF). Of course, different random variables have different cumulative distribution functions. The distribution function captures all the information we need for a random variable. For example, the probability that a random variable X has values between x and y is given by

$$\mathbb{P}(x < X \leq y) = F(y) - F(x).$$

We could also recover the probability mass function by noting, in our example above, that

$$\mathbb{P}(X = k) = \mathbb{P}(k - 1 < X \leq k).$$

A legitimate distribution always has the property that the probability of being smaller than infinity (or greater than $-\infty$) is 1:

$$\lim_{y \to \infty} \mathbb{P}\left(X \leq y\right) = 1 \text{ and } \lim_{y \to -\infty} \mathbb{P}\left(X \leq y\right) = 0.$$

This is nothing mind blowing. It just means that ∞ or $-\infty$ are not an option, but also means that when you sum up the probabilities of all possible outcomes they must sum to 1: for a random variable X taking values $x_1, x_2, \ldots$,

$$\sum_k \mathbb{P}(X = x_k) = 1.$$

4.1 Some Important Distributions

Later we will also cover expectation and variance. In the list below I include those properties. I will use the common abbreviation w.p. for "with probability".

1. **Bernoulli:** $X \sim Bernoulli(p)$ is a RV where

$$X = \begin{cases} 1 & w.p. \;\; p \\ 0 & w.p. \;\; 1 - p \end{cases}$$

Hence, we have $\mathbb{P}(X=1)=p$ and $\mathbb{P}(X=0)=1-p$.
Mean and Variance: $\mathbb{E}[X]=p, \mathbb{V}\text{ar}(X)=p(1-p)$.

2. **Geometric:** $X \sim Geom(p)$ is the number of independent trials until a success, where p is the probability of success in each trial.

$$\mathbb{P}(X=1)=p, \mathbb{P}(X=2)=(1-p)\,p, \cdots, \mathbb{P}(X=k)=(1-p)^{k-1}\,p.$$

Mean and Variance: $\mathbb{E}[X]=1/p, \mathbb{V}\text{ar}(X)=(1-p)/p^2$.

3. **Binomial:** $X \sim Bin(n,p)$ is the number of successes in n independent trials, each having success probability p. $\mathbb{P}(X=k)$ is the probability of seeing k successes out of n Bernoulli trials.

$$\mathbb{P}(X=k)=\binom{n}{k}p^k\,(1-p)^{n-k}.$$

Mean and Variance: $\mathbb{E}[X]=np, \mathbb{V}\text{ar}(X)=np(1-p)$.

4. **Poisson:** $X \sim Poisson\,(\lambda)$ means

$$\mathbb{P}(X=k)=\frac{e^{-\lambda}\lambda^k}{k!} \text{ for } k \in \mathbb{Z}^+.$$

Mean and Variance: $\mathbb{E}[X]=\mathbb{V}\text{ar}(X)=\lambda$.

The binomial and Poisson distributions are closely related through the following result.

Poisson approximation to the binomial: Let $X_{n,p} \sim Bin(n,p)$. Consider a sequence of these random variables such that n is increasing along the sequence but p is decreasing so as to keep $np=\lambda$ constant. Then,

$$\mathbb{P}(X_{n,p}=k) \to \frac{e^{-\lambda}\lambda^k}{k!}.$$

The right hand side is the $Poisson(\lambda)$ distribution. What do we practically do with such a limit result? It says that *with a large number of independent experiments n, each with a small probability of success p*, it is reasonable to approximate the likelihood of k successes by $\mathbb{P}(Y=k)$ where $Y \sim Poisson(np)$.

4.2 Expectation

Definition 1 *Let X be a discrete r.v. taking on values $x_1, x_2, \ldots$. The expectation of X is defined as*

$$\mathbb{E}[X]=\sum_{k=1}^{\infty} x_k \mathbb{P}(X=x_k).$$

For a Binomial r.v, we have

$$\mathbb{E}[X]=\sum_{k=0}^{n} k\binom{n}{k}p^k\,(1-p)^{n-k}=np.$$

For a Geometric r.v, we have

$$\mathbb{E}[X]=\sum_{k=1}^{n} k\,(1-p)^{k-1}\,p=\frac{1}{p}.$$

Expectation of a function of a random variable: If X takes on values $x_1, x_2, \ldots$ and $g(X)$ is a function of the random variable (in our single-order, multi-period inventory example, D was the random variable

representing demand, Q was the inventory level (or quantity of units stocked), and we were interested in part in $g(D) = \max(Q - D, 0)$), it then obtains values $g(x_1), g(x_2), \ldots$ with respective probabilities $\mathbb{P}(X = x_1), \mathbb{P}(X = x_2), \ldots$. Thus,

$$\mathbb{E}[g(X)] = \sum_{k=1}^{\infty} g(x_k)\mathbb{P}(X = x_k).$$

We had examples in class of the application of expectations: For example, the single-order, multi-period inventory model, a.k.a. the newsvendor model (example for function of RVs and optimization).

The *variance* of a random variable X is also just the expectation of a function of the RV. Specifically, let $g(X) = (X - \mathbb{E}[X])^2$. Then,

$$\mathbb{V}\text{ar}(X) = \mathbb{E}[g(X)] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

(Convince yourself that the last inequality is correct by opening up the quadratic).

4.3 Joint distribution, independence and conditional probabilities for Discrete RVs

When we have two (or more) random variables, we can talk about their joint distribution. Consider two RVs X and Y taking values in $\mathcal{X} = \{x_1, \ldots, x_n, \ldots\}$ and $\mathcal{Y} = \{y_1, \ldots, y_n, \ldots\}$, respectively. Then, we can talk about the intersection of the events $A_i = \{X = x_i\}$ and $B_j = \{Y = y_j\}$. Recall that the two events would be independent if

$$\mathbb{P}(A_i \bigcap B_j) = \mathbb{P}(A_i)\mathbb{P}(B_j).$$

We would say that the random variables X and Y are independent if all events A_i and B_j are independent. That is, if

$$\mathbb{P}(\{X = x_i\} \bigcap \{Y = y_j\}) = \mathbb{P}(X = x_i)\mathbb{P}(Y = y_j), \text{ for all } x_i, y_j.$$

To simplify notation we write $\mathbb{P}\{X = x_i, Y = y_j\}$ instead of $\mathbb{P}\{\{X = x_i\} \bigcap \{Y = y_j\}\}$. Then, the collection of joint probabilities

$$\mathbb{P}\{X = x_i, Y = y_j\}, \quad \text{for all } x_i, y_j$$

is the *joint distribution* of (X, Y).

Notice that from the joint distribution we can recover the *marginal distribution*, meaning that of X or Y. Indeed, this is again like with events and the connection might be useful. Before we had that if $B_1, \ldots, B_n$ is a partition, then we can write

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \bigcap B_i).$$

The same holds here. Consider the event $A = \{X = x\}$ and let $B_j = \{Y = y_j\}$. Then,

$$\mathbb{P}(X = x) = \sum_j \mathbb{P}(X = x, Y = y_j).$$

Now (again similar to events) we can ask what is the likelihood of $X = x_i$ given that $Y = y_j$:

$$\mathbb{P}\left(X = x_i | Y = y_j\right) = \frac{\mathbb{P}\left(X = x_i, Y = y_j\right)}{\mathbb{P}\left(Y = y_j\right)}.$$

This follows from the identity for events $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \bigcap B)}{\mathbb{P}(B)}$. Independence then implies that $\mathbb{P}\left(X = x_i | Y = y_j\right) = \mathbb{P}\left(X = x_i\right)$ for all x_i and y_j.

Just like we compute expectation from probability, we can compute conditional expectations from conditional probabilities:

$$\mathbb{E}[X|Y = y] = \sum_x x\mathbb{P}\{X = x|Y = y\}.$$

Notice that we can remove the conditions by multiplying then by the probability of Y. Meaning, we have the *de-conditioning*

$$\mathbb{P}\{X = x\} = \sum_y \mathbb{P}\{X = x|Y = y\}\mathbb{P}\{Y = y\},$$

and

$$\mathbb{E}[X] = \sum_y \mathbb{E}[X|Y = y]\mathbb{P}\{Y = y\}.$$

The following example illustrates how cleverly using conditional expectation (and de-conditioning) can help answer some non-trivial questions.

Example 1 Toss a coin with probability p of getting Heads (H). How many tosses, in expectation, do you need until you get k consecutive H? Let us define two things:

$N_k = \#$ of trials until k consecutive H, and

$M_k = \mathbb{E}[N_k]$.

Our question concerns what is the value of M_k? It is not trivial to determine how to answer this, but it seems we should be able to build a recursion. Specifically, if we know that it took n tosses until we saw a sequence of H of length $k-1$, then it should take at least $n+1$ until we see one of k but possibly more; that is, if the $(n+1)^{st}$ toss is a tail (which happens with probability $1-p$) then we have to re-start all over again so that we have N_k to add. More formally. we have that

$$\mathbb{E}[N_k|N_{k-1} = n] = n + 1 + (1-p)\mathbb{E}[N_k].$$

Upon de-conditioning, we obtain

$$\begin{aligned}
\mathbb{E}[N_k] &= \sum_n \mathbb{E}[N_k|N_{k-1} = n]\mathbb{P}\{N_{k-1} = n\} \\
&= \sum_n (n + 1 + (1-p)\mathbb{E}[N_k])\mathbb{P}\{N_{k-1} = n\} \\
&= \sum_n n\mathbb{P}\{N_{k-1} = n\} + \sum_n (1 + (1-p)\mathbb{E}[N_k])\mathbb{P}\{N_{k-1} = n\} \\
&= \mathbb{E}[N_{k-1}] + (1 + (1-p)\mathbb{E}[N_k]).
\end{aligned}$$

In the last line I just used the fact that $\sum_n \mathbb{P}\{N_{k-1} = n\} = 1$. So we have arrived at the recursion $M_k = M_{k-1} + 1 + (1-p)M_k$, which we can write as

$$M_k = \frac{1 + M_{k-1}}{p}.$$

Now let us apply this recursion with the initial condition $M_0 = 0$. Then, we have $M_1 = \frac{1}{p}$, $M_2 = \frac{(1+1/p)}{p} = \frac{1}{p} + \frac{1}{p^2}$, and so on to get the result for general k:

$$M_k = \frac{1}{p} + \frac{1}{p^2} + \ldots \frac{1}{p^k}.$$

■

Another useful fact is that, if X and Y are independent, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y].$$

This is because

$$\begin{aligned}\mathbb{E}[XY] =& \sum_{(i,j)} x_i y_j \times \mathbb{P}(X = x_i, Y = y_j) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} x_i y_j \times \mathbb{P}(X = x_i, Y = y_j) \\ =& \sum_{i=0}^{\infty} x_i \sum_{j=0}^{\infty} y_j \mathbb{P}(X = x_i)\,\mathbb{P}(Y = y_j) = \sum_{i=0}^{\infty} x_i \mathbb{P}(X = x_i) \sum_{j=0}^{\infty} y_j \mathbb{P}(Y = y_j) \\ =& \mathbb{E}[X]\,\mathbb{E}[Y].\end{aligned}$$

Definition 1 *The covariance between two random variables X and Y is defined as*

$$\mathbb{C}ov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Covariance is not always an informative quantity because it does not allow for comparisons of relationship. Say you have a quantity Y that you are trying to predict and two dependent variables X_1, X_2, and suppose that $\mathbb{C}\text{ov}(Y, X_1) > 0$ and $X_2 = 10 \cdot X_1$. Then,

$$\mathbb{C}\text{ov}(Y, X_2) = \mathbb{C}\text{ov}(Y, 10 \cdot X_1) = 10 \cdot \mathbb{C}\text{ov}(Y, X_1).$$

Do you want to conclude that X_2 is a much better predictor of Y. Probably not. After all they are the same variable just scaled. Correlation corrects for this scale effect and allows you to compare normalized relationship.

Definition 2 *The correlation between two random variables X and Y is defined as*

$$\rho(X,Y) = \frac{\mathbb{C}ov(X,Y)}{\sqrt{\mathbb{V}ar(X)}\sqrt{\mathbb{V}ar(Y)}}.$$

Notice (check) that ρ is scale insensitive. Meaning that for $a, b > 0$

$$\rho(aX, bY) = \rho(X, Y).$$

So, going back to the example above, $\rho(Y, X_2) = \rho(Y, X_1)$ captures the fact that X_1 and X_2 would equally be good predictors of Y.

Further observe that if X and Y are independent, then $\mathbb{C}\text{ov}(X, Y) = 0$ and $\rho(X, Y) = 0$.

4.4 Some useful inequalities

The beauty of the inequalities below is that the more information you use the better bounds you get. These bounds can be extremely useful when the underlying structure of the problem is too complicated for us to actually derive the distribution, yet we can derive certain quantities like the expectation or standard deviation. Below are some basic examples we covered, followed by a more elaborate example. I will use the common abbreviation i.i.d. for "independent and identically distributed".

Theorem 1 (Markov inequality) *For a positive RV X,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \text{ for any } a > 0.$$

Proof: Note that

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k\mathbb{P}(X = k) = \sum_{k=1}^{a-1} k\mathbb{P}(X = k) + \sum_{k=a}^{\infty} k\mathbb{P}(X = k)$$
$$\geq \sum_{k=a}^{\infty} k\mathbb{P}(X = k) \geq a \sum_{k=a}^{\infty} \mathbb{P}(X = k) = a\mathbb{P}(X \geq k).$$

Rearranging the terms, we obtain the desired inequality. ■

Example 2 *Let $X \sim Bin(n, p)$. We have $\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ and $\mathbb{E}[X] = np$. Suppose we have $n = 200$ and $p = 0.1$. Then, $\mathbb{E}[X] = np$ and, by the Markov inequality,*

$$\mathbb{P}(X \geq 120) \leq \frac{\mathbb{E}[X]}{120} = \frac{1}{6}.$$

Theorem 2 (Chebyshev Inequality) *If X is a RV with variance $\mathbb{V}ar(X)$, then*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathbb{V}ar(X)}{a^2}.$$

Proof: Since $(X - \mathbb{E}[X])^2$ is a nonnegative RV, applying Markov's inequality with $a = k^2$ renders $\mathbb{P}[(X - \mathbb{E}[X])^2 \geq k^2] \leq \frac{\mathbb{E}[(X-\mathbb{E}[X])^2]}{k^2}$. But since $(X - \mathbb{E}[X])^2 \geq k^2$ if and only if $|(X - \mathbb{E}[X])| \geq k$, this is equivalent to

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq k] \leq \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}{k^2} = \frac{\mathbb{V}\text{ar}[X]}{k^2}.$$

■

Example 3 *For a random variable $X \sim Bin(n, p)$, $\mathbb{V}ar(X) = np(1-p)$. Instead of the $\frac{1}{6}$ in the previous example, we get a sharper bound this time:*

$$\mathbb{P}(X \geq 120) = \mathbb{P}(X - 20 \geq 100) \leq \mathbb{P}(|X - 20| \geq 100) \leq \frac{\mathbb{V}ar(X)}{100^2} = \frac{18}{100^2} = 0.0018.$$

Theorem 3 (Chernoff Bound) *Let X be a RV.*

$$\mathbb{P}(X \geq a) = \mathbb{P}(g(X) \geq g(a)) \leq \frac{\mathbb{E}[g(X)]}{g(a)}.$$

Let $g(x) = e^x$. We have

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}\left[e^X\right]}{e^a},$$

which is the Chernoff bound.

Example 4 *Let $X \sim Bin(n, p)$. Then*

$$\mathbb{E}\left[e^X\right] = \mathbb{E}[e^{\sum_{i=1}^n Z_i}] = \prod_{i=1}^{n} \mathbb{E}\left[e^{Z_i}\right] = (pe + 1 - p)^n.$$

where the Z_i are i.i.d. Bernoulli(p).

By Chernoff's bound, for $n = 200$ and $p = 0.1$, we have

$$\mathbb{P}(X \geq 120) \leq \frac{(0.1 \times e + 0.9)^2 00}{e^{120}} \approx 1.4 \times 10^{-44},$$

which is much sharper than the previous 2 bounds.

Using Chernoff's to evaluate random algorithms (an example)

The original formulation, which I used in class, concerns a set of n nodes and m subsets of the set of nodes. But you can think of the nodes as individuals and the subsets as groups or teams of individuals, in addition to many other applications.

You have a collection of n nodes U and m subsets of U: $S_1, \ldots, S_m$. The subsets can have non-empty intersections (there might be nodes that belong to multiple subsets). Now suppose I want to "paint" each of the n nodes in one of two colors, say Blue or Red, while making sure that the number of blue and red nodes in each subset S_i is as balanced as possible.

Specifically, the imbalance in subset S_i is given by

$$Disc(S_i) = |\#redsinS_i - \#blueinS_i|,$$

and I am interested in minimizing the largest discrepancy in my set of nodes. That is, I want an algorithm to minimize

$$\max_i Disc(S_i).$$

A deterministic algorithm could be very complicated. Instead, we use a "stupid" random algorithm that paints a node blue with probability 1/2 and red with probability 1/2. The question we then want to ask is *how bad/good is this algorithm?*

Analyzing this algorithm exactly is very complicated because there can be overlap between subsets, in which case there is dependency, and the structure of overlap can be arbitrarily crazy. **Chernoff's bound comes to the rescue here!!!**

With the random algorithm the outcome is going to be, well, random. Let us say I want to get a number so that I can claim with a certainty of $1 - 1/m$ that the max discrepancy $\max_i Disc(S_i)$ is below this number. In other words, find a number $K = 2a$, so that you can guarantee that, under the randomized algorithm,

$$\mathbb{P}\{\max_i Disc(S_i) > 2a\} \leq \frac{1}{m}.$$

What is this number m? I claim that $K = 2a = \sqrt{12n \log m}$ is this number. Namely, the following is true:

$$\mathbb{P}\{\max_i Disc(S_i) > \sqrt{12n \log m}\} \leq \frac{1}{m}.$$

What follows is a proof of this claim.

If $X \sim Bin(n, p)$, then $X = \sum_i Z_i$ where Z_i are i.i.d. Bernoulli RVs, and we apply the Chernoff bound to obtain

$$\mathbb{P}(X \geq a) \leq \frac{(pe + (1-p))^n}{e^a}.$$

It is useful, instead, to think of deviations from the mean: $|X - \mathbb{E}[X]|$. Using similar ideas we can derive the bound

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \delta \mathbb{E}[X]) \leq e^{\frac{-\delta^2 \mathbb{E}[X]}{3}}.$$

Let us focus on a single subset, say S_1, and suppose k is the number of nodes in S_1. Notice that (since $\#red = k - \#blue$), we have

$$Disc(S_1) = |\#red - \#blue| = 2|\#red - k/2|.$$

Thus, $Disc(S_1) \geq 2a$ if and only if $|\#red - k/2| \geq a$.

Let us define the random variable R_1 to be the number of red nodes in subset 1. This is a $Bin(k, 1/2)$ random varaible and $\mathbb{E}[R] = k/2$. Then,

$$\mathbb{P}(Disc(S_1) \geq 2a) = \mathbb{P}(|R_1 - k/2| \geq a) = \mathbb{P}(|R - \mathbb{E}[R]| \geq a) \leq e^{-\frac{a^2}{3\mathbb{E}[R]}} = e^{-\frac{2a^2}{3k}} \leq e^{-\frac{2a^2}{3n}}.$$

In the first inequality we used the Chernoff bound with $\delta = (a)/\mathbb{E}[R]$; in the second to last equality we used $\mathbb{E}[R] = k/2$; and in the last inequality we used the fact that $k \leq n$. We can repeat the same arguments for any subset. Thus, we have established that, for any subset i,

$$\mathbb{P}(Disc(S_i) \geq 2a) \leq e^{-\frac{2a^2}{3n}}.$$

Finally, notice that the event $\{\max_i Disc(S_i) \geq 2a\}$ is the union of the events $\{Disc(S_i) \geq 2a\}$ for $i = 1, \ldots, m$. This is a good point to recall as a useful fact. Next recall that, if you take the union of two events A and B, then $\mathbb{P}(A \bigcup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$; this also applies to muliple events. So we have

$$\mathbb{P}(\max_i Disc(S_i) > 2a) \leq \sum_{i=1}^{m} \mathbb{P}(Disc(S_i) > 2a) \leq m \max_i \mathbb{P}(Disc(S_i) > 2a) \leq me^{-\frac{2a^2}{3n}}.$$

Now choose $a = \sqrt{3n \log m}$ (so that $2a = \sqrt{12n \log m}$). Then, you can verify that $e^{-\frac{2a^2}{3n}} = \frac{1}{m^2}$ and we have the desired bound.

4.5 Introduction to Decision Trees

The example covered in class is provided in the file Decision_Tree_Example.pdf. There are some minor differences in notation and terms as follows. The decisions at hand are whether to build a small-sized complex (denoted B_S in class and d_1 in the handout), a medium-sized complex (denoted B_M in class and d_2 in the handout), and a large-sized complex (denoted B_L in class and d_3 in the handout). The sources of uncertainty (called "states of nature" in the handout) are with respect to demand, where two (discrete) possibilities were identified for the model: strong demand (denoted D_S in class and s_1 in the handout); and weak demand (denoted D_W in class and s_2 in the handout). Table 4.1 provides the estimated rewards (profits) for the various possible model outcomes.

Decision	Strong Demand (D_S)	Weak Demand (D_W)
B_S	8	7
B_M	14	5
B_L	20	-9

Table 4.1: Expected Rewards for Decision Tree Example.

For the application of Bayes rule, you can recover the original formulation in terms of $\mathbb{P}[D_S|F]$, $\mathbb{P}[D_S|U]$, $\mathbb{P}[D_W|F]$, $\mathbb{P}[D_W|U]$ upon being given $\mathbb{P}[F|D_S] = 0.9$, $\mathbb{P}[U|D_S] = 0.1$, $\mathbb{P}[F|D_W] = 0.25$, $\mathbb{P}[U|D_W] = 0.75$ and recalling $\mathbb{P}[F] = 0.77$ and $\mathbb{P}[U] = 0.23$. (Convince yourself that this is indeed the case).