2 Discrete Random Variables

Big picture: We have a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Let X be a real-valued function on Ω. Each time we do the experiment we get some outcome ω. We can then evaluate the function on this outcome to get a real number $X(\omega)$. So $X(\omega)$ is a random real number. It is called a *random variable*, or just RV. We can define events in terms of X, e.g., $X \geq 4$, $X = 2$, ... We would like to compute the probability of such events.

Example: Roll a dice 10 times. We can take Ω to be all 10-tuples whose entries are $1, 2, 3, 4, 5$ or 6. Let $X =$ sum of the 10 rolls, $Y =$ the number of rolls with a 3. There are two random varibles.

In this course RV's will come in two flavors - discrete and continuous. For purposes of this course, a RV is discrete if its range is finite or countable, and is continuous otherwise.

Very important idea: The sample space Ω may be quite large and complicated. But we may only be interested in one or a few RV's. We would like to be able to extract all the information in the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ that is relevant to our random varialbe(s), and forget about the rest of the information contained in the probability space.

2.1 Probability mass function

Definition 1. *A discrete random variable X on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ is a function $X \to \mathbb{R}$ such that the range of X is finite or countable and for $x \in \mathbb{R}$, $\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F}$. The probability mass function (pmf) $f(x)$ of X is the function on $\mathbb{R}$ given by*

$$f(x) = \mathbf{P}(X = x) = \mathbf{P}(\{\omega \in \Omega : X(\omega) = x\})$$

Notation/terminology: If we have more than one RV, then we have more than one pmf. To distinguish them we use $f_X(x)$ for the pmf for X, $f_Y(x)$ for the pmf for Y, etc. Sometimes the pmf is called the "density function" and sometimes the "distribution of X." The latter is really confusing as the term "distribution function" refers to something else.

Example: Roll two four sided dice. Let X be their sum. It is convenient to give the pmf of X in table.

x	2	3	4	5	6	7	8
$f_X(x)$	1/16	2/16	3/16	4/16	3/16	2/16	1/16

The next theorem says that the probabilty mass function realizes our goal of capturing all the information in the probability space that is relevant to X. The theorem says that we can compute the probability of an event defined just in terms of X from just the pmf of X. We don't need $\mathbf{P}$.

Theorem 1. *Let X be a discrete RV. Let $A \subset \mathbb{R}$. (Note that A is not an event, but $X \in A$ is.) Then*

$$\mathbf{P}(X \in A) = \sum_{x \in A} f(x)$$

The sum above merits some comment. We might as well just sum over the values x which are in A and in the range of X since if they are not in the range, then $f(x) = 0$. So the sum of the nonzero terms in the above is countable or finite.

Proof. The proof of the theorem is trivial. First note that if replace A by its intersection with the range of X, then the event $X \in A$ does not change and the sum in the theorem does not change since $f_X(x) = 0$ when x is not in the range of X. So we might as well assume that A is a subset of the range of X. In particular this means A is finite or countable. We can write the event $X \in A$ as the disjoint union over $x \in A$ of the events $X = x$. These events are disjoint, so by the countable additivity property

$$\mathbf{P}(X \in A) = \mathbf{P}(\cup_{x \in A} \{X = x\}) = \sum_{x \in A} \mathbf{P}(X = x) = \sum_{x \in A} f(x)$$

□

Another very important idea Suppose we have two completely different probability spaces $(\omega_1, \mathcal{F}_1, \mathbf{P}_1)$ and $(\omega_2, \mathcal{F}_2, \mathbf{P}_2)$, and RV's X_1 on the first and X_2 on the second. Then it is possible that X_1 and X_2 have the same range and identical pmf's, i.e., $f_{X_1}(x) = f_{X_2}(x)$ for all x. If we only look at X_1 and X_2 when we do the two experiments, then we won't be able to tell the experiements apart.

Definition 2. *Let X_1 and X_2 be random variables, not necessarily defined on the same probability space. If $f_{X_1}(x) = f_{X_2}(x)$ for all x, then we say X_1 and X_2 are identically distributed.*

If you are a mathematician, a natural question is what functions can be pmf's? The following theorem gives an answer

Theorem 2. *Let $x_1, x_2, x_3, \cdots$ be a finite or countable set of real numbers. Let $p_1, p_2, p_3, \cdots$ be positive numbers with $\sum_n p_n = 1$. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and an random variable X on Ω such that the range of X is $\{x_1, x_2, \cdots\}$ and the pmf of X is given by $f_X(x_i) = p_i$.*

Approximate end of September 2 lecture

2.2 Discrete RV's - catalog

Since different experiments and random variables can give rise to the same probability mass functions, it is possible that certain pmf's come up a lot in applications. This is indeed the case, so we begin to catalog them.

Bernoulli RV (one parameter $p \in [0, 1]$) This is about as simple as they get. The RV X only takes on the values 0 and 1.

$$p = \mathbf{P}(X = 1), \quad 1 - p = \mathbf{P}(X = 0)$$

We can think of this as coming from a coin with probability p of heads. We flip it only once, and $X = 1$ corresponds to heads, $X = 0$ to tails.

Binomial RV (two parameters: $p \in [0, 1]$, positive integer n) The range of the random variable X is $0, 1, 2, \cdots, n$.

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

Think of flipping an unfair coin n times. p is the probability of heads on a single flip and X is the number of head we get out of the n flips. The parameter n is often called the "number of trials."

We review some stuff. The notation $\binom{n}{k}$ is read "n choose k". It is defined by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}$$

It gives the number of ways of picking a subset of k objects out of a set of n distinguishable objects. (Note that by saying a "subset" of k objects, we mean that we don't care about the ordering of the k objects.) The binomial theorem is the algebraic identity

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k \, y^{n-k}$$

We derive the formula for $\mathbf{P}(X = k)$ as follows. An outcome that contributes to the event $X = k$ must have k heads and $n-k$ tails. The probability of any one such sequence of flips is $p^k(1-p)^{n-k}$. We need to figure out how many such sequences there are. The is the same as the following counting problem. We have k H's and $n-k$ T's and we have to arrange them in a line. There are $\binom{n}{k}$ ways to choose the positions for the H's and then the T's have no freedom - they go into the remaining empty slots.

Poisson RV (one parameter: $\lambda > 0$) The range of the random variable X is $0, 1, 2, \cdots$.

$$\mathbf{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Note that

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$$

which implies that the sum of the $\mathbf{P}(X = k)$ is 1 as it should be. There is no simple experiment that produces a Poisson random variable. But it is a limiting case of the binomial distribution and it occurs frequently in applications.

Geometric (one parameter: $p \in [0,1]$) The range of the random variable X is $1, 2, \cdots$.

$$\mathbf{P}(X = k) = p(1-p)^{k-1}$$

Check that the sum of these probabilities is 1. Think of flipping an unfair coin with p being the probability of heads until we gets heads for the first time. Then X is the number of flips (including the flip that gave heads.)

Caution: Some books use a different convention and take X to be the number of tails we get before the first heads. In that case $X = 0, 1, 2, ...$ and the pmf is different.

Negative binomial (two parameters: $p \in [0, 1]$ and a positive integer n) The range of the random variable X is $n, n+1, n+2, n+3, \cdots$.

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

Think of an unfair coin with probability p of heads. We flip it until we get heads a total of n times. Then we take X to be the total number of flips including the n heads. So X is at least n.

We derive the formula as follows. If $X = k$ then there are a total of k flips. Out of them, exactly n are heads. One of these heads must occur on the last (kth) flip. For a particular such sequence the probability is $p^n(1-p)^{k-n}$. We need to count how many such sequences there are. The kth flip must be heads. The first $k-1$ flips contain $n-1$ heads and $k-n$ tails. They can be in any arrangment. So there are $\binom{k-1}{n-1}$ such sequences of them.

2.3 Functions of discrete RV's

Recall that if X is a RV, then it is a function from the sample space Ω to the real numbers $\mathbb{R}$. Now let $g(x)$ be a function from $\mathbb{R}$ to $\mathbb{R}$. Then $Y = g(X)$ is a new random variable. Note that what we are doing is composing two functions. The notation hides the arguments of the functions. We have $Y(\omega) = g(X(\omega))$. What is the probability mass function of Y? As we will see, this is a relatively simple computation. When we come to continuous random variables it will be more involved.

Proposition 1. *Let X be a random variable, g a function from $\mathbb{R}$ to $\mathbb{R}$. Define a new random varialbe by $Y = g(X)$. Then the pmf function of Y is given by*

$$f_Y(y) = \sum_{x: g(x)=y} f_X(x) = \sum_{x \in g^{-1}(y)} f_X(x)$$

Proof. By definition $f_Y(y)$ is $\mathbf{P}(Y = y)$. The event is the disjoint union of the events $X = x$ where x ranges over x such that $g(x) = y$. More verbosely,

$$\{\omega : Y(\omega) = y\} = \bigcup_{x:g(x)=y} \{\omega : X(\omega) = x\}$$

The formula in the proposition follows. □

Example Roll a four-sided die twice. Let X be the first roll minus the second roll. Let $Y = X^2$. Find the pmf of X and use it to find the pmf of Y.

2.4 Expected value

We start with a really simple example. Suppose that a RV X only takes on the three values $1, 2, 3$ and the pmf is given in the table. We do the

x	1	2	3
$f_X(x)$	0.2	0.5	0.3

experiment a million times and record the one million values of X that we get. Then we average these million numbers. What do we get? In our list of one million values of X, we will get approximately $200,000$ that are 1, approximately $500,000$ that are 2, and approximately $300,000$ that are 3. So the average will be approximately

$$\frac{0.2 \times 10^6 \times 1 + 0.5 \times 10^6 \times 2 + 0.3 \times 10^6 \times 3}{10^6} = 0.2 \times 1 + 0.5 \times 2 + 0.3 \times 3$$

More generally, if we have a discrete RV X and we repreat the experiment N times, we will get $X = x$ approximately $f_X(x)N$ times. So the average will be

$$\frac{\sum_x x\, f_X(x) N}{N} = \sum_x x\, f_X(x)$$

So we make the following definition.

Definition 3. *Let X be a discrete RV with probability mass function $f_X(x)$. The expected value of X, denoted* $\mathbf{E}[X]$ *is*

$$\mathbf{E}[X] = \sum_x x f_X(x)$$

provided that

$$\sum_x |x| f_X(x) < \infty$$

Terminology/notation The expected value is also called the mean of X. Sometimes $\mathbf{E}[X]$ is just written $\mathbf{E}X$. When the above does not converge absolutely, we the mean in not defined.

Example: Roll a six-sided die and let X be the number you get. Compute $\mathbf{E}X$. Compute $\mathbf{E}[X^2]$.

Next we compute the means of the random variables in our catalog.

Bernoulli

$$\mathbf{E}[X] = 0 \times (1-p) + 1 \times p = p$$

Poisson

$$\begin{aligned}\mathbf{E}[X] &= \sum_{n=0}^{\infty} n \frac{\lambda^n}{n!} e^{-\lambda} = \sum_{n=1}^{\infty} \frac{\lambda^n}{(n-1)!} e^{-\lambda} = e^{-\lambda}\lambda \sum_{n=1}^{\infty} \frac{\lambda^{n-1}}{(n-1)!} \\ &= e^{-\lambda}\lambda \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^{-\lambda}\lambda e^{\lambda} = \lambda\end{aligned}$$

Geometric One of the homework problems will be to compute the mean of the geometric distribution. You should find $\mathbf{E}[X] = 1/p$.

Binomial We will show that $\mathbf{E}[X] = np$. We have

$$\begin{aligned}\mathbf{E}[X] &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=0}^{n} k \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\ &= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \\ &= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \\ &= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k} = np\end{aligned}$$

The last sum is 1 since this is just the normalization condition for the binomial RV with $n-1$ trials.

Suppose X is a RV and $g : \mathbb{R} \to \mathbb{R}$. As before we define a new random variable by $Y = g(X)$. Suppose we know the probability mass function of X and we want to compute the mean of Y. The long way to do this is to first work out the probability mass function of Y and then compute the mean of Y. However, there is a shortcut.

Theorem 3. *(Law of the unconscious statistician) Let X be a discrete RV, g a function from $\mathbb{R}$ to $\mathbb{R}$. Define a new RV by $Y = g(X)$. Let $f_X(x)$ be the pmf of X. Then*

$$\mathbf{E}[Y] = \mathbf{E}[g(X)] = \sum_x g(x) f_X(x)$$

Proof. We start with the definition of $\mathbf{E}Y$:

$$\mathbf{E}Y = \sum_y y\, f_Y(y)$$

By a previous theorem we can write the pmf for Y in terms of the pmf for X:

$$\sum_y y\, f_Y(y) = \sum_y y \sum_{x:g(x)=y} f_X(x) = \sum_y \sum_{x:g(x)=y} g(x)\, f_X(x)$$

Every x in the range of X appears in the right side exactly once. So this can be written as

$$\sum_x g(x)\, f_X(x)$$

□

Example We continue a previous example. Roll a four-sided die twice. Let X be the first roll minus the second roll. Let $Y = X^2$. Find $\mathbf{E}[Y]$.

Definition 4. *We also call $\mathbf{E}X$ the first moment of X. The kth moment of X is $\mathbf{E}[X^k]$. The variance of X is*

$$var(X) = \mathbf{E}[(X-\mu)^2]$$

where $\mu = \mathbf{E}X$. The standard deviation of X is $\sqrt{var(X)}$.

The expected value has a lot of useful properties.

Theorem 4. *Let X be a discrete RV with finite mean. Let $a, b \in \mathbb{R}$.*

1. $\mathbf{E}[aX + b] = a\mathbf{E}[X] + b$
2. *If* $\mathbf{P}(X = b) = 1$, *then* $\mathbf{E}[X] = b$.
3. *If* $\mathbf{P}(a \leq X \leq b) = 1$, *then* $a \leq \mathbf{E}[X] \leq b$.
4. *If* $g(X)$ *and* $h(X)$ *have finite mean, then* $\mathbf{E}[g(X) + h(X)] = \mathbf{E}[g(X)] + \mathbf{E}[h(X)]$

Proof. **GAP !!!!!!!!!!!!!!!!!!!!!!!!!!** □

Remarks:

1. The above properties will also hold for the expected value of continuous random variables.
2. The expected value acts like a definite integral on $[0, 1]$. Compare the following properties for definite integrals.

$$\int_0^1 [af(x) + b]\, dx = a \int_0^1 f(x)\, dx + b$$

If $a \leq f(x) \leq b$, then

$$a \leq \int_0^1 f(x)\, dx \leq b$$

Proposition 2. *If X has finite first and second moments, then*

$$var(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$$

and

$$var(cX) = c^2 var(X), \quad c \in \mathbb{R}$$

Proof. **GAP !!!!!!!!!!!!!!!!!!!!!!!!!!** □

2.5 Conditional expectation

Recall the definition of conditional probability. The probability of A given that B happens is

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}$$

Fix an event B. If we define a function Q on events by $Q(A) = \mathbf{P}(A|B)$, then we showed before that this defines a new probability measure. So if we have a RV X, then we can consider its probability mass function with respect to the probability measure Q. And so we can compute its expected value with repect to this new pmf. This is called the conditional expectation of X given B. The formal definition follows.

Definition 5. *Let X be a discrete RV. Let B be an event with $\mathbf{P}(B) > 0$. The conditional probability mass function of X given B is*

$$f(x|B) = \mathbf{P}(X = x|B)$$

The conditional expectation of X given B is

$$\mathbf{E}[X|B] = \sum_x x\, f(x|B)$$

(provided $\sum_x |x|\, f(x|B) < \infty$).

Example: Roll a six-sided die. Let X be the number on the die. Find $\mathbf{E}[X]$ and $\mathbf{E}[X|X \, is \, odd]$.

Recall that the partition theorem gave a formula for the probability of an event A in terms of conditional probabilities of A given the events in a partition. There is a similar partition theorem for the expected value of a RV. It is useful when it is hard to compute the expected value of X directly, but it is relatively easy if we know something about the outcome of the experiment.

Theorem 5. *Let $B_1, B_2, B_3, \cdots$ be a finite or countable partition of Ω. (So $\cup_k B_k = \Omega$ and $B_k \cap B_l = \emptyset$ for $k \neq l$.) We assume also that $\mathbf{P}(B_k) > 0$ for all k. Let X be a discrete random variable. Then*

$$\mathbf{E}[X] = \sum_k \mathbf{E}[X|B_k]\, \mathbf{P}(B_k)$$

provided that all the expected values are defined.

Remark Note that if B is an event with $0 < \mathbf{P}(B) < 1$, then the theorem applies to the partition with two events : B and B^c. So we have

$$\mathbf{E}[X] = \mathbf{E}[X|B]\,\mathbf{P}(B) + \mathbf{E}[X|B^c]\,\mathbf{P}(B^c)$$

Example: Roll a die until we get a 6. Let X be the number of 1's we got before the 6 came up. Find $\mathbf{E}[X]$. If we know how many rolls it took to get the 6, then this is a pretty easy expected value to compute. So we define our partition by looking at the number of rolls. Let N be the number of rolls (including the final 6). Note that N has a geometric distribution with $p = 1/6$. Consider $\mathbf{E}[X|N = n]$. This amount to computing the expected value of X in a modified experiment. In the modified experiment it takes exactly n rolls to get the first 6. So the first $n-1$ rolls can be $1, 2, 3, 4$ or 5, and the nth roll is a 6. The pmf for X given $N = n$ is a binomial distribution with $n-1$ trials and $p = 1/5$. So $\mathbf{E}[X|N = n] = (n-1)/5$. Thus using the partition theorem

$$\mathbf{E}[X] = \sum_{n=1}^{\infty} \mathbf{E}[X|N = n]\,\mathbf{P}(N = n) = \sum_{n=1}^{\infty} \frac{n-1}{5}\,\mathbf{P}(N = n)$$

We know that

$$\mathbf{P}(N = n) = \left(\frac{5}{6}\right)^{n-1} \frac{1}{6}$$

We could plug this into the above and try to compute the series. Instead we do something a bit more clever.

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{n-1}{5}\,\mathbf{P}(N = n) &= \frac{1}{5}\sum_{n=1}^{\infty} n\,\mathbf{P}(N = n) - \frac{1}{5}\sum_{n=1}^{\infty} \mathbf{P}(N = n) \\
&= \frac{1}{5}\mathbf{E}[N] - \frac{1}{5} = \frac{1}{5}(6-1) = 1
\end{aligned}$$

Example Insects and eggs example.

The following example is done in the language of gambling and is typically referred to as the "gambler's ruin" problem. But is comes up in a variety of settings and is an important example.

Example (Gambler's ruin): There is a game with two players A and B. Each time they play, the probability that A wins is p, the probability B wins is $1 - p$. The player that loses pays the winner 1$. At the start, player A has a dollars and player B has b dollars. They play until one player is broke. (No ATM.) We want to find the expected number of games they play.

Let X be the number of games they play. Let A_1 be the event that player A wins the first game. By the partition theorem,

$$\mathbf{E}[X] = \mathbf{E}[X|A_1]\,\mathbf{P}(A_1) + \mathbf{E}[X|A_1^c]\,\mathbf{P}(A_1^c)$$

Look at $\mathbf{E}[X|A_1]$. We know that A won the first game. After this first game, A has $a+1$ dollars and B has $b-1$ dollars. So $\mathbf{E}[X|A_1]$ is 1 plus the expected number of games when A starts with $a+1$ dollars and B starts with $b-1$ dollars. So we consider a bunch of experiments, indexed by $k = 0, 1, 2, \cdots, a+b$. In experiment k, player A starts with k dollars and player B starts with $a+b-k$ dollars. Each experiment has a different probability measure and so random variables have different expected values in the different experiments. So we use $\mathbf{E}_k[X]$ to denote the expected value of X in the kth experiment. Then we have

$$\mathbf{E}_k[X|A_1] = 1 + \mathbf{E}_{k+1}[X]$$

Similarly,

$$\mathbf{E}_k[X|A_1^c] = 1 + \mathbf{E}_{k-1}[X]$$

So (1) becomes

$$\mathbf{E}_k[X] = (1 + \mathbf{E}_{k+1}[X])\,p + (1 + \mathbf{E}_{k-1}[X])(1-p)$$

We let $m_k = \mathbf{E}_k[X]$. Then we have

$$m_k = 1 + pm_{k+1} + (1-p)m_{k-1}$$

This equation is true for $0 < k < a+b$. Note that $m_0 = 0$ since this is the experiment where A starts off broke, $m_{a+b} = 0$ since this is the experiment where B starts off broke.

We now have a big system of linear equations in the unknowns m_k, $k = 0, 1, \cdots, a+b$.