15-251: Great Theoretical Ideas in Computer Science

Notes on Probability

One should always be a little improbable. —Oscar Wilde

I am giddy; expectation whirls me round.
Th' imaginary relish is so sweet
that it enchants my sense. —William Shakespeare
Troilus and Cressida, act 3, scene 2

1. Probability Spaces

In the dice and coin flipping examples we've looked at, the probabilistic elements are fairly intuitive. Probability is formalized in terms of sample spaces, events, and probability distributions. There is also a very powerful but elementary calculus for probability that allows us to compute and approximate event probabilities.

A probabilistic model or *probability space* is comprised of

1. A *sample space* Ω. The sample space is thought of as the set of possible *outcomes* of an "experiment" appropriate for the problem being modeled.
2. A *probability distribution* $\mathbb{P} : 2^{\Omega} \to [0, 1]$ is a function which assigns to every set A of possible outcomes a *probability* $\mathbb{P}(A)$, which is a number between zero and one. Subsets of Ω are called *events*.

It is often quite helpful to think of Ω as the outcomes of some experiment, which can often be thought of as a physical process. In many cases, the "experiment" is artificial, and can simply be thought of as the roll of a (many-sided) die. In other cases, the physical process is natural to the problem, for example in the case of arrivals to a web server.

The possible outcomes of the experiment must be chosen to be non-overlapping; that is, mutually exclusive. On the other hand, events can be non-exclusive, since they are arbitrary subsets of possible outcomes.

1.1. The Axioms of Probability

A probability distribution $\mathbb{P}$ for a sample space Ω is required to satisfy the following axioms.

1. $\mathbb{P}(A) \geq 0$ for every event A

2. If A and B are events with $A \cap B = \emptyset$, then

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$$

More generally, if $\{A_i\}_{i=1}^n$ is a finite sequence of disjoint events, then

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mathbb{P}(A_i)$$

3. The probability of the sample space Ω is one: $\mathbb{P}(\Omega) = 1$.

The most common, important and intuitive way of constructing probability models is when the sample space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ is finite. In this case a probability distribution is simply a set of numbers $p_i = \mathbb{P}(\omega_i)$ satisfying $0 \leq p_i \leq 1$ and $\sum_{i=1}^n p_i = 1$. The probability of an event is obtained by simply adding up the probabilities of the outcomes contained in that event; thus

$$\mathbb{P}(\{a_1, a_2, \ldots, a_m\}) \;=\; \sum_{j=1}^m \mathbb{P}(a_j) \tag{1}$$

Such a distribution is easily seen to satisfy the axioms.

Example 1.1. *Bernoulli trials*

For a sample space of size two, we can set $\Omega = \{H, T\}$, where the experimental outcomes correspond to the flip of a coin coming up heads or tails. There are only four events, $A = \emptyset, \{H\}, \{T\}, \{H, T\}$. The probability distribution is determined by a single number, the probability of heads $\mathbb{P}(H)$. This is because of the third axiom, which requires that $\mathbb{P}(H) + \mathbb{P}(T) = 1$.

The flip of a coin is often referred to as a *Bernoulli trial* in probability, after Jacob Bernoulli, who studied some of the foundations of probability theory between 1685 and 1689.

Example 1.2. *Multinomial probabilities*

This is the crucial example where the experiment is a single roll of an n-sided die[1] It includes the case of Bernoulli trials as a special case.

Consider the case of rolling a standard die. In this case $\Omega = \{\omega_1, \ldots, \omega_6\}$ where ω_i corresponds to i dots showing on the top of the die. There are $2^6 = 64$ possible events, one being, for example, the event "the number of dots is even." A "fair" die corresponds to the distribution $\mathbb{P}(\omega_i) = \frac{1}{6}$ for each $i = 1, \ldots, 6$.

[1]While such an experiment may be difficult to execute physically, because of the difficulty of constructing such a die for large n, it is conceptually useful to think in these terms.

2. Random Variables

A *random variable* is a numerical value associated with each experimental outcome. Thus, an r.v. can be thought of as a mapping

$$X : \Omega \longrightarrow \mathbb{R}$$

from the sample space Ω to the real line. If you like, it can be thought of as a "measurement" associated with an experiment. Thus, random variables are really random *functions*; but what's random is the input to the function.

Example 2.1. *Max of a roll*

Consider rolling a pair of (distinct) dice, so the sample space is $\Omega = \{(i,j) \,|\, 1 \leq i,j \leq 6\}$. Let $X(i,j) = max(i,j)$.

If X is a random variable, and $S \subset \mathbb{R}$, define

$$\begin{aligned} \mathbb{P}(X \in S) &= \mathbb{P}(\{\omega \,|\, X(\omega) \in S\}) \\ &= \mathbb{P}(X^{-1}(S)) \end{aligned}$$

where

$$X^{-1}(S) = \{\omega \in \Omega \,|\, X(\omega) \in S\}.$$

In particular,

The *probability mass function* (pmf) of a random variable X is

$$\mathbb{P}_X(x) = \mathbb{P}(\{\omega \,|\, X(\omega) = x\})$$

Some of its basic properties are:

- A discrete (finite) r.v. takes on only a discrete (finite) set of values.
- The probability of a set S is given by

  $$\mathbb{P}_X(S) = \sum_{x \in S} \mathbb{P}_X(x)$$

  since $X^{-1}(S) = \bigcup_{x \in S} X^{-1}(s)$.
- The total probability is one:

  $$\sum_x \mathbb{P}_X(x) = 1$$

  (by the normalization axiom).

So, we can picture a probability mass function as a "bar graph" with a bar over each of the possible values of the random variable, where the sum of the heights of the bars is one.

For a given random variable X, we can just work with the p.m.f. $\mathbb{P}_X(x)$. But when we do this hides the sample space! We need to remember that the sample space is really there in the background.

Example 2.2. *Tetrahedral die*

Consider the roll of a pair of (distinguishable) tetrahedral die. The sample space is $\Omega = \{(i,j) \mid 1 \leq i,j \leq 4\}$. Let X be the random variable $X(i,j) = \max(i,j)$.

Then

$$\begin{aligned} \mathbb{P}_X(1) &= \frac{1}{16} \\ \mathbb{P}_X(2) &= \frac{3}{16} \\ \mathbb{P}_X(3) &= \frac{5}{16} \\ \mathbb{P}_X(4) &= \frac{7}{16} \end{aligned}$$

3. Some Important RVs

One of the most central random variables is the Bernoulli.

Example 3.1. *Bernoulli*

The *Bernoulli random variable* is simply the indicator function for a coin flip:

The Bernoulli random variable is

$$X(\{H\}) = 1 \qquad X(\{T\}) = 0$$

for the sample space $\Omega = \{H, T\}$, with probability distribution $\mathbb{P}(H) = 1 - \mathbb{P}(T) = p$.

Thus the probability mass function of the random variable is

$$\mathbb{P}_X(1) = 1 - \mathbb{P}_X(0) = p.$$

A sum of Bernoullis is called a *binomial random variable*. Here the sample space is the collection of all sequences such as

$$\underbrace{(HHTH\cdots T)}_{n \text{ times}};$$

that is, the set of n flips of a (biased) coin, assuming the flips are independent. Let $X(\omega)$ be the random variable

$$X(\omega) = \text{number of heads in the sequence } \omega$$

> The pmf of a binomial random variable, for n flips of a coin, is
>
> $$\mathbb{P}_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$
>
> where p is the probability of heads.

Note that by the binomial theorem

$$\begin{aligned} \sum_{k=0}^{n} \mathbb{P}_X(k) &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^k \\ &= (p+1-p)^n \\ &= 1 \end{aligned}$$

The Boston Museum of Science has a wonderful *Galton machine*, also known as a *Quincunx machine* (after the Roman design of five dots that you see on a die), that illustrates the binomial distribution as balls drop into bins after deflecting to the right or left (a coin flip) off of several levels of pegs. As more and more balls fall, the pattern in the bins resembles a "bell curve."

A *geometric random variable* corresponds to tossing a coin with bias p repeatedly, until a "heads" comes up for the first time. The probability mass function is

$$\mathbb{P}_X(k) = (1-p)^{k-1} p, \qquad k = 1, 2, 3, \ldots$$

We'll work with this random variable quite often.

To verify that this is a valid pmf we appeal to the geometric series:

$$\begin{aligned} \sum_{k=1}^{\infty} \mathbb{P}_X(k) &= \sum_{k=1}^{\infty} (1-p)^{k-1} p \\ &= p \sum_{k=0}^{\infty} (1-p)^k \\ &= p \frac{1}{1-(1-p)} \\ &= 1 \end{aligned}$$

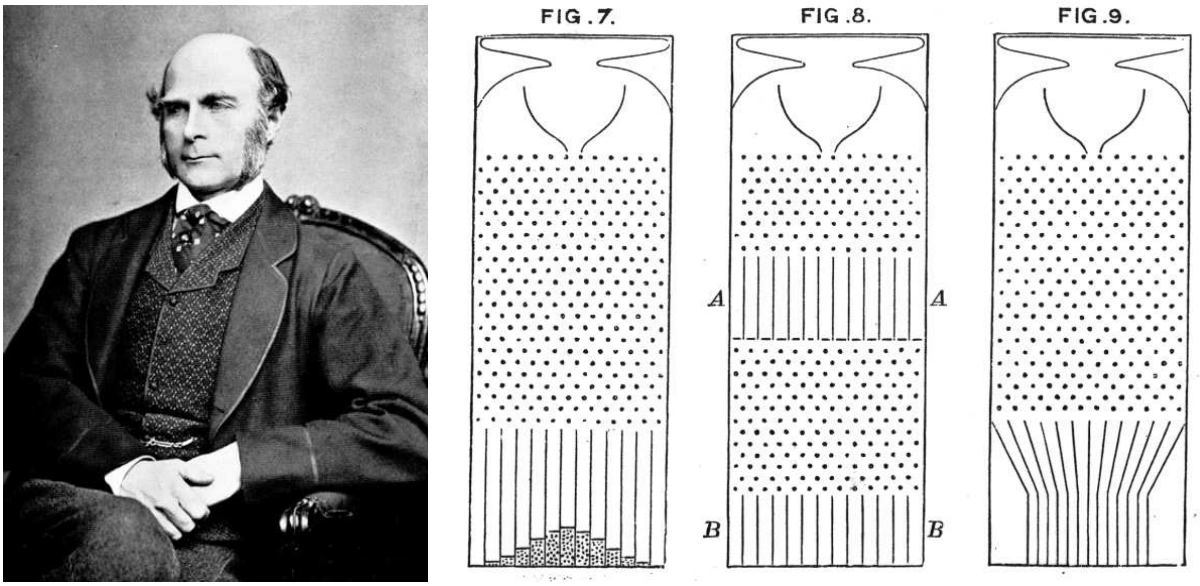

Figure 1: Francis Galton (l); Galton's original drawing of the "quincunx" machine (r). (public domain)

The sample space underlying this random variable is the collection of sequences

$$\Omega = \{H, TH, TTH, TTTH, TTTTH, \ldots\}$$

The pmf of the *Poisson random variable* is

$$\mathbb{P}_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

for a parameter $\lambda > 0$.

This is a valid pmf since

$$\begin{aligned}\sum_{k=0}^{\infty} \mathbb{P}_X(k) &= \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\, e^{\lambda} \\ &= 1\end{aligned}$$

It's used to model things like the counts of physical events like radioactive decay, traffic accidents, arrivals to a Web server, etc.

Some notation: We write

$$X \sim F(\theta)$$

if the random variable X has distribution $F(\theta)$, where the distribution is parameterized by some parameter θ. So, for example, if $X \sim \text{Poisson}(3)$ then

$$\mathbb{P}_X(k) = e^{-3}\frac{3^k}{k!}$$

4. Independent Random Variables

If X and Y are independent random variables, we say that that they are *independent* in case the events

$$\begin{aligned} A_x &= \{\omega \in \Omega \,:\, X(\omega) = x\} \\ B_y &= \{\omega \in \Omega \,:\, Y(\omega) = y\} \end{aligned}$$

are independent events for every pair of values x and y; thus,

$$\mathbb{P}(A_x \cap A_y) = \mathbb{P}(A_x)\,\mathbb{P}(A_y)$$

The statement involves going back to the original probability space; a random variable is just a function on the sample space.

Alternatively two random variables X, Y are independent if their joint pmf factors:

$$\mathbb{P}_{X,Y}(x, y) = \mathbb{P}_X(x)\mathbb{P}_Y(y)$$

In this case

$$\begin{aligned} \mathbb{P}_{X,Y}(A, B) &= \sum_{x\in A, y\in B} \mathbb{P}_{X,Y}(x, y) \\ &= \sum_{x\in A}\sum_{y\in B} \mathbb{P}_X(x)\mathbb{P}_Y(y) \\ &= \left(\sum_{x\in A} \mathbb{P}_X(x)\right)\left(\sum_{y\in B} \mathbb{P}_Y(y)\right) \\ &= \mathbb{P}_X(A)\mathbb{P}_Y(B) \end{aligned}$$

Similarly, for any event A, X and Y are *conditionally independent given* A if

$$\mathbb{P}(X = x,\, Y = y \,|\, A) = \mathbb{P}(X = x \,|\, A)\mathbb{P}(Y = y \,|\, A)$$

As before, conditional independence does not imply independence, and vice-versa.

Example 4.1. *Weighted Tetrahedral Dice*

Suppose we roll two strangely crafted tetrahedral dice such that the probabilities of getting particular pairs of rolls are given in the table below:

Y \ X	1	2	3	4
4	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	0
3	$\frac{2}{20}$	$\frac{4}{20}$	$\frac{1}{20}$	$\frac{2}{20}$
2	0	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{1}{20}$
1	0	$\frac{1}{20}$	0	0

Are they independent? No. Look at the event of rolling $(X = 1, Y = 1)$.

$$\mathbb{P}_{X,Y}(1,1) = 0 \neq \mathbb{P}_X(1)\mathbb{P}_Y(1) = \frac{3}{20}\frac{1}{20}$$

Are the independent given the event $A = \{X \leq 2\} \cap \{Y \geq 3\}$? Yes, you can normalize the values and see that

$$\mathbb{P}(X \mid Y, A) = \mathbb{P}(X \mid A)$$

for $Y = 3, 4$.

Random variables $X_1, X_2, \ldots, X_n$ are independent in case

$$\mathbb{P}_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}_{X_1}(x_1)\mathbb{P}_{X_2}(x_2)\ldots\mathbb{P}_{X_n}(x_n)$$

For *independent* random variables, expectation of a product factors as the product of the expectations:

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

The proof is straightforward:

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{x,y} xy\,\mathbb{P}_{X,Y}(x,y) \\ &= \sum_{x} x \sum_{y} y\,\mathbb{P}_Y(y)\mathbb{P}_X(x) \\ &= \sum_{x} x\,\mathbb{P}_X(x) \sum_{y} y\,\mathbb{P}_Y(y) = \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

So, for sums we don't require independence, but for products we do. Similarly,

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

5. Sums of Poissons

Now, a nice property of the Poisson distribution is that

> $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$ are independent random variables then
>
> $$X_1 + X_2 \;\sim\; \text{Poisson}(\lambda_1 + \lambda_2)$$

To show this,

$$\begin{aligned}\mathbb{P}_{X_1+X_2}(k) &= \sum_{j=0}^{k} \mathbb{P}_{X_1}(j)\mathbb{P}_{X_2}(k-j) \\ &= \sum_{j=0}^{k} e^{-\lambda_1}\,\frac{\lambda_1^j}{j!}\,e^{-\lambda_2}\,\frac{\lambda_2^{k-j}}{(k-j)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!}\sum_{j=0}^{k}\binom{k}{j}\lambda_1^j\,\lambda_2^{k-j} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!}(\lambda_1+\lambda_2)^k\end{aligned}$$

which is the pmf of the Poisson$(\lambda_1 + \lambda_2)$ random variable.

Note that we haven't made explicit the sample space for the the Poisson random variable. We are working with it purely in terms of its pmf. But you can think of the sample space as the natural numbers $\mathbb{N} = \{0, 1, 2, 3, \ldots, \}$.

6. Poisson Approximation to the Binomial

The Poisson can be used to approximate the binomial for large n (numbers of flips):

$$e^{\lambda}\,\frac{\lambda^k}{k!} \;\approx\; \binom{n}{k} p^k (1-p)^{n-k}$$

where $pn = \lambda$, for n large and p small.

Why is this? Well, note that

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{(n-k)!}\frac{1}{k!}\left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}.$$

Now, fix k and let $n \to \infty$. Then this becomes

$$\begin{aligned}\binom{n}{k} p^k (1-p)^{n-k} &\approx \frac{n(n-1)\cdots(n-k+1)}{n^k}\,\frac{n^k}{k!}\,\frac{e^{-\lambda}}{\left(1-\frac{\lambda}{n}\right)^k} \\ &\approx \frac{\lambda^k}{k!}\,e^{-\lambda}\end{aligned}$$

where we have used the facts that

$$\frac{n(n-1)\cdots(n-k+1)}{n^k} \longrightarrow 1$$

and

$$\left(1-\frac{\lambda}{n}\right)^k \longrightarrow 1, \qquad \left(1-\frac{\lambda}{n}\right)^n \longrightarrow e^{-\lambda}.$$

Note that λ is the parameter of the Poisson, and (n, p) are the parameters of the binomial. These are fixed constants, and are not random.

7. Functions of RVs

If X is a random variable, then we can apply a function to it to get another random variable $Y = g(X)$. Here we are just viewing the random variable as a function on the sample space. What is the pmf of Y? A moment's thought shows that

If $Y = g(X)$ then

$$\mathbb{P}_Y(y) = \sum_{x\,:\,g(x)=y} \mathbb{P}_X(x)$$

This is because

$$\begin{aligned}
\mathbb{P}_Y(y) &= \sum_{\omega\,:\,Y(\omega)=y} \mathbb{P}(\omega) \\
&= \sum_{\omega\,:\,g(X(\omega))=y} \mathbb{P}(\omega) \\
&= \sum_{x\,:\,g(x)=y} \; \sum_{\omega\,:\,X(\omega)=x} \mathbb{P}(\omega) \\
&= \sum_{x\,:\,g(x)=y} \mathbb{P}_X(x)
\end{aligned}$$

In other words, the pmf of $Y = g(X)$ at a value y is just the sum of the pmf of X on the "inverse image of y".

8. Expectation

Expectation is a very simple concept, but a very powerful one. We've already introduced it and looked at some of its key properties in the first probability lecture. To review, the expectation (or *mean*) of a random variable X can be thought of as the "center of mass" of a probability mass function $\mathbb{P}_X$.

The *expectation* of a random variable X is the weighted average of its values:

$$\mathbb{E}[X] \;=\; \sum_x \mathbb{P}_X(x)\, x.$$

It's sometimes useful to do a mental "type check"; expectation only makes sense for random variables, that is numeric functions of the sample spaces. Expectation doesn't make sense on an abstract probability space.

Example 8.1. *Mean of a Bernoulli*

If X is a Bernoulli random variable, corresponding to a coin flip with bias p, then the expectation is

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

Example 8.2. *Mean of a Geometric*

If $X \sim \text{Geometric}(p)$, then one way to calculate the mean is to use a little calculus; the following differentiation trick for series is very handy:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=1}^{\infty} (1-p)^{k-1} p\, k \\
&= p \sum_{k=1}^{\infty} k\, (1-p)^{k-1} \\
&= p \sum_{k=1}^{\infty} k\, q^{k-1} \\
&= p \left. \frac{d}{dq} \sum_{k=0}^{\infty} q^k \right|_{q=1-p} \\
&= p \left. \frac{d}{dq} \frac{1}{1-q} \right|_{q=1-p} \\
&= p \left. \frac{1}{(1-q)^2} \right|_{q=1-p} \\
&= \frac{1}{p}
\end{aligned}$$

So, for example, when we flip a fair coin, the expected number of flips until a "heads" appears is 2. We'll see a more elegant way of calculating this, using conditioning, later on.

Example 8.3. *Mean of a Poisson*

The same differentiation trick can be used for the Poisson. If $X \sim \text{Poisson}(\lambda)$ then

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{\infty} k\, e^{-\lambda} \frac{\lambda^k}{k!} \\
&= \lambda\, e^{-\lambda} \sum_{k=0}^{\infty} k\, \frac{\lambda^{k-1}}{k!} \\
&= \lambda\, e^{-\lambda} \frac{d}{d\lambda} e^{\lambda} \\
&= \lambda
\end{aligned}$$

Alternatively, and more simply, we calculate that

$$\begin{aligned}
\mathbb{E}[X] &= \lambda\, e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{k-1} \\
&= \lambda\, e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k}}{k} \\
&= \lambda
\end{aligned}$$

9. Linearity of Expectation

The most important property of expectation is that it is *linear*; that is, the expectation of the sum of two random variables is the sum of their expectations:

For any pair of random variables X and Y

$$\mathbb{E}[X+Y] \quad = \quad \mathbb{E}[X]+\mathbb{E}[Y]$$

This holds whether or not the random variables are independent. In addition,

$$\mathbb{E}[aX] \quad = \quad a\mathbb{E}[X]$$

for any constant a. Linearity is very easy to prove, just starting from the definition of expectation:

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{\omega} \mathbb{P}(\omega)\,(X+Y)(\omega) \\
&= \sum_{\omega} \mathbb{P}(\omega)\,(X(\omega)+Y(\omega)) \\
&= \sum_{\omega} \mathbb{P}(\omega)X(\omega) + \sum_{\omega} \mathbb{P}(\omega)Y(\omega) \\
&= \mathbb{E}[X]+\mathbb{E}[Y]
\end{aligned}$$

Example 9.1. *Expectation of a Binomial*

As an example, we can write a binomial random variable as a sum of (independent) Bernoulli random variables. That is, if $X \sim \text{Binomial}(n,p)$ then

$$X = X_1 + X_2 + \cdots + X_n$$

where $X_i \sim \text{Bernoulli}(p)$. Therefore

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[X_1 + X_2 + \cdots + X_n] \\
&= \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n] \\
&= p + p + \cdots p \\
&= np
\end{aligned}$$

This could also be proved using the differentiation trick, but this is unnecessary.

10. Expectation of Functions of RVs

For a function of a random variable X it is easy to show that

If $Y = g(X)$ is a function of X then

$$\mathbb{E}[Y] = \sum_x g(x)\,\mathbb{P}_X(x)$$

This follows from

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_y y\,\mathbb{P}_Y(y) \\
&= \sum_y y \sum_{x\,:\,g(x)=y} \mathbb{P}_X(x) \\
&= \sum_y \sum_{x\,:\,g(x)=y} y\,\mathbb{P}_X(x) \\
&= \sum_y \sum_{x\,:\,g(x)=y} g(x)\,\mathbb{P}_X(x) \\
&= \sum_x g(x)\,\mathbb{P}_X(x)
\end{aligned}$$

As a special case of this, we have

$$\begin{aligned}
\mathbb{E}[aX+b] &= \sum_x (ax+b)\,\mathbb{P}_X(x) \\
&= a\mathbb{E}[X] + b
\end{aligned}$$

11. Jointly Distributed RVs

Suppose we have two random variables X and Y. Implicitly, this means that they are defined with respect to the same probability space $(\Omega, \mathbb{P})$. The *joint probability mass function* of X and Y is

$$\begin{aligned}
\mathbb{P}_{X,Y}(x,y) &= \mathbb{P}(X=x, Y=y) \\
&= \mathbb{P}\left(\{\omega : X(\omega)=x\} \cap \{\omega : Y(\omega)=y\}\right)
\end{aligned}$$

Thus, $\mathbb{P}_{X,Y}(x,y)$ is the probability of the event "$X=x$ and $Y=y$". We can view a pair of random variables as a single function from Ω to $\mathbb{R}^2$, or in terms of the probability mass function, which assigns a weight to each of the finite (for now) possible (x,y) pairs.

If we arrange the (x,y) pairs in a grid, then the columns sum to $\mathbb{P}_X(x)$ and the rows sum to $\mathbb{P}_Y(y)$. That is,

$$\begin{aligned}
\mathbb{P}_X(x) &= \mathbb{P}\left(\{\omega : X(\omega)=x\}\right) \\
&= \sum_y \mathbb{P}\left(\{\omega : X(\omega)=x\} \cap \{\omega : Y(\omega)=y\}\right) \\
&= \sum_y \mathbb{P}_{X,Y}(x,y)
\end{aligned}$$

Thus, we have that

For jointly distributed random variables X and Y,

$$\mathbb{P}_X(x) = \sum_y \mathbb{P}_{X,Y}(x,y) \quad \text{and} \quad \mathbb{P}_Y(y) = \sum_x \mathbb{P}_{X,Y}(x,y)$$

Example 11.1. *Two die*

As a simple example let $\Omega = \{(i,j)\}$ be the sample space for a roll of a pair of (distinguishable) dice. Let $X =$ "first roll is odd" and $Y =$ "second roll is odd". That is, we take two Bernoulli random variables corresponding to indicator functions for two events $A =$ "first roll is odd" and $B =$ "second roll is odd". The joint pmf is given by

$$\mathbb{P}_{X,Y}(0,0) = \mathbb{P}_{X,Y}(0,1) = \mathbb{P}_{X,Y}(1,0) = \mathbb{P}_{X,Y}(1,1) = \frac{1}{4}$$

Now, we have from before that

$$\mathbb{E}[g(X,Y)] \ = \ \sum_{x,y} g(x,y)\, \mathbb{P}_{X,Y}(x,y)$$

for jointly distributed random variables and any function $g : \mathbb{R}^2 \longrightarrow \mathbb{R}$.

From this we recover the basic fact that expectation is linear: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$. Why? Simply let $g(x,y) = ax + by$ and apply the above fact.

12. The Multinomial

Suppose we roll the die n times, and side one comes up x_1 times, side two comes up x_2 times, and so on: side i comes up x_i times. If side i comes up with probability p_i, then the *multinomial* probability distribution is

$$\mathbb{P}(x_1, x_2, \ldots, x_k) = \binom{n}{x_1 x_2 \cdots x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

where we have that $\sum_{i=1}^k x_i = n$ and $\sum_{i=1}^k p_i = 1$.

Then the *marginal distribution* of X_j is Binomial(n, p_j). The expectation of the multinomial is

$$\mathbb{E}(X) = n(p_1, p_2, \ldots, p_k).$$

If $X \sim Mult(n, \vec{p})$ what is $\mathbb{E}(X)$?

$$\mathbb{E}(X) = n\vec{p} = n(p_1, p_2, \ldots, p_k)$$

Why? Because $X_i \sim Binomial(n, p_i)$.

13. Sampling and Expectation

Here's another way to think about expectation. Suppose $X \sim F$, where F is some distribution. Let

$$X_i \sim F \quad i = 1, 2, \ldots, n$$

be independent draws from F. Then

$$\mathbb{E}[X] \approx \frac{1}{n} \sum_{i=1}^{n} X_i.$$

This is called the law of averages or the law of large numbers.

14. Conditioning

Conditional expectation just uses the notions of conditional probability that we're already familiar with, so there are no really new concepts.

Let A be an event. We define the conditional pmf for a random variable as

$$\mathbb{P}_{X \,|\, A}(x) = \mathbb{P}(X = x \,|\, A) = \frac{\mathbb{P}(\{X = x\} \cap A)}{\mathbb{P}(A)}$$

assuming that $\mathbb{P}(A) > 0$.

Example 14.1. *Truncated geometric*

A student takes a test repeatedly up to a maximum of n times, each time passing with probability p; the success or failure in one try is independent of the success or failure in the other tries. What is the pmf of the number of attempts, given that she passes the test?

We have that

$$\begin{aligned} \mathbb{P}(\text{passes} \cap \{X = k\}) &= (1-p)^{k-1}p \\ \mathbb{P}(\text{passes}) &= \sum_{k=1}^{n} (1-p)^{k-1}p \end{aligned}$$

Then the conditional distribution is

$$\begin{aligned} \mathbb{P}_{X \,|\, A}(k) &= \frac{(1-p)^{k-1}p}{\sum_{j=1}^{n}(1-p)^{j-1}p} \\ &= \frac{(1-p)^{k-1}p}{1-(1-p)^n} \end{aligned}$$

using the fact that $1 + q + \cdots q^{n-1} = (1-q^n)/(1-q)$.

In this way we can condition one random variable on another:

$$\begin{aligned}\mathbb{P}_{X\,|\,Y}(x\,|\,y) &= \frac{\mathbb{P}(\{X=x\}\cap\{Y=y\})}{\mathbb{P}(\{Y=y\})}\\ &= \frac{\mathbb{P}_{X,Y}(x,y)}{\mathbb{P}_Y(y)}\end{aligned}$$

As for conditional probability, we then have the chain rule for conditional pmfs:

$$\mathbb{P}_{X\,|\,Y}(x\,|\,y)\,\mathbb{P}_Y(y) = \mathbb{P}_{X,Y}(x,y).$$

Normally we use abbreviated notation, and just write

$$\mathbb{P}(x\,|\,y)\,\mathbb{P}(y) = \mathbb{P}(x,y).$$

Inductively, we get the chain rule

$$\mathbb{P}(x_1,x_2,\ldots,x_n) = \mathbb{P}(x_1)\prod_{j=2}^{n-1}\mathbb{P}(x_j\,|\,x_1,\ldots,x_{j-1}).$$

In addition, we have an analogue of the rule of total probability in the form

$$\mathbb{P}_X(x) = \sum_y \mathbb{P}_Y(y)\,\mathbb{P}_{X\,|\,Y}(x\,|\,y).$$

15. Conditional expectation

Now, once we have the distribution $\mathbb{P}_{X\,|\,Y}$ or $P_{X\,|\,A}$ we can define expectation. This is called *conditional expectation*.

$$\mathbb{E}[X\,|\,A] = \sum_x x\mathbb{P}_{X\,|\,A}(x)$$

Conditional expectation of X given Y takes value y is

$$\mathbb{E}[X\,|\,Y=y] = \sum_x x\mathbb{P}_{X\,|\,Y}(x|y)$$

and then

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

You should prove this as an exercise.

Example 15.1. *Mean of a Geometric with probability p*

$$\begin{aligned}\mathbb{E}[X] &= 1 \cdot \mathbb{P}(X = 1) + \mathbb{E}[X \mid X > 1]\mathbb{P}(X > 1) \\ &= p + (1-p)(1 + \mathbb{E}[X])\end{aligned}$$

which when solved for $\mathbb{E}[X]$ gives $\mathbb{E}[X] = 1/p$.

Recall that we define

$$\mathbb{P}_{X \mid Y}(x \mid y) = \frac{\mathbb{P}_{X,Y}(x, y)}{\mathbb{P}_Y(y)}$$

using this we define the *conditional expectation of* X *given* $Y = y$ as

$$\mathbb{E}[X \mid Y = y] = \sum_x x \, \mathbb{P}_{X \mid Y}(x \mid y).$$

This is the average value of X when we know that the value of Y is y. Similarly

$$\mathbb{E}[r(X, Y) \mid Y = y] = \sum_x r(x, y) \, \mathbb{P}_{X \mid Y}(x \mid y)$$

Now here is an important point: $\mathbb{E}[X]$ is a *number* while $\mathbb{E}[X \mid Y = y]$ is a function of y. Before we observe the value of Y, it is unknown so the value of this expectation is a random variable $\mathbb{E}[X \mid Y]$.

Example 15.2. *Coins with different bias*

Suppose I have two coins, one with bias 1/4 and one that has bias 1/10. I use either coin with equal probability $p = 1/2$. I keep flipping until I get a head. What is the expected number of flips of the second coin?

$$\mathbb{E}[Y \mid X = 1] = 4\mathbb{E}[Y \mid X = 0] = 10$$

Conditioning on the choice of coin, we have

$$\mathbb{E}[Y \mid X = 1] \cdot \frac{1}{2} + \mathbb{E}[Y \mid X = 0] \cdot \frac{1}{2} = 2 + 5 = 7$$

In general, $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]]$. The inner expectation is the average over randomness in Y given that $X = x$ and the outer expectation represents the average randomness of this function over X.

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[Y \mid X]] &= \sum_x \mathbb{E}[Y \mid X = x]\, \mathbb{P}_X(X = x) \\
&= \sum_x \left(\sum_y y\, \mathbb{P}_{Y \mid X}(y \mid x) \right) \mathbb{P}_X(X = x) \\
&= \sum_x \sum_y y\, \mathbb{P}_{X,Y}(x, y) = \sum_y y\, \mathbb{P}_Y(y) = \mathbb{E}[Y]
\end{aligned}$$

This is called the *total expectation theorem.*

In words: we can condition on the value of any other (jointly distributed r.v.) and then average over the values of that other random variable.

Example 15.3. *Mean of the Geometric*

Let $X \sim$ Geometric(p). We computed previously that $\mathbb{E}[X] = 1/p$. Let's use conditioning to compute this.

$$\mathbb{E}[X] = \sum_i \mathbb{P}(A_i)\, \mathbb{E}[X \mid A_i]$$

where $\{A_i\}$ is a partition of the sample spaces. Let $A = \{X > 1\},\ A^c = \{X = 1\}$. Then when we condition on A,

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{P}(A^c) \cdot \mathbb{E}[X \mid \{X = 1\}] + \mathbb{P}(A) \cdot \mathbb{E}[X \mid \{X > 1\}] \\
&= p \cdot 1 + (1 - p) \cdot \mathbb{E}[X \mid \{X > 1\}] \\
&= p + (1 - p) \cdot (\mathbb{E}[X] + 1) \\
p \cdot \mathbb{E}[X] &= 1 \\
\mathbb{E}[X] &= \frac{1}{p}
\end{aligned}$$