4 Random Variables

4.1 Discrete Random Variables

Many of our examples involve *functions* of standard outcomes. For instance:

Example 4.1. Roll a fair die twice and compute the *sum*. We consider the sample space

$$S = \{(a,b) : 1 \le a+b \le 6\}$$

of all possible outcomes of rolling the dice twice. Since each element of S is equally likely, we can compute by counting. For instance, if $X = a+b$ is the sum of the dice rolls, then

$$\mathbb{P}\{X=4\} = \mathbb{P}\{(a,b) \in S : X = 4\} = \mathbb{P}\{(1,3),(2,2),(3,1)\} = \frac{3}{36}$$

Note that we are taking the probability of an event: $\{X=4\}$ is the *set* of possible outcomes $\{(a,b)\}$ for which $a+b=4$, a *subset* of the sample space S.

The function $X = a+b$ in the example gets a name.

Definition 4.2. Let S be a sample space. A *random variable* is a function $X : S \to \mathbb{R}$.
The *(cumulative) distribution function* $F : \mathbb{R} \to \mathbb{R}$ is defined by

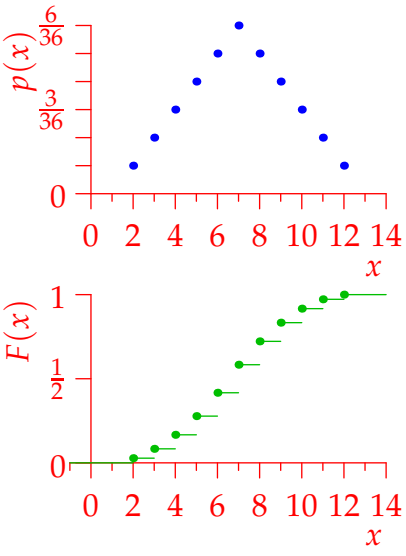
$$F(x) = \mathbb{P}\{X \le x\}$$

If the range of X is (finite or) countable, we call X a *discrete random variable*. In such a case, the *(probability) mass function* is $p(x) = \mathbb{P}\{X = x\}$. If $X = \{x_n\}$ where $x_n < x_{n+1}$ for all n, then

$$F(x_n) = \sum_{k \le n} p(x_k) \quad \text{and} \quad p(x_n) = F(x_n) - F(x_{n-1})$$

The above random variable $X = a+b$ takes finitely many values $2, \ldots, 12$ and is therefore discrete. Its distribution and mass functions are

x	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
$F(x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$



Observe how the distribution function $F(x)$ *increases* to 1. Since its domain is $\mathbb{R}$, it is really a *step function*; for example

$$F(\pi) = \mathbb{P}\{X \le \pi\} = F(3) = \frac{3}{36}$$

We'll typically just plot discrete dots for the graph of $F(x)$.

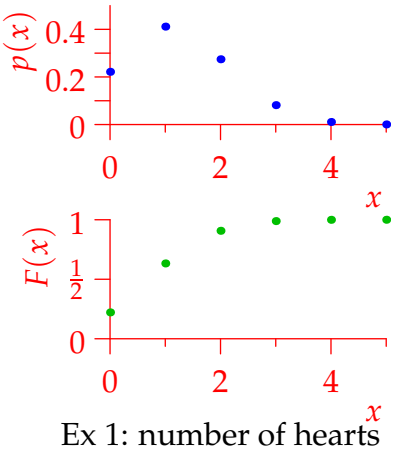
Here are three further examples of random variables. Notice in all cases how the distribution function increases from 0 to 1. For clarity, in the first two examples where X is discrete, we only evaluate the distribution function $F(x)$ at values for which $p(x) > 0$.

Examples 4.3. 1. Deal 5 cards from a standard pack and let X be the number of hearts in the hand. With the aid of a calculator/computer, we can find the mass and distribution functions directly:

$$p(x) = \frac{\binom{13}{x}\binom{39}{5-x}}{\binom{52}{5}}$$

$$F(x) = \sum_{y=0}^{x} p(y)$$

x	$p(x)$	$F(x)$
0	22.15%	22.15%
1	41.14%	63.30%
2	27.42%	90.72%
3	8.15%	98.88%
4	1.07%	99.95%
5	0.05%	100%



Ex 1: number of hearts

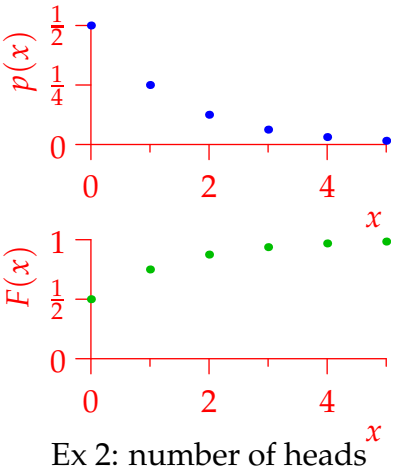
2. We toss a fair coin repeatedly and count the number X of heads appearing before the first tail. Since

$$\text{range } X = \mathbb{N}_0 = \{0, 1, 2, 3, 4, \ldots\}$$

is *countably* infinite, X is a discrete random variable. Its mass and distribution functions are

$$p(n) = \mathbb{P}(\underbrace{H \cdots H}_{n \text{ heads}} T) = \frac{1}{2^{n+1}}$$

$$F(n) = \sum_{k=0}^{n} p(k) = 1 - \frac{1}{2^{n+1}}$$



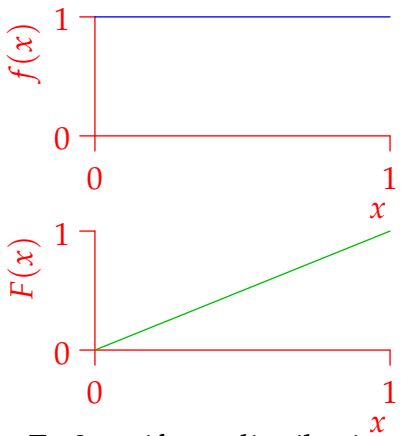
Ex 2: number of heads

3. A random number generator selects a real number X between 0 and 1 according to the distribution function

$$F(x) = \mathbb{P}\{X \le x\} = x \quad \text{whenever} \quad 0 \le x \le 1$$

Since X can take any value in the interval $[0, 1]$ (an *uncountably infinite* set), it is *not* a discrete random variable.

We will later describe X as a *uniformly distributed, continuous* random variable. While there is no mass function, the *derivative* $f(x) = F'(x) = 1$ on the open interval $(0, 1)$ will be seen to play a similar role; we'll call it a *density*. Instead of $\sum p(x) = 1$, we have $\int f(x)\,dx = 1$.



Ex 3: uniform distribution

While the mass function (for a discrete distribution) feels more natural, it is really the distribution

function that tells you everything you need to know about a random variable. For completeness, we gather some of its basic properties.

Theorem 4.4. *Let $F(x)$ be the distribution function of a random variable X. Then*

1. *$F(x)$ is non-decreasing on $\mathbb{R}$*
2. $\lim_{x\to\infty} F(x) = 1$
3. $\lim_{x\to-\infty} F(x) = 0$
4. *$F(x)$ is right-continuous; $\forall r \in \mathbb{R}$, $\lim_{x\to r^+} F(x) = F(r)$*

Proof. The first part is immediate from

$$x \le y \implies \{X \le x\} \subseteq \{X \le y\} \implies F(x) \le F(y)$$

since the events $\{X \le x\}$ and $\{x < X \le y\}$ are mutually exclusive with union $\{X \le y\}$.

Now suppose $x_0 \ge x_1 \ge x_2 \ge \cdots$ is a non-increasing sequence; by the above, so also is $F(x_n)$. Since the latter is bounded below (by zero!), it *converges* to some limit $L = \lim F(x_n) \in [0, 1]$. Let $r = \lim x_n$ (it could be $-\infty$!), and consider a sequence of mutually exclusive events and their union

$$E_n = \{x_n < X \le x_{n-1}\}, \qquad \{X > r\} = \{X > x_0\} \cup \bigcup_{n=1}^{\infty} E_n$$

By mutual exclusivity,

$$\begin{aligned}
\mathbb{P}\{X > r\} &= \mathbb{P}\{X > x_0\} + \sum_{n=1}^{\infty} \mathbb{P}(E_n) = 1 - F(x_0) + \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(E_k) \\
&= 1 - F(x_0) + \lim_{n\to\infty} \sum_{k=1}^{n} (F(x_{n-1}) - F(x_n)) \\
&= 1 - \lim_{n\to\infty} F(x_n) = 1 - L
\end{aligned}$$

There are two cases, which together prove parts 3 and 4.

- If $r = -\infty$, then $\mathbb{P}\{X < r\} = \mathbb{P}(\emptyset) = 0 \implies L = 0$.
- If r is finite, then $\mathbb{P}\{X < r\} = 1 - F(r) \implies L = F(r)$.

Part 2 is similar: if x_n is non-decreasing with limit r, then we have events

$$E_n = \{x_{n-1} < X \le x_n\}, \qquad \{X < r\} = \{X \le x_0\} \cup \bigcup_{n=1}^{\infty} E_n \implies \mathbb{P}\{X < r\} = \lim_{n\to\infty} F(x_n)$$

If $r = \infty$, this proves part 2. ∎

Note the reason we don't get left-continuity: if r is finite, then $\mathbb{P}\{X < r\}$ need not be equal to $\mathbb{P}\{X \le r\} = F(r)$.

4.2 Expectation and Variance of Discrete Random Variables

After students take a test, the first question asked is often, 'What's the average?' This is perhaps the single simplest number which describes the overall performance of the class. The question can be rephrased in terms of a probability distribution.

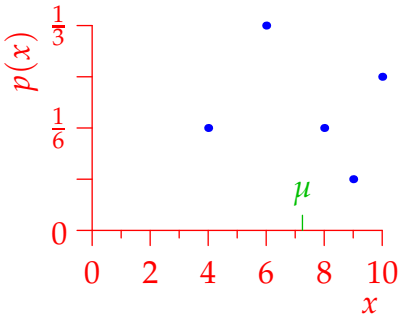
Example 4.5. Twelve students take a quiz out of ten points and the scores are as follows:

4, 4, 6, 6, 6, 6, 8, 8, 9, 10, 10, 10

If a student is chosen at random, the probability $p(x)$ that they scored x points is immediate:

$$p(x) = \frac{\text{number of students scoring } x}{12}$$

x	0	1	2	3	4	5	6	7	8	9	10
$p(x)$	0	0	0	0	$\frac{2}{12}$	0	$\frac{4}{12}$	0	$\frac{2}{12}$	$\frac{1}{12}$	$\frac{3}{12}$



The average score for the class is

$$\mu = \frac{1}{12}\sum \text{scores} = \frac{1}{12}(4\cdot 2 + 6\cdot 4 + 8\cdot 2 + 9\cdot 1 + 10\cdot 3)$$
$$= 7.25 = \sum_{x=0}^{12} xp(x)$$

There are two crucial observations from the example:

1. $p(x)$ forms the mass function of a random variable X "a randomly chosen student's score."
2. The average is the sum of the possible scores *weighted* by the mass function $\sum xp(x)$.

More generally, if n trials are conducted, where each produces a numerical value x_i, then we have a distribution X: "how many trials produce the value x." This has mass function

$$p(x) = \frac{\text{number of trials resulting in } x}{n}$$

and average

$$\mu = \frac{1}{n}\sum_{k=1}^{n} x_i = \frac{1}{n}\sum_{\text{values } x} x\cdot(\text{number of trials resulting in } x) = \sum_x xp(x)$$

We make this a definition for any discrete random variable, even if there are infinitely many trials!

Definition 4.6. Let X be a discrete random variable with mass function $p(x)$. The *expectation, mean value* or *expected value* of X, is

$$\mathbb{E}[X] = \sum_x xp(x)$$

Since the expectation represents a mean, it is very common to denote this by the Greek letter μ.

Examples 4.7. We start with the two of the simplest situations imaginable, before revisiting our examples from the previous section.

1. If X is the number of heads obtained when a fair coin is tossed once, then $X = 0, 1$ with equal probabilities. It follows that the expected number of heads is

$$\mathbb{E}[X] = 0p(0) + 1p(1) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$$

2. A single fair die is rolled and X is the number appearing. The expected value of the roll is then

$$\mathbb{E}[X] = 1 \cdot p(1) + 2 \cdot p(2) + \cdots + 6p(6) = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$$

3. (4.1) For the experiment of rolling the sum of two dice, we have

$$\mathbb{E}[X] = \sum_{x=2}^{12} xp(x) = \frac{2 \cdot 1}{36} + \frac{3 \cdot 2}{36} + \frac{4 \cdot 3}{36} + \cdots + \frac{12 \cdot 1}{36} = \frac{1}{36} = 7$$

   If you want a more generalizable way to calculate this, try the following

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x=2}^{6} x\frac{x-1}{36} + \sum_{y=7}^{12} y\frac{13-y}{36} = \frac{1}{36}\left[\sum_{n=1}^{6} n^2 - n + \sum_{n=1}^{6} 13n - n^2\right] \\ &= \frac{1}{36}\sum_{n=1}^{6} 12n = \frac{1}{3} \cdot \frac{1}{2} \cdot 6 \cdot (6+1) = 7\end{aligned}$$

   where we substituted $x = n$ and $y = 13 - n$ in the first line.

4. (4.3, part 1) It is easy to approximate the expectation for the number of hearts in a 5 card hand just be summing the data in the table:

$$\mathbb{E}[X] = p(1) + 2p(2) + 3p(3) + 4p(4) + 5p(5) = 1.25$$

   In fact the expectation is *precisely* 1.25 (see the hypergeometric distribution later).

5. (4.3, part 2) This time we have to compute an infinite sum. To do this, recall a little geometric series from calculus: whenever $|x| < 1$, we have

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n \implies \frac{1}{(1-x)^2} = \frac{\mathrm{d}}{\mathrm{d}x}\sum_{n=0}^{\infty} x^n = \sum_{n=0}^{\infty} nx^{n-1}$$

   The expected number of heads before a tail is therefore

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} np(n) = \sum_{n=0}^{\infty} \frac{n}{2^{n+1}} = \frac{1}{4} \cdot \frac{1}{(1-\frac{1}{2})^2} = 1$$

For the final part of Example 4.3, we'd really like to say that $\mathbb{E}[X] = \frac{1}{2}$: this is indeed true, and we'll discuss why in the next chapter.

If a random variable takes infinitely many values, is possible for the expectation to behave somewhat surprisingly.

Example 4.8. Suppose that a random variable X has range $\{2, 4, 8, 16, \ldots\}$ with probability mass function

$$p(2^n) = \frac{1}{2^n}$$

The expectation of this variable is infinite!

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} \frac{2^n}{2^n} = \infty$$

It is even possible do design a random variable which doesn't have an expectation, by constructing an infinite series which diverges by oscillation!

Expectations of Functions

If X is a random variable X, and $g : \mathbb{R} \to \mathbb{R}$ is a function, then $g(X)$ is also a random variable, whose expectation can be computed using the definition.

Example 4.9. Suppose X has mass function

x	-2	-1	0	1	2
$p_X(x)$	0.1	0.3	0.3	0.1	0.2

$$\implies \mathbb{E}[X] = -0.1$$

Then X^2 is also a random variable with mass function

x	0	1	4
$p_{X^2}(x)$	0.3	0.4	0.3

It follows that X^2 has expectation

$$\mathbb{E}[X^2] = 0p_{X^2}(0) + 1p_{X^2}(1) + 4p_{X^2}(4) = 0.4 + 1.2 = 1.6$$

Note that this is *not* the same thing as $(\mathbb{E}[X])^2 = 0.01$!

Rather than continuing in this manner, consider computing the probability mass function $q(y)$ of $g(X)$ abstractly. Since g could be non-injective ($g(x_i) = g(x_j)$ for some $x_i \neq x_j$), we might need to add several terms:

$$q(y) = \sum_{x : g(x) = y} p(x)$$

Now use the definition:

$$\mathbb{E}[g(X)] = \sum_{y \in \text{range}(g)} yq(y) = \sum_{y \in \text{range}(g)} y \sum_{x : g(x) = y} p(x) = \sum_x g(x)p(x)$$

To summarise:

Lemma 4.10. *If X is a discrete random variable and $g : \mathbb{R} \to \mathbb{R}$ is a function, then the expectation of the random variable $g(X)$ is the sum*

$$\mathbb{E}[g(X)] = \sum_x g(x)p(x)$$

The moments *of X are the expectations $\mathbb{E}[X^n]$ where $n \in \mathbb{N}$.*

Note that $\mathbb{E}[1] = \sum_x p(x) = 1$ and moreover that expectations play nicely with linear functions

$$\mathbb{E}[aX + b] = \sum_x (ax + b)p(x) = a\sum_x xp(x) + b\sum_x p(x) = a\mu + b$$

We've already seen that this is false for squares!

Examples 4.11. 1. Revisit the previous example. We have

$$\mathbb{E}[X^2] = 4 \cdot \frac{1}{10} + 1 \cdot \frac{3}{10} + 0 \cdot 0.3 + 1 \cdot \frac{1}{10} + 4 \cdot \frac{2}{10} = \frac{16}{10} = 1.6$$

the same answer as before, though we didn't have to compute the mass function for X^2.

2. The daily return X on a highly volatile stock is either $\pm 10\%$. Suppose that the value either increases or decreases for two days consecutively, each with equal probabilty $\frac{1}{2}$. If you invest \$100, then the expected value of your investment after two days will be

$$\mathbb{E}[100(1+X)^2] = 100\left[1.1^2 p(\uparrow\uparrow) + 0.9^2 p(\downarrow\downarrow)\right] = 100 \cdot \frac{1}{2} \cdot 2.02 = \$101$$

That this expectation is *not* your original investment is very important in financial mathematics!

3. Two biased coins have, respectively, probabilities a and b of landing heads. Let X be the random variable "number of heads obtained by randomly choosing one of the coins and tossing it twice."

The probability mass function for X is

$$\begin{aligned} p(0) &= \frac{1}{2}(1-a)^2 + \frac{1}{2}(1-b)^2, \\ p(1) &= \frac{1}{2}\left(a(1-a) + (1-a)a\right) + \frac{1}{2}(b(1-b) + (1-b)b) = a(1-a) + b(1-b) \\ p(2) &= \frac{1}{2}a^2 + \frac{1}{2}b^2 \end{aligned}$$

The expected number of heads is therefore

$$\mathbb{E}[X] = p(1) + 2p(2) = a(1-a) + b(1-b) + a^2 + b^2 = a + b$$

Hopefully this seems totally reasonable!
Now we compute the expectation of the *square* of the number of heads:

$$\mathbb{E}[X^2] = p(1) + 4p(2) = a(1-a) + b(1-b) + 2a^2 + 2b^2 = a^2 + b^2 + a + b$$

Variance

Of particular importance is the expectation of the function $g(X) = (X - \mu)^2$ where $\mu = \mathbb{E}[X]$ is the expectation of the random variable itself.

Definition 4.12. The *variance* of a discrete random variable is

$$\operatorname{Var} X := \mathbb{E}[(X - \mu)^2]$$

The variance is often denoted σ^2, where σ itself is the *standard deviation*

$$\sigma = \sqrt{\operatorname{Var} X}$$

The variance measures the extent to which X is spread out from its mean. The standard deviation is useful because it has the same units as X and is therefore easier to visualize. If you really wanted to measure the average deviation of X from its mean, it might seem more sensible to compute $\mathbb{E}[|X - \mu|]$ rather than the standard deviation. The reason we choose the variance and standard deviation is purely mathematical; they behave more nicely!

Example 4.13. Here are the probability mass functions $p_X(x), p_Y(x)$ for two random variables X, Y with the same expectation but different variances.

$$p_X(0) = \frac{1}{2} = p_X(4) \qquad p_Y(1) = \frac{1}{2} = p_Y(3)$$

$$\mu_X = 0 \cdot \frac{1}{2} + 4 \cdot 12 = 2 \qquad \mu_Y = 1 \cdot \frac{1}{2} + 3 \cdot 12 = 2$$

$$\operatorname{Var} X = (0-2)^2 \cdot \frac{1}{2} + (4-2)^2 \cdot \frac{1}{2} = 4 \qquad \operatorname{Var} Y = (1-2)^2 \cdot \frac{1}{2} + (3-2)^2 \cdot \frac{1}{2} = 1$$

$$\sigma_X = 2 \qquad \sigma_Y = 1$$

The variables X, Y have the same mean but the first has twice the standard deviation; intuitively X is twice as spread out as Y.

Before computing any other variances, we collect a couple of useful facts.

Lemma 4.14. *For any (discrete) random variable X,*

1. $\operatorname{Var} X = \mathbb{E}[X^2] - \mu^2$
2. $\operatorname{Var}(aX + b) = a^2 \operatorname{Var} X$

When we remarked above that the variance behaves nicely, this result is partly what we me. By contrast, observe that in general

$$\mathbb{E}[|X - \mu|] \neq \mathbb{E}[|X|] - |\mu|$$

Indeed if X only takes positive values, then the right side is zero!

Proof. Both follow straightforwardly from the definition:

$$\begin{aligned}\mathbb{E}[(X-\mu)^2] &= \sum_x (x-\mu)^2 p(x) = \sum_x (x^2 - 2\mu x + \mu^2) p(x) \\ &= \sum_x x^2 p(x) - 2\mu \sum_x x p(x) + \mu^2 \sum_x p(x) \\ &= \mathbb{E}[X]^2 - 2\mu \mathbb{E}[X] + \mu^2 \cdot 1 = \mathbb{E}[X]^2 - \mu^2\end{aligned}$$

and

$$\begin{aligned}\operatorname{Var}(aX+b) &= \mathbb{E}[(aX+b)^2] - (a\mu + b)^2 = \mathbb{E}[a^2X^2 + 2abX + b^2] - a^2\mu^2 - 2ab\mu - b^2 \\ &= a^2\mathbb{E}[X^2] + 2ab\mathbb{E}[X] + b^2 - a^2\mu^2 - 2ab\mu - b^2 = a^2 \operatorname{Var} X\end{aligned}$$

■

Examples (4.11, mark II). We compute the variances of our previous examples.

1. $\operatorname{Var} X = \mathbb{E}[X^2] - \mu^2 = 1.6 - 0.01 = 1.59$
2. $\operatorname{Var} X = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] = \frac{1}{2}(0.1^2 + (-0.1)^2) = 0.01$
3. $\operatorname{Var} X = a^2 + b^2 + a + b - (a+b)^2 = a + b - 2ab$

4.3 Bernoulli and Binomial Random Variables

We have already encountered examples of a general family of distributions.

Definition 4.15. Let $p \in [0,1]$ be a fixed constant representing the probability of 'success' at a single trial. Given a positive integer n, let X count the number of successes resulting from n *independent* trials. We say that X has a *binomial distribution with parameters* (n,p) and write

$$X \sim B(n,p)$$

If there is only a single trial ($n = 1$), this is known as the *Bernoulli distribution*. The binomial distribution therefore models n independent Bernoulli trials.

Since there are precisely $\binom{n}{k}$ different ways to obtain exactly k successes from n trials, the probability mass function for $X \sim B(n,p)$ is

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

This explains why we call it the *binomial* distribution! Indeed the binomial theorem guarantees the basic property of the mass function

$$\sum_{k=0}^{n} \mathbb{P}\{X = k\} = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + 1 - p)^n = 1$$

Note also that the binomial distribution is *symmetric*: if $Y = n - X$ counts the number of *failures* from n independent trials, then $Y \sim B(n, 1-p)$. Because of this symmetry, and because it turns up so often in formulae, we write $q = 1 - p$.

Examples 4.16. 1. The first of Examples 4.7 where we tossed a fair coin once and counted the number of heads was a Bernoulli distributed random variable with parameter $p = \frac{1}{2}$.

2. A biased coin has probability $p = \frac{1}{3}$ of coming up heads. If we flip it four times, then the number of heads has a $B(4, \frac{1}{3})$ distribution. More specifically, we have the mass function

$$\mathbb{P}\{X=0\} = \binom{4}{0}\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^4 = \frac{16}{81} \qquad \mathbb{P}\{X=1\} = \binom{4}{1}\left(\frac{1}{3}\right)^1\left(\frac{2}{3}\right)^3 = \frac{32}{81}$$
$$\mathbb{P}\{X=2\} = \binom{4}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^2 = \frac{24}{81} \qquad \mathbb{P}\{X=3\} = \binom{4}{3}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^1 = \frac{8}{81}$$
$$\mathbb{P}\{X=4\} = \binom{4}{4}\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)^0 = \frac{1}{81}$$

We can easily compute the expected number of heads and their variance

$$\mathbb{E}[X] = 0\cdot\frac{16}{81} + 1\cdot\frac{32}{81} + 2\cdot\frac{24}{81} + 3\cdot\frac{8}{81} + 4\cdot\frac{1}{81} = \frac{108}{81} = \frac{4}{3}$$
$$\operatorname{Var} X = \left(-\frac{4}{3}\right)^2\frac{16}{81} + \left(-\frac{1}{3}\right)^2\frac{32}{81} + \left(\frac{2}{3}\right)^2\frac{24}{81} + \left(\frac{5}{3}\right)^2\frac{8}{81} + \left(\frac{8}{3}\right)^2\frac{1}{81} = \frac{648}{9\cdot 81} = \frac{8}{9}$$

3. Continuing the previous example. Suppose you play a game where you bet \$1 and toss the biased coin 4 times. You win nothing if you toss less than three heads, \$5 for three heads, and \$10 for four. Is this game worth playing?

We need to compute the expectation of the function $g(X)$, where

x	0	1	2	3	4
$g(x)$	-1	-1	-1	5	10

We have

$$\mathbb{E}[g(X)] = -\frac{16}{81} - \frac{32}{81} - \frac{24}{81} + 5\cdot\frac{8}{81} + 10\cdot\frac{1}{81} = -\frac{22}{81}$$

If you play the game, you should expect to lose an average of $\frac{22}{81} \approx 27¢$ per game.

Expectation and Variance

It is incredibly easy to compute the expectation and variance of the Bernoulli $B(1,p)$ distribution:

$$\mu = \mathbb{E}[X] = 0\cdot\mathbb{P}\{X=0\} + 1\cdot\mathbb{P}\{X=1\} = p$$

$$\begin{aligned}\operatorname{Var} X = \mathbb{E}\left[(X-\mu)^2\right] &= (0-p)^2\mathbb{P}\{X=0\} + (1-p)^2\mathbb{P}\{X=1\} \\ &= p^2q + q^2p = pq\end{aligned}$$

where we follow the convention that $q = 1-p$. It seems entirely reasonable that if a single trial produces a expected p successes, repeating the trial n times should result in an average of np successes. This is indeed the case, as confirmed in the above Example 4.16.2 where $\mathbb{E}[X] = 4\cdot\frac{1}{3}$. What

is perhaps more surprising is that the same is true for the *variance.*

Theorem 4.17. *If $X \sim B(n,p)$, then $\mathbb{E}[X] = np$ and $\operatorname{Var} X = npq$*

Proof. To motivate the general approach, we compute $\mu = \mathbb{E}[X]$ directly

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

The critical observation was that[1] $k\binom{n}{k} = n\binom{n-1}{k-1}$. Now we repeat with the square: observe that if $Y \sim B(n-1,p)$, then

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{k=0}^{n} k^2\binom{n}{k}p^k q^{n-k} = np\sum_{k=1}^{n} k\binom{n-1}{k-1}p^{k-1}q^{n-k} \\ &= np\sum_{j=0}^{n-1}(j+1)\binom{n-1}{j}p^k q^{n-1-j} && \text{(let } k = j-1\text{)} \\ &= np\mathbb{E}[Y+1] = np\big((n-1)p+1\big)\end{aligned}$$

In particular

$$\begin{aligned}\operatorname{Var} X &= \mathbb{E}[X^2] - \mu^2 = np\big((n-1)p+1\big) - n^2p^2 \\ &= np\big((n-1)p+1-np\big) = np(1-p) = npq\end{aligned}$$

■

As a corollary, the argument gives something of a recurrence for the moments of the binomial distributions

$$X \sim B(n,p),\ Y \sim B(n-1,p) \implies \mathbb{E}[X^n] = np\mathbb{E}[(Y+1)^{n-1}]$$

Example 4.18. You deal a card from a deck and record whether it is a ace. You then replace the card, shuffle and repeat until you've done this 26 times. The number of aces dealt is a random variable $X \sim B(48, \frac{1}{13})$ with expectation and standard deviation

$$\mathbb{E}[X] = \frac{48}{13} \approx 3.692, \qquad \sigma = \sqrt{\operatorname{Var} X} = \sqrt{\frac{48\cdot 12}{13^2}} = \frac{24}{13} \approx 1.846$$

The Shape of the Binomial Distribution and its Computation

The previous example illustrates a problem. Suppose we wanted to compute the probability

$$\mathbb{P}\{X = 10\} = \binom{48}{10}\frac{12^{38}}{13^{48}} = \frac{48!12^{38}}{10!38!13^{48}}$$

[1] Recall the two ways to count how many ways we can choose a committee of size k *with a chair* from a group of n people.

These are some very large numbers (48! has 2 digits!), numbers which it is not practical for a computer to work with directly. Instead, it is easier to develop a recurrence. Suppose $X \sim B(n,p)$, then

$$\frac{\mathbb{P}\{X=k\}}{\mathbb{P}\{X=k-1\}} = \frac{\frac{n!}{k!(n-k)!}p^k q^{n-k}}{\frac{n!}{(k-1)!(n-k+1)!}p^{k-1}q^{n-k+1}} = \frac{(n-k+1)p}{kq} \qquad (*)$$

Armed with this, a computer can rapidly compute the sequence

$$\begin{aligned}
\mathbb{P}\{X=0\} = q^n \implies \mathbb{P}\{X=1\} &= \frac{np}{q}q^n = npq^{n-1} \\
\implies \mathbb{P}\{X=2\} &= \frac{(n-1)p}{2q}npq^{n-1} = \frac{1}{2}n(n-1)p^2q^{n-2} \\
\implies \mathbb{P}\{X=3\} &= \frac{(n-2)p}{3q}\cdot\frac{1}{2}n(n-1)p^2q^{n-2} = \frac{1}{6}n(n-1)(n-2)p^3q^{n-3}
\end{aligned}$$

These expressions are precisely those used by Newton to compute his binomial series. For estimating by hand, these are not particularly useful. Instead, we can appeal to a famous approximation.

Theorem 4.19 (Stirling's Formula). *If n is large, then $n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$*

Example 4.20. A fair coin is flipped 100 times. The probability that 50 heads appear is

$$\mathbb{P}\{X=50\} = \binom{100}{50}\frac{1}{2^{100}} = \frac{100!}{(50!)^2 2^{100}} \approx \frac{100^{100}e^{-100}\sqrt{2\pi\cdot 100}}{50^{100}e^{-100}2\pi\cdot 50\cdot 2^{100}} = \frac{10}{50\sqrt{2\pi}} \approx 7.96\%$$

For really large n, the individual probabilities are usually tiny, and an estimation of $\mathbb{P}\{a \le X \le b\}$ is typically desired. For this an approximation in terms of the normal distribution can be used (later).

A second nice effect of our recurrence $(*)$ is that it can easily be rearranged: observe that

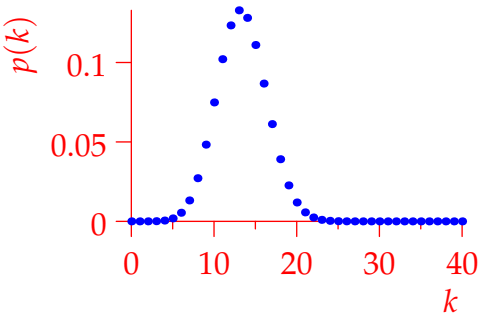
$$\mathbb{P}\{X=k\} > \mathbb{P}\{X=k-1\} \iff (n-k+1)p > kq \iff (n+1)p > k(q+p) = k$$

We conclude:

Theorem 4.21. *The probability mass function for the distribution $B(n,p)$ increases to a maximum at $X = \lfloor (n+1)p \rfloor$, then decreases. In the exceptional case that $(n+1)p$ is an integer, the outcomes $X = (n+1)p$ and $(n+1)p - 1$ have equal maximum likelihood.*

This single outcome with greatest likelihood is called the *mode*.

Example 4.22. You roll a fair die 40 times and count how often you roll a 1 or a 2. The outcome with the greatest likelihood is that you succeed $\lfloor \frac{41}{3} \rfloor = 13$ times.



4.4 The Poisson Distribution

The binomial distribution involves *finitely many* trials. As we've already seen, if there are a very large number of trials, then the binomial distribution can be very difficult to work with. Here we consider an approximation to a distribution $X \sim B(n,p)$ where n is large and p is small. Let $\lambda = np$, then

$$\begin{aligned}\mathbb{P}\{X=k\} &= \binom{n}{k} p^k q^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!}\left(\frac{\lambda}{n}\right)^k q^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n^k q^k}\frac{\lambda^k}{k!}q^n \\ &= \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)}{q^k}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\end{aligned}$$

If we take the limit as $n \to \infty$ while holding $\lambda = np$ constant, then

$$p \to 0, \quad q \to 1, \quad \left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda} \implies \mathbb{P}\{X=k\} \to \frac{\lambda^k}{k!}e^{-\lambda}$$

Thanks to the familiar fact that $\sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^{\lambda}$, we can define a new distribution.

Definition 4.23. A random variable X taking values $0, 1, 2, \ldots$ is said to have a *Poisson distribution* with parameter λ, written $X \sim \text{Poisson}(\lambda)$, if its probability mass function is

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

Example 4.24. The number of bottles of wine X sold by a small store on a given day obeys a Poisson distribution with parameter $\lambda = 3$.

(a) The probability that the store sells 5 bottles of wine in a day is

$$\mathbb{P}\{X=12\} = \frac{3^5}{5!}e^{-3} = \frac{81}{40}e^{-3} \approx 10.1\%$$

(b) The probability that the store sells at least one bottle of wine is

$$\mathbb{P}\{X \geq 1\} = 1 - \mathbb{P}\{X=0\} = e^{-3} \approx 5.0\%$$

(c) The probability that fewer than 4 bottles are sold is

$$\mathbb{P}\{0 \leq X \leq 3\} = \sum_{k=0}^{3} \frac{3^k}{k!}e^{-3} = 13e^{-3} \approx 64.7\%$$

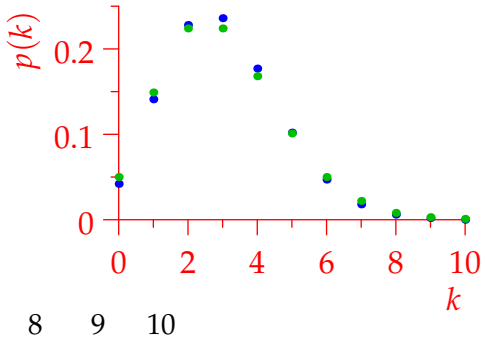
As seen above, the Poisson distribution is a reasonable approximation for the binomial distribution $B(n,p)$ provided n is large and p is small, $\lambda = np$ is of reasonable size and k is also fairly small. Even in the example, the real situation is probably closer to that of a binomial distribution: there will

be a finite (if large) number n of bottles of wine in the store, each of which has a small probability $p = \frac{\lambda}{n} = \frac{3}{n}$ of being purchased.

Continuing the example, suppose the store has 30 bottles of wine, each of which has a probability $\frac{1}{10}$ of being purchased in a given day. The number of bottles sold is binomially distributed. We compare the first few terms of the two distributions

$$X \sim \text{Poisson}(3) \qquad p_Y(k) = \frac{3^k}{k!}e^{-3}$$

$$Y \sim B(30, \tfrac{1}{10}) \qquad p_X(k) = \binom{30}{k}\left(\frac{1}{10}\right)^k\left(\frac{9}{10}\right)^{30-k}$$



In the table, all values are quoted as percentages to 1 d.p.

k	0	1	2	3	4	5	6	7	8	9	10
$p_X(k)(\%)$	5.0	14.9	22.4	22.4	16.8	10.1	5.0	2.2	0.8	0.3	0.1
$p_Y(k)(\%)$	4.2	14.1	22.8	23.6	17.7	10.2	4.7	1.8	0.6	0.2	0.0

Hopefully the example convinces you that the two distributions are very similar. This is an entirely typically application of the Poisson distribution. Given that the distribution is used as an approximation to the binomial distribution, it seems reasonable that the expectation and variance of the two distributions correspond:

$$X \sim \text{Poisson}(\lambda) \approx B(n,p) \overset{?}{\Longrightarrow} \mathbb{E}[X] \approx np = \lambda, \quad \text{Var}\, X \approx npq \approx \lambda$$

Indeed this is the case.

Theorem 4.25. *If $X \sim \text{Poisson}(\lambda)$, then $\mathbb{E}[X] = \text{Var}\, X = \lambda$.*

Proof. Recall that $e^\lambda = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$, whence

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k\mathbb{P}\{X=k\} = \sum_{k=1}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} = \lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}e^{-\lambda} = \lambda\sum_{j=0}^{\infty}\frac{\lambda^j}{j!}e^{-\lambda} = \lambda$$

where $j = k-1$ assisted in the computation. Similarly

$$\mathbb{E}[X^2] = \sum_{k=0}^{\infty} k^2\mathbb{P}\{X=k\} = \lambda\sum_{k=1}^{\infty} k\frac{\lambda^{k-1}}{(k-1)!}e^{-\lambda} = \lambda\sum_{j=0}^{\infty}(1+j)\frac{\lambda^j}{j!}e^{-\lambda} = \lambda\left(1+\mathbb{E}[X]\right) = \lambda+\lambda^2$$

from which

$$\text{Var}\, X = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \lambda$$

■

Example 4.26. A quantity of carbon emits electrons (β-radiation) according to a Poisson distribution with parameter $\lambda = 0.2$ per minute.[2]Find the expected number of electrons emitted in one minute given that at least one is detected.

First compute the required probabilities

$$\mathbb{P}\{X = k | X \geq 1\} = \frac{\mathbb{P}\{X = k \cap X \geq 1\}}{\mathbb{P}\{X \geq 1\}}$$

Plainly this is 0 if $k = 0$. Otherwise

$$k \geq 1 \implies \mathbb{P}\{X = k | X \geq 1\} = \frac{\mathbb{P}\{X = k\}}{1 - \mathbb{P}\{X = 0\}} = \frac{\lambda^k e^{-\lambda}}{k!(1 - e^{-\lambda})}$$

The expectation is therefore

$$\mathbb{E}[X | X \geq 1] = \sum_{k=1}^{\infty} \frac{k\lambda^k e^{-\lambda}}{k!(1 - e^{-\lambda})} = \frac{\lambda e^{-\lambda}}{1 - e^{-\lambda}} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \frac{\lambda}{1 - e^{-\lambda}} \approx 1.103 \text{ particles/min}$$

Poisson Processes

A common application of the Poisson distribution is to count how often how many instances of a recurring event occur in a given time interval. For Californians, the obvious example is a major earthquake; this provides our motivating example.

Example 4.27. It is estimated that a very dangerous quake (magnitude ≥ 7) occurs somewhere in California on average every 10 years. We would like to estimate the probability of k such earthquakes happening over an interval of t years.

To analyse this, we make some assumptions.

- Over a small interval of length Δt, assume that the probability of one event occurring is approximately $\lambda \Delta t$, and the probability of more than one event occurring is approximately zero.
- Over adjacent intervals, $[t_0, t_1] \cup [t_1, t_2]$ the probabilities of events occurring are independent.

Split the interval $[0, t]$ of length t into n equal subintervals of length $\Delta t = \frac{t}{n}$. The number of events $N(t)$ occurring over $[0, t]$ is then binomially distributed

$$N(t) \sim B(n, \lambda \Delta t) = B\left(n, \frac{\lambda t}{n}\right)$$

which, as $n \to \infty$, approaches the distribution Poisson(λt).

[2]In practice, this means that there is a very large number n of radioactive carbon atoms, each of which has a very small probability $p = \frac{0.2}{n}$ of decaying per minute. To put some flesh on this, 12 grams of naturally occurring carbon contains $\approx 6 \times 10^{23}$ (Avogadro's number), of which ≈ 1.2 in 10^{12} will be radioactive carbon 14. For such a sample, $n \approx 5 \times 10^{11}$ and $p \approx 4 \times 10^{-13}$. For numbers like this, the Poisson distribution will be very accurate!

Definition 4.28. Events occur according to a *Poisson process with rate* λ if the number of events $N(t)$ occurring over an interval of length t satisfies

$$\mathbb{P}\{N(t) = k\} = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

To revisit our example, we model the occurrence of dangerous quakes by a Poisson process with rate $\lambda = \frac{1}{10}$ (i.e. one quake every 10 years). If someone lives in California for 25 years, the number of dangerous quakes they might expect to occur during their tenure will follow a Poisson(2.5) distribution. For instance, the chance that no such quakes occur would be

$$\mathbb{P}\{N(25) = 0\} = e^{-2.5} \approx 8.2\%$$

Time to Next Event Given events occurring according to a Poisson process with rate λ, consider the random variable T measuring the *time* from now until the next event occurs. Note that T is *not* a discrete distribution. We can still compute its cumulative probability function however.

$$\mathbb{P}\{T \leq t\} = \mathbb{P}\{N(t) \geq 1\} = 1 - \mathbb{P}\{N(t) = 0\} = 1 - e^{-\lambda t}$$

For obvious reasons, we say that T has an *exponential distribution*; we shall return to this later.

Example 4.29. Customers arrive at a store according to a Poisson process with rate $\lambda = 0.5$ per minute.

(a) Find the probability that at least 5 customers arrive within 4 minutes.

Over the 4 minute interval, the number of customers arriving is distributed according to $N(4) \sim \text{Poisson}(2)$. We therefore wish to compute

$$\begin{aligned}
\mathbb{P}\{N(4) \geq 5\} &= 1 - \mathbb{P}\{N(4) \leq 4\} = 1 - \sum_{k=0}^{4} \frac{2^k}{k!} e^{-2} \\
&= 1 - \left(1 + 2 + \frac{4}{2} + \frac{8}{6} + \frac{16}{24}\right) e^{-2} \\
&= 1 - 7e^{-2} \approx 5.3\%
\end{aligned}$$

(b) Find the probability that the store will be empty for the first 10 minutes after opening.

We need to compute

$$\begin{aligned}
\mathbb{P}\{T > 10\} &= 1 - \mathbb{P}\{T \leq 10\} = 1 - \mathbb{P}\{N(10) = 0\} = 1 - 1 + e^{-10\lambda} \\
&= e^{-5} \approx 0.67\%
\end{aligned}$$

4.5 Other Common Discrete Distributions

We quickly mention several other examples and compute their means and variances. Some of these we have seen before. These come in two flavors, the first are related to independent Bernoulli trials and ask the question of how many trials are necessary for a given number of successes.

Definition 4.30. The *negative binomial distribution* $Y \sim \text{NegativeBinomial}(r, p)$ counts how many independent Bernoulli trials are required to achieve r successes. Its mass function is

$$\mathbb{P}\{X = n\} = \binom{n-1}{r-1} p^r (1-p)^{n-r} = \binom{n-1}{r-1} p^r q^{n-r}, \quad r \leq n$$

since the first $n-1$ trials must contain precisely $r-1$ successes and $n-r$ failures.

The *geometric distribution* $\text{Geometric}(p)$ is simply $\text{NegativeBinomial}(1, p)$; how many trials to achieve a first success.

$$\mathbb{P}\{X = n\} = p(1-p)^{n-1} = pq^{n-1}, \quad 1 \leq n$$

Example 4.31. In the United States approximately 35% of over-25 year olds have completed a Bachelor's degree. Since the population is massive (≈330 million!), provided we are only looking for a relatively small number of successes, we may assume that each sample is, essentially, an independent Bernoulli trial, whether or not a person is replaced into the population before resampling.

1. How many over-25's do we expect to have to sample before we find our first graduate?

   If X is the required random variable, then we have $X \sim \text{Geometric}(p)$ where $p = 0.35$. This problem is very similar to Example 4.7, part 5 and may be computed the same way. We therefore require the expectation:

   $$\mathbb{E}[X] = \sum_{n=1}^{\infty} npq^{n-1} = \frac{p}{(1-q)^2} = \frac{1}{p} = \frac{20}{7} \approx 2.857 \text{ people}$$

2. A polling company needs to interview 100 college graduates. How many over-25's would expect to have to sample?

   This time we require the expectation of a negative binomial distribution $X \sim \text{NegativeBinomial}(r, p)$ where $(r, p) = (100, 0.35)$.

   $$\begin{aligned}
   \mathbb{E}[X] &= \sum_{n=r}^{\infty} n \binom{n-1}{r-1} p^r q^{n-r} = \sum_{n=r}^{\infty} r \binom{n}{r} p^r q^{n-r} \\
   &= r \sum_{m=s}^{\infty} \binom{m-1}{s-1} p^{s-1} q^{m-s} && (r = s-1,\ n = m-1) \\
   &= \frac{r}{p} \sum_{m=s}^{\infty} \binom{m-1}{s-1} p^s q^{m-s} = \frac{r}{p} = \frac{100}{0.35} = \frac{2000}{7} \approx 285.7 \text{ people}
   \end{aligned}$$

   since the last sum is that of the probability mass function of a $\text{NegativeBinomial}(s, p)$ random variable.

A similar argument computes the variance. Indeed the moments satisfy a recurrence, just as with the standard binomial (Theorem 4.17); simply replace the first n with n^k to prove the following.

Theorem 4.32. *If $X \sim \text{NegativeBinomial}(r, p)$, then for any $k \geq 1$,*

$$\mathbb{E}[X^k] = \frac{r}{p}\mathbb{E}\big[(Y-1)^{k-1}\big]$$

where $Y \sim \text{NegativeBinomial}(r+1, p)$. In particular,

$$\mathbb{E}[X] = \frac{r}{p}\mathbb{E}[1] = \frac{r}{p}, \qquad \mathbb{E}[X^2] = \frac{r}{p}\mathbb{E}[Y-1] = \frac{r}{p}\left(\frac{r+1}{p} - 1\right) = \frac{r(r+q)}{p^2}$$
$$\operatorname{Var} X = \frac{r(r+q)}{p^2} - \frac{r^2}{p^2} = \frac{rq}{p^2} = \frac{r(1-p)}{p^2}$$

Following our example, the standard deviation in the size of sample required to find 100 college graduates is

$$\sigma = \sqrt{\operatorname{Var} X} = \sqrt{\frac{100 \cdot 0.65}{0.35^2}} = \frac{\sqrt{65}}{0.35} = \frac{20}{7}\sqrt{65} \approx 23 \text{ people}$$

It is common to quote this as a estimate: we expect to sample 285 ± 23 people.

Sampling with Replacement

Our previous distributions concerned independent trials: a repeated coin toss, or drawing balls from a bag *with replacement*. Our final two distributions concern drawing balls *without replacement*.

Definition 4.33. We draw balls without replacement from a bag of N, m of which are 'successes.'

- If X counts the number of successes when we withdraw n balls, we say that X has a *hypergeometric distribution* with parameters (N, m, n). Its mass function is[a]

$$\mathbb{P}\{X = k\} = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \qquad 0 \leq k \leq n$$

  since we must have k successes from m and $n-k$ failures from $N-m$.

- If Y counts how many draws are necessary to obtain r successes, then X has a *negative hypergeometric distribution* with parameters (N, m, r). Its mass function is

$$\mathbb{P}\{X = k\} = \frac{\binom{m}{r-1}\binom{N-m}{k-r}}{\binom{N}{k-1}} \frac{m-r+1}{N-k+1} \qquad k \geq r$$

  since the first $k-1$ draws produce $r-1$ successes, after which we multiply by the probability of a single success ($m-(r-1)$ of the remaining $N-(k-1)$ balls).

[a] Follow the convention that $\binom{a}{b} = 0$ if $b < 0$ or $b > a$. Strictly $\max\{0, m+n-N\} \leq k \leq \min\{m, n\}$

Example 4.34. A population of squirrels in a forest comprises 50 red and 100 grey squirrels. Squirrels are captured, where it is assumed that each individual is equally likely to be so.

(a) If 20 squirrels are captured, then the number of red squirrels in the sample has a hypergeometric distribution X with parameters $(N, m, n) = (150, 50, 20)$. The probability that half are red is

$$\mathbb{P}\{X = 10\} = \frac{\binom{50}{10}\binom{100}{10}}{\binom{150}{20}} \approx 4.9\%$$

(b) If squirrels are captured until 10 red squirrels have been found, then the number required Y has a negative hypergeometric distribution with parameters $(N, m, r) = (150, 50, 10)$. The probability that we'd need to capture exactly 20 squirrels in order to find 10 reds is then

$$\mathbb{P}\{Y = 20\} = \frac{\binom{50}{9}\binom{100}{10}}{\binom{150}{19}} \cdot \frac{50 - 9}{150 - 19} \approx 2.4\%$$

Similarly to Theorem 4.32, if X is hypergeometric with parameters (N, m, n), then

$$\mathbb{E}[X^k] = \frac{mn}{N}\mathbb{E}\left[(Y + 1)^{k-1}\right]$$

where Y is hypergeometric with parameters $(N - 1, m - 1, n - 1)$. Setting $k = 1$, we obtain

$$\boxed{\mathbb{E}[X] = \frac{mn}{N} = np} \quad \text{where } p = \frac{m}{N}$$

This should be intuitive since $p = \frac{m}{N}$ is the chance that any randomly drawn ball is a success. It can similarly be shown that

$$\boxed{\operatorname{Var} X = \frac{N - n}{N - 1} np(1 - p)}$$

Both expressions confirm our intuition that if the sample size n is small relative to the number of balls N, then the hypergeometric distribution is approximately binomial $B(n, p)$.

Returning to our squirrel example, we see that the expectation and standard deviation of the number of red squirrels captured in our sample of 20 are easily computed:

$$p = \frac{m}{N} = \frac{50}{150} = \frac{1}{3} \implies \mathbb{E}[X] = \frac{20}{3} \approx 6.67 \text{ squirrels}$$
$$\sigma = \sqrt{\operatorname{Var} X} = \sqrt{\frac{150 - 20}{150 - 1} \cdot 20 \cdot \frac{1}{3} \cdot \frac{2}{3}} \approx 1.97 \text{ squirrels}$$

Indeed it is easy to check that the mode (most likely outcome) of the distribution X is that the sample contains 7 squirrels ($\mathbb{P}\{X = 7\} \approx 19.6\%$).

It is not worth trying to memorize the corresponding expressions for the negative hypergeometric

distribution; look them up if you ever need them! For sanity's sake, the values for this example are

$$\mathbb{E}[Y] = r\frac{N+1}{m+1} \approx 29.6, \qquad \sigma = \sqrt{\operatorname{Var} Y} = \sqrt{\frac{(N-m)r(m+1-r)(N+1)}{(m+1)^2(m+2)}} \approx 6.77$$

The (negative) hypergeometric distributions have less utility than (negative) binomials. The former are harder to calculate with and are approximated very well by the latter when working with large data sets. Even for moderate sizes, the approximation can be quite good. For instance, if we modelled Example 4.34 binomially with $p = \frac{1}{3}$ of the squirrels being red (i.e. sampling with replacement), then

$$\mathbb{P}\{X = 10\} = 5.4\%$$

which isn't far from the 4.9% given by the hypergeometric distribution. Similarly, Example 4.31 is really a negative hypergeometric distribution where $N \approx 330$ million, though in practice there is no point treating it as such since the required sample size will be miniscule compared to N: indeed the expected sample size required when using the two distributions agree to *five* decimal places!

Summary

Here are the four distributions we discussed relative to drawing balls from a bag with or without replacement.

Random Variable	Without Replacement	With Replacement
# Successes in n trials	Binomial	Hypergeometric
# trials to obtain r success(es)	Negative Binomial	Negative Hypergeometric

There are many other finite distributions, but these are the most commonly encountered.

4.6 Expectations of Sums of Random Variables

We often wish to work with several random variables simultaneously, in particular with their sum. The expectations of sums is particularly easy to calculate.

Theorem 4.35. *Let S be a countable sample space.*

1. *The expectation of a discrete random variable X is a weighted average over S:*

$$\mathbb{E}[X] = \sum_{s \in S} X(s)p(s)$$

where $p(s) = \mathbb{P}(\{s\})$ to be the probability of the singleton event $s \in S$ occurring.

2. *If $X_i : S \to \mathbb{R}$ are discrete random variables, then their expectations sum*

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$

Note that this holds even if the random variables X_i are not independent!

Proof. 1. $\mathbb{E}[X] = \sum_k k\mathbb{P}\{X = k\} = \sum_k k \sum_{s\in S: X(s)=k} p(s) = \sum_{s\in S}\sum_{k:X(s)=k} kp(s) = \sum_{s\in S} X(s)p(s)$

2. Let $Y = X_1 + \cdots + X_n$, then

$$\mathbb{E}[Y] = \sum_{s\in S} Y(s)p(s) = \sum_{s\in S}\big(X_1(s) + \cdots + X_n(s)\big)p(s) = \sum_{i=1}^{n}\mathbb{E}[X_i]$$

■

Examples 4.36. 1. Let Y be the number of successes from n trials when the i^{th} trial has probability of success p_i. Compute $\mathbb{E}[Y]$ and $\operatorname{Var} Y$.

If X_i counts the 'number' of successes at the i^{th} trial, then $\mathbb{E}[X_i] = 0(1-p_i) + 1\cdot p_i = p_i$, whence

$$\mathbb{E}[Y] = \sum \mathbb{E}[X_i] = \sum p_i$$

For this example, the variance also satisfies the same formula!

$$\begin{aligned}\operatorname{Var} Y &= \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = \sum \mathbb{E}[X_i^2] + 2\sum_{i<j}\mathbb{E}[X_iX_j] - \sum p_i^2 - 2\sum_{i<j} p_ip_j\\ &= \sum \operatorname{Var} X_i + 2\sum_{i<j}\big[\mathbb{E}[X_iX_j] - p_ip_j\big]\\ &= \sum \operatorname{Var} X_i = \sum p_i(1-p_i)\end{aligned}$$

since $X_iX_j = 1$ if and only if both are successes.

In particular, if the distributions X_i are all identical, then their sum is the binomial distribution $Y \sim B(n,p)$ and we obtain the usual expressions for its expectation and variance.

2. A similar approach an be used to compute the expectation and variance of the negative binomial distribution $Y \sim \text{NagativeBinomial}\, N(r,p)$. Simply view this as a sum of identically distributed geometric distributions $X_i \sim \text{Geometric}(p)$.

$$\mathbb{E}[Y] = \sum_{i=1}^{r}\mathbb{E}[X_i] = \frac{r}{p}$$

The variance also satisfies the same additive formula.[3]

[3] In general, the variance is not additive and satisfies the formula

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\operatorname{Var} X_i + \sum_{i\neq j}\Big(\mathbb{E}[X_iX_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]\Big) = \sum_{i=1}^{n}\operatorname{Var} X_i + \sum_{i\neq j}\operatorname{Cov}(X_i, X_j)$$

where the extra terms are called *covariances*.

However, if the random variables X_i are independent ($\mathbb{P}\{X_i = x_i \text{ and } X_j = x_j\} = \mathbb{P}\{X_i = x_i\}\mathbb{P}\{X_j = x_j\}$ for discrete variables) then the covariances are zero and the variance is additive. This explains why the binomial and negative binomial distributions have variances which are simply the sums of those of the Bernoulli and Geometric distributions respectively. It also suggests why the (negative) hypergeometric distributions have much nastier expressions! The details of this will be covered in a future course.