3.2.1 Bernoulli distribution

Definition 3.6. X has the **Bernoulli distribution** if $\text{Range}(X) = \{0, 1\}$ and

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

for some $p \in [0, 1]$. We denote this $X \sim \text{Bern}(p)$.

The Bernoulli random variable is the simplest possible example of a random variable, and it serves as our canonical model of a binary trial: yes/no, true/false, success/failure, heads/tails, win/lose, 1/0. The parameter p is the *probability of success*:

$$P(X = 1) = p^1(1-p)^{1-1} = p^1(1-p)^0 = p.$$

And so $1 - p$ is the probability of failure by the complement rule. To compute the expected value of this distribution, we simply apply the definition:

$$E(X) = \sum_{i=0}^{1} iP(X = i) = 0(1-p) + 1p = p.$$

In most cases of interest, $0 < p < 1$, and so we see that $E(X) \notin \text{Range}(X) = \{0, 1\}$. So it is a strange property of the expectation that this "typical value" need not actually be one of the values that X could take on when the random phenomenon is finally realized. Curious!

Example 3.2. Here is a simple construction that shows how the Bernoulli distribution can arise from first principles. Recall the formal construction of a random variable:

- (**base space**) an underlying probability space $(S, A \subseteq S, P_0)$;
- (**random variable**) a function $X : S \to \mathbb{R}$ that takes outcomes s and returns real numbers $X(s)$;
- (**pushforward space**) a new probability space $(\mathbb{R}, B \subseteq \mathbb{R}, P)$ induced by the function X "pushing forward" the randomness of the base space to the real numbers. The new probability measure P is called the probability distribution of X.

Consider some fixed event of interest $A \subseteq \mathbb{R}$. We can define the *indicator function*:

$$I_A(s) = \begin{cases} 0 & s \notin A \\ 1 & s \in A. \end{cases} \tag{3.5}$$

So this is the function that *indicates* whether or not the event A occurs. $I_A : S \to \mathbb{R}$ is an **indicator random variable** with

- (**range**) $\text{Range}(I_A) = \{0, 1\}$
- (**distribution**) $P(I_A = 1) = P_0(A)$ and $P(I_A = 0) = P_0(A^c) = 1 - P_0(A)$.

So we see that $I_A \sim \text{Bern}(p = P_0(A))$.

3.2.2 Binomial distribution

Definition 3.7. X has the **binomial distribution** if $\text{Range}(X) = \{0, 1, 2, ..., n\}$ and

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k = 0, 1, 2, ..., n,$$

for some $p \in [0, 1]$. We denote this $X \sim \text{Binom}(n, p)$.

The binomial distribution arises when we define a random variable X that counts the number of successes that occur in a sequence of independent binary trials, each with the same probability of success $0 \leq p \leq 1$. We can use $X \sim \text{Binom}(n, p)$ to model various phenomena:

- the number of heads you observe in $n = 100$ flips of a fair ($p = 1/2$) coin;
- an assembly line manufactures $n = 200$ light bulbs. How many actually work?
- you swipe right on $n = 1000$ dating profiles. How many swipe right on you?
- you text a survey to $n = 50$ eligible voters. How many actually respond?
- you attempt $n = 40$ soufflés. How many come out of the oven light, risen, fluffy, and delectable?
- so in general, in n independent attempts, each with probability of success p, how many attempts turn out successful?

This is the sort of random variable that we build up from first principles. The underlying sample space S is the set of all length-n binary strings. Each digit is 0 or 1 indicating whether or not the trial was a failure or a success. Given a string, X is essentially counting the number of ones: $X(011010) = 3$. To see where the pmf of the binomial comes from, consider the case of $n = 3$. We start by enumerating the underlying sample space, and then seeing how the underlying probabilities get pushed forward:

P_0	S	$X = k$	$P(X = k)$
$(1-p)^3$	000	0	$(1-p)^3$
$p(1-p)^2$	100	1	$3p(1-p)^2$
$p(1-p)^2$	010		
$p(1-p)^2$	001		
$p^2(1-p)$	110	2	$3p^2(1-p)$
$p^2(1-p)$	101		
$p^2(1-p)$	011		
p^3	111	3	p^3

First, we see that $\text{Range}(X) = \{0, 1, 2, 3\}$, because in three trials you could have anywhere from no successes to all successes. Next, we see that, to compute $P(X = 1)$ for instance, we recognize that the event "$X = 1$" only happens if the original sequence is 100, or 010, or 001. This is a disjoint union, so

$$P(X = 1) = P_0(100 \cup 010 \cup 001) = 3p(1-p)^2.$$

$p(1-p)^2$ is the probability of observing exactly one success and exactly two failures in a sequence of three *independent* trials, and so $P(X = 1)$ is equal to this baseline probability multiplied by the total number of ways it can happen. In the general n case, the event $X = k$ only occurs is we observe exactly k successes and exactly $n-k$ failures in our n trials. The individual probability of any one such outcome is $p^k(1-p)^{n-k}$, and so the overall probability of $X = k$ is this number multiplied by the total number of ways we could construct a length-n binary string with exactly k ones in it. This is a "select k from n" counting problem, where we have k slots set aside for successful trials, and we are selecting which of the n trials to designate as the successes. We are drawing without replacement, and order does not matter, so the total number of ways our length-n string could have exactly k ones is $\binom{n}{k}$, thus giving the final probability

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, ..., n.$$

Next, let us compute the expected value. For now, we will do it by proceeding directly from the definition, but later we will develop tools that allow us to perform the calculation with much less pain:

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

At the end of the computation, we invoked the *binomial theorem*: $(x+y)^m = \sum_{j=0}^{m} \binom{m}{j} x^j y^{m-j}$ with $m = n-1$, $x = p$, and $y = 1-p$. Indeed, it is due to their role in this theorem that the binomial coefficients get their name, and that the distribution we study gets *its* name. In any case, we see that this was a rather clumsy and inartful computation, and we will have cleaner methods for doing it later. The final result, though, is intuitive. It implies that the probability of success p has the interpretation as the proportion of the total number of trials that we expect to be a success: $p = E(X)/n$.

Remark 3.1. Figure 3.6 displays the pmf of the binomial distribution for various choices of $n \in \mathbb{N}$ and $p \in (0, 1)$. We see that, regardless what p is, the pmf resembles a bell-shaped curve more and more as n increases. This is an example of the **central limit theorem**, which we will study later.

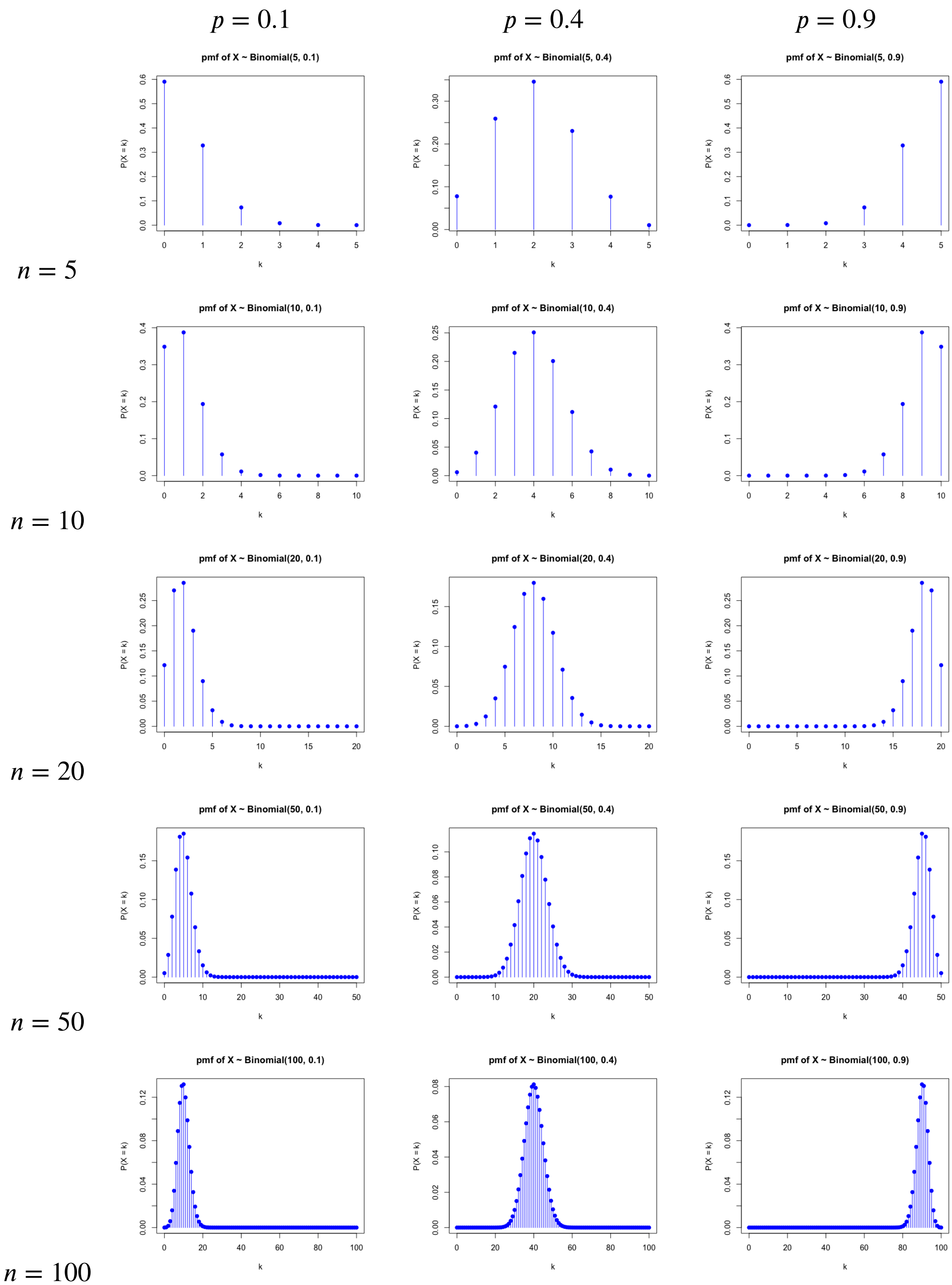


Figure 3.6: As n increases the pmf of Binom(n, p) looks more like a bell curve.

3.2.3 Geometric distribution

Definition 3.8. X has the **geometric distribution** if Range$(X) = \{1, 2, 3, ...\}$ and

$$P(X = k) = (1 - p)^{k-1} p \quad k = 1, 2, 3, ...,$$

for some $p \in (0, 1)$. We denote this $X \sim \text{Geom}(p)$.

The geometric distribution arises when we define a random variable X that counts the number of independent binary trials we must sit through until we observe the first success. The first success could occur on the first attempt, or the second, or the third, or the billionth, so Range$(X) = \{1, 2, 3, ...\}$. The event that $X = k$ is equivalent to the event that the first $k - 1$ trials are failures (each occurs with probability $1 - p$) and the kth trial is a success (occurs with probability p). Since the trials are independent, the probability of $k - 1$ failures followed by a success is the product of these probabilities, like so:

X	Trial 1	Trial 2	Trial 3	Trial 4	$\cdots$	Trial k	$\cdots$	$P(X = k)$
1	1							p
2	0	1						$(1-p)^1 p$
3	0	0	1					$(1-p)^2 p$
4	0	0	0	1				$(1-p)^3 p$
$\vdots$								$\vdots$
k	0	0	0	0	$\cdots$	1		$(1-p)^{k-1} p$
$\vdots$								$\vdots$

To see where this distribution gets its name, let us verify that the pmf sums to one:

$$\begin{aligned}
\sum_{k=1}^{\infty} P(X = k) &= \sum_{k=1}^{\infty} (1-p)^{k-1} p \\
&= p \sum_{k=1}^{\infty} (1-p)^{k-1} \\
&= p \sum_{i=0}^{\infty} (1-p)^{i} \\
&= p \frac{1}{1 - (1-p)} && \text{geometric series, since } |1-p| < 1 \\
&= p \frac{1}{p} \\
&= 1.
\end{aligned}$$

Next let us calculate the expected value:

$$
\begin{aligned}
E(X) &= \sum_{k=1}^{\infty} kP(X=k) \\
&= \sum_{k=1}^{\infty} k(1-p)^{k-1}p \\
&= \sum_{k=1}^{\infty} (k+0)(1-p)^{k-1}p \\
&= \sum_{k=1}^{\infty} (k-1+1)(1-p)^{k-1}p \\
&= \sum_{k=1}^{\infty} [(k-1)(1-p)^{k-1}p + (1-p)^{k-1}p] \\
&= \sum_{k=1}^{\infty} (k-1)(1-p)^{k-1}p + \sum_{k=1}^{\infty} (1-p)^{k-1}p \\
&= \sum_{i=0}^{\infty} i(1-p)^{i}p + 1 \\
&= 0 + \sum_{i=1}^{\infty} i(1-p)^{i}p + 1 \\
&= (1-p)\sum_{i=1}^{\infty} i(1-p)^{i-1}p + 1 \\
&= (1-p)E(X) + 1.
\end{aligned}
$$

Solving $E(X) = (1-p)E(X) + 1$, we have that $E(X) = 1/p$.

Example 3.3. The first probability problem we encountered was "how many flips of a fair coin does it take on average until you flip the first head?" A sequence of fair coin flips is a sequence of independent binary trials with probability of success $p = 1/2$, and so the number of flips until the first head has $X \sim \text{Geom}(1/2)$. We see then that the expected number of flips until the first head is $E(X) = 1/(1/2) = 2$. So on average, the first head occurs on the second flip.

3.2.4 Poisson distribution

Definition 3.9. X has the **Poisson distribution** if $\text{Range}(X) = \mathbb{N} = \{0, 1, 2, 3, ...\}$ and

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad k \in \mathbb{N} \tag{3.6}$$

for some *rate* $\lambda > 0$. We denote this $X \sim \text{Poisson}(\lambda)$.

The Poisson distribution is often used to model the number of random *arrivals* in a given window of time: the number of emails you receive in an hour, the number of claims an insurance company receives in a month, the number of α-particles discharged from a radioactive material, etc.

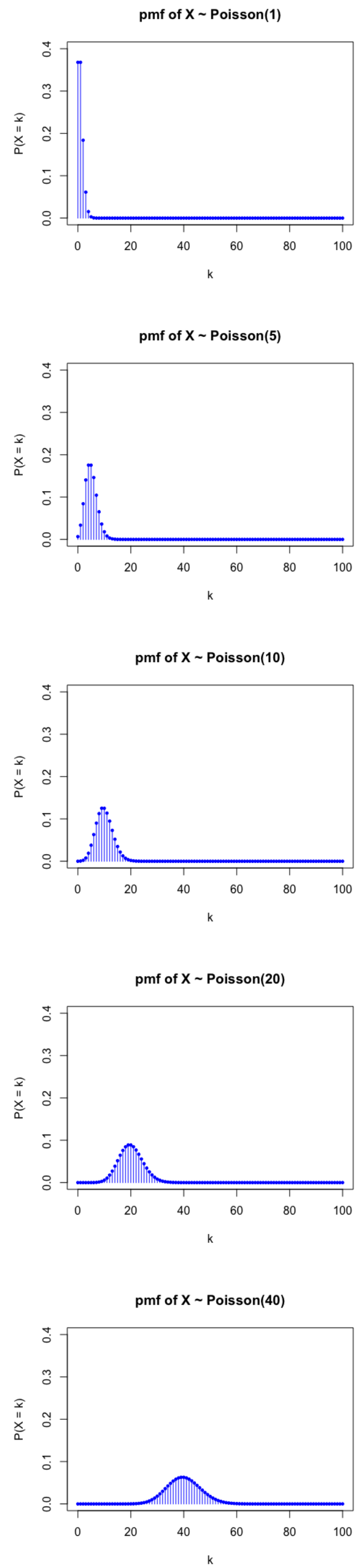
We first check that the pmf is valid:

$$\sum_{k=0}^{\infty} P(X = k) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1.$$

Next we compute the expected value:

$$\begin{aligned}
E(X) &= \sum_{n=0}^{\infty} nP(X = n) \\
&= \sum_{n=0}^{\infty} n\frac{\lambda^n}{n!}e^{-\lambda} \\
&= e^{-\lambda}\sum_{n=0}^{\infty} n\frac{\lambda^n}{n!} && \text{Pull out constant} \\
&= e^{-\lambda}\sum_{n=1}^{\infty} n\frac{\lambda^n}{n!} && n = 0 \text{ term is equal to } 0 \\
&= e^{-\lambda}\sum_{n=1}^{\infty} \frac{\lambda^n}{(n-1)!} && n \text{ cancels} \\
&= \lambda e^{-\lambda}\sum_{n=1}^{\infty} \frac{\lambda^{n-1}}{(n-1)!} && \text{Pull out } \lambda \\
&= \lambda e^{-\lambda}\sum_{j=0}^{\infty} \frac{\lambda^j}{j!} && \text{Reindex} \\
&= \lambda e^{-\lambda}e^{\lambda} && \text{Recall Taylor series: } e^x = \sum_{k=0}^{\infty}\frac{x^k}{k!}\ \forall x \in \mathbb{R} \\
&= \lambda.
\end{aligned}$$

Remark 3.2. Figure 3.7 displays the pmf of the Poisson for different values of the rate parameter λ. As λ increases, we see that the pmf shifts rightward, which makes sense given that $E(X) = \lambda$. We also see that it gets wider, which we will make sense of later when we show that $\text{var}(X) = \lambda$ as well. Lastly, we see that the pmf looks more and more bell-like as λ grows, similar to what we observed with the binomial distribution in Figure 3.6. These of course are not coincidences. Both are instances of the same general phenomenon: the central limit theorem.



$\lambda = 1$

$\lambda = 5$

$\lambda = 10$

$\lambda = 20$

$\lambda = 40$

Figure 3.7: As λ increases, the Poisson pmf shifts right, widens, and becomes more bell-like.

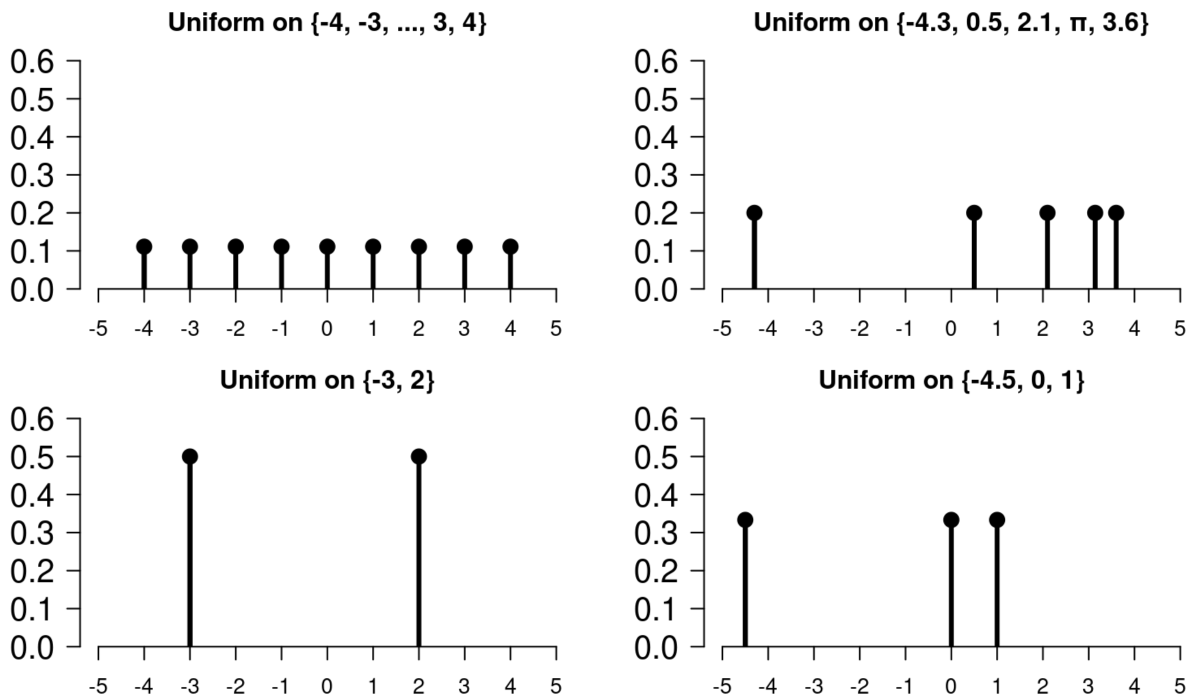


Figure 3.8: Uniform distributions on different sets of real numbers.XXXShould add CDFs as well

3.2.5 Discrete uniform distribution

Definition 3.10. X has the (**discrete**) **uniform distribution** if it has a finite range and all values are equiprobable. So $\text{Range}(X) = \{x_1, x_2, ..., x_n\} \subseteq \mathbb{R}$ for some $n \in \mathbb{N}$ and

$$P(X = x_i) = \frac{1}{n}, \quad \forall i = 1, 2, ..., n.$$

We denote this $X \sim \text{Unif}(x_1, x_2, ..., x_n)$.

Figure 3.8 displays some examples of what the pmf might look like. Because all of the probabilities are the same, it has none of the peaks and valleys and tails we usually expect from a distribution plot. It's just *uniform* across the range. Furthermore, note that this is our first discrete random variable that is supported on a set of arbitrary real numbers. The Bernoulli, binomial, geometric, and Poisson distributions are all supported on a subset of $\mathbb{N}$, but this is not a requirement for a discrete random variable. Any finite (or countably infinite) set of real numbers will serve, and here we have our first.

The expected value of this distribution is

$$E(X) = \sum_{i=1}^{n} x_i P(X = x_i) = \sum_{i=1}^{n} x_i \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

This is the familiar **average**. So whenever we compute the average or *mean* of a set of n numbers, we are implicitly treating that set of numbers as the range of a discrete random variable with uniform $(1/n)$ probabilities, and then calculating the expected value.