Random Variables

Discrete random variables, probability mass functions, cumulative distribution functions

10/5/23

Suppose we want to simulate tossing a coin 10 times, and compute the proportion of times that the coin lands heads. We might use the function `sample()` to sample from `(0, 1)`, where the outcome "Heads" is represented by the number `1` and the outcome "Tails" is represented by the number `0`. We can save the result as `tosses_10`. Our code might be `tosses_10 <- sample(c(0,1), 10, replace = TRUE)`. `tosses_10` might look like `0 0 1 1 0 1 1 1 1 0`, and if we compute its mean (`mean(tosses_10)`), we would get the proportion of times the coin landed heads (in this example it is 0.6).

Random variables

What we did, in fact, was define a **function** that assigned a real number to each possible outcome in Ω. In our simulation above, if Ω is the set of outcomes {"Heads", "Tails"}, we assigned the outcome "Heads" to the real number 1, and the outcome "Tails" to the real number 0. By sampling over and over again from `(0,1)`, we got a sequence of 0's and 1's that was *randomly generated* by our sampling, and then we could do arithmetic on this sequence, such as compute the proportion of times we sampled 1. To put it in mathematical notation:

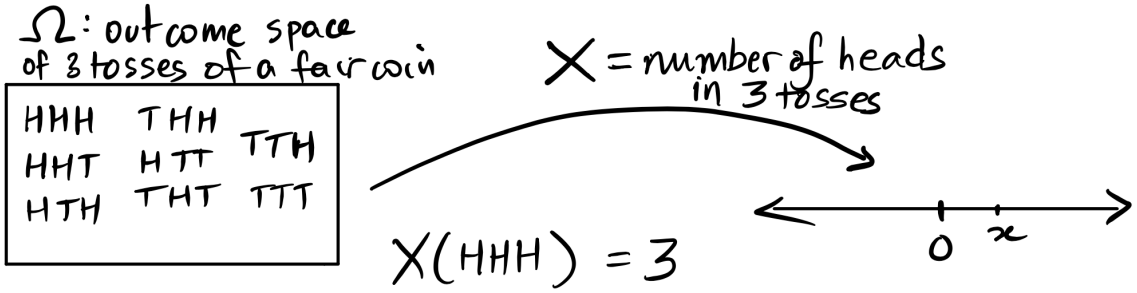
$$X : \Omega \to \mathbb{R}$$

X is called a *random variable*: variable, because it takes different values on the real line, and random, because it inherits the randomness from the generating process (in this case, the process is tossing a coin).

Random variable A random variable is a function that associates real numbers with outcomes from a random experiment which are in an outcome space Ω.

These assigned numbers have probabilities coming from the probability distribution on Ω. The range of the random variable is the set of all the possible values that X can take. We usually denote random by $X, Y, ...$ We write statements about the values X takes, such as $X = 1$ or $X = 0$. Note that $X = 1$ is an *event* and it may be true or not. The probability of such events is written as $P(X = x)$, where x is a real number.

Probability distribution of a random variable X The set of possible values of X, along with the associated probabilities, is called the *probability distribution* for the random variable X.



Discrete and continuous random variables Discrete random variables are restricted to take *particular* values in an interval, they cannot take just any value, as opposed to continuous random variables which can take any value in some specified interval. Note the similarity to discrete and continuous quantitative data types.

Examples of discrete random variables

- The number of heads in 3 tosses of a fair coin: The assignment is similar to the outcomes from a single toss, except now we have the possible outcome from tossing a coin three times. For example, the outcome `HHH` is assigned the number 3, the outcomes `HHT, HTH, THH` are all assigned the number 2 etc. Note that even though we should write $X(\texttt{HHH}) = 3$, it is common to write just $X = 3$.
- The number of tosses until the coin lands heads for the first time: If X is the random variable representing the number of tosses until a coin lands heads, the smallest value X can take is 1 (you need at least 1 toss), and there is no upper bound, since in theory, one could keep tossing the coin forever and it could land tails every single time.
- The number of people that arrive at an ATM in a day: This is also a counting random variable, as described.

Examples of continuous random variables

In all of the following, we do not restrict the value taken by the random variable.

- Time between consecutive people arriving at an ATM
- Price of a stock
- Height of a randomly selected stat 20 student

Example: Making bets on red in Roulette

The roulette wheels used in Las Vegas have 38 numbered slots, numbered from 1 to 36, of which 18 are colored red, and 18 black. There are two green slots numbered with zero and a double zero. As the wheel spins, a ball is sent spinning in the opposite direction. When the wheel slows the ball will land in one of the numbered slots. Players can make various bets on where the ball lands, such as betting on whether the ball will land in a red slot or a black slot. If a player bets one dollar on red, and the ball lands on red, then they win a dollar, in addition to getting their stake of one dollar back. If the ball does not land on red, then they lose their dollar to the casino. Suppose a player bets six times on six consecutive spins, betting on red each time. Their *net gain* can be defined as the amount they won minus the amount they lost. Is net gain a random variable? What are its possible values (write down how much the player can win or lose in one spin of the wheel, then two, and so on)?

Check your answer

Yes, net gain *is* a random variable, and its possible values are: $-6, -4, -2, 0, 2, 4, 6$. (Why?)

The probability distribution of a random variable X

We can list the values taken by a random variable in a table, along with the probability that the variable takes a particular value. This table is called the *distribution table* of the random variable. For example, let X be the number of heads in 3 tosses of a fair coin. The probability

distribution table for X is shown below. The first column should have the possible values that X can take, denoted by x, and the second column should have $P(X = x)$. We should make sure that the probabilities add up to 1: $\sum_x P(X = x) = 1$.

x	$P(X = x)$
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

The probability mass function or pmf of a discrete random variable

The probability mass function (pmf) of a discrete random variable X The **pmf** of a discrete random variable X is defined to be the function $f(x) = P(X = x)$.

We can write down the definition of the function $f(x)$ and it gives the same information as in the table:

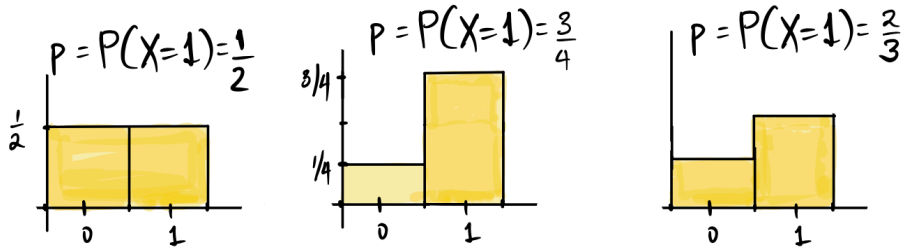
$$f(x) = \begin{cases} \frac{1}{8}, & x = 0, 3 \\ \frac{3}{8}, & x = 1, 2 \end{cases}$$

We see here that $f(x) > 0$ for only 4 real numbers, and is 0 otherwise. We can think of the total probability mass as 1, and $f(x)$ describes how this mass of 1 is distributed among the real numbers. It is often easier and more compact to define the probability distribution of X using f rather than the table.

Special distributions

Bernoulli distribution

This is the simplest discrete distribution: X be a random variable that takes the value 1 with probability p and the value 0 with probability $1 - p$, then X is called a *Bernoulli* random variable. We say that X is Bernoulli with parameter p, because if we know p, we can calculate the probabilities associated with X. We can visualize the distribution and the pmf using a probability histogram. This is similar to the plots of distributions that we have seen before. It is called "histogram'' for a couple of reasons. Firstly, we want to think about areas, with the total area of the bins being 1. Secondly, we will use the same type of plot for continuous random variables, and it can easily generalize.



Parameter of a probability distribution A constant(s) associated with the distribution. If you know the parameters of a probability distribution, then you can compute all the values of $f(x)$.

Discrete uniform random distribution

Let's suppose X takes the values $1, 2, 3, \ldots, n$ with $P(X = k) = \dfrac{1}{n}$ for each of the k from 1 to n. We call X a *discrete uniform* random variable, and the probability distribution is called the discrete uniform probability distribution: $P(X = k) = \dfrac{1}{n}$ for $1 \leq k \leq n$. Think of die rolls, where $n = 6$ for a standard die. If X is the outcome when we roll a die, then X is called the discrete uniform random variable with $n = 6$. In fact, the only thing we need to know, in order to compute the probability that X will take a particular value is n. We call n the *parameter* of the discrete uniform distribution.

Rolling a pair of dice and summing the spots

Suppose we roll a pair of dice and sum the spots, and let X be the sum. Is X a discrete uniform random variable?

Check your answer

No. X takes discrete values: $2, 3, 4, \ldots, 12$, but these are **not** equally likely. Can you compute their probabilities?

Binomial distribution

Suppose a particular random experiment has two possible outcomes which we designate as a ''success" or a "failure' ', where the probability of a success on any trial is p, regardless of what happens on any other trial. Suppose we repeat this experiment n times, and let X count the number of successes in n **independent** trials of this random experiment (think tossing a coin

n times, and counting the number of heads). The independence of the trials is a consequence of the probability of success p staying the same for every trial. Then

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

where k takes the values $0, 1, 2, \ldots, n$, and $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ and is read as "n *choose* k". These are called the binomial coefficients.[1]

We say that X has the *binomial* distribution with parameters n and p and write this as $X \sim Bin(n, p)$. Note that the Bernoulli distribution is a special case of the binomial, with $n = 1$.

Example: Proportion of Californians that have had at least one positive COVID test.

Based on the December 2022 report "State of the COVID-19 Pandemic"[2], about 35% of Americans reported having at least one positive COVID test. Suppose we survey 10 California residents, sampling them *with* replacement from a database of voters, what is the probability that more than 3 of the individuals in our sample would have had at least one positive COVID test?

In this example, since we are counting the number of individuals in our sample that have had at least one positive COVID test, such an individual would count as a "success", because a success is whatever outcome we are counting. We can now set up our random variable X:

Let X be the number of individuals in our sample who have had at least one positive COVID test.

Then $X \sim Bin(10, p = 0.35)$. (Why are these the parameters?) Using the complement rule,

$P(X > 3) = 1 - P(X \leq 3) = 1 - (P(X = 0) + P(X = 1) + P(X = 2))$.

Note that since these events are mutually exclusive, we can use the addition rule. This gives us:

$$P(X > 3) = 1 - \left(\binom{10}{0} (0.35)^0 (0.65)^{10} + \binom{10}{1} (0.35)^1 (0.65)^9 + \binom{10}{2} (0.35)^2 (0.65)^8 \right) \approx 0.74$$

In `R` we can use a special function to compute the binomial probabilities $f(k) = P(X = k)$. It is called `dbinom(x, size, prob)` and takes as input the k that we want (`x`), the number of trials n (`size`), and p (`prob`). In this example, `x` are the values of interest, or 0, 1, 2; `size` is the parameter n, which is 10, and `prob` is the probability of a success.

[1] https://en.wikipedia.org/wiki/Binomial_coefficient
[2] https://www.covidstates.org/reports/state-of-the-covid-19-pandemic

```r
1 - dbinom(0, size = 10, prob = 0.35) -
  dbinom(1, size = 10, prob = 0.35) -
  dbinom(2, size = 10, prob = 0.35)
```

```
[1] 0.7383926
```

As stated above, we can define *events* by the values of random variables. For example, let $X = Bin(10, 0.4)$. In words X counts the number of successes in 10 trials. Given this X, what are the following events in words?

- $X = 5$
- $X \leq 5$
- $3 \leq X \leq 8$

What are these events in words? What are their probabilities?

Check your answer

- X is the number of successes in ten trials, where the probability of success in each trial is 40%. $X = 5$ is the event that we see exactly five successes in the ten trials, while $X \leq 5$ is the event of seeing *at most* five successes in ten trials. The last event, $3 \leq X \leq 8$ is the event of at least three successes, but not more than eight, in ten trials. We will use `dbinom()` to compute the probabilities.

```r
# P(X=5)
dbinom(x = 5, size = 10, prob = 0.4)
```

```
[1] 0.2006581
```

```r
# P(X <= 5)
dbinom(x = 0, size = 10, prob = 0.4) + dbinom(x = 1, size = 10, prob = 0.4) +
  dbinom(x = 2, size = 10, prob = 0.4) + dbinom(x = 3, size = 10, prob = 0.4) +
  dbinom(x = 4, size = 10, prob = 0.4) + dbinom(x = 5, size = 10, prob = 0.4)
```

```
[1] 0.8337614
```

```r
# P(3 <= X <= 8)
dbinom(x = 3, size = 10, prob = 0.4) + dbinom(x = 4, size = 10, prob = 0.4) +
```

```r
dbinom(x = 5, size = 10, prob = 0.4) + dbinom(x = 6, size = 10, prob = 0.4) +
dbinom(x = 7, size = 10, prob = 0.4) + dbinom(x = 8, size = 10, prob = 0.4)
```

```
[1] 0.8310325
```

Hypergeometric distribution

In the example above, we sampled from the population of California *with* replacement. Usually we sample *without* replacement. Suppose our population consists of N units. If, each time we draw a unit for our sample, all the units are equally likely to be drawn (just like in the box model), a sample drawn without replacement is called a *simple random sample*. Suppose our population consists of just two types of units that we call "successes" and "failures" (as usual, a "success" is whatever outcome we are interested in), and we draw a sample of size n without replacement. Now, the probability of success will *not* stay the same on each draw. If we let X be the number of successes in n draws, then we have that

$$P(X = k) = \frac{\binom{G}{k} \times \binom{N-G}{n-k}}{\binom{N}{n}}$$

where N is the size of the population, G is the total number of successes in the population, and n is the sample size (so k can take the values $0, 1, \ldots, n$ or $0, 1, \ldots, G$, if the number of successes in the population is smaller than the sample size.)

Example: Gender discrimination at large supermarket?

A large (with 1,000 employees) supermarket chain in Florida occasionally selects employees to receive management training. A group of women there claimed that female employees were passed over for this training in favor of their male colleagues. The company denied this claim. (A similar complaint of gender bias was made about promotions and pay for the 1.6 million women who work or who have worked for Wal-Mart. The Supreme Court heard the case in 2011 and ruled in favor of Wal-Mart, in that it rejected the effort to sue Wal-Mart.)[3] If we set this up as a probability problem, we might ask the question of how many women have been selected for executive training in the last 10 years. Suppose no women had ever been selected in 10 years of annually selecting one employee for training. Further suppose that the number of men and women were equal, and suppose the company claims that it draws employees at random for the training, from the 1,000 eligible employees. If X is the number of women that have been picked for training in the past 10 years, what is $P(X = 0)$?

[3] https://www.latimes.com/world/la-xpm-2011-jun-20-la-naw-wal-mart-court-20110621-story.html

Since there are 1,000 employees, and half are women, we have $G = N - G = 500$. Of the 10 picked, none are women. We are picking a sample of size 10 without replacment, therefore we have that:

$$P(X = 0) = \frac{\binom{500}{0} \times \binom{500}{10}}{\binom{1000}{10}} \approx 0.0009$$

The function that we can use in `R` to compute hypergeometric probabilities is called `dhyper(x, m, n, k)`, where `x` is the number of successes in the sample that we are counting (what we call k), `m` is G or the number of successes in the population, and n is $N - G$. k is the sample size.

```
dhyper(0, 500, 500, 10)
```

```
[1] 0.0009331878
```

Note that the function used in `R` to compute the binomial coefficient $\binom{n}{k}$ is `choose(n,k)`.

```
choose(500, 0)*choose(500, 10)/choose(1000, 10)
```

```
[1] 0.0009331878
```

Binomial vs Hypergeometric distributions

Both these distributions deal with the counting the number of successes in a *fixed* number of *trials* (where each instance of the random experiment that generates a success or a failure is called a trial, for example each toss of a coin, or each card dealt from a deck, and we count a heart to be a success). The difference is that for a binomial random variable, the probability of a success stays the *same* for each trial, and for a hypergeometric random variable, the probability *changes* with each trial. If we use a box of tickets to describe these random variables, both distributions can be modeled by sampling from boxes with each ticket marked with 0 or 1, but for the binomial distribution, we sample n times *with* replacement and count the number of successes by summing the draws; and for the hypergeometric distribution, we sample n times *without* replacement, and count the number of successes by summing the draws.

Note that when the sample size is small relative to the population size, there is not much difference between the probabilities if we use a binomial distribution vs using a hypergeometric distribution. Let's pretend in the gender discrimination example above that we are sampling *with* replacement. Now suppose we sample 10 times with replacement, let's use the binomial distribution to compute the chance of never picking a woman.

```r
# n = 10
# p = 0.5

dbinom(0, 10, 0.5)
```

```
[1] 0.0009765625
```

You can see that the values are very close to each other. (Recall that the probability using the hypergeometric distribution was `0.0009331878`.

The cumulative distribution function $F(x)$

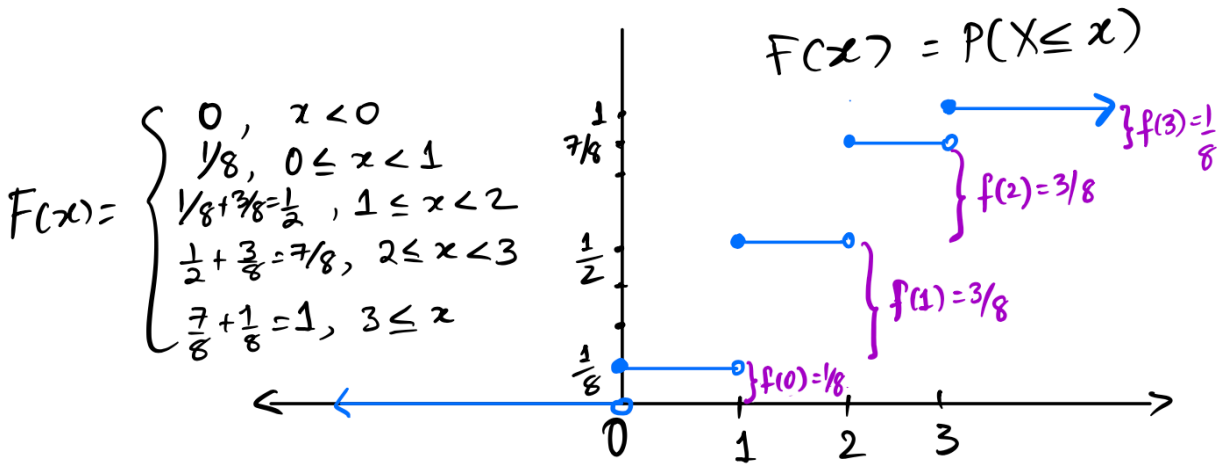
The cumulative distribution function (cdf) $F(x)$ of a random variable X is defined for *every* real number, and gives, for each x, the amount of probability or mass that has been *accumulated* up to (and including) the point x, that is, $F(x) = P(X \leq x)$.

We usually abbreviate this function to cdf. It is a very important function since it also describes the probability distribution of X.

For example, if X is the number of heads in 3 tosses of a fair coin, recall that:

$$f(x) = \begin{cases} \frac{1}{8}, & x = 0, 3 \\ \frac{3}{8}, & x = 1, 2 \end{cases}$$

In this case, $F(x) = P(X \leq x) = 0$ for all $x < 0$ since the first positive probability is at 0. Then, $F(0) = P(X \leq 0) = 1/8$ after which it stays at $1/8$ until $x = 1$. Look at the graph below:



Notice that $F(x)$ is a step function, and right continuous. The jumps are at exactly the values for which $f(x) > 0$. We can get $F(x)$ from $f(x)$ by adding the values of f up to and including x, and we can get $f(x)$ from $F(x)$ by looking at the size of the jumps.

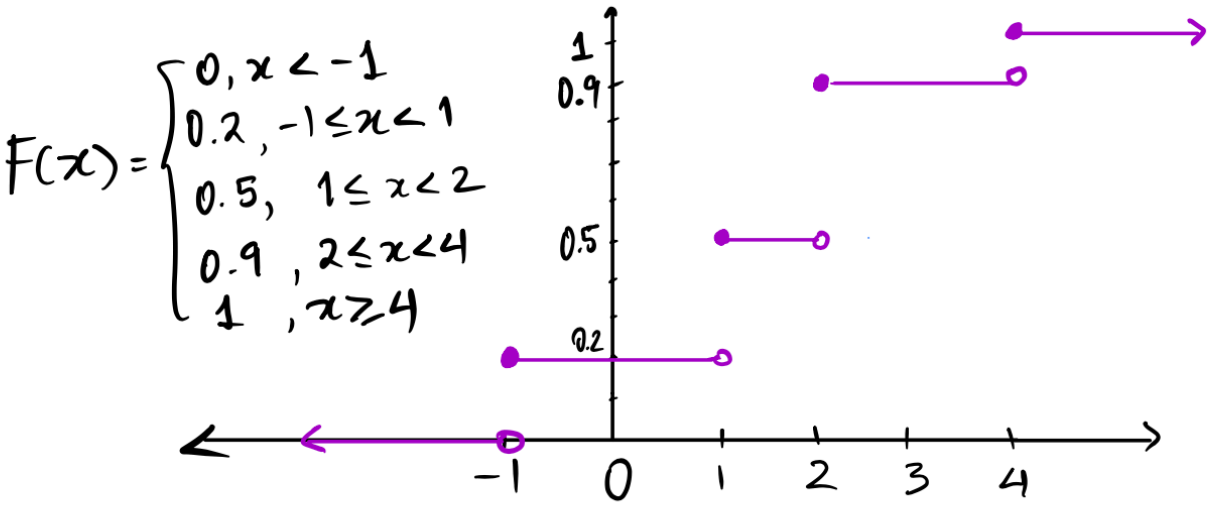
Example: Drawing the graph of the cdf

Let X be the random variable defined by the distribution table below. Find the cdf of X, and draw the graph, making sure to define $F(x)$ for all real numbers x. Before you do that, you will have to determine the value of $f(x)$ for $x = 4$.

x	$P(X = x)$
-1	0.2
1	0.3
2	0.4
4	??

Check your answer

Since $\sum_x P(X = x) = \sum_x f(x) = 1$, $f(4) = 1 - (0.2 + 0.3 + 0.4) = 0.1$. Therefore $F(x)$ is as shown below.



In R, we have functions that calculate $F(x)$ for some special distributions, including the binomial and hypergeometric distributions. For a binomial distribution, we use the function `pbinom(x, size, prob)` Similarly, for the hypergeometric distribution, we use `phyper(x, m, n, k)`.

Going back to the example above, let's compute the probabilities using `pbinom(x, 10, 0.4)` and compare them to the probabilities computed earlier using `dbinom(x, size, prob)`

```
# P(X = 5)
dbinom(x = 5, size = 10, prob = 0.4)
```

```
[1] 0.2006581
```

```
# P(X = 5)
pbinom(5, 10, 0.4) - pbinom(4, 10, 0.4)
```

```
[1] 0.2006581
```

```
# P(X <= 5)
dbinom(x = 0, size = 10, prob = 0.4) + dbinom(x = 1, size = 10, prob = 0.4) +
```

```
  dbinom(x = 2, size = 10, prob = 0.4) + dbinom(x = 3, size = 10, prob = 0.4) +
  dbinom(x = 4, size = 10, prob = 0.4) + dbinom(x = 5, size = 10, prob = 0.4)
```

```
[1] 0.8337614
```

```
# P(X <= 5)
pbinom(5, 10, 0.4)
```

```
[1] 0.8337614
```

```
# P(3 <= X <= 8)
dbinom(x = 3, size = 10, prob = 0.4) + dbinom(x = 4, size = 10, prob = 0.4) +
  dbinom(x = 5, size = 10, prob = 0.4) + dbinom(x = 6, size = 10, prob = 0.4) +
  dbinom(x = 7, size = 10, prob = 0.4) + dbinom(x = 8, size = 10, prob = 0.4)
```

```
[1] 0.8310325
```

```
# P(3 <= X <= 8)
pbinom(8, 10, 0.4) - pbinom(2, 10, 0.4)
```

```
[1] 0.8310325
```

What is going on in the last expression? Why is $P(3 <= X <= 8) = F(8) - F(2)$?

Check your answer

$P(3 <= X <= 8)$ consists of all the probability at the points $3, 4, 5, 6, 7, 8$. $F(8) = P(X 8)$ $ is all the probability up to 8, including any probability at 8. We subtract off all the probability up to and including 2 from $F(8)$ and are left with the probability at the values 3 up to and including 8, which is what we want.

Connections between discrete random variables and draws from a box of tickets

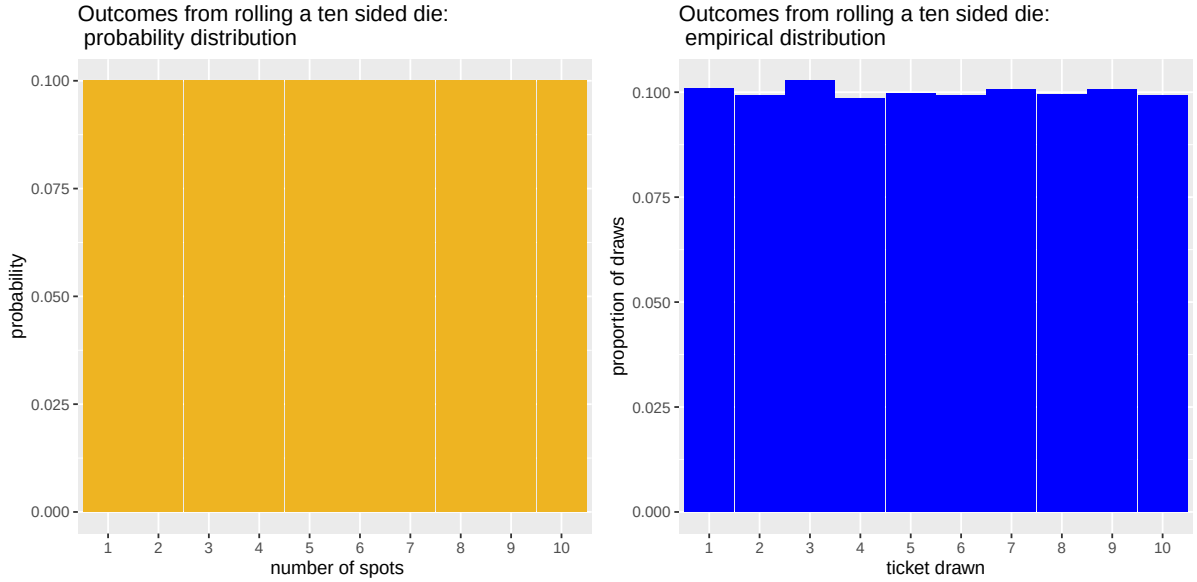
Example: the discrete uniform random variable

Say we have a ten sided die. How can we set up a simulation to model for rolling this die once? (We can simulate it using a box of tickets, or in R).

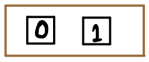
Check your answer

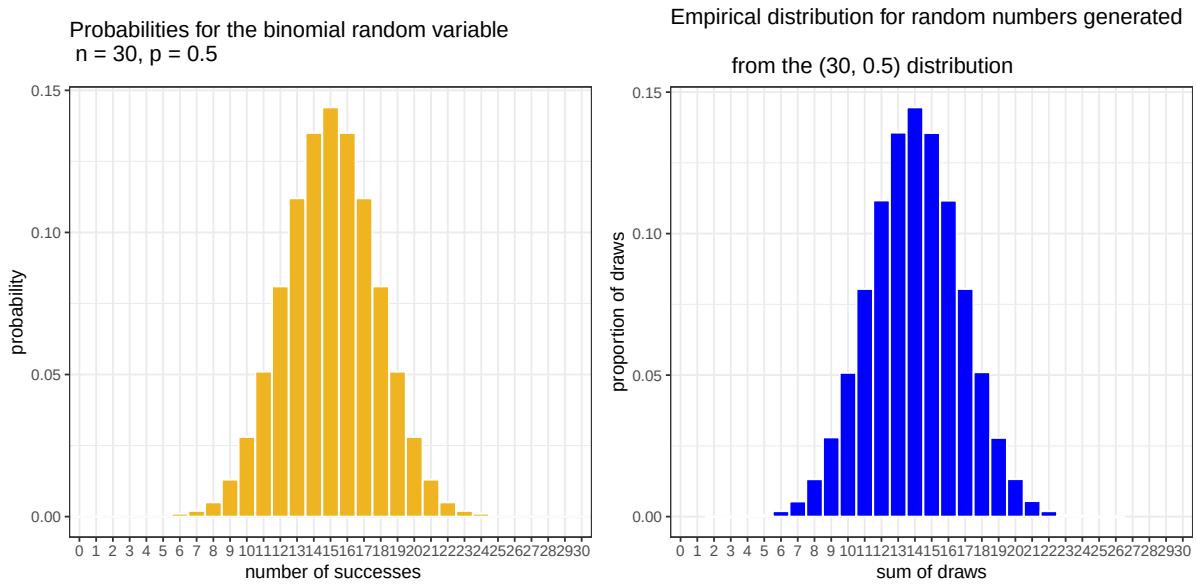
A box with ten tickets, marked $\boxed{1}, \boxed{2}, \ldots \boxed{10}$; and draw once from this box.

We can also simulate the probability distribution by drawing over and over again, and looking at the empirical distribution. You can see below that it closely matches the actual probability distribution.



Example: the binomial(n, p) random variable

Let $X \sim Bin(30, 0.5)$. We can simulate the values of this random variable by drawing from the box  30 times, and summing the draws. This will simulate counting the number of successes in n trials. As in the example above, we will plot the probability distribution of X on the left, and the empirical distribution on the right.



Summary

- In these notes, we defined random variables, and described discrete and continuous random variables.
- For any random variable, there is an associated probability distribution, and this is described by the probability mass function or pmf $f(x)$. We also defined a function that, for a random variable X, and any real number x, describes all the probability that is to the left of x. This function is called the cumulative distribution function (cdf) of X and is denoted $F(x)$.
- We looked at some special distributions (Bernoulli, binomial, discrete uniform, and hypergeometric)
- Finally, we looked at connections between random variables and boxes with tickets, and saw how to simulate some empirical distributions.