Chapter 3

We can calculate probabilities for common discrete random variables easily in R. There are two parts to finding probabilities using the PMF or CDF.
R uses four prefixes to reference difference elements of a random variable. These are:

p for "probability", the cumulative distribution function (CDF)
q for "quantile", the inverse CDF
d for "density", the probability mass function (PMF)
r for "random", a random variable having the specified distribution

In addition, we have suffixes for common random variables: **binom** (binomial), **pois** (Poisson), and **geom** (Geometric).

For instance, suppose that $X \sim Binomial(n = 10, p = 0.5)$, that is, X is a binomial random variable that counts the number of heads in 10 fair coin flips. We can calculate P(X=5), the probability of exactly 5 heads in 10 fair coin flips, as follows.

```
dbinom(x=5,size=10,prob=0.5)

## [1] 0.2460938
```

If we want the probability of at most five heads in 10 coin flips, $P(X \leq 5)$, we can use **pbinom**.

```
pbinom(q=5,size=10,prob=0.5)

## [1] 0.6230469
```

Now, suppose that $Y \sim Poisson(\lambda = 2)$ and we want to know $P(Y > 3)$. First, we note that $P(Y > 3) = 1 - P(Y \leq 3)$.

```
1-ppois(q=3,lambda=2)

## [1] 0.1428765
```

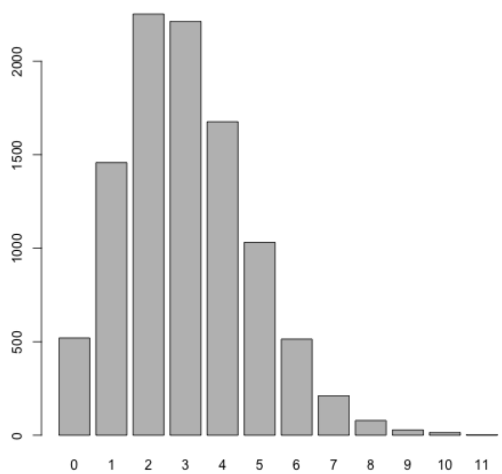
If we use the prefix **r**, we can generate simulated data according to a particular random variable. Let's generate 10,000 observations from a Poisson distribution with $\lambda = 3$, look at the first few values, and then calculate some summary statistics. This will require us to use the **rpois** function to generate random Poisson data.

```
rand.data = rpois(n=10000,lambda=3)
head(rand.data) # look at the first few values
```

```
## [1] 3 6 2 4 1 4

mean(rand.data) # calculate the mean

## [1] 3.0088

var(rand.data) # calculate the variance

## [1] 3.047827
```

From theory, we know that if $Y \sim Poisson(\lambda = 3)$, then $E[Y] = Var(Y) = \lambda = 3$. Our simulated results are pretty close to the theoretical values.

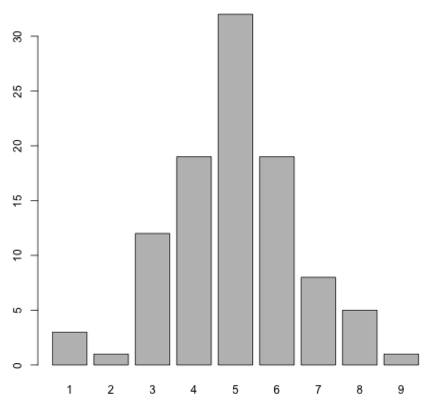
```
table(rand.data)

## rand.data
##    0    1    2    3    4    5    6    7    8    9   10   11
##  520 1458 2252 2213 1676 1031  514  211   79   28   15    3

barplot(table(rand.data))
```



How would you describe the distribution above? It seems to be asymmetric, skewed right (tail on the right), and unimodal.

Example: Let's simulate some data and calculate the probabilities in the sample. Suppose we flip 10 fair coins and we count the number of heads. Let X be the number of heads. We can model X as a binomial random variable with parameters n=10 (number of coin flips) and p=0.5 (fair coin). Let's simulate running this experiment 100 times.

```
set.seed(2020)
coins = rbinom(n=100,size=10,prob=0.5)
table(coins)

## coins
##  1  2  3  4  5  6  7  8  9
##  3  1 12 19 32 19  8  5  1

barplot(table(coins))
```



In our simulation, what was the probability that 5 heads occurred out of 10 flips? In the above table, we can see we got 5 heads on 32 of the 100 simulations, so $P(5) = 32/100 = 0.32$.

We could also calculate as follows:

```
p5 = sum(coins == 5)/100; p5

## [1] 0.32
```

What is the probability that we got more than 7 heads?

```
p = sum(coins > 7)/100; p

## [1] 0.06
```