Chapter 6

The Binomial Distribution

References: Pruim 2.3.2

In this chapter we turn our attention to an important family of discrete distributions that is important for its ability to describe real world problems.

6.1 Binomial random variable

A **binomial** experiment arises in a situation where a random process can be conceptualized as a sequence of smaller random processes called trials and

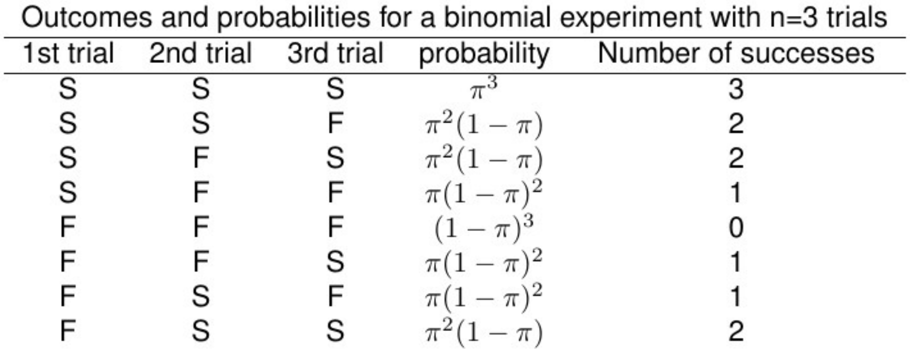
a. the number of trials (usually denoted by n) is specified in advance,

b. there are two outcomes (traditionally called success and failure) for each trial,

c. the probability of success (frequently denoted p or π) is the same in each trial, and

d. each trial is independent of the other trials.

The sample space for a binomial experiment consists of all the sequences of successes and failures that result across the n trials. As an example, the following table lists the set of 8 possible outcomes when $n = 3$ along with their associated probabilities using the assumption of independence across trials.



Outcomes and probabilities for a binomial experiment with n=3 trials

1st trial	2nd trial	3rd trial	probability	Number of successes
S	S	S	π^3	3
S	S	F	$\pi^2(1-\pi)$	2
S	F	S	$\pi^2(1-\pi)$	2
S	F	F	$\pi(1-\pi)^2$	1
F	F	F	$(1-\pi)^3$	0
F	F	S	$\pi(1-\pi)^2$	1
F	S	F	$\pi(1-\pi)^2$	1
F	S	S	$\pi^2(1-\pi)$	2

The number of successes in a binomial experiment is called a **binomial random variable**.

There are actually many different binomial distributions, one for each positive integer n and probability π. Collectively, we refer to these distributions as the binomial family. Members of the family are distinguished by the values of the **parameters** n and π, but are otherwise very similar.

Definition 6.1. Suppose that n independent trials, each of which results in a success with probability π and in a failure with probability $1 - \pi$, are to be performed. If X represents the number of successes that occur in the n trials, then X is said to be a binomial random variable with parameters (n, π).

We write $X \sim Binom(n, \pi)$. When $n = 1$, X is commonly referred to as a **Bernoulli** random variable.

Example 6.1. Consider the following random variables. Which ones are binomial?

a. X is the number of black marbles in a sample of 2 chosen randomly (with replacement) from two black (B_1, B_2) and one red (R).

b. X is the number of black marbles in a sample of 2 chosen randomly (without replacement) from two black (B_1, B_2) and one red (R).

c. Roll a die four times. We call a roll a “good” if the number rolled is greater than the roll number. So the first roll is good if it is a 2 or higher. The fourth roll is good if it is a 5 or a 6. Let X count the number of good rolls.

d. Independent trials consisting of the flipping of a fair coin are performed until a head is obtained. Let X denote the number of flips.

The tools we have developed in the previous chapters make derivation of a formula for the PMF a binomial random variable straightforward.

Theorem 6.1 (Binomial PMF)**.** *Let $X \sim Binom(n, \pi)$. Then the PMF of X is given by*

$$f(x) = P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, \quad x = 0, 1, 2, \dots, n$$

Proof. The sample space corresponding to the binomial experiment consists of all possible sequences of success and failure that that result from n independent trials:

$$\underset{\text{trial 1}}{\underline{S}} \times \underset{\text{trial 2}}{\underline{F}} \times \underset{\text{trial 3}}{\underline{F}} \cdots \times \underset{\text{trial n}}{\underline{S}} .$$

The number of successes x can equal any integer from 0 to n. By independence of the trials, any sequence with x successes will have probability $\pi^x(1 - \pi)^{n-x}$ and the number of such sequences is $\binom{n}{x}$ since we must select x of the n trials to be successful. So

$$\begin{aligned} P(X = x) &= \left(\begin{array}{c} \text{number of ways} \\ \text{to select } x \\ \text{of the } n \text{ trials} \end{array} \right) \cdot \left(\begin{array}{c} \text{probability of any} \\ \text{particular sequence of} \\ x \text{ successes and} \\ (n - x) \text{ failures} \end{array} \right), \\ &= \binom{n}{x} \pi^x (1 - \pi)^{n-x}. \end{aligned}$$

□

Let's now use the binomial PMF to do a probability calculation.

Example 6.2. Free throw Freddy is a 80% shooter which means the probability he makes a shot is 0.8. Which is more likely, that he makes at least 9 out of 10 shots or at least 18 out of 20 shots? What assumption are you making?

R offers functions to calculate the binomial PMF, CDF as well as a function that will make random draws from a binomial distribution.

```
####  Functions for working with X ~ Binom(size, prob)   ########

dbinom(x=9, size=10, prob=0.8)        #returns P(X=x) (the PMF)
## [1] 0.2684355

pbinom(q=8, size=10, prob=0.8)        #returns P(X<=q) (the CDF)
## [1] 0.6241904

pbinom(q=17, size=20, prob=0.8, lower.tail=F)     #returns P(X > q)
## [1] 0.2060847
```

```
#smallest x such that P(X <= x) >= p
qbinom(p = 0.9, size = 20, prob = 0.8)
## [1] 18
# makes 10 random draws of X.
set.seed(9898)            #set random number seed for reproducibility
rbinom(n = 10, size = 20, prob = 0.8)
##  [1] 19 16 15 15 14 15 18 17 16 17
```

There are similar functions in R for many of the distributions we will encounter and they all follow a similar naming scheme. We simply replace `binom` with the R-name for a different distribution.

Returning to example 6.2, we can calculate the probability that he will make at least 90% of his shots using R as shown below.

```
pbinom( q = 8, size = 10, prob = 0.8, lower.tail = F)  #P(X >= 8)
```

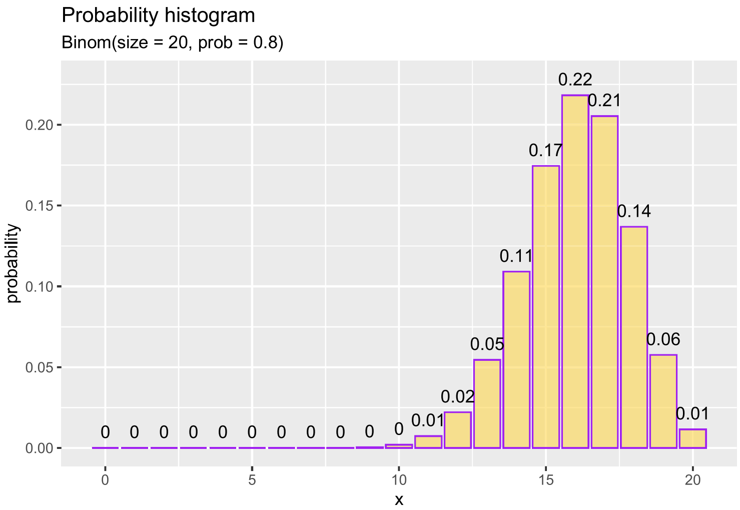
```
## [1] 0.3758096
```

```
pbinom(q = 17, size = 20, prob = 0.8, lower.tail = F)  #P(Y >= 18)
```

```
## [1] 0.2060847
```

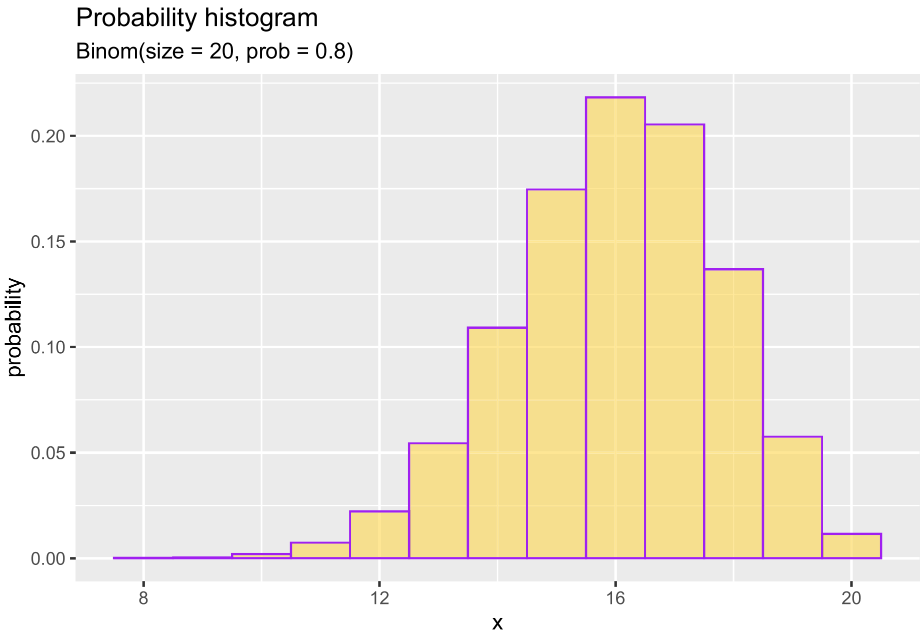
We can create a probability histogram for the binomial random variable using `ggplot` as we did in the previous chapter.

```
library(tidyverse)
bball <- tibble(
   x = 0:20,
   f = dbinom(x, size = 20, prob = 0.8)
)

# make probability histogram

ggplot(data = bball,
       mapping = aes(x=x, y=f) ) +
  geom_col(fill = "gold", color = "purple", alpha = 0.5) +
  geom_text(   #requires x, y, label
       mapping = aes(label = round(f,2),
                     y = f + 0.01) )+
  labs(x = "x",
       y = "probability",
       title = "Probability histogram",
       subtitle = "Binom(size = 20, prob = 0.8)")
```



The function `gf_dist()` from the **fastR2** package provides a nice canned alternative for creating the histogram layer for displaying a probability distribution corresponding to a named distribution. For example, we could use it to recreate the plot we made earlier as shown below.

```r
library(fastR2)
gf_dist( dist = "binom",
         kind = "histogram",
         binwidth = 1,
         fill = "gold",
         color = "purple",
         params = list(size = 20, prob = 0.8) ) +
labs(x="x",
       y="probability",
       title = "Probability histogram",
       subtitle = "Binom(size = 20, prob = 0.8)")
```



6.2 Practice Problems

1. Let $X \sim binomial(100, \frac{1}{3})$. Calculate using R.

   a. $f(40)$
   b. $F(40)$
   c. $P(30 \leq X \leq 40)$
   d. $P(X > 27)$

2. Suppose that a symmetrical six-sided die is rolled 20 independent times, and each time we record whether or not the event $\{2, 3, 5, 6\}$ has occurred.

   a. What is the distribution of the number of times this event occurs in 20 rolls?
   b. Calculate the probability that this event occurs 10 times.

3. It is known that diskettes produced by a certain company will be defective with probability 0.01, independently of each other. The company sells the diskettes in packages of 10 and offers a money back guarantee that at most 1 of the 10 diskettes is defective.
   What proportion of packages sold must the company replace?