# Introduction to Random Variables

*Chapter 3, Lab 1: Solutions*

*OpenIntro Biostatistics*

**Topics**

- Distributions of random variables
- Mean and variance of a random variable
- Binomial distribution

A random variable numerically summarizes the possible outcomes of a random experiment. Formally, a random variable assigns numerical values to the outcomes of a random phenomenon. This lab introduces random variables by exploring the clinical trial example described at the beginning of Chapter 3 and by discussing the binomial distribution.

The material in this lab corresponds to Sections 3.1 and 3.2 of *OpenIntro Biostatistics*.

### Clinical Trial Simulation

1. Suppose that a clinical trial to test a new drug will be conducted on 8 patients, in which the probability of a good response to the drug is thought to be 0.15. The following code simulates the trial with 500 replicates.

```
#define parameters
number.patients = 8
response.prob = 0.15
number.replicates = 500

#create empty vectors to store results
number.responses.replicate = vector("numeric", number.patients)
number.responses = vector("numeric", number.replicates)

#set the seed for a pseudo-random sample
set.seed(2018)

#simulate the trials
for(k in 1:number.replicates){

  number.responses.replicate = sample(c(0,1), size = number.patients,
                                      prob = c(1 - response.prob, response.prob),
                                      replace = TRUE)

  number.responses[k] = sum(number.responses.replicate)

}
```

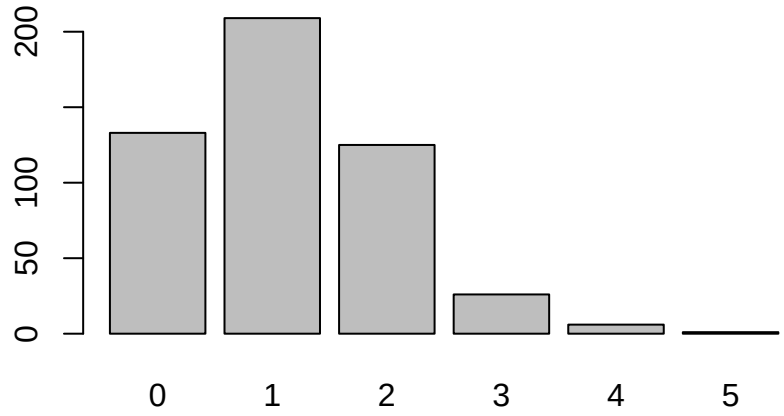
a) Run the simulation and view the results (*Hint*: make a plot). Describe the distribution of the number of good responses to the drug.

The distribution is right-skewed; there are very few replicates where more than 3 patients respond well to the drug. In most of the replicates, 1 patient out of 8 responds well.

```
table(number.responses)
```

```
## number.responses
##   0   1   2   3   4   5
## 133 209 125  26   6   1
```

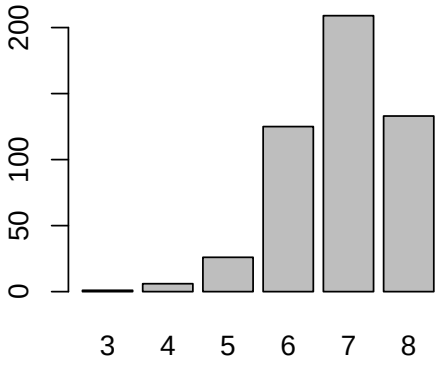
```
barplot(table(number.responses))
```

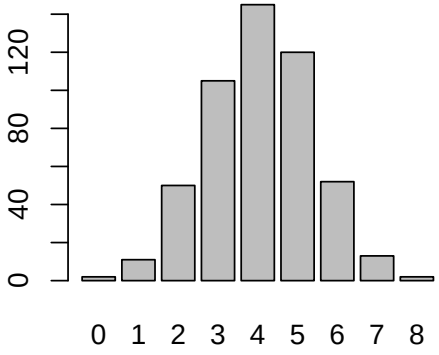



b) What value(s) for response probability would produce a left-skewed distribution? What value(s) would produce a symmetric distribution?

A response probability close to 1, such as 0.85, produces a left-skewed distribution; in this case, almost all of the 8 patients would be expected to respond well to the drug.

A response probability near 0.5 produces a symmetric graph. In this case, typically half of the patients would be expected to respond well to the drug.




c) Based on the results of the simulation, estimate the probability that 0 patients respond well to the new drug.

The probability that 0 patients respond well is 133/500 = 0.266.

```
sum(number.responses == 0)/number.replicates
```

```
## [1] 0.266
```

d) Let $X$ be a random variable defined as the number of patients who respond well to the experimental drug. Based on the results of the simulation, …

i. Construct a probability distribution for $X$.

$X$ can take on values 0 through 8, inclusive.

| $x_i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| $P(X = x_i)$ | 0.266 | 0.418 | 0.250 | 0.052 | 0.012 | 0.002 | 0 | 0 | 0 | = 1.00 |

ii. Calculate $E(X)$, where

$$E(X) = x_1 P(X = x_1) + \cdots + x_k P(X = x_k) = \sum_{i=1}^{k} x_i P(X = x_i)$$

$$\begin{aligned} E(X) &= \sum_{i=1}^{k} x_i P(X = x_i) \\ &= (x_1)P(X = x_1) + (x_2)P(X = x_2) + \cdots + (x_k)P(X = x_k) \\ &= (0)(0.266) + (1)(0.418) + (2)(0.250) + (3)(0.052) \\ &\quad + (4)(0.012) + (5)(0.002) + (6)(0) + (7)(0) + (8)(0) \\ &= 1.132 \end{aligned}$$

iii. Calculate $\text{Var}(X)$, where

$$\text{Var}(X) = (x_1 - \mu)^2 P(X = x_1) + \cdots + (x_k - \mu)^2 P(X = x_k) = \sum_{j=1}^{k} (x_j - \mu)^2 P(X = x_j)$$

$$\begin{aligned}
\text{Var}(X) &= \sum_{j=1}^{k} (x_j - \mu)^2 P(X = x_j) \\
&= (x_1 - \mu)^2 P(X = x_1) + \cdots + (x_k - \mu)^2 P(X = x_k) \\
&= (0 - 1.132)^2(0.266) + (1 - 1.132)^2(0.418) + (2 - 1.132)^2(0.250) + (3 - 1.132)^2(0.052) \\
&\quad + (4 - 1.132)^2(0.012) + (5 - 1.132)^2(0.002) + (6 - 1.132)^2(0) + (7 - 1.132)^2(0) \\
&\quad (8 - 1.132)^2(0) \\
&= 0.847
\end{aligned}$$

```r
#using r as a calculator
prop.table(table(number.responses))  #print probability distribution table
```

```
## number.responses
##     0     1     2     3     4     5
## 0.266 0.418 0.250 0.052 0.012 0.002
```

```r
#calculate E(X)
x.i = 0:8
prob.x.i = prop.table(table(number.responses))
e.x = sum(x.i[1:6]*prob.x.i)
e.x
```

```
## [1] 1.132
```

Note that when using R as a calculator to calculate $E(X)$ based on the simulation, it is necessary to account for unobserved values of $x_i$. Since there were no instances of 6, 7, or 8 good responses observed in the simulation, the calculation should specifically be made only from the first 6 values of the `x.i` vector (0, 1, 2, 3, 4, 5).

**Binomial Distribution**

Let $X$ represent the number of successes in a series of independent Bernoulli trials. Suppose the probability of a single trial being a success is $p$. The probability of observing exactly $k$ successes in $n$ independent trials is given by

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

For a binomial random variable, $E(X) = np$ and $\text{Var}(X) = np(1-p)$.

Binomial probabilities are calculated in R with the use of `dbinom` and `pbinom`.

The following code shows how to calculate $P(X = 5)$, $P(X \leq 5)$, and $P(X > 5)$ for $X \sim \text{Bin}(10, 0.35)$.

```
#probability X equals 5
dbinom(5, 10, 0.35)
```

```
## [1] 0.1535704
```

```
#probability X is less than or equal to 5
pbinom(5, 10, 0.35)
```

```
## [1] 0.9050659
```

```
#probability X is greater than 5
pbinom(5, 10, 0.35, lower.tail = FALSE)
```

```
## [1] 0.09493408
```

2. Why is the value returned by `pbinom(5, 10, 0.35)` greater than the value from `dbinom(5, 10, 0.35)`?

   The value returned by `pbinom(5, 10, 0.35)` represents $P(X \leq 5)$, while `dbinom(5, 10, 0.35)` returns $P(X = 5)$. $P(X \leq 5)$ is equivalent to $P(X = 0) + P(X = 1) + \cdots + P(X = 5)$; since the probability of 0, 1, 2, 3, or 4 events occurring is nonzero, $P(X \leq 5)$ is greater than $P(X = 5)$.

3. The hypothetical clinical trial discussed in Question 1 can be modeled with a binomial distribution.

   a) Confirm that the clinical trial satisfies the conditions for a binomial experiment.

      There is a fixed number of trials, $n$, at 8 patients; it is reasonable to assume that the response of each patient is independent of the others. Each outcome can be classified as either success or failure. The probability of a good response, $p$, is assumed to be the same for each patient.

   b) Calculate $E(X)$ and $\text{Var}(X)$ using the formulas specific to the binomial distribution; compare the results to the answers from part d) of Question 1. Explain any observed discrepancies.

      According to the binomial formulas, $E(X) = np = (8)(0.15) = 1.2$ and $\text{Var}(X) = np(1-p) = (8)(0.15)(1-0.15) = 1.02$. The values are a bit different from the answers in part d), with a lower mean and higher variance. The probability distribution from simulation will approach the theoretical distribution as the number of replicates increases, and the formulas will give identical results.

4. Approximately 12,500 stocks of *Drosophila melanogaster* flies are kept at the Bloomingdale *Drosophila* Stock Center for research purposes. A 2006 study examined how many stocks were infected with Wolbachia, an intracellular microbe that can manipulate host reproduction for its own benefit. About 30% of stocks were identified as infected. Researchers working with infected stocks should be cautious of the potential confounding effects that Wolbachia infection may have on experiments.

   Consider a random sample of 250 stocks. Let $X$ represent the number of infected stocks in the sample.

   a) Calculate the probability that exactly 60 stocks are infected.

      The probability that exactly 60 stocks are infected is $P(X = 60) = 0.0063$.

```
dbinom(60, 250, 0.30)
```

```
## [1] 0.006301219
```

   b) Calculate the probability that at most 60 stocks are infected.

      The probability that at most 60 stocks are infected is $P(X \leq 60) = 0.021$.

```
pbinom(60, 250, 0.30)
```

```
## [1] 0.02103864
```

   c) Calculate the probability that at least 80 stocks are infected.

      The probability that at least 80 stocks are infected is $P(X \geq 80) = 1 - P(X \leq 79) = 0.223$.

```
pbinom(79, 250, 0.30, lower.tail = FALSE)
```

```
## [1] 0.2654606
```

   d) Assume that a researcher will use all of the stocks sampled for an experiment. If the researcher wants to be sure that no more than 40% of the stocks used for an experiment are infected, does it seem reasonable to take a random sample of 250 stocks?

      40% of 250 stocks is 100. What is the probability that out of 250 stocks, no more than 100 stocks are infected? This is $P(X \leq 100) \approx 1$. It seems reasonable to expect that from a random sample of 250, no more than 40% are infected.

```
pbinom(100, 250, 0.30)
```

```
## [1] 0.9997024
```

   e) Demonstrate how parts a) through d) can be approached through simulation.

```
#define parameters
sample.size = 250
prob.infection = 0.30
number.replicates = 1000

#create empty vectors to store results
number.infected.replicate = vector("numeric", sample.size)
number.infected = vector("numeric", number.replicates)
```

```r
#set the seed for a pseudo-random sample
set.seed(2018)

#simulate the samples
for(k in 1:number.replicates){

  number.infected.replicate = sample(c(0,1), size = sample.size,
                                     prob = c(1 - prob.infection, prob.infection),
                                     replace = TRUE)

  number.infected[k] = sum(number.infected.replicate)

}

#part a)
sum(number.infected == 60)/number.replicates
```

```
## [1] 0.006
```

```r
#part b)
sum(number.infected <= 60)/number.replicates
```

```
## [1] 0.023
```

```r
#part c)
sum(number.infected >= 80)/number.replicates
```

```
## [1] 0.249
```

```r
#part d)
sum(number.infected <= 100)/number.replicates
```

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

5. The National Vaccine Information Center estimates that 90% of Americans have had chickenpox by the time they reach adulthood.

   a) Calculate the probability that exactly 97 out of 100 randomly sampled American adults had chickenpox during childhood.

      Let $X$ represent the number of individuals in a sample who had chickenpox during childhood. $P(X = 97) = 0.006$; the probability that exactly 97 out of 100 randomly sampled American adults had chickenpox during childhood is 0.006.

```r
dbinom(97, size = 100, prob = 0.90)
```

```
## [1] 0.005891602
```

   b) Calculate the probability that exactly 3 out of a new sample of 100 American adults have not had chickenpox in their childhood.

      The event that exactly 3 out of 100 adults did not have chickenpox during childhood

is equivalent to the event that exactly 97 out of 100 did have chickenpox during childhood; thus, the probability is 0.006 from calculations in part a).

c) What is the probability that at least 1 out of 10 randomly sampled American adult has had chickenpox?

$P(X \geq 1) = P(X > 0)$; the probability that at least 1 out of 10 randomly sampled American adults have had chickenpox is essentially 1.

```
1 - dbinom(0, size = 10, prob = 0.90)
```

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

```
pbinom(0, size = 10, prob = 0.90, lower.tail = FALSE)
```

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

d) What is the probability that at most 3 out of 10 randomly sampled American adults have not had chickenpox?

$P(X \leq 3)$, where the probability of not having had chickenpox during childhood is 1 - 0.90 = 0.10; the probability that at most 3 out of 10 randomly sampled American adults have not had chickenpox is 0.99.

```
pbinom(3, size = 10, prob = 0.10)
```

```
## [1] 0.9872048
```

e) Suppose that the adults were sampled from the same city. Would it be appropriate to use the binomial distribution to calculate the probabilities from parts a) through d)?

It might not be appropriate to use the binomial distribution. If the city is small enough that 10 or 100 people may have come into contact with each other, then there is potentially some dependence between each person's contracting chickenpox as a child. In contrast, it is reasonable to assume independence if 10 or 100 adults are sampled from the entire United States.