

Introduction to Random Variables

Chapter 3, Lab 1

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.
- b) What value(s) for response probability would produce a left-skewed distribution? What value(s) would produce a symmetric distribution?
- c) Based on the results of the simulation, estimate the probability that 0 patients respond well to the new drug.
- 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 .
 - ii. Calculate $E(X)$, where

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

- iii. Calculate $Var(X)$, where

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

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 $Var(X) = np(1 - p)$.

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)`?
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.
 - b) Calculate $E(X)$ and $Var(X)$ using the formulas specific to the binomial distribution; compare the results to the answers from part d) of Question 1. Explain any observed discrepancies.
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.
 - b) Calculate the probability that at most 60 stocks are infected.
 - c) Calculate the probability that at least 80 stocks are infected.
 - d) If a researcher wants to make 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?
 - e) Demonstrate how parts a) through d) can be approached through simulation.
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.
 - b) Calculate the probability that exactly 3 out of a new sample of 100 American adults have not had chickenpox in their childhood.
 - c) What is the probability that at least 1 out of 10 randomly sampled American adult have had chickenpox?
 - d) What is the probability that at most 3 out of 10 randomly sampled American adults have not had chickenpox?
 - 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)?