Conditional Probability

Chapter 2, Lab 2

OpenIntro Biostatistics

Topics

- Definition of conditional probability
- Simulation

The previous lab demonstrated the use of `for` loops to simulate many repetitions of an experiment. This lab illustrates the use of `if` statements, another element of programming that is helpful when estimating probability via simulation.

The material in this lab corresponds to Sections 2.2.1 - 2.2.4 of *OpenIntro Biostatistics*.

Counting successes with `if` statements

A bag contains 3 red and 3 white balls. Two balls are drawn from the bag, one at a time; the first ball is not replaced before the second ball is drawn.

1. What is the probability of drawing a white ball on the first pick and a red on the second?

   Run the following code to simulate the results for 10 sets of two draws from the bag, where red and white balls are represented by `R` and `W`, respectively.

```
#define parameters
balls = rep(c("R", "W"), c(3,3))
number.draws = 2
replicates = 10

#create empty vector to store results
successes = vector("numeric", replicates)

#set the seed for a pseudo-random sample
set.seed(5011)

#simulate the draws
for(k in 1:replicates){

  draw = sample(balls, size = number.draws, replace = FALSE)

  if(draw[1] == "W" & draw[2] == "R"){
successes[k] = 1
  }

}
```

```
#view the results
successes
table(successes)
```

a) The `rep(x, times)` command is a generic way to replicate elements of a vector `x` a certain number of times. Describe what the vector `balls` contains, and explain how the code to create the vector could be modified for a bag that contains 5 red balls and 2 white balls.

b) Explain the code used to generate `draw`.

c) An `if` statement has the basic structure `if( condition ) { statement }`; if the condition is satisfied, then the statement will be carried out. The `if` statement in the loop records when a "success" occurs; if a particular replicate k is considered a success, then a `1` is recorded as the k^{th} element of the vector `successes`.

Examine the condition in the `if` statement and explain how the condition specifies when a success occurs. What is considered a success, in the context of this problem?

d) For these 10 replicates, when did a set of two draws result in a white ball on the first pick and a red on the second?

e) Set the number of replicates to 10,000 and re-run the simulation. What is the estimated probability of drawing a white ball on the first pick and a red on the second?

f) Using algebraic methods, calculate the probability of drawing a white ball on the first pick and a red on the second. Confirm that the answer matches the one from part e).

2. What is the probability of drawing exactly one red ball?

   a) Using algebraic methods, calculate the probability of drawing exactly one red ball.

   b) Using simulation, estimate the probability of drawing exactly one red ball. Hint: remember that the logical operator for "or" is the `|` symbol.

   c) Bonus: How can the condition in the `if` statement used in part b) be written more simply, so as to directly express the event of drawing exactly one red ball? Hint: use the `sum()` function.

Simulating a population with `if` statements

3. In the United States population, approximately 20% of men and 3% of women are taller than 6 feet (72 inches). Let F be the event that a person is female and T be the event that a person is taller than 6 feet. Assume that the ratio of males to females in the population is 1:1.

   Consider the following questions.

   - What is the probability that the next person walking through the door is female and taller than 6 feet?
   - What is the probability that the next person walking through the door is taller than 6 feet?

   One approach to estimating the probabilities is to simulate a large population based on the known information. Essentially, the idea is to randomly assign sex and height status, then count the number of individuals that represent a "successful" outcome.

   The code for simulating a population of 10,000 individuals is shown below, with some missing pieces. Examine the code, then answer the following questions.

```
#define parameters
p.female =
p.tall.if.female =
p.tall.if.male =
population.size = 10000

#create empty vectors to store results
sex = vector("numeric", population.size)
tall = vector("numeric", population.size)

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

#assign sex
sex = sample()

#assign tall or not
for (k in 1:population.size){

  if (sex[k] == 0) {
tall[k] = sample(, size = 1)
  }

  if (sex[k] == 1) {
tall[k] = sample(, size = 1)
  }
}

#view results
```

```
addmargins(table(sex, tall))
```

a) The first three parameters in the code refer to $P(F)$, $P(T|F)$, and $P(T|M)$, respectively. Identify their values and enter in the parameters.

b) The results will be stored in two vectors, `sex` and `tall`. Write the code for filling in the `sex` vector. Let `0` represent males and `1` represent females.

c) Explain why filling in the `tall` vector requires the use of `if` statements in addition to the `sample()` command.

d) Write the code for filling in the `tall` vector. Let `0` represent individuals who are not taller than 6 feet and `1` represent individuals who are taller than 6 feet.

e) Run the simulation and estimate the desired probabilities.

f) Using algebraic methods, calculate the desired probabilities.

4. Suppose a disease is caused by a single gene, with alleles A and a; the alleles have frequency 0.90 and 0.10, respectively.

   a) Calculate the genotype frequencies for AA, Aa, and aa. Assume independent inheritance of alleles and independent mating.

   b) Suppose the disease is not fully penetrant, so that the probability of developing the disease is 0.8 for genotype AA, 0.4 for genotype Aa, and 0.1 for genotype aa. Simulate a population of 10,000 individuals, recording their genotype and disease status.

      i. What is the prevalence (overall probability) of disease in the population?

      ii. Given that an individual is known to have the disease, what is the probability they are genotype AA?

      iii. Using algebraic methods, confirm your answers to parts i. and ii.