Week 3 Lab

Akanksha Tipparti

09-12-2024

Topics

- Defining probability
- Rules of probability
- Simulation

Probabilities for events can be calculated via simulation by simply repeating an experiment a large number of times and then counting the number of times the event of interest occurs. According to the Law of Large Numbers, as the number of repetitions increase, the proportion $\hat{p}_n$ of occurrences converges to the probability p of that event (where n is the number of repetitions).

Introduction

```
#define parameters
prob.heads = 0.6
number.tosses = 5

#simulate the coin tosses
outcomes = sample(c(0, 1), size = number.tosses,
                      prob = c(1 - prob.heads, prob.heads), replace = TRUE)

#view the results
table(outcomes)
```

```
## outcomes
## 0 1
## 1 4
```

```
#store the results as a single number
total.heads = sum(outcomes)
total.heads
```

```
## [1] 4
```

```r
#define parameters
prob.heads = 0.6
number.tosses = 5
number.replicates = 50

#create empty vector to store outcomes
outcomes = vector("numeric", number.replicates)

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

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

  outcomes.replicate = sample(c(0, 1), size = number.tosses,
                      prob = c(1 - prob.heads, prob.heads), replace = TRUE)

  outcomes[k] = sum(outcomes.replicate)
}

#view the results
outcomes
```

```
##  [1] 5 3 1 2 4 4 4 4 2 3 3 3 4 3 4 4 4 3 3 2 4 4 2 3 3 3 2 3 3 2 4 3 3 1 3 2 2 5
## [39] 0 2 3 3 3 4 3 4 4 3 5 4
```

```r
addmargins(table(outcomes))
```

```
## outcomes
##   0   1   2   3   4   5 Sum
##   1   2   9  20  15   3  50
```

```r
heads.3 = (outcomes == 3)
table(heads.3)
```

```
## heads.3
## FALSE  TRUE
##    30    20
```

```r
outcomes[4]
```

```
## [1] 2
```

```
## heads.3
## FALSE  TRUE
##  6479  3521
```

```r
prob.heads = 0.6

10*prob.heads^3*(1-prob.heads)^2
```

```
## [1] 0.3456
```

Mandatory Drug Testing

The simulation framework illustrated in the previous section can easily be adapted for other scenarios that may seem more complicated than coin tossing. The drug testing scenario examined in this section appears in *OI Biostat* as Example 2.30.

Mandatory drug testing in the workplace is common practice for certain professions, such as air traffic controllers and transportation workers. A false positive in a drug screening test occurs when the test incorrectly indicates that a screened person is an illegal drug user. Suppose a mandatory drug test has a false positive rate of 1.2% (i.e., has probability 0.012 of indicating that an employee is using illegal drugs when that is not the case). Given 150 employees who are in reality drug free, what is the probability that at least one will (falsely) test positive? Assume that the outcome of one drug test has no effect on the others.

```
prob.false.positive = 0.012
prob.true.negative = 1 - prob.false.positive

1 - prob.true.negative^150
```

```
## [1] 0.836491
```

```
#define parameters
prob.false.positive = 0.012
number.employees = 150

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

#simulate the tests
results = sample(c(0,1), size = number.employees,
                 prob = c(1 - prob.false.positive, prob.false.positive),
                 replace = TRUE)

#view the results
table(results)
```

```
## results
##   0   1
## 148   2
```

```
sum(results)
```

```
## [1] 2
```

```r
#define parameters
prob.false.positive = 0.012
number.employees = 150
number.replicates = 100000

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

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

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

  results.replicate = sample(c(0,1), size = number.employees,
                prob = c(1 - prob.false.positive, prob.false.positive),
                replace = TRUE)

  results[k] = sum(results.replicate)

}

#view the results
table(results)
```

```
## results
##     0     1     2     3     4     5     6     7     8     9    10
## 16282 29847 27015 16201  7183  2524   720   169    49     7     3
```

```r
at.least.1.pos = (results >= 1)
table(at.least.1.pos)
```

```
## at.least.1.pos
## FALSE  TRUE
## 16282 83718
```

Mammograms

5. The specificity of a diagnostic test refers to the probability that a test is negative in the absence of disease. Mammograms have a specificity of 95% for detecting breast cancer.

a) Define the relationship between the specificity of a test and the probability of a false positive.

```
#define parameters
specificity = 0.95
number.women = 50
number.replicates = 100000

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

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

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

  results.replicate = sample(c(0,1), size = number.women,
                             prob = c(specificity, 1 - specificity),
                             replace = TRUE)

  results[k] = sum(results.replicate)

}

#view the results
table(results)
```

```
## results
##     0     1     2     3     4     5     6     7     8     9    10    11
##  7858 20162 26036 21929 13560  6716  2568   855   237    66    11     2
```

```
at.most.1.pos = (results <= 1)
table(at.most.1.pos)
```

```
## at.most.1.pos
## FALSE  TRUE
## 71980 28020
```

```
#define parameters
specificity = 0.99
number.women = 50
```

```r
number.replicates = 100000

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

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

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

  results.replicate = sample(c(0,1), size = number.women,
                             prob = c(specificity, 1 - specificity),
                             replace = TRUE)

  results[k] = sum(results.replicate)

}

#view the results
table(results)
```

```
## results
##     0     1     2     3     4     5     6
## 60444 30595  7548  1249   150    12     2
```

```r
at.most.1.pos = (results <= 1)
table(at.most.1.pos)
```

```
## at.most.1.pos
## FALSE  TRUE
##  8961 91039
```

#Answer: When the specificity of the test increases, it means that fewer people without breast cancer will receive a false positive result. So, the test with 99% specificity is better because it significantly reduces the number of false positives, meaning fewer women without breast cancer are incorrectly told they might have it.