BDA - Assignment 1

Anonymous

Contents

Basic probability theory notation and terms

Probability describes the chance of an event happening.

Probability mass is a value ranging from 0 to 1 that describes the probability of a discrete event

The probability density of a range of values describes the probability that a continuous random variable will fall into that range.

A probability mass function is a function that gives the probability of a discrete event

A probability density function describes the probability that a continuous random variable will fall into a range of values.

A probability distribution is a function that gives the probabilities of different possible outcomes for an event.

A discrete probability distribution is a probability distribution for discrete values.

A continuous probability distribution is a probability distribution for continuous values.

A cumulative distribution function cdf(x) describes the probability that a random value will be less or equal to x.

Likelihood describes how well a statistical model fits a set of data.

Basic computer skills

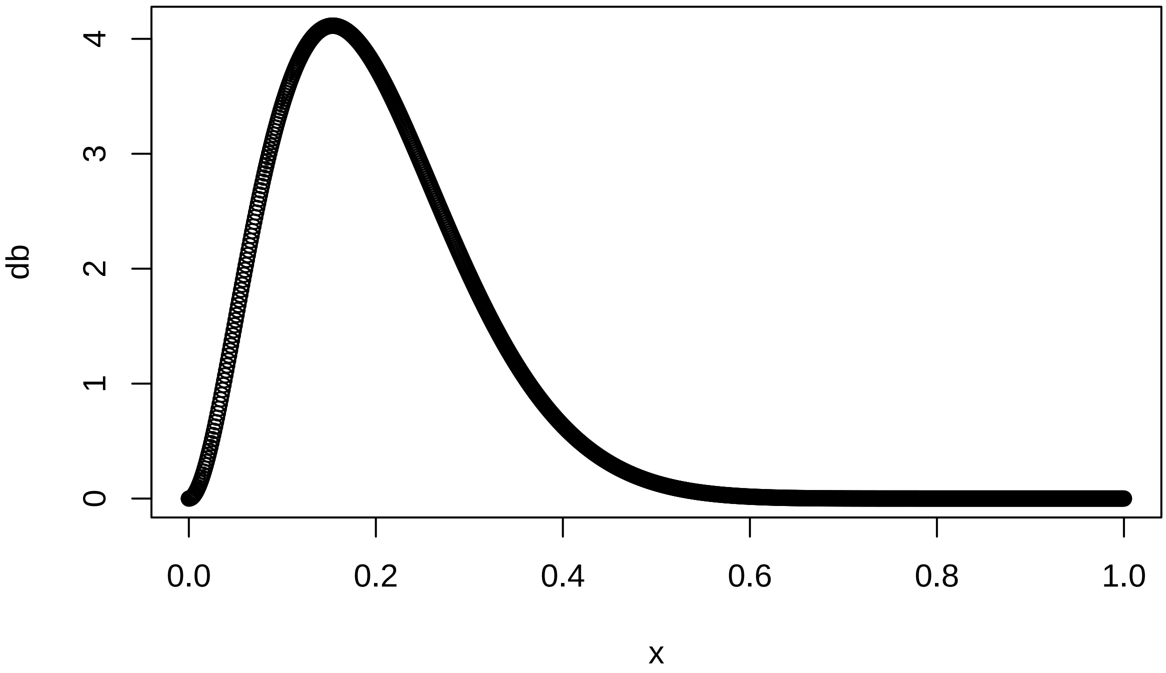
a)

Let's plot the density function of a Beta-distribution with $\mu = 0.2$ and $\sigma^2 = 0.01$.

```
x <- seq(0,1,length=1001)
mu = 0.2
sigma2 = 0.01

a = mu*(((mu*(1-mu))/sigma2) - 1)
b = (a*(1-mu))/mu

db = dbeta(x,a,b)
plot(x,db)
```

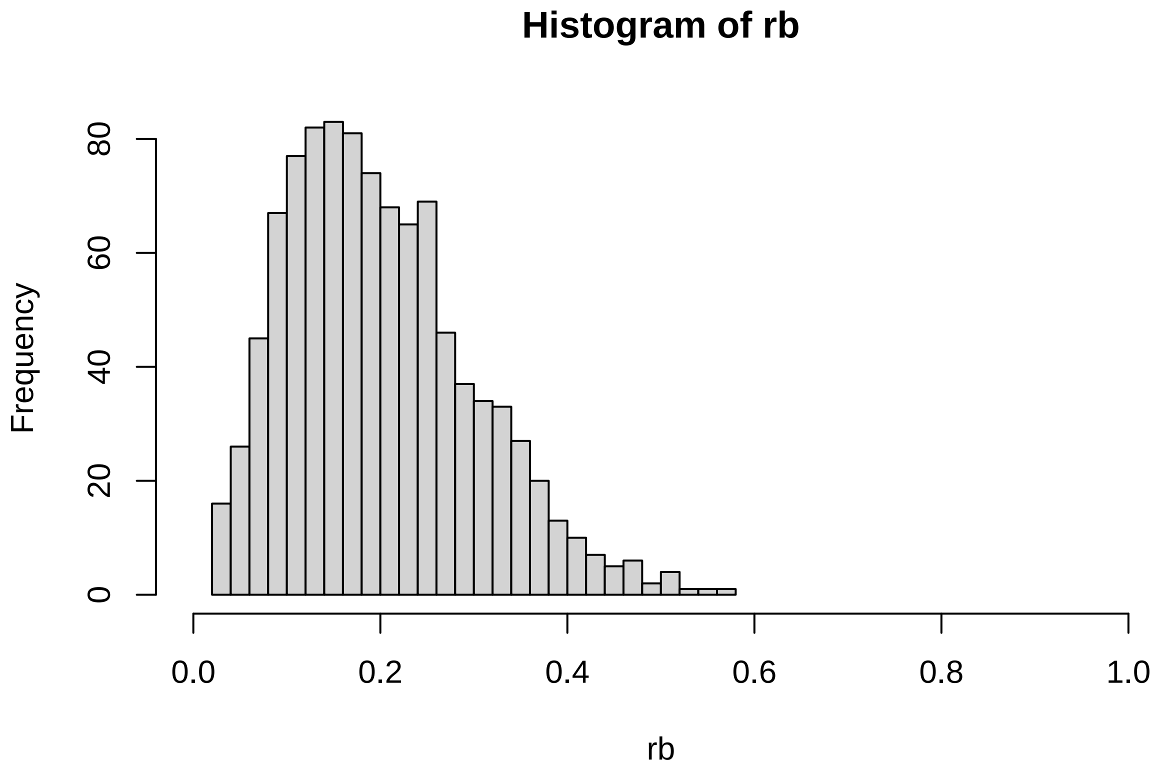


b)

Now we'll sample 1000 random numbers from the distribution and plot them as a histogram:

```
rb = rbeta(1000, a, b)

hist(rb, xlim=c(0,1), breaks=20)
```



The histogram matches the density function visually.

c)

The sample mean and variance are easy to compute with R.

```
sample_mean = mean(rb)
sample_variance = var(rb)

paste("Sample mean", sample_mean)
```

```
## [1] "Sample mean 0.200250451612957"
```

```
paste("Sample variance", sample_variance)
```

```
## [1] "Sample variance 0.00980457283007887"
```

```
paste("True mean", mu)
```

```
## [1] "True mean 0.2"
```

```
paste("True variance", sigma2)
```

```
## [1] "True variance 0.01"
```

The sample mean and variance match the true mean and variance very closely.

d)

Finally, to compute the central 95% probability interval of the distribution:

```
quantile(rb, probs=c(0.025, 0.975))
```

```
##       2.5%      97.5% 
## 0.04661606 0.42461454
```

The interval is [0.05, 0.42].

Bayes' theorem 1

To see how well the test would work in practice, we'll compute the false negative and positive rates using Bayes' formula.

```
test_true_positive = 0.98
test_true_negative = 0.96
test_false_negative = 1 - test_true_positive
test_false_positive = 1 - test_true_negative
cancer = 1/1000
no_cancer = 1 - cancer

# Apply Bayes' theorem to get the false negative/positive values

#p(has cancer|test result is negative)
false_negatives = (cancer * test_false_negative) /
  (cancer * test_false_negative + no_cancer * test_true_negative)

#p(does not have cancer|test result is positive)
false_positives = (no_cancer * test_false_positive) /
  (no_cancer * test_false_positive + cancer * test_true_positive)


false_negatives
```

```
## [1] 2.085375e-05
```

```
false_positives
```

```
## [1] 0.9760625
```

The test doesn't have a problem with false negatives. As lung cancer is rare, almost all (99.998%) of negatives are true negatives.

The number of people who don't have lung cancer is much larger than those who do. As a result, even with a false positive rate of 4%, most (97.6%) of positive results will be false positives. A positive result will be a false alarm for the vast majority of the time, so the test should not be used in its current form.

Bayes' theorem 2

a)

The function p_red computes the probability of picking a red ball from one of the boxes.

```
boxes = matrix(c(2,4,1,5,1,3), ncol=2, dimnames = list(c("A", "B", "C"), c("red", "white")))
boxes
```

```
##   red white
## A   2     5
## B   4     1
## C   1     3
```

```
p_red = function(boxes) {
  #Probabilities of selecting a box
  a = 0.4
  b = 0.1
  c = 0.5
  #Probability of picking a red ball from given box
  ar = a*boxes[1,1]/(boxes[1,1]+boxes[1,2])
  br = b*boxes[2,1]/(boxes[2,1]+boxes[2,2])
  cr = c*boxes[3,1]/(boxes[3,1]+boxes[3,2])
  #Total probability of picking a red ball
  ar+br+cr
}
p_red(boxes)
```

```
## [1] 0.3192857
```

The probability of picking a red ball is 31.93%.

b)

The function p_box calculates the most probable box using Bayes' rule.

```
p_box = function(boxes) {
  #Probabilities of selecting a box
  a=0.4
  b=0.1
  c=0.5
  #Ratio of red balls in a given box
  ar = boxes[1,1]/(boxes[1,1]+boxes[1,2])
  br = boxes[2,1]/(boxes[2,1]+boxes[2,2])
  cr = boxes[3,1]/(boxes[3,1]+boxes[3,2])
  #Probability that a red ball is from given box. P(ball is from box)/P()
  ax = (a*ar)/(a*ar + b*br + c*cr)
  bx = (b*br)/(a*ar + b*br + c*cr)
  cx = (c*cr)/(a*ar + b*br + c*cr)
  c(ax, bx, cx)

}
p_box(boxes)
```

```
## [1] 0.3579418 0.2505593 0.3914989
```

The ball came most probably from box C.

Bayes' theorem 3

The probability of Elvis being an identical twin is given by Bayes' rule We have to calculate the ratio of identical twins to all twins and take fraternal gender into account, so the formula comes out to $P(is\_identical) = \frac{identical\_prob}{\frac{fraternal\_prob}{2}+identical\_prob}$

```
p_identical_twin = function(fraternal_prob, identical_prob) {
  #Probability that a twin is identical.
  #Divide fraternal by 2 to account for different sex twins.
  identical_prob/(fraternal_prob/2 + identical_prob)
}
p_identical_twin(1/150, 1/400)
```

```
## [1] 0.4285714
```

The probability that Elvis had an identical twin brother is ~ 43%.