BDA - Assignment 1

Anonymous

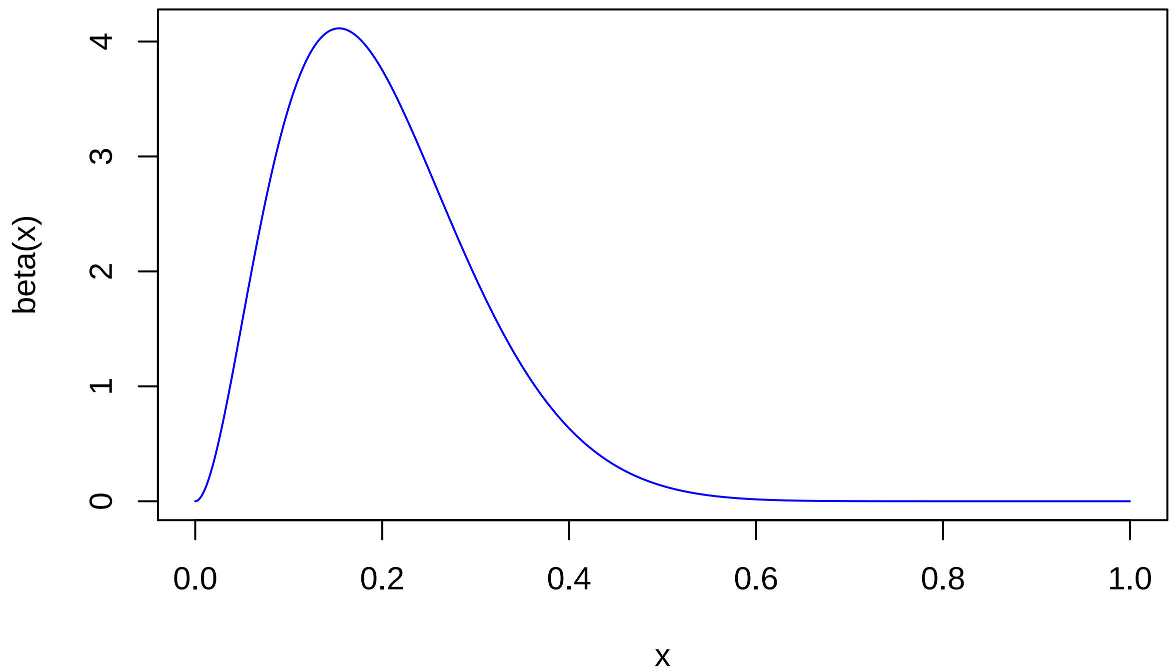
1 Basic probability theory notation and terms

- probability: a number that reflects the chance or likelihood that a particular event will occur.
- probability mass: probability of happening a discrete random variable is probability mass.
- probability density: probability distribution for a continuous random variable.
- probability mass function (pmf): a function that gives the probability that a discrete random variable is exactly equal to some value. It should sum to one over its inputs space.
- probability density function (pdf): a probability function to describe a continuous probability distribution. It should integrate to one over its input space.
- probability distribution: a function that describes all the densities/mass that a continious/discrete random variable can take in the sample space.
- discrete probability distribution: the probability of occurrence of each value of a discrete random variable that has countable values
- continuous probability distribution: describes the probabilities of the possible values of a continuous random variable. A continuous random variable is a random variable with a set of possible values (known as the range) that is infinite and uncountable.
- cumulative distribution function (cdf): the probability that a real-valued random variable X will take a value less than or equal to x.
- likelihood: the number that is the probability of some observed outcomes given a set of parameter values.

2 Basic computer skills

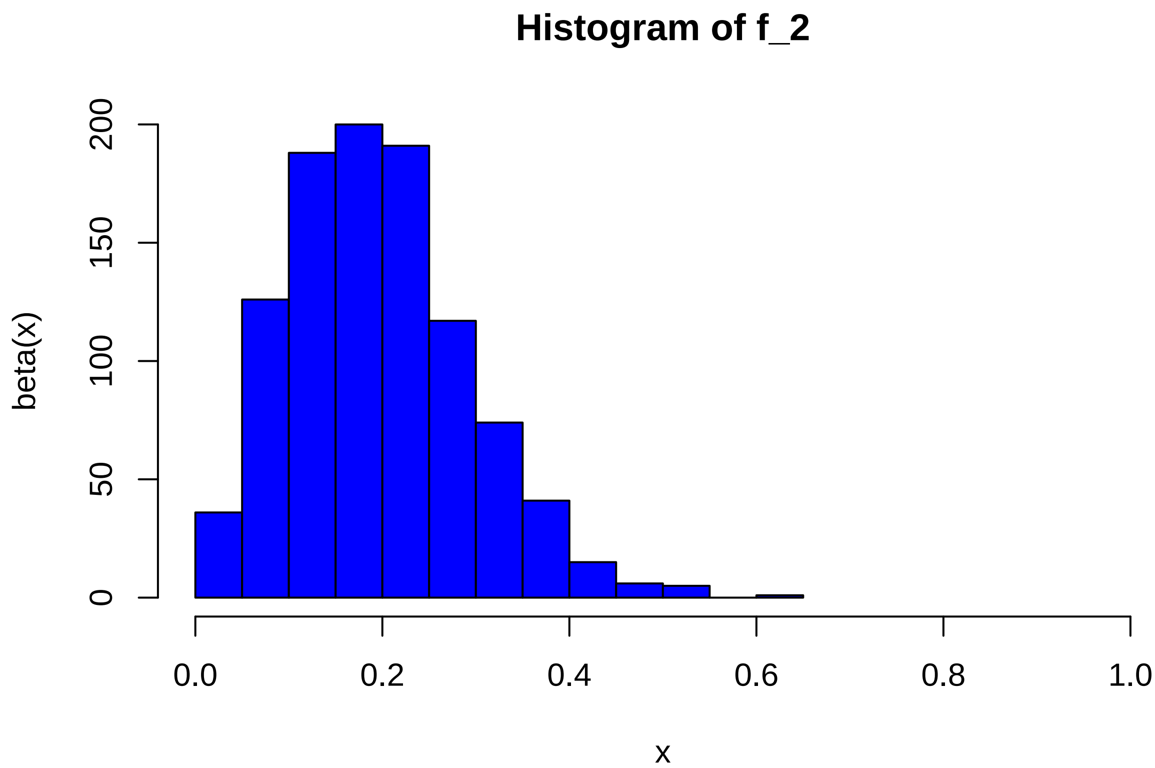
a)

```
x <- seq(0, 1, length = 1000)
mu = 0.2
sigma <- 0.1
alpha <- mu*((mu*(1-mu))/(sigma^2) - 1)
beta <- alpha*(1-mu)/mu
f_1 <- dbeta(x, alpha, beta)
plot(x, f_1, type='l',col = 'blue', xlab='x', ylab='beta(x)')
```



b)

```
f_2 <- rbeta(1000,alpha, beta)
hist(f_2, xlim=c(0,1), col = 'blue', xlab='x', ylab='beta(x)')
```



As we can see the plots are similar.

c)

```
sample_mean <- mean(f_2)
sample_var <- var(f_2)
threshold=0.01
if(mu-threshold <sample_mean  | sample_mean< mu+threshold) {
  cat('mean of sample matches the true mean with threshold ', threshold)
  } else {
    cat('mean of sample do not match the true mean with threshold ', threshold)
  }
```

```
## mean of sample matches the true mean with threshold  0.01
```

```
if(sigma^2-threshold<sample_var  | sample_var<sigma^2+threshold) {
  cat('\n variance of sample matches the true variance with threshold ',threshold)
  } else {
    cat('\n variance of sample do not match the true variance with threshold ',threshold)
  }
```

```
##
##  variance of sample matches the true variance with threshold  0.01
```

d)

```r
quantile(f_2, c(0.025,0.975))
```

```
##       2.5%      97.5% 
## 0.04370219 0.40267213
```

3 Bayes' theorem

```
h : hypothesis \quad  D : data\\
```

$$h = \{cancer, \neg cancer\}, \quad D = \{+, -\}$$

\

$$p(cancer) = 0.001, \quad p(\neg cancer) = 1 - 0.001 = 0.999$$

\

$$p(+|cancer) = 0.98, \quad p(-|cancer) = 1 - 0.98 = 0.02$$

\

$$p(-|\neg cancer) = 0.96, \quad p(+|\neg cancer) = 1 - 0.96 = 0.04$$

\

$$p(h|D) = \frac{p(D|h)p(h)}{p(D)}$$

\

$$p(D = -) = p(-|cancer)p(cancer) + p(-|\neg cancer)p(\neg cancer) = 0.02 \times 0.001 + 0.96 \times 0.999 = 0.95906$$

\

$$p(h = cancer|D = -) = \frac{p(-|cancer)p(cancer)}{p(-)} = \frac{0.02 \times 0.001}{0.95906} = 0.00002$$

\

$$p(D = +) = p(+|cancer)p(cancer) + p(+|\neg cancer)p(\neg cancer) = 0.98 \times 0.001 + 0.04 \times 0.999 = 0.04094$$

\

$$p(h = \neg cancer|D = +) = \frac{p(+|\neg cancer)p(\neg cancer)}{p(+)} = \frac{0.04 \times 0.999}{0.04094} = 0.976$$

\
According to this probability $p(h = \neg cancer|D = +)$ in 97.6% of time that the test result is positive, the subjects does not have cancer. Since positive results would be follwed up immidetly by expensive treatments, it is not cost effective.

4 Bayes' theorem

$$p(A) = 0.4, \quad p(B) = 0.1, \quad p(C) = 0.5$$

\

$$p(r|A) = \frac{2}{7}, \quad p(r|B) = \frac{4}{5}, \quad p(r|C) = \frac{1}{4}$$

\

$$p(r) = p(r|A)p(A) + p(r|B)p(B) + p(r|C)p(C) = \frac{2}{7} \times 0.4 + \frac{4}{5} \times 0.1 + \frac{1}{4} \times 0.5$$

```r
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
```

```r
prob_box = c(0.4, 0.1, 0.5)
p_r_box <- function(boxes){
  p = c()
  for(i in 1:nrow(boxes)){
    p[i] <- boxes[i,1]/sum(boxes[i,])
  }
  return(p)
}
p_r_box(boxes)
```

```
## [1] 0.2857143 0.8000000 0.2500000
```

```r
p_red <- function(boxes){
  p = 0
  for(i in 1:nrow(boxes)){
   p <- p + (p_r_box(boxes)[i]*prob_box[i])
  }
  return(p)
}

cat("the probability of picking the red ball = ", p_red(boxes),"\n")
```

```
## the probability of picking the red ball =  0.3192857
```

```r
p_box <- function(boxes){
  p = c()
  for(i in 1:nrow(boxes)){
    p[i] <- (p_r_box(boxes)[i]*prob_box[i])/p_red(boxes)
  }
  return(p)
}

cat("the probability of each boxes = ", p_box(boxes))
```

```
## the probability of each boxes =  0.3579418 0.2505593 0.3914989
```

5 Bayes' theorem

$p(i) = \frac{1}{400}$ and $p(f) = \frac{1}{150}$. When they are identical twins, both must be from the same gender (bb or gg). However in fraternal twins there are possibilities of bb, gg, bg, gb. Therefore $p(bb|i) = \frac{1}{2}$ and $p(bb|f) = \frac{1}{4}$.

$$p(i|bb) = \frac{p(bb|i)p(i)}{p(bb)}$$

\

$$p(bb) = p(bb|i)p(i) + p(bb|f)p(f)$$

```r
p_identical_twin <- function(fraternal_prob = 1/125, identical_prob = 1/300){
  p_identical_twin <- 0.5*identical_prob / (0.5*identical_prob + 0.25*fraternal_prob)
  return(p_identical_twin)
}

cat("the probability of Elvis being an identical twin is ", p_identical_twin(1/150, 1/400))
```

```
## the probability of Elvis being an identical twin is  0.4285714
```