BDA - Assignment 1

Anonymous

Table of Contents

Exercise 1 (Basic probability theory notation and terms)

Explain each of the following terms with one sentence:
- probability is a way of quantifying the belief of something being true or false.
- probability mass is the probability of observing a single discrete value point
- probability density is the probability of observing a single continuous value range
- probability mass function is the function of all combinations of the probability mass for all value points
- probability density function is the function whose integral defines the probability of all value ranges
- probability distribution is the distribution of probability of observing all possible values
- discrete probability distribution is the probability distribution of discrete variables
- continuous probability distribution is the probability distribution ofcontinuous variables
- cumulative distribution function is the function f(x) that returns the probability of observing values less than or equal to x
- likelihood is the probability of observing the data that has been observed assuming that the data came from a specific scenario

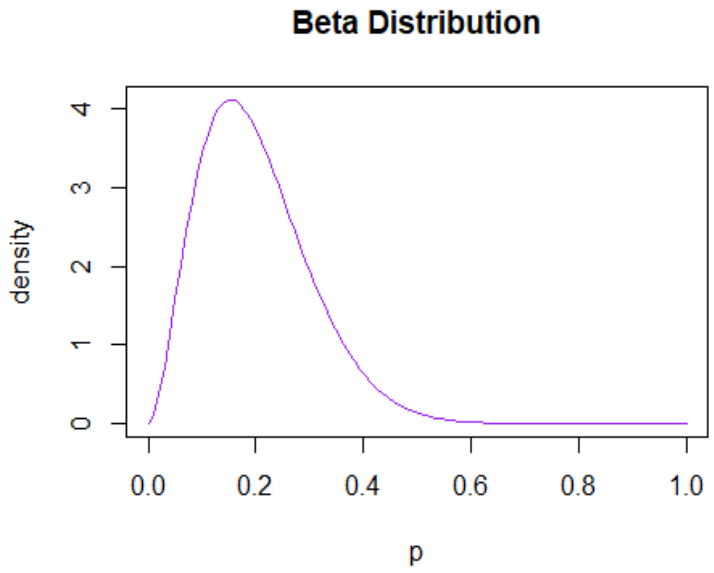
Exercise 2 (Basic computer skills)

a) Plot the density function of Beta-distribution, with mean = 0.2 and variance = 0.01. The parameters α and β of the Beta-distribution are related to the mean and variance
Hint! Useful R functions: seq(), plot() and dbeta()

Create plot of Beta distribution with mean = 0.2 and variance = 0.01

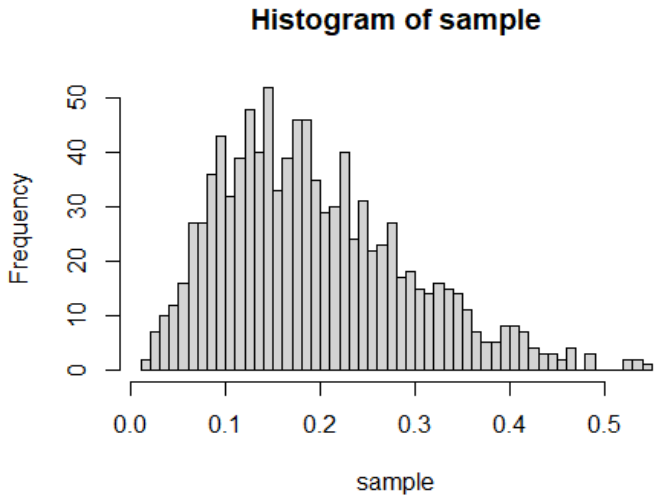
```
# define the range
p = seq(0, 1, length=100)
```

```r
mu = 0.2 # mean
sigma2 = 0.01 # variance
# alpha and beta calculated according to the formulas
alpha = mu * ( ((mu * (1 - mu))/sigma2) - 1)
beta = (alpha * (1 - mu))/mu
# Plotting the Beta distribution
plot(p, dbeta(p, alpha, beta), ylab='density',
     type ='l', col='purple', main='Beta Distribution')
```



b) Take a sample of 1000 random numbers from the above distribution and plot a histogram of the results. Compare visually to the density function.
Hint! Useful R functions: rbeta() and hist()

```r
sample = rbeta(1000, alpha, beta)
hist(sample, breaks=50)
```



The histogram of random sampled points closely resemble the Beta distribution above

c) Compute the sample mean and variance from the drawn sample. Verify that they match (roughly) to the true mean and variance of the distribution. Hint! Useful R functions: mean() and var()

The sample mean is:

```
mean(sample)

## [1] 0.1933702
```

The sample mean is close to the true mean = 0.2

The sample variance is:

```
var(sample)

## [1] 0.009717383
```

The sample variance is close to the true variance = 0.01

d) Estimate the central 95% probability interval of the distribution from the drawn samples. Hint! Useful R functions: quantile()

The central 95% probability interval is:

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

##       2.5%      97.5%
## 0.04442173 0.41821598
```

Exercise 3 (Bayes' theorem)

A group of researchers has designed a new inexpensive and painless test for detecting lung cancer. The test is intended to be an initial screening test for the population in general. A positive result (presence of lung cancer) from the test would be followed up immediately with medication, surgery or more extensive and expensive test. The researchers know from their studies the following facts:
- Test gives a positive result in 98% of the time when the test subject has lung cancer.
- Test gives a negative result in 96% of the time when the test subject does not have lung cancer.
- In general population approximately one person in 1000 has lung cancer.
The researchers are happy with these preliminary results (about 97% success rate), and wish to get the test to market as soon as possible. How would you advise them? Base your answer on Bayes' rule computations.

Let + sign be positive test result and - sign be negative test result. We know that:
- $P(+|have.cancer) = 0.98 => P(-|have.cancer) = 0.02$
- $P(-|no.cancer) = 0.96 => P(+|no.cancer) = 0.04$
- $P(have.cancer) = 0.001 => P(no.cancer) = 0.999$

Let the number of patients who truly have cancers in the test as A, and number of patients who do not as B. The success rate is 97% => 0.98A + 0.96B = 0.97(A+B) => A = B
=> There are equal number of patients having and not having the cancer in the test
From these identities, we can calculate the evidence:

- $P(+) = P(+|have.cancer) * P(have.cancer) + P(+|no.cancer) * P(no.cancer) = 0.98 * 0.001 + 0.04 * 0.999 = 0.04094$

- $P(-) = P(-|have.cancer) * P(have.cancer) + P(-|no.cancer) * P(no.cancer) = 0.02 * 0.001 + 0.96 * 0.999 = 0.95906$

The posteriors can be calculated as:
- $P(have.cancer|+) = \frac{P(+|have.cancer)P(have.cancer)}{P(+)} = \frac{0.98 \times 0.001}{0.04094} \approx 0.023937$
- $P(no.cancer|-) = \frac{P(-|no.cancer)P(no.cancer)}{P(-)} = \frac{0.96 \times 0.999}{0.95906} \approx 0.999979$
- $P(have.cancer|-) = 1 - P(no.cancer|-) = 0.000021$
- $P(no.cancer|+) = 1 - P(have.cancer|+) = 0.976063$

The false negative probability (cancer doesn't get detected) is 0.000021 (almost uncertainty) and the false positive probability (unnecessarily administered medication) is 0.976063 (almost certainty).

The purpose of the test is to detect cancer. Since the false negative probability is very low, this test is effective to detect cancer in the general population if it reports negative. However, the false positive probability is very high, suggesting that this test almost involves unnecessary medication when it reports positive of cancer. Therefore, it is advisable that these doctors should not use this new testing technique for the larger market and try to reduce the false positive probability for the testing methods.

Exercise 4 (Bayes' theorem)

We have three boxes, A, B, and C. There are
- 2 red balls and 5 white balls in the box A,
- 4 red balls and 1 white ball in the box B, and
- 1 red ball and 3 white balls in the box C.

Consider a random experiment in which one of the boxes is randomly selected and from that box, one ball is randomly picked up. After observing the color of the ball it is replaced in the box it came from. Suppose also that on average box A is selected 40% of the time and box B 10% of the time (i.e. P(A) = 0.4).

a) What is the probability of picking a red ball?
Probability of picking a red a ball:
$P(R) = P(A) * P(R|A) + P(B) * P(R|B) + P(C) * P(R|C) = 0.4 * 2/7 + 0.1 * 4/5 + 0.5 * 1/4 = 447/1400 \approx 0.319285$

b) If a red ball was picked, from which box it most probably came from?
According to the Bayesian Theorem:

$$P(A|R) = \frac{P(R|A) * P(A)}{P(R)} = \frac{2/7 * 0.4}{447/1400} = \frac{160}{447} \approx 0.357941$$
$$P(B|R) = \frac{P(R|B) * P(B)}{P(R)} = \frac{4/5 * 0.1}{447/1400} = \frac{160}{447} \approx 0.250559$$
$$P(C|R) = \frac{P(R|C) * P(C)}{P(R)} = \frac{1/4 * 0.5}{447/1400} = \frac{160}{447} \approx 0.391498$$

Therefore, the red ball mostly likely came from the box C

Implement two functions in R that computes the probabilities.

```
p_red <- function(boxes){
  probABC <- c(0.4, 0.1, 0.5)
  reds <- boxes[,1]
  whites <- boxes[,2]
  redRatio <- reds/(reds + whites)
  # The dot product between the red ratio and probability of choosing the box
  prob_red <- redRatio %*% probABC
  return(prob_red[1,1])
}

p_box <- function(boxes){
  probABC <- c(0.4, 0.1, 0.5)
  reds <- boxes[,1]
  whites <- boxes[,2]
  redRatio <- reds/(reds + whites)
  prob_red <- redRatio %*% probABC
  # Element-wise product
  unnormalizedBoxProb = redRatio * probABC
  prob_box <- unnormalizedBoxProb/prob_red[1,1]
  return(prob_box)
}

boxes <- matrix(c(2,2,1,5,5,1), ncol = 2, dimnames = list(c("A", "B", "C"), c
("red", "white")))
prob_red = p_red(boxes=boxes)
prob_box = p_box(boxes=boxes)
print("The p_red for exercise 3 is: ")

## [1] "The p_red for exercise 3 is: "

print(prob_red)

## [1] 0.3928571

print("The p_box for exercise 3 is: ")

## [1] "The p_box for exercise 3 is: "

print(prob_box)
```

```
##          A          B          C
## 0.29090909 0.07272727 0.63636364

boxes <- matrix(c(2,4,1,5,1,3), ncol = 2, dimnames = list(c("A", "B", "C"), c
("red", "white")))
answer = p_red(boxes=boxes)
prob_red = p_red(boxes=boxes)
prob_box = p_box(boxes=boxes)
print("The p_red for the example is: ")

## [1] "The p_red for the example is: "

print(prob_red)

## [1] 0.3192857

print("The p_box for the example is: ")

## [1] "The p_box for the example is: "

print(prob_box)

##         A         B         C
## 0.3579418 0.2505593 0.3914989

# To check your code/functions, just run
# mark_my_assignment()
```

Exercise 5 (Bayes' theorem)

Assume that on average fraternal twins (two fertilized eggs and then could be of different sex) occur once in 150 births and identical twins (single egg divides into two separate embryos, so both have the same sex) once in 400 births (Note! This is not the true value, see Exercise 1.6, page 28, in BDA3). American male singer-actor Elvis Presley (1935 - 1977) had a twin brother who died in birth. Assume that an equal number of boys and girls are born on average. What is the probability that Elvis was an identical twin? Show the steps how you derived the equations to compute that probability.

Because it is known that equal number of boys and girls are born on average, it should mean that the pairwise twins should also be balanced in gender. If the twin is fraternal, there are four different cases:
- MM: both babies are male
- FF: both babies are female
- MF: first baby is male while second baby is female
- FM: first baby is female while second baby is male

Each case occurs equally likely, so they all share 1/4 probability given that the twin is fraternal If the twin is identical, there are two different cases:
- MM: both babies are male
- FF: both babies are female

Each case occurs equally likely, so they all share $1/2$ probability given that the twin is identical

From these information, we can calculate the likelihood of being both males for twins: $P(MM) = P(identical) * P(MM|identical) + P(fraternal) * P(MM|fraternal) = 1/400 * 1/2 + 1/150 * 1/4 = 7/2400$ Let's call Elvis Presley as the first male baby and his brother the second male baby. From the Bayesian Theorem, the probability that Elvis was an identical twin is:

$$P(identical|MM) = \frac{P(MM|identical)P(identical)}{P(MM)} = \frac{1/2 \times 1/400}{7/2400} = 3/7 \approx 0.428571$$

Implement this as a function in R that computes the probability.

```
p_identical_twin <- function(fraternal_prob, identical_prob){
  p_mm = fraternal_prob * 1/4 + identical_prob * 1/2
  p_mmGivenIden = 1/2
  p_IdenGivenmm = (p_mmGivenIden * identical_prob)/p_mm
  return(p_IdenGivenmm)
}
p_IdenGivenmm <- p_identical_twin(fraternal_prob = 1/150, identical_prob = 1/
400)
print("The probability of exercise 5 is: ")

## [1] "The probability of exercise 5 is: "

print(p_IdenGivenmm)

## [1] 0.4285714

p_IdenGivenmm <- p_identical_twin(fraternal_prob = 1/125, identical_prob = 1/
300)
print("The probability of example 1 is: ")

## [1] "The probability of example 1 is: "

print(p_IdenGivenmm)

## [1] 0.4545455

p_IdenGivenmm <- p_identical_twin(fraternal_prob = 1/100, identical_prob = 1/
500)
print("The probability of example 2 is: ")

## [1] "The probability of example 2 is: "

print(p_IdenGivenmm)

## [1] 0.2857143

# mark_my_assignment()
```