BSDA: Assignment 1

Anonymous student

Contents

General Information to include

- **Time used for reading and self-study exercises**: ~10 hours.
- **Time used for the assignment**: ~8 hours
- **Good with assignment**: I liked the last two questions, they were a little challenging and a good refresher of some concepts. The last question was tricky and made me think which was cool.
- **Things to improve in the assignment**: It's always difficult to me as a non-native-english speaker questions like 1 where I need to define technical concepts in few words. I never know the amount of detail expected for each definition.

1. Basic probability theory notation and terms

- probability: is a measure between 0 and 1 of one's belief in the occurrence of an event.
- probability mass: the probability of a discrete random variable taking a specific value.
- probability density: the probability of a continuous random variable taking a value within a specific interval.
- probability mass function (pmf): is a function that assigns a probability to each value of a discrete random variable (r.v.), which represents the probability that the r.v. takes that value in an observation.

- probability density function (pdf): is a function that assigns a probability density to each value of a continuous r.v., since there are infinite values in an interval it provides a form of measuring probabilities over intervals.
- probability distribution: is a description of how probabilities are distributed in a random variable.
- discrete probability distribution: a formula, table or graph that provides $P(Y = y)\ \forall y$, Y is a r.v.
- continuous probability distribution: is a description of how probabilities are distributed in a continuous random variable.
- cumulative distribution function (cdf): the cumulative probability that a random variable is less than or equal to a given value.
- likelihood: measure of how probable it is that observed data are generated by a specific statistical model.
- aleatoric uncertainty: the inherent variability in data that cannot be eliminated or reduced with additional information.
- epistemic uncertainty: uncertainty associated with the lack of knowledge or complete information about a phenomenon, which can be reduced with additional information.

2. Basic computer skills

a) Plot the density function of Beta-distribution, with mean $\mu = 0.2$ and variance $\sigma^2 = 0.01$. The parameters α and β of the Beta-distribution are related to the mean and variance according to the following equations

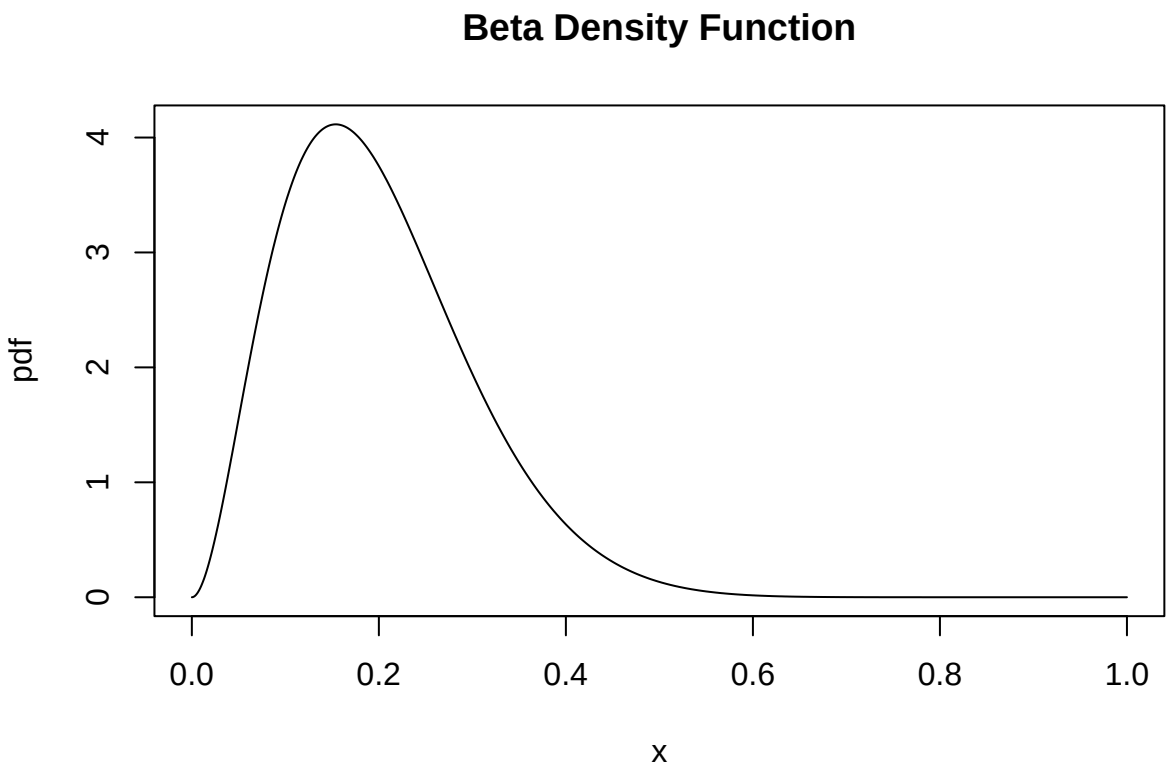
$$\alpha = \mu\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right), \quad \beta = \frac{\alpha(1-\mu)}{\mu}.$$

```r
# define parameters
mu <-  0.2
sigma_2 <- 0.01
a <- mu*(mu*(1-mu)/sigma_2 - 1)
b <- a*(1-mu)/mu

# sequence of values
x <- seq(0, 1, by = 0.001)

# density
beta <- dbeta(x, shape1 = a, shape2 = b)

# plot
plot(x, beta, type = "l", ylab = "pdf", main = "Beta Density Function")
```



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.

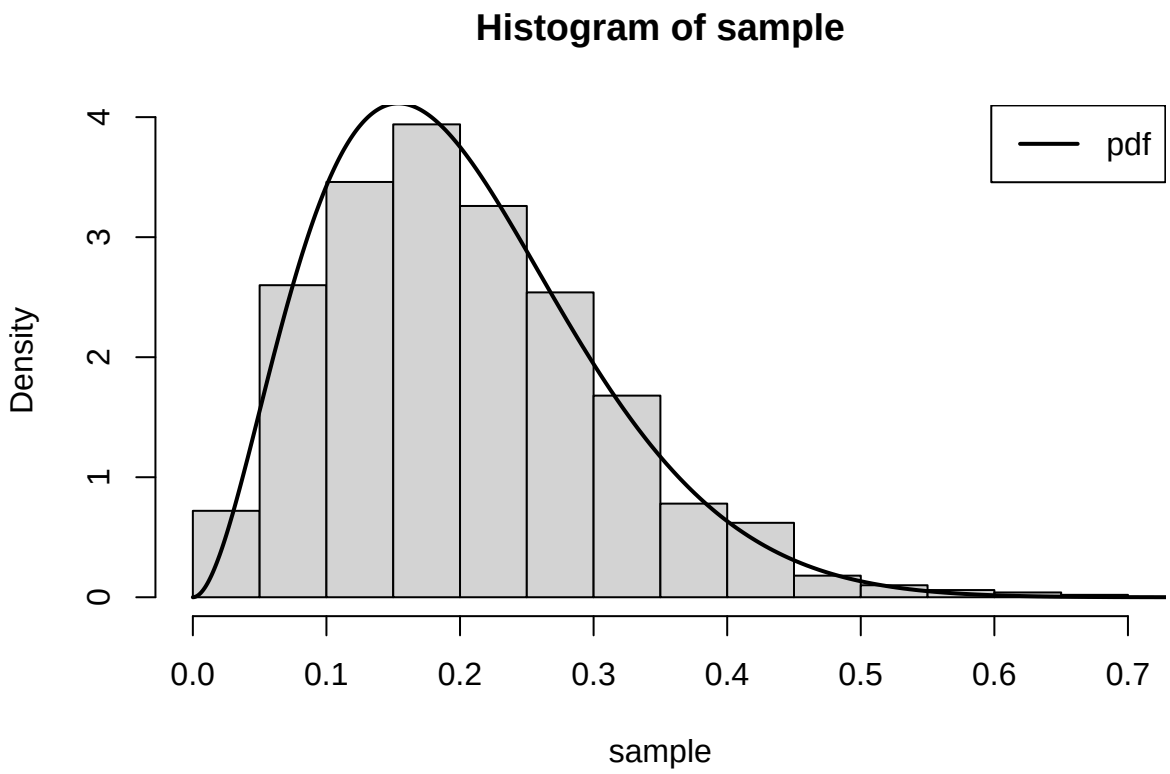
```r
# set seed to replicate always same plot
set.seed(100)

# random sample of 1,000 numbers
sample <- rbeta(1000, a, b)

# plot histogram with densities instead of frequencies
hist(sample, freq = FALSE)

# plot beta density function in same plot to compare
lines(x, beta, lwd = 2)

# add legend
legend("topright", legend = c("pdf"), lwd = 2)
```



The density function does not fit the sample perfectly, but it is a convenient approximation (specially since we already know the underlying distribution of the sample).

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.

If $X \sim B(\alpha, \beta)$ then:

$$E[X] = \frac{\alpha}{\alpha + \beta}$$

$$Var[X] = \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}$$

```
# sample mean and variance
s_mean <- mean(sample)
s_var <- var(sample)

# true mean and variance of the beta distribution
mean <- a / (a + b)
var <- (a * b) / ((a + b + 1)*(a + b)^2)


cat("sample mean:", s_mean)
```

```
## sample mean: 0.2040357
```

```r
cat("sample variance:", s_var)
```

```
## sample variance: 0.0110002
```

```r
cat("true mean:", mean)
```

```
## true mean: 0.2
```

```r
cat("true variance:", var)
```

```
## true variance: 0.01
```

The sample mean and variance mean are 0.204 and 0.011 respectively. Which is fairly close to the true mean and variance, which are 0.2 and 0.01 respectively.

d) Estimate the central 95% probability interval of the distribution from the drawn samples.

```r
q <- quantile(sample, probs = c(0.025, 0.975))

q
```

```
##       2.5%      97.5%
## 0.04375745 0.43232832
```

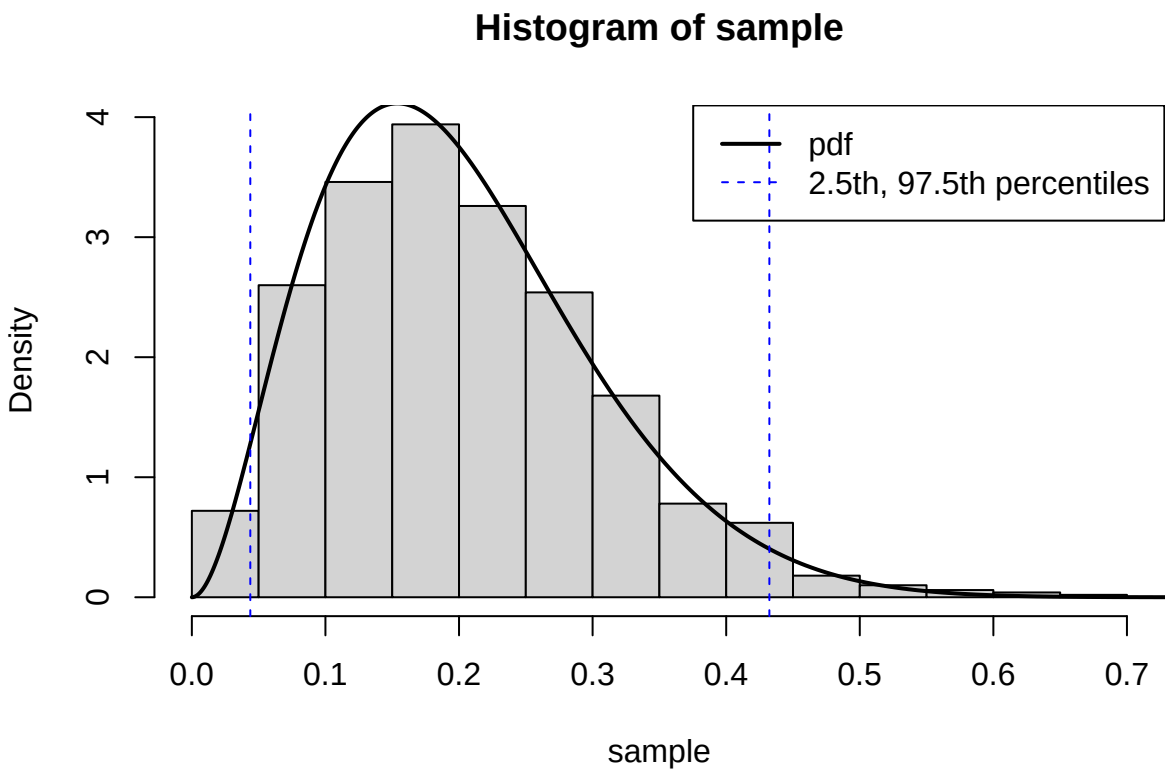
Central 95% probability interval: [0.044, 0.432]

```r
# plot histogram with densities instead of frequencies
hist(sample, freq = FALSE)

# plot beta density function in same plot to compare
lines(x, beta, lwd = 2)

# add vertical lines marking the central 95% probability interval
abline(v = q, col = "blue", lty = 2)

# add legend
legend("topright", legend = c("pdf", "2.5th, 97.5th percentiles"),
       col = c("black", "blue"), lwd = c(2, 1), lty = c(1, 2))
```



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

```
p_positive_1 = (.98*(1/1000))
p_positive_0 = (.04*(1-1/1000))
p_negative_1 = (.02*(1/1000))
p_negative_0 = (.96*(1-1/1000))

p_positive = 0.98*(1/1000) + 0.04*(1-1/1000)
p_negative = 0.02*(1/1000) + 0.96*(1-1/1000)

p_1_p_bayes = p_positive_1/p_positive
p_0_p_bayes = p_positive_0/p_positive
p_1_n_bayes = p_negative_1/p_negative
p_0_n_bayes = p_negative_0/p_negative
```

Let T be the result of the test for detecting lung cancer. $T \in \{positive, negative\}$.

Let X be the status of having or not having lung cancer.

$$\mathrm{X} = \begin{cases} 1, & \text{if patient has lung cancer} \\ 0, & \text{if patient doesn't have lung cancer} \end{cases}$$

We know:

$$\begin{aligned} P(T = positive|X = 1) &= .98 \\ P(T = negative|X = 1) &= .02 \\ P(T = negative|X = 0) &= .96 \\ P(T = positive|X = 0) &= .04 \\ P(X = 1) &= \frac{1}{1000} \\ P(X = 0) &= 1 - \frac{1}{1000} \end{aligned}$$

By the multiplicative law of probability:

$$P(T = positive \cap X = 1) = P(T = positive|X = 1)P(X = 1) = 0.00098 \tag{1}$$

By the law of total probability:

$$P(T = positive) = P(T = positive|X = 1)P(X = 1) + P(T = positive|X = 0)P(X = 0) = 0.0409 \tag{2}$$

Using Bayes' Theorem, 1, and 2, we have :

$$P(X = 1|T = positive) = \frac{P(T = positive|X = 1)P(X = 1)}{P(T = positive)} = 0.0239 \tag{3}$$

Following similar steps as the ones we followed to get 3, we can obtain:

$$\begin{aligned} P(X = 0|T = positive) &= \frac{P(T = positive|X = 0)P(X = 0)}{P(T = positive)} = 0.9761 \\ P(X = 1|T = negative) &= \frac{P(T = negative|X = 1)P(X = 1)}{P(T = negative)} = 0.000021 \\ P(X = 0|T = negative) &= \frac{P(T = negative|X = 0)P(X = 0)}{P(T = negative)} = 0.999979 \end{aligned}$$

The test can identify true positives and true negatives relatively well, however if we check the predictive values of the test we can get a sense of the accuracy of the test, because they depend also on the prevalence rate of cancer (0.001).

The positive predictive value of the test is 0.0239, meaning that of all patients who receive a positive result, approximately 2.39% actually have cancer, while the remaining 97.61% do not have cancer. The test has a really low predictive power on the population, therefore I wouldn't recommend to get the test to the market. A lot of people could start the costly treatment even though they don't have cancer.

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. $P(A) = 0.4$).

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

# define functions
p_red <- function(boxes) {

  p_boxes <- c(0.4, 0.1, 0.5)
  prob <- (boxes[1]/sum(boxes[1,]))*p_boxes[1] +
    boxes[2]/sum(boxes[2,])*p_boxes[2] +
    boxes[3]/sum(boxes[3,])*p_boxes[3]

  return(prob)
}


p_box <- function(boxes) {

  p_boxes <- c(0.4, 0.1, 0.5)
  p_A <- ((boxes[1]/sum(boxes[1,]))*p_boxes[1])/p_red(boxes)
  p_B <- ((boxes[2]/sum(boxes[2,]))*p_boxes[2])/p_red(boxes)
  p_C <- ((boxes[3]/sum(boxes[3,]))*p_boxes[3])/p_red(boxes)

  return(c(p_A, p_B, p_C))
}
```

a) What is the probability of picking a red ball?

```
p_red(boxes)
```

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

We know that:

$$P(red|A) = \frac{2}{7}$$
$$P(red|B) = \frac{4}{5}$$
$$P(red|C) = \frac{1}{4}$$

By the law of total probability

$$P(red) = P(red|A)P(A) + P(red|B)P(B) + P(red|C)P(C) = 0.3193$$

b) If a red ball was picked, from which box it most probably came from?

```
p_box(boxes)
```

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

by Bayes' theorem:

$$P(A|red) = \frac{P(red|A)P(A)}{P(red)} = 0.3579$$
$$P(B|red) = \frac{P(red|B)P(B)}{P(red)} = 0.2506$$
$$P(C|red) = \frac{P(red|C)P(C)}{P(red)} = 0.3915$$

Therefore, if a red ball was picked it is most probable it came from C.

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. 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.

```
p_identical_twin <- function(fraternal_prob, identical_prob) {

  p_ss_ft <- 1/2
  p_ss <- p_ss_ft * fraternal_prob + 1 * identical_prob
  p_it_ss <- (1*identical_prob)/p_ss

  return(p_it_ss)
}

p_identical_twin(1/150, 1/400)
```

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

We know that:

$$P(\text{fraternal twins}) = \frac{1}{150}$$
$$P(\text{identical twins}) = \frac{1}{400}$$
$$P(\text{same sex}|\text{fraternal twins}) = \frac{1}{2}$$
$$P(\text{same sex}|\text{identical twins}) = 1$$

Using law of total probability:

$$P(\text{same sex}) = P(\text{same sex}|\text{fraternal twins})P(\text{fraternal twins}) + P(\text{same sex}|\text{identical twins})P(\text{identical twins})$$
$$= \frac{1}{2} * \frac{1}{150} + 1 * \frac{1}{400}$$

Using Bayes' Theorem:

$$P(\text{identical twins}|\text{same sex}) = \frac{P(\text{same sex}|\text{identical twins})P(\text{identical twins})}{P(\text{same sex})} = 0.4286$$