Assignment 1 - BSDA

September 2, 2023

List of packages used

Below I am stating some necessary administrative regarding solving the assignment, eg the time it took to solve the assignment and pros with the assignment.

Times used for reading and self-study exercise: 6
Time used for the assignment: 17
Good with the assignment: several hints, both conditional and unconditional probabilities
Things to improve in the assignment: No points for being able to compile to PDF

Introduction about the assignment

This assignment consist of 5 questions and can be seen as a introduction of the course. The first question consist of basic probability theory notations and terms and the second question is being based on programming/computer skills where we will be plotting a density function among other things. The last three question will be solved using Bayes theorem and functions in R. Notice: The result from R package markmyassignment will not be included in the assignment.

Question 1 - Probability theory notation and terms

Probability - how likely an event is to occur which is a value between zero and one.

Probability mass - probability that a discrete random variable takes on an exact value

Probability density - the relative likelihood of a continuous random variable

Probability mass function - a function that gives the probability that a discrete random variable takes on a specific value within the domain of the variable.

Probability density function - a function that gives the relative likelihood that a continuous random variable for any given sample in the domain of the variable.

Probability distribution - function for all different outcomes and probabilities of a random variable within a sample space

Discrete probability distribution - a function for outcomes and respective probabilities of all different values of a random variables within the sample space.

Continuous probability distribution - a function for outcomes and respective probabilities of all different values of a continuous random variable within the sample space.

Cumulative distribution - the aggregated density / probability from lowest possible value of a random variable up to a certain value.

Likelihood - the plausibly of different parameter values, often using the joint function.

Aleatoric uncertainty - an uncertainty due to lack of randomness regarding the stochastic variable with a probability function of $P(y|\theta)$

Epistemic uncertainty - due to lack of knowledge and we gain information when doing experiment and having probability function of $P(\theta)$

Question 2 - Basic computing skills

Q2a; Plot the density function of Beta-distribution

We will plot the density function for β-distribution given $\mu = 0.2$ and $\sigma^2 = 0.01$. The parameters α and β of the β-distribution is related to the mean and variance according to below:

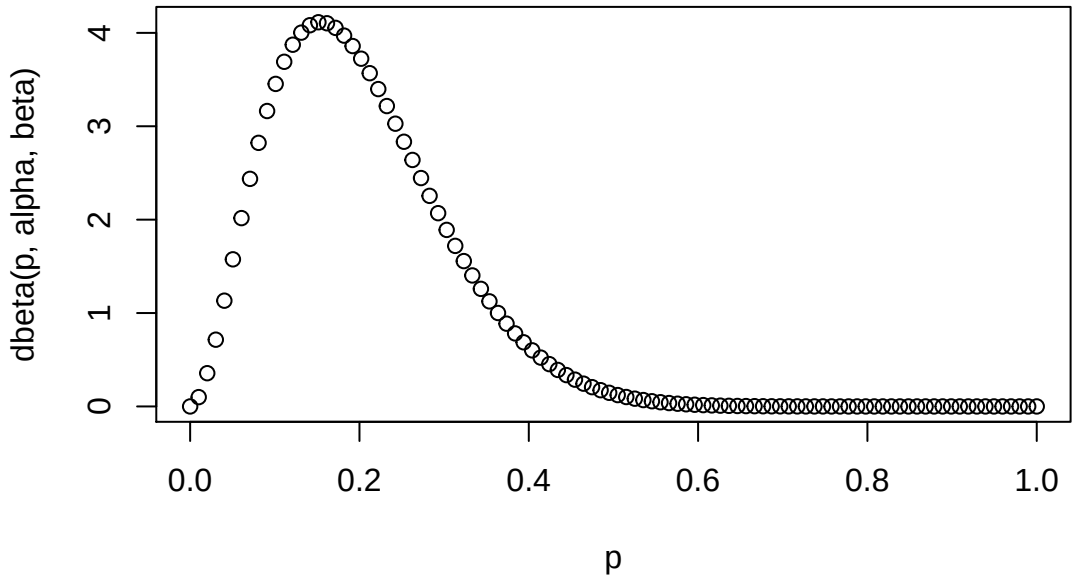
$$\alpha = \mu\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right), \quad \beta = \frac{\alpha(1-\mu)}{\mu}$$

and when we include the values of the parameter mean and variance we get

$$\alpha = 0.2\left(\frac{0.2(1-0.2)}{0.01} - 1\right) = 3, \quad \beta = \frac{3(1-0.2)}{0.2} = 12$$

Hence, we have $X \sim \beta(\alpha = 3, \beta = 12)$ distribution.

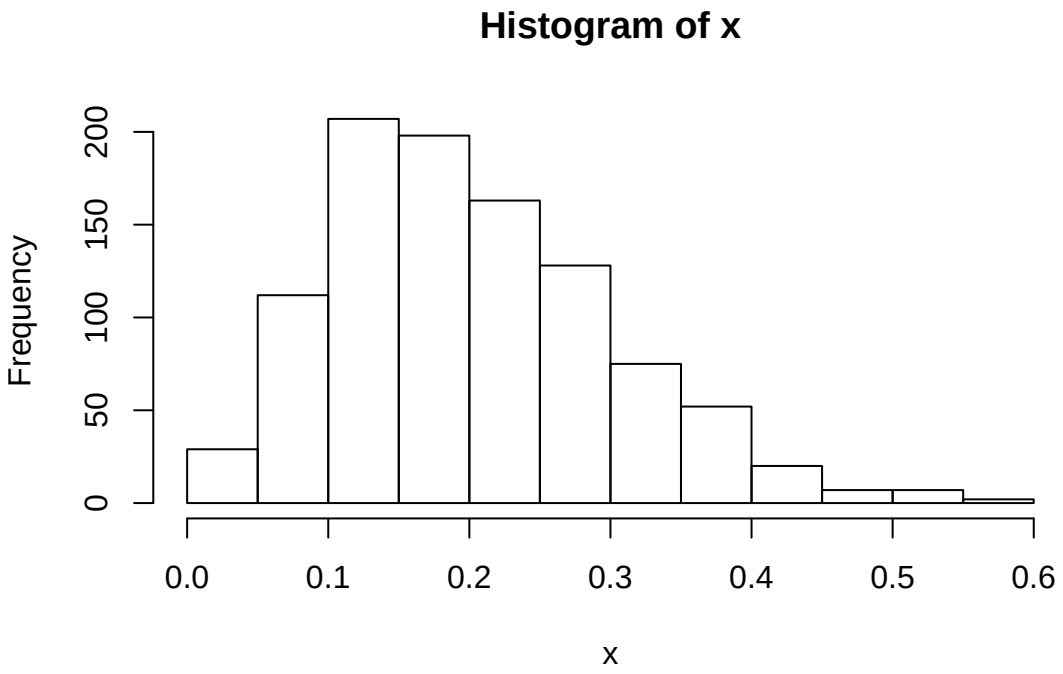
```
mu = 0.2
sigma2 = 0.01
alpha = mu * ( mu*(1- mu)/sigma2-1)
beta = alpha*(1-mu) / mu

# the domain for the beta distribution is from 0 to 1 so we
# use the sequence function to get a sequence of values within the domain.
p = seq(0, 1, length=100)
plot(p, dbeta(p, alpha, beta))
```



Q2b; Plot the density function of Beta-distribution

For this subtask we will take a sample of 1000 random numbers from the $X \sim \beta(\alpha = 3, \beta = 12)$ distribution and plot a histogram. This will be done using rbeta-function and hisT-function in R.

```
x <- rbeta(1000, shape1 = alpha, shape2 = beta)
hist(x)
```



The histogram of 1000 random numbers and the plotted density function looks very similar.

Q2c; Compute the sample mean and variance

We will compute the sample mean and variance by using the mean-function and var-function, respective. The sample mean and variance for the generated 1000 random numbers is

```
mean(x) # 0.2016153

## [1] 0.20238

var(x) # 0.01086038

## [1] 0.009935242
```

which is approximate to the population mean and variance i.e $\bar{X} \approx \mu = 0.2$ and $\hat{\sigma}^2 \approx \sigma^2 = 0.01$. This is consistent with the asymptotic theory that $\bar{X} \to_p \mu$ and $S^2 \to_p \sigma^2$ as $n \to \infty$.

Hence, because parameters α and β of the Beta-distribution is related to mean and variance we get that the estimated α and β is roughly matched with the true parameters.

Q2d; Estimate the 95 % probability interval

We will compute the central 95 % probability interval using the quantile-function in R and we get

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

##       2.5%      97.5%
## 0.04554328 0.41852371
```

So we have that $(Q_{0.025}, Q_{0.975}) \approx (0.044, 0.44) = 0.95$.

Question 3, Calculating false positive and false negative

Two binary random variables has been defined in this question. Random variable T stands for testing positive or not and random variable C stands for individual having cancer or not.

Now, to calculate the false positive, i.e $P(C = 0|T = positive)$, we are using the Bayes theorem where $P(C = 1) = \frac{1}{1000}$. Hence, $P(C = 0) = \frac{999}{1000}$ due to law of total probability.
Before, need to calculate the probability of testing positive.

$P(T = positive) = \sum_{i=1}^{2} P(T = positive|C_i) * P(C_i)$

$= P(T = positive|C = 1) * P(C = 1) + P(T = positive|C = 0) * P(C = 0)$

$= 0.98 * \frac{1}{1000} + (1 - P(T = negative|C = 0)) * \frac{999}{1000} = \frac{1}{1000}(0.98 + 0.04 * 999) = \frac{1}{1000} * 40.94$

Now when the probability of testing positive has been calculated we can calculate the

$P(C = 0|T = positive) = \frac{P(T=positive|C=0)*P(C=0)}{P(T=positive)}$

$= \frac{(1-P(T=negative|C=0))*\frac{999}{1000}}{P(T=positive)} = \frac{(1-0.96)*\frac{999}{1000}}{P(T=positive)} = \frac{(0.04)*\frac{999}{1000}}{\frac{1}{1000}*40.94} = \frac{0.04*999}{40.94} = 0.97606$

So the probability of false positive equals 0.97606. Now the steps to calculate the false negative is the same as above so we start with calculating $P(T = negative)$.
$P(T = negative) = \sum_{i=1}^{2} P(T = negative|C_i) * P(C_i)$

$= P(T = negative|C = 1) * P(C = 1) + P(T = negative|C = 0) * P(C = 0)$

$= (1 - P(T = positive|C = 1)) * \frac{1}{1000} + 0.96 * \frac{999}{1000}$

$= (1 - 0.98) * \frac{1}{1000} + 0.96 * \frac{999}{1000} = \frac{1}{1000}(0.02 + 0.96 * 999) = \frac{1}{1000} * 959.06$

Now when the unconditional probability of testing negative has been calculated we can calculate the false negative
$P(C = 1|T = negative) = \frac{P(T=negative|C=1)*P(C=1)}{P(T=negative)} = \frac{(1-P(T=positive|C=1))*\frac{1}{1000}}{\frac{1}{1000}*959.06} = \frac{0.02}{959.06} = 2.98 * 10^{-5}$

and the probability of false negative equals $2.98 * 10^{-5}$. The cancer screening test had a high false positive but a low false negative probability. The $2.98 * 10^{-3}$% false negative result indicates that the test gives true negative in almost all patients. However, the 97.606% false positive results indicates that the test gives true positive in about 2.3% of the patients. So this cancer screening test is not good at detecting cancer patients. I would advise them not to get the test out on the market with this results.

Question 4; Bayes theorem

This question consist of 2 sub-question, the first sub-question is to calculate the probability of picking a red and the second sub-question is to compute conditional probability.

We have three boxes consisting of different amounts of red and while balls. The data set below shows the number of red and white balls in respective box, ie box A, box B and box C.

```
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
```

Q4a; Probability of picking up a red ball

We can calculate the probability of picking up a red ball by using the Bayes theorem. Hence, we use the conditional probability of picking up a red ball given a specific box to solve the problem. Random variable A is the color of the ball and random variable B is the box

$A \in (red, white)$

$B \in (boxA, boxB, boxC)$

so we get the probability of picking up a red ball, i.e marginal probability, by the following formula:

$$P(A = red) = \sum_{i=1}^{3} P(A = red|B_i)P(B_i) = P(A = red|B_1)P(B_1) + P(A = red|B_2)P(B_2) + P(A = red|B_3)P(B_3)$$

and the R function below computes the probability of picking up a red ball

```
p_red <- function(boxes) {

  ProbVector <- c(0.4, 0.1, 0.5)

  Prob.red.i <- c()

  for ( i in 1:nrow(boxes)) {

    Prob.red.i[i] <- boxes[i,1]   / (  boxes[i,1] + boxes[i, 2]  ) * ProbVector[i]

  }
  output <- sum(Prob.red.i)
  return(output)
}
probred <- p_red(boxes = boxes)
probred  # 0.3192857

## [1] 0.3192857
```

The probability of picking up a red ball is $P(A = red) = 0.3192857$.

Q4b, Conditional probability

For this sub-question we want to know which box that is most likely that a red ball comes from given that a red ball has been picked up. Definition for random variable A and random variable B can be found in the section Q4.
Also in this case we will use Bayes theorem to solve $P(B|A = red), \quad \forall B \in (A, B, C)$. Hence, we will get three different conditional probabilities, one for each box. By using the Bayes theorem we get

$$P(B|A = red) = \frac{P(A = red|B) * P(B)}{P(A = red)}$$

where $P(A = red) = 0.3192857$. If we apply the data into the formula we get the following probabilities:

$P(B = A|A = red) = \frac{\frac{2}{7} * 0.4}{0.3192857} = 0.3579418$

$P(B = B|A = red) = \frac{\frac{4}{5} * 0.1}{0.3192857} = 0.2505593$

$P(B = C|A = red) = \frac{\frac{1}{4} * 0.5}{0.3192857} = 0.3914989$

Below we implement a R function to compute these probabilities. Worth mentioning is that the function only consist of one function argument which is for the data set.

```
ProbVector <- c(0.4, 0.1, 1-(0.4+0.1))

p_box <- function(boxes) {
  Prob.red.i <- c()
  Prob.box.i <- c()

  for ( i in 1:nrow(boxes)) {
    Prob.red.i[i] <- boxes[i,1]   / (  boxes[i,1] + boxes[i, 2]  ) * ProbVector[i]
  }

  Prob.box.i <- Prob.red.i / sum(Prob.red.i)

  return(Prob.box.i)
}
p_box(boxes = boxes)

## [1] 0.3579418 0.2505593 0.3914989
```

It is most likely that the red ball was picked from Box C.

Question 5; Bayes theorem and Twin brother

Three random binary variables is defined for this problem. We have variable I, G and T for identical or not, gender and twin or not, respective. All of these variables is binary so it can easy be described in a equation system

Hence, the unconditional joint probability for identical twin brother is
$P(I = 1, T = 1, G = 1) = P(I = 1) * P(T = 1, G = 1|I = 1)$ while
$P(I = 0, T = 1, G = 1) = P(I = 0) * P(T = 1, G = 1|I = 0)$ is the unconditional joint probability for fraternal twin brother.

We also have that
$P(identical) = P(I = 1) = 1/400$ and
$P(fraternal) = P(I = 0) = 1/150$.

Also, So having a twin brother given being identical is $P(T = 1, G = 1|I = 1) = \frac{1}{2}$ and having a twin brother given being fraternal is $P(T = 1, G = 1|I = 0) = \frac{1}{4}$. So we get
$P(I = 1, T = 1, G = 1) = \frac{1}{400} * \frac{1}{2} = \frac{1}{800}$
$P(I = 0, T = 1, G = 1) = \frac{1}{150} * \frac{1}{4} = \frac{1}{600}$

Now the conditional probability that being identical given having a twin brother

$$P(I = 1|T = 1, G = 1) = \frac{P(I = 1, T = 1, G = 1)}{P(T = 1, G = 1)} = \frac{\frac{1}{800}}{\frac{1}{600} + \frac{1}{800}} = 0.4285714$$

Below is the R function that computes the probability

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

  # unconditional probability for the identical and fraternal joint with same gender
  prob_iden_twinbrother <- identical_prob * 1/2
  prob_frat_twinbrother <- fraternal_prob * 1/4
```

```r
  # conditional probability for the identical given having a twin brother
  condprob <- prob_iden_twinbrother / ( prob_iden_twinbrother + prob_frat_twinbrother)
  return(condprob)
}
p_identical_twin(fraternal_prob = 1/150, identical_prob = 1/400)

## [1] 0.4285714
```

and the probability of being identical when having a twin brother equals 0.4285714.