BDA - Assignment 1

Anonymous

Contents

Exercise 1) Basic probability theory notation and terms

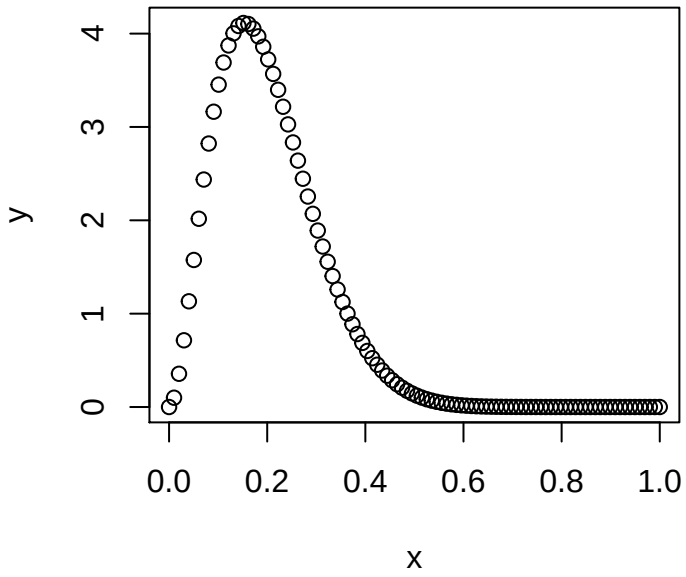
- **probability**: represents how likely an event is to happen
- **probability mass**: the likelihood of a discrete event to happen
- **probability density**: the likelihood of a continuous random variable to lie within a certain range of values.
- **probability mass function (pmf)**: function that describes the probability of discrete random variable.
- **probability density function (pdf)**: function that describes the relative likelihood of a random continuous variable being equal to any given sample in the sample space.
- **discrete probability distribution**: probability distribution that describes the likelihood of the values in a discrete random variable.
- **continuous probability distribution**: probability distribution that describes the possible values of a continuous random variable
- **cumulative distribution function (cdf)**: represents the probability that a given random variable will take a value of less than or equal to a certain value.
- **likelihood**: $p(y|\theta)$ as a function of θ given fixed y provides information about the epistemic uncertainty observed from the data.

Exercise 2) Basic computer skills

a)

Density function of Beta-distribution, with mean $\mu = 0.2$ and variance $\sigma^2 = 0.01$.

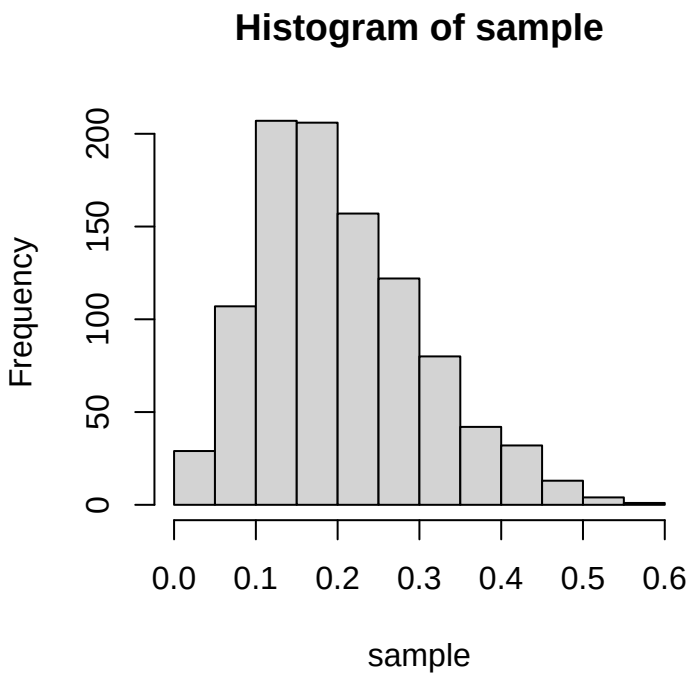
```
mu <- 0.2
var <- 0.01
alpha <- mu * (((mu*(1-mu))/var) - 1)
beta <- (alpha*(1-mu))/mu
x <- seq(0, 1, length=100)
y <- dbeta(x, alpha, beta)
plot(x, y)
```



b)

Histogram of a sample of 1000 random numbers from the above distribution.

```
n <- 1000
sample <- rbeta(n, alpha, beta)
hist(sample)
```



As we can see, the histogram of the samples that were drawn from the beta distribution follows the shape of the density function above.

c)

Sample mean and variance from the drawn sample

```
mu_bar <- mean(sample)
var_bar <- var(sample)
print(paste("Sample mean:", format(mu_bar)))
```

```
## [1] "Sample mean: 0.2034272"
```

```
print(paste("Sample variance:", format(var_bar)))
```

```
## [1] "Sample variance: 0.01027089"
```

As we can see both values are approximately the same as the true values (i.e. $\mu = 0.2$ and $\sigma^2 = 0.01$).

d)

Central 95% probability interval of the distribution from the drawn samples

```
lower <- quantile(sample, probs = 0.025)
upper <- quantile(sample, probs = 0.975)
print(paste("The central 95% probability interval of the distribution from the drawn samples"))
```

```
## [1] "The central 95% probability interval of the distribution from the drawn samples"
```

```
print(paste("is [", format(lower), ",", format(upper), "]"))
```

```
## [1] "is [ 0.04679795 , 0.4302923 ]"
```

Exercise 3) Bayes' theorem

The facts given by the researchers are the following:

- $P(positive|cancer) = 0.98$,
- $P(negative|healthy) = 0.96$,
- $P(cancer) = 0.001$

To simplify notation, we assume that not cancer equals being healthy. In order to know the recommendation that we would issue, we must compute what is the probability that a person has cancer given that the test predicts a positive result. We can compute this using the Bayes' rule:

$$P(cancer|positive) = \frac{P(positive|cancer)P(cancer)}{P(positive)} \tag{1}$$

$$= \frac{P(positive|cancer)P(cancer)}{P(positive|cancer)P(cancer) + P(positive|healthy)P(healthy)} \tag{2}$$

$$= \frac{0.001 * 0.98}{0.98 * 0.001 + 0.04 * 0.999} \tag{3}$$

$$= 0.0239 \tag{4}$$

where

$$P(positive|healthy) = 1 - P(negative|healthy) = 1 - 0.96 = 0.04$$

and

$$P(healthy) = 1 - P(cancer) = 1 - 0.001 = 0.999.$$

Thus, we now know that for a random patient that is examined with this method, there is only a 2.4% probability of the patient actually having cancer when the test gives a positive result. This is very alarming because the consequences of a positive test are very drastic: medication, surgery or more testing. That would mean that most of the patients that would be given these treatments would not actually need them.

Based on this information, I would recommend decreasing the number of false positives (i.e. healthy predicted as positive), because that would increase $P(negative|healthy)$. This would mean that $P(positive|healthy)$ would decrease which is key to improve $P(cancer|positive)$ because 99.9% of the population do not have cancer, which gives this probability a large weight in the computation.

Exercise 4) Bayes' theorem

a)

The probability of picking a red ball is computed as

$$P(red) = P(A) * P(red|A) + P(B) * P(red|B) + P(C) * P(red|C) \tag{5}$$

$$= 0.5 * \frac{2}{2+5} + 0.1 * \frac{4}{4+1} + 0.5 * \frac{1}{1+3} \tag{6}$$

$$= 0.3193 \tag{7}$$

```
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) {
  p_A <- 0.4
  p_B <- 0.1
  p_C <- 0.5
  p <- p_A*boxes[1,1]/sum(boxes[1,]) + p_B*boxes[2,1]/sum(boxes[2,]) + p_C*boxes[3,1]/sum(boxes[3,])
  return(p)
}
p_red(boxes = boxes)
```

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

b)

To know the most probable box a red ball would be picked from, we need to compute

- $P(A|red) = \frac{P(A) * P(red|A)}{P(red)}$,
- $P(B|red) = \frac{P(B) * P(red|B)}{P(red)}$ and
- $P(C|red) = \frac{P(C) * P(red|C)}{P(red)}$.

The function definition would be the following

```
p_box <- function(boxes) {
  p_A <- 0.4
  p_B <- 0.1
  p_C <- 0.5
  p_A.red <- p_A*boxes[1,1]/sum(boxes[1,])/p_red(boxes)
  p_B.red <- p_B*boxes[2,1]/sum(boxes[2,])/p_red(boxes)
  p_C.red <- p_C*boxes[3,1]/sum(boxes[3,])/p_red(boxes)
  return(c(p_A.red, p_B.red, p_C.red))
}
p_box(boxes = boxes)
```

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

Looking at the output of the function we can conclude that if a red ball was picked, it most probably came from box C.

Exercise 5) Bayes' theorem

We know that $P(\text{identical twins}) = \frac{1}{150}$, $P(\text{fraternal twins}) = \frac{1}{400}$ and $P(\text{male}) = P(\text{female}) = 0.5$. Then, we can compute

$$P(\text{identical twins \& twin brother}) = P(\text{identical twins}) * P(\text{male twins} \mid \text{identical twins})$$

and

$$P(\text{fraternal twins \& twin brother}) = P(\text{fraternal twins}) * P(\text{male twins} \mid \text{fraternal twins}).$$

Knowing this, we can obtain the probability of Elvis being an identical twin given that he had a twin brother by

$$P(\text{identical twins} \mid \text{twin brother}) = \frac{P(\text{identical twins \& twin brother})}{P(\text{twin brother})}$$

This can be computed in R as

```
p_identical_twin <- function(fraternal_prob, identical_prob) {
  p_it.and.tb <- identical_prob * 0.5
  p_ft.and.tb <- fraternal_prob * 0.5 * 0.5
  return(p_it.and.tb/(p_it.and.tb + p_ft.and.tb))
}
p_identical_twin(fraternal_prob = 1/150, identical_prob = 1/400)
```

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

Thus, there is a 42.86% probability that Elvis was an identical twin.