BDA - Assignment 1

Anonymous

5/2/2020

Q1 - Definitions of:

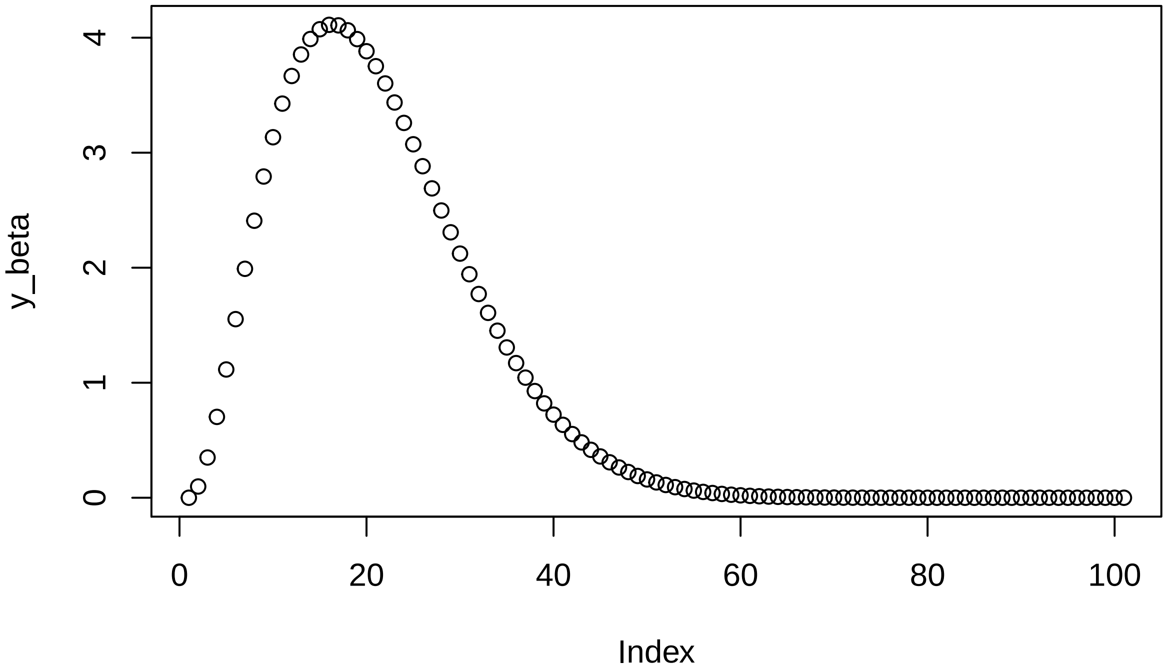
1. Probability: How likely that an event will occur
2. Probability mass: The chance for a discrete random variable is excactly the same as a given value
3. Probability density: a comparable value of the probability distribution of a continuous random variable
4. Probability mass function (pmf): The function that maps each possible event to the corresponding probability mass, for discrete events
5. Probability density function (pdf): The density of a continuous random variable, used to used to specify the probability of the random variable falling within a particular range of values, as opposed to taking on any one value
6. Probability distribution: The probability structure of a random variable
7. Discrete probability distribution: When visualizing, instead of a curve like the continuous probability distribution, discrete distribution shows the probabilitits of outcomes with finite values.
8. Continuous probability distribution: Built from outcomes that potentially have infinute measurable values, which will be visualized by a continus curve in a graph.
9. Cumulative distribution function (cdf): The distribution function of the random variable, The probability that the random variable (x) is less than or equal to x
   - $F(x) = P[X \leq x]$ for all $-\infty < x < \infty$
10. Likelihood: Describes how well a model fits to a sample of data, which can be said to be the y-axis values for fixed data points with distribution being non-fixed, hereby giving information about epistemic uncertainty

Q2 - Computing and plotting

a) Plot of the density function for $\mu = 0.2$ and σ^2=0.01 , where α and β are related to μ and σ^2 by:

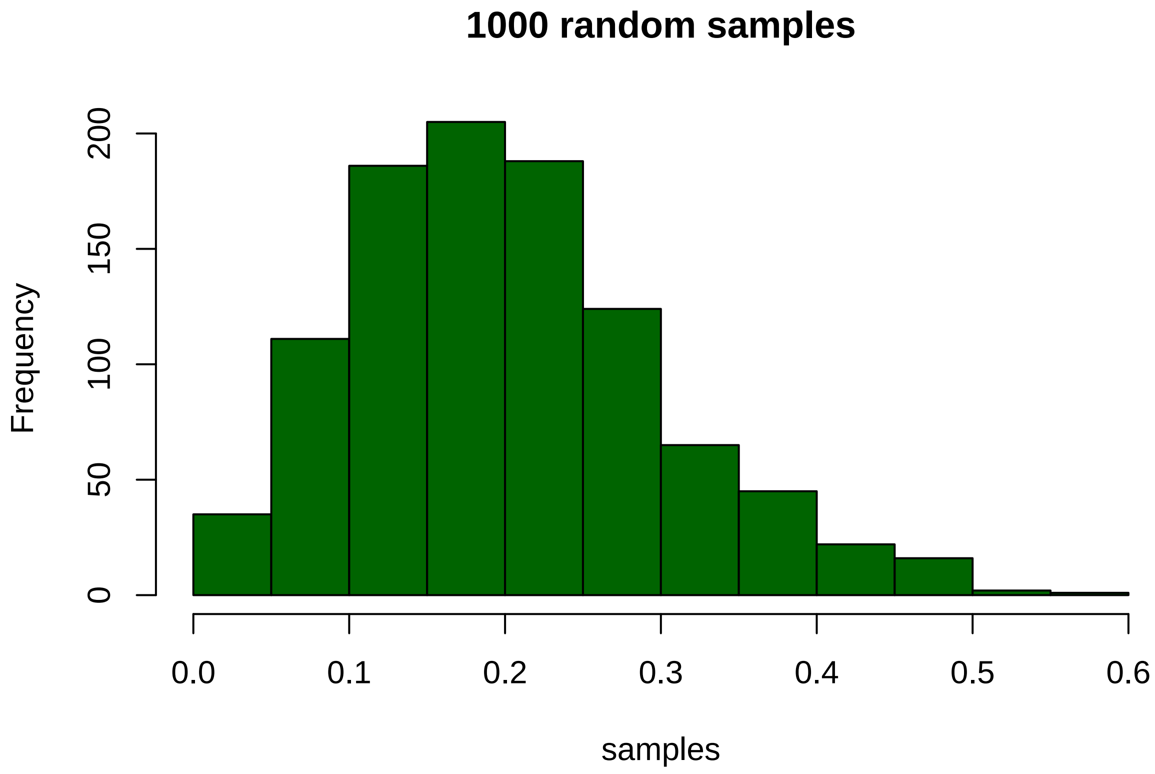
$$\alpha = \mu(\frac{\mu(1-\mu)}{\sigma^2} - 1), \quad \beta = \frac{\alpha(1-\mu)}{\mu}$$

```r
mu = 0.2
var = 0.01
a = mu * (mu * (1 - mu) / var - 1)
b = a * (1 - mu) / mu
x_beta <- seq(0, 1, by = 0.01) # linear space bewteen 0 and 1
y_beta <- dbeta(x_beta, shape1 = a, shape2 = b)
plot(y_beta)
```



b) Plot of 1000 random numbers from above distribution:

```
random1000 <- rbeta(1000, shape1=a, shape2=b)
hist(random1000, main="1000 random samples", xlab="samples", col="darkgreen")
```



c) sample mean and variance for random1000:

```
mean(random1000) # theory: 0.2
```

```
## [1] 0.2012012
```

```
var(random1000) # theory: 0.01
```

```
## [1] 0.009659045
```

d) 95% probability interval for random1000:

```
quantile(random1000, probs = c(0.05, 0.95))
```

```
##        5%       95%
## 0.0580681 0.3766058
```

Q3 - Bayes' theorem on lung cancer

From the assignment we have:

- P(C) Population with cancer = 0.001
- P(C | POS) = 0.98
- P(H | NEG) = 0.96

From this we find:

- P(H) Population with no cancer = 1 - P(c) = 0.999
- P(C | NEG) Negative tests for cancer population = 1 - P(C | POS) = 0.02
- P(H | POS) Positive test for healthy population = 1 - P(H | NEG) = 0.04

True and False, positive and negative tests should also be presented:

$$\text{True Positive}: \ P(C)\,P(C \mid POS) = 0.001 * 0.98 = 0.00098 = 0.098\%$$

$$\text{False Negative}: \ P(C)\,P(C \mid NEG) = 0.001 * 0.02 = 0.00002 = 0.002\%$$

$$\text{False Positive}: \ P(H)\,P(H \mid POS) = 0.999 * 0.04 = 0.03996 = 3.996\%$$

$$\text{True Negative}: \ P(H)\,P(H \mid NEG) = 0.999 * 0.96 = 0.95804 = 95.804\%$$

Summarizing this in a table:

Result	C(P)=0.001	C(H)=0.999
POS	0.00098	0.03996
NEG	0.00002	0.95804

So with 0.002%, we find that in a population of 6.000.000 people, with 0.1% having lung cancer, there is a chance that 120 people with lung cancer will be tested negative for having lung cancer. The test should therefore be improved to lower the false negative number.

We can also look at the probability for actually having cancer, if the test is negative. For this we need to include all the negative test results for both Cancer and Healthy. Applying Bayes theorem gives us:

$$P(NEG \mid C) = \frac{P(C)\,P(C \mid NEG)}{P(C)\,P(C \mid NEG) + P(H)\,P(H \mid NEG)} = \frac{0,001 * 0,02}{0,001 * 0,02 + 0,999 * 0,96} \approx 0,000021$$

The result is a bit higher here, with 0.0021% chance of having cancer, if your test result is negative.

Q4 Bayes' Theorem

Given is a 3 different boxes A, B and C, and 2 possible colours of balls red and white, placed in the boxes in a given order, so that:

```
p_A = .4 # P(A) = 40% (given from assignment)
p_B = .1 # P(B) = 10% (given from assignment)
p_C = 1-(p_A + p_B) # P(C)= 50%

p_red_A = 2/(2+5) # P(red|A) = 2/7 (given from assignment)
p_red_B = 4/(4+1) # P(red|B) = 4/5 (given from assignment)
p_red_C = 1/(1+3) # P(red|C) = 1/4 (given from assignment)
```

a) Probability of picking a red ball, selceting box and ball randomly:

```
p_red = p_A*p_red_A + p_B*p_red_B + p_C*p_red_C
p_red # P(red)
```

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

P(red)= 31.93 %

b) Using Bayes Theorem to find which box it's most likely a red ball is picked from:

```
p_A_red = p_A*p_red_A / p_red
p_B_red = p_B*p_red_B / p_red
p_C_red = p_C*p_red_C / p_red
p_A_red # Probability it's from box A
```

```
## [1] 0.3579418
```

```
p_B_red # Probability it's from box b
```

```
## [1] 0.2505593
```

```
p_C_red # Probability it's from box C
```

```
## [1] 0.3914989
```

The results are:
P(A|red) = 35.79 %
P(B|red) = 25.06 %
P(C|red) = 31.15 %

Q5 Bayes' Theorem on identical twins:

From the assignment is given the following probabilities of twins, where we assume an equal number of boys and girls are born on average

```
p_fraternal = 1/150
p_identical = 1/400
p_twin = p_fraternal + p_identical
```

From this we can find the possibility of Elvis Presley having and identical twin:

```
p_identical_twin = p_identical / p_twin
p_identical_twin
```

```
## [1] 0.2727273
```

P(identical|twin) = 27.27 %