Worksheet 3

1.a) > x = c(1, 2, 3, 4, 5, 6)

This command is defining a vector from 1 to 6

> probx = c(1/4, 1/8, 1/8, 1/8, 1/8, 1/4)

This command defines a vector probx which has the probabilities for each value of x.

> F16 = sample(x, size = 1500, replace = T, prob = probx)

This command creates a vector of size = 1500 with elements from the vector x & these elements being sampled have the probability (weights) as assigned in the vector probx.

(b) > mean (F16)

> 3.507333

The mean of F16 is 3.507333.

> var (F16)

> 3.679066

The variance of f16 is 3.679066

Range (F16) = $(\mu - \tau, \mu + \tau)$

where μ = mean (F16) = 3.507333, τ = S.D(f16) = $\sqrt{3.679066}$ = 1.918089

∴ Range (F16) = (1.589244, 5.425422)

(c) True mean = $1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{8} + 3 \cdot \frac{1}{8} + 4 \cdot \frac{1}{8} + 5 \cdot \frac{1}{8} + 6 \cdot \frac{1}{4}$

= 3.5

True variance = $(1-3.5)^2 \cdot \frac{1}{4} + (2-3.5)^2 \cdot \frac{1}{8} + (3-3.5)^2 \cdot \frac{1}{8}$

$+ (4-3.5)^2 \cdot \frac{1}{8} + (5-3.5)^2 \cdot \frac{1}{8} + (6-3.5)^2 \cdot \frac{1}{4}$ = 3.75.

The mean & variance are close to the true mean & true variance with an error of 0.007333 & 0.070934 resp.

2) > b1 = rbinom(100, 10, 0.5)
> b2 = rbinom(100, 10, 0.25)
> b3 = rbinom(100, 10, 0.75)

(a) > ?rbinom

It generates a random sample following binomial distribution.

for b1, we have 10 observations i.e. variables from 1 to 10 and a sample of size 100 is produced where the probability of success in each trial is 0.5.

for b2, a random sample of size 100 is produced with 10 observations but now the probability of success is 0.25.

for b3, a random sample of size 100 is produced with 10 observations with probability of success = 0.75.

(b) > mean(b1)
[1] 4.88
The mean of b1 = 4.88

> mean(b2)
[1] 2.59
The mean of b2 = 2.59

> mean(b3)
[1] 7.74
The mean of b3 = 7.74.

> var(b1)
[1] 2.692525
The variance of b1 = 2.692525

> var(b2)
[1] 1.961515
The variance of b2 = 1.961515

> var(b3)
[1] 1.871111
The variance of b3 = 1.871111

for b1, n = 10, p = 0.5 ∴ mean = np = 5.
variance = npq = 10 × 0.5 × 0.5 = 2.5

For b2, n = 10, p = 0.25, mean = np = 2.5.
variance = npq = 2.5 × 0.75 = 1.875

for b3 of $n = 10$, $p = 0.75$, mean $= np = 7.5$

variance $= npq = 1.875$.

We can see for b1, |true mean − mean(b1)| = 0.12 i.e. the average of the sample is little far from the actual average.

|true variance − var(b1)| = 0.19 i.e. the spread of the data of the sample is also different from the actual spread.

For b2, |true mean − mean(b2)| = 0.09 i.e. the average of the sample is almost equal to the actual average.

|true variance − var(b2)| = 0.09 i.e. the spread of the sample is also close to the actual spread.

For b3, |true mean − mean(b3)| = 0.24 i.e. the avrge of the sample is different from the actual one.

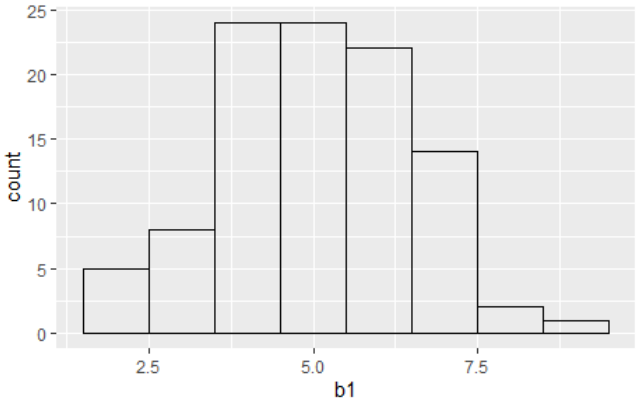
|true variance − var(b3)| = 0.004 i.e. the sample data has the same spread as the actual data.

3.9) p11 plots a histogram which displays the frequency of the sample b1 by taking the class intervals to be 0–2.5, 2.5–3.5, 3.5–4.5, ..., 7.5–8.5, 8.5–10.

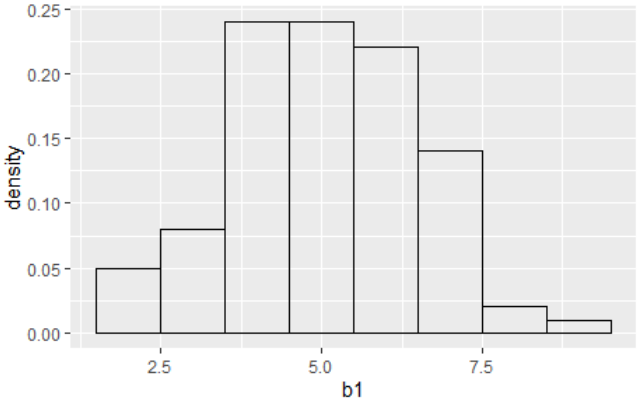
p21 plots a histogram with the values of the sample data on the x axis and the density i.e. the ratio of the frequency of a particular class interval to the total number of observations, on the y axis. y: ..density.. makes the sum of the heights of each rectangle of the histogram = 1.

Both the graphs have color = 'black' because in which the boundary lines of the rectangles are black.

Solution 3.a)



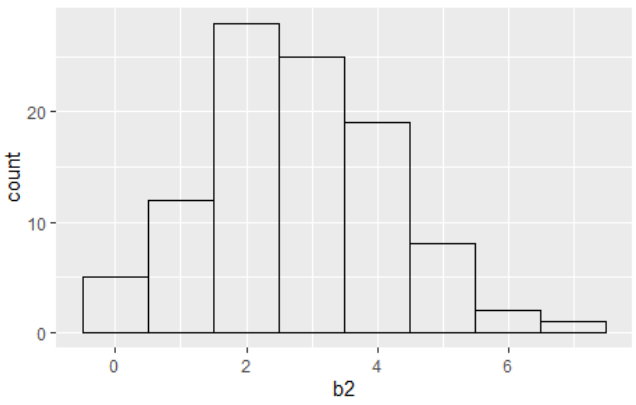
Output of p11



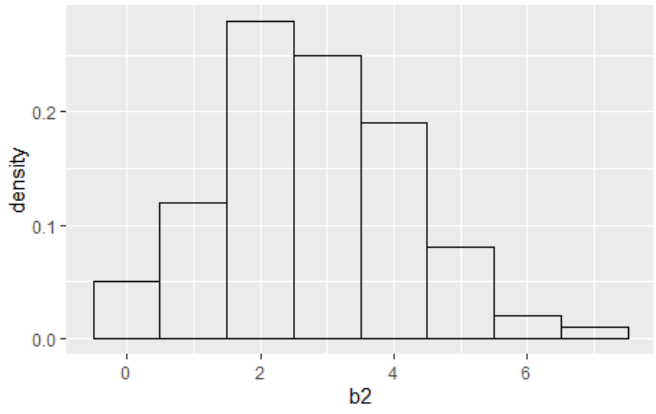
Output of p21

Solution 3.b)

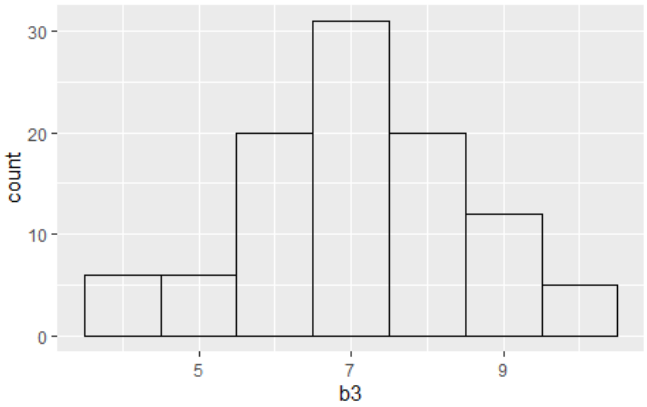
```
Console   Terminal   Jobs
R 4.1.1 · ~/
> #Solution to Question 3.a)
>
>
> b1 = rbinom(100,10,0.5)
> df1=data.frame(b1)
> p11= ggplot(df1) + geom_histogram(mapping=aes(x=b1), color="black", fill="NA", binwidth=1)
> p21= ggplot(df1) + geom_histogram(mapping=aes(x=b1, y=..density..), color="black", fill="NA", binwidth=
1)
> p11
> p21
>
>
>
>
> #Solution to Question 3.b)
>
> b2 = rbinom(100,10,0.25)
> b3 = rbinom(100,10,0.75)
> df2 = data.frame(b2)
> df3 = data.frame(b3)
> p12= ggplot(df2) + geom_histogram(mapping=aes(x=b2), color="black", fill="NA", binwidth=1)
> p22= ggplot(df2) + geom_histogram(mapping=aes(x=b2, y=..density..), color="black", fill="NA", binwidth=
1)
> p12
> p22
>
>
> p13= ggplot(df3) + geom_histogram(mapping=aes(x=b3), color="black", fill="NA", binwidth=1)
> p23= ggplot(df3) + geom_histogram(mapping=aes(x=b3, y=..density..), color="black", fill="NA", binwidth=
1)
> p13
> p23
>
>
```



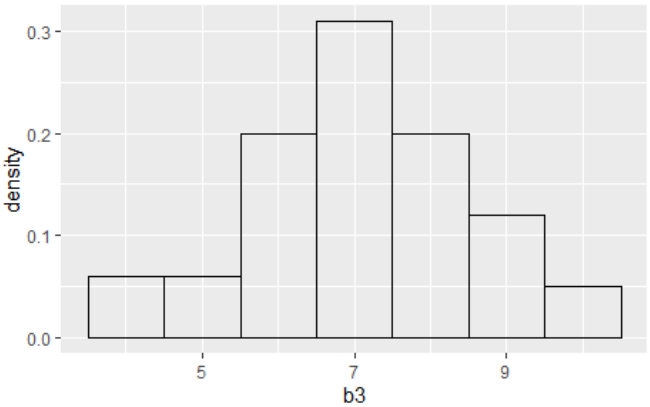
Output of p12



Output of p22



Output of p13



Output of p23

3.b) p11, p12 & p13 are plotting the variable b1, b2 & b3 against the count of each class-interval respectively.

For p11, the highest frequency is observed between 3.5 – 5.5. The same observation is seen in p21, just this time we can conclude that the values between 3.5 and 5.5 have the highest density. These graphs are symmetric

For p12, the highest frequency is in between 1.5 to 2.5 i.e. the lower range of values, near the 1st quartile, hence in p22 also the values between 1.5 to 2.5 are the most dense. This graph is skewed towards the left.

For p13, the highest frequency is in between 6.5 to 7.5 i.e. near the 3rd quartile. Hence in p23 also the values between 6.5 to 7.5 have the highest density. If the scaling would have been the same like p12 then these graphs would have been skewed towards the right.

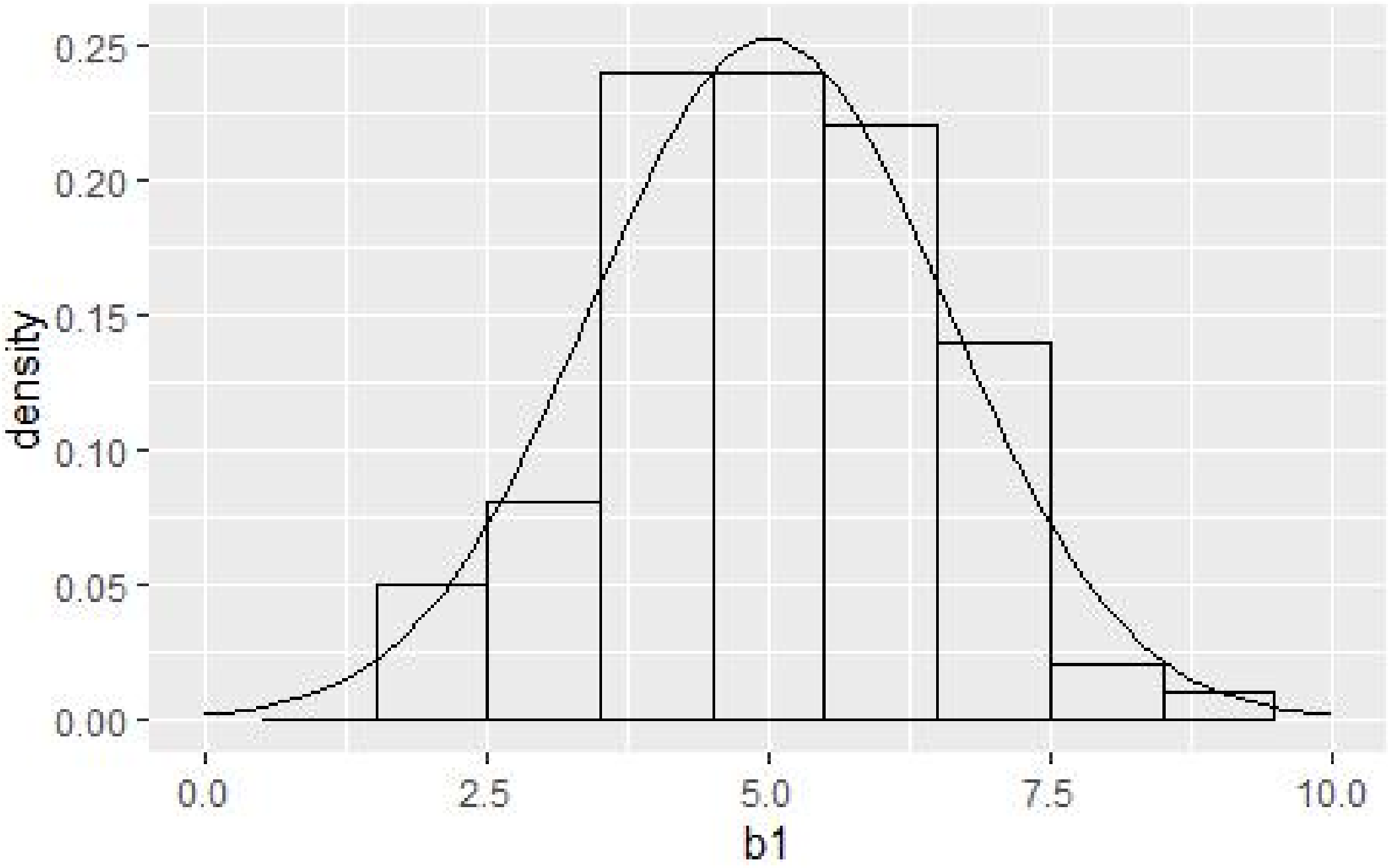
The above plots are showing how changing the probability changes the values obtained in our sample.

4.a) $\int_3^6 \text{density}(x)\,dx$ gives us the area below the curve between $x = 3$ and $x = 6$. We can approximate this by adding the heights of the rectangles in this range.

$$\int_3^6 \text{density}(x)\,dx \sim 0.85 + 0.24 + 0.24 + 0.215$$

$$= 0.78.$$

i.e. there is a 78% chance that the values lie between 3 & 6. $\Rightarrow \int_3^6 \text{density}(x)\,dx = P(3 \le x \le 6)$


0.25
0.20
0.15
0.10
0.05
0.00
density
0.0
2.5
5.0
7.5
10.0
b1

4.b) area under the histogram $= \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$
between 3 & 7.

Originally we have taken,

$\frac{1}{\sqrt{2\pi}\,s} \times e^{-(x-a)^2/2s^2}$ which gives us the curve

and here we have standardized our x.

i.e. $\frac{x-\mu}{\sigma}$ has been substituted for x where μ is the mean & σ is the standard deviation.

$\therefore \left.\frac{x-\mu}{\sigma}\right|_{\text{at } x=3} = a$

$\Rightarrow \frac{3-5}{2.5} = a \Rightarrow a = -0.8$

$\left.\frac{x-\mu}{\sigma}\right|_{\text{at } x=7} = b \Rightarrow \frac{7-5}{2.5} = b$

$\Rightarrow b = 0.8$.

$\therefore a = -0.8, \; b = 0.8$.

(c) Since as calculated before the true mean and variance of b2 are 2.5 and 1.875. When we found the graph for b1, we took a = 5 and $s = \sqrt{2.5}$ which are the true mean and standard deviation for b1. In order to do the same thing for b2 we need to substitute $a = 2.5$, $s = \sqrt{1.875}$.

Similarly for b3 we need to substitute, $a = 7.5$, $s = \sqrt{1.875}$.

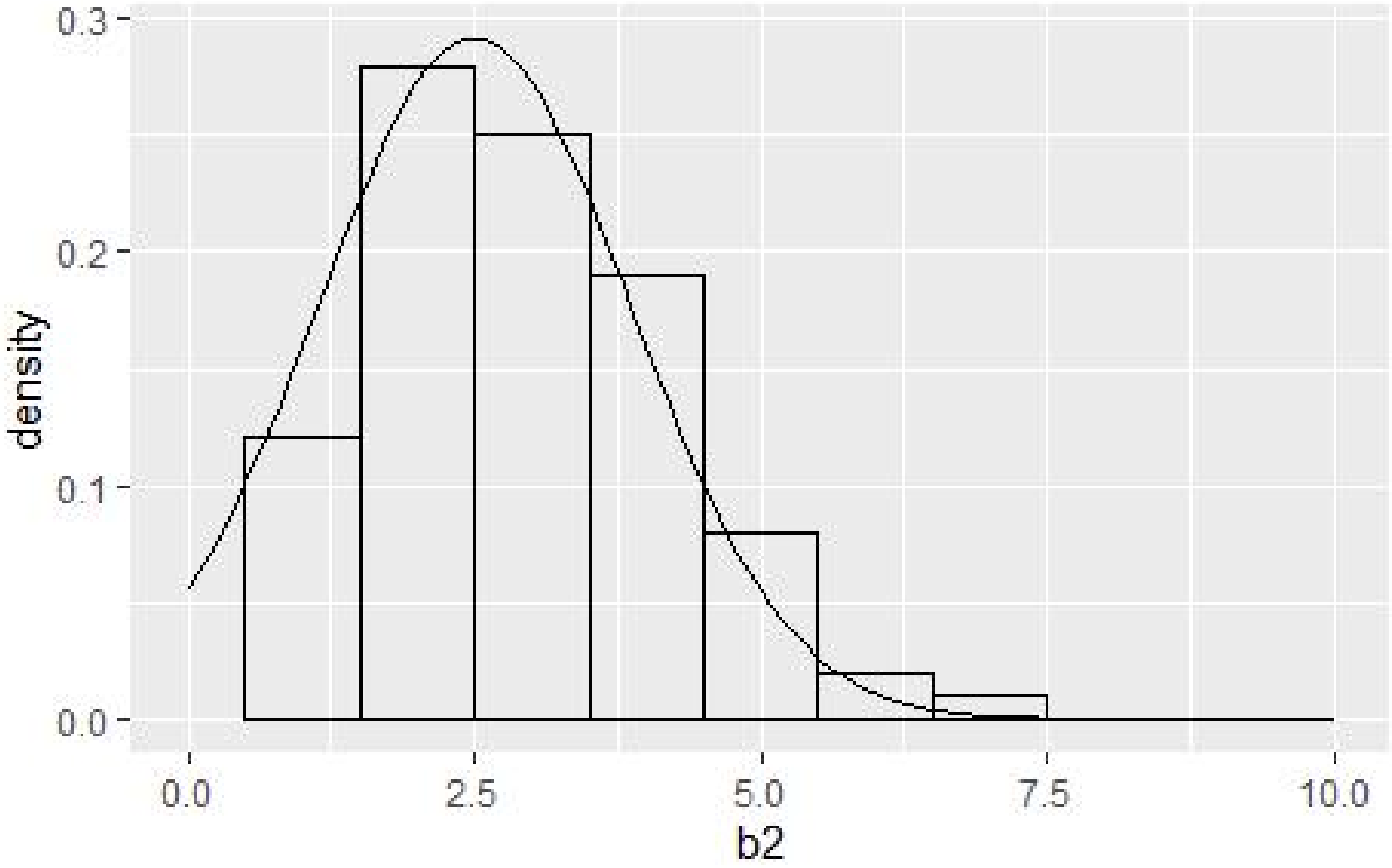
So we have,

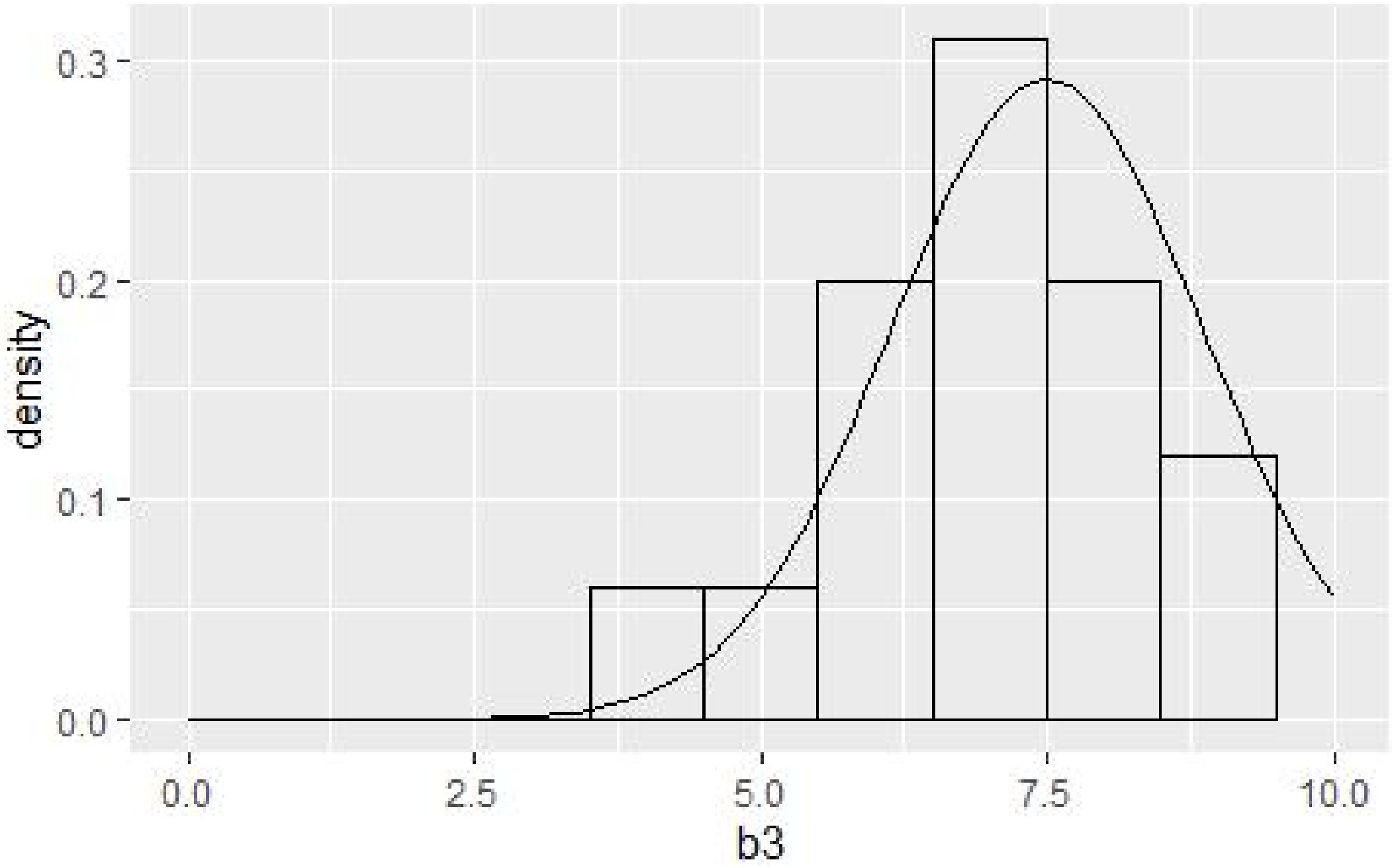
```
> p22 = ggplot(df2) + geom_histogram(mapping = aes(x = b2,
  y = ..density..), color = 'black', fill = 'NA', binwidth = 1)
+ xlim(0, 10) + geom_function(fun = dnorm, args = list(a = 2.5, s = (1.875)^(0.5)))
> p22
```

This graph is skewed towards the left and $\int_3^6 \text{density}(x)\,dx$ for p22 is approximately more than 0.80.

```
> p23 = ggplot(df3) + geom_histogram(mapping = aes(x = b3,
y = ..density..), color = 'black', fill = 'NA', binwidth = 1) +
xlim(0, 10) + geom_function(fun = dnorm, args = list(a = 7.5,
                                    s = (1.875)^(0.5)))
```

This graph is skewed towards the right with most of the area of the graph being covered by values from 6 to 9.


0.3
0.2
0.1
0.0
density
0.0
2.5
5.0
7.5
10.0
b2


0.3
0.2
0.1
0.0
density
0.0
2.5
5.0
7.5
10.0
b3

5.a) ? matrix

matrix creates a matrix from the given data i.e. from Rolls. Rolls is a sample data of size 1500 & matrix command creates a matrix with 5 rows and 300 columns like each element of the matrix coming from Rolls.

? apply

apply commands creates a vector of size 300 which has the row column wise sum of the elements of the matrix Rollsm. For example, the first element of Rollsums is the sum of all the elements in the first column of Rollm = 13

6.a) $\int_{12}^{21}$ density(x, mu, sigma)dx approximates to the area covered by the histogram i.e. it gives us the probability that the values lie between 12 & 21.

i.e. $\int_{12}^{21}$ density(x, mu, sigma)dx = $P(12 \leq x \leq 21)$

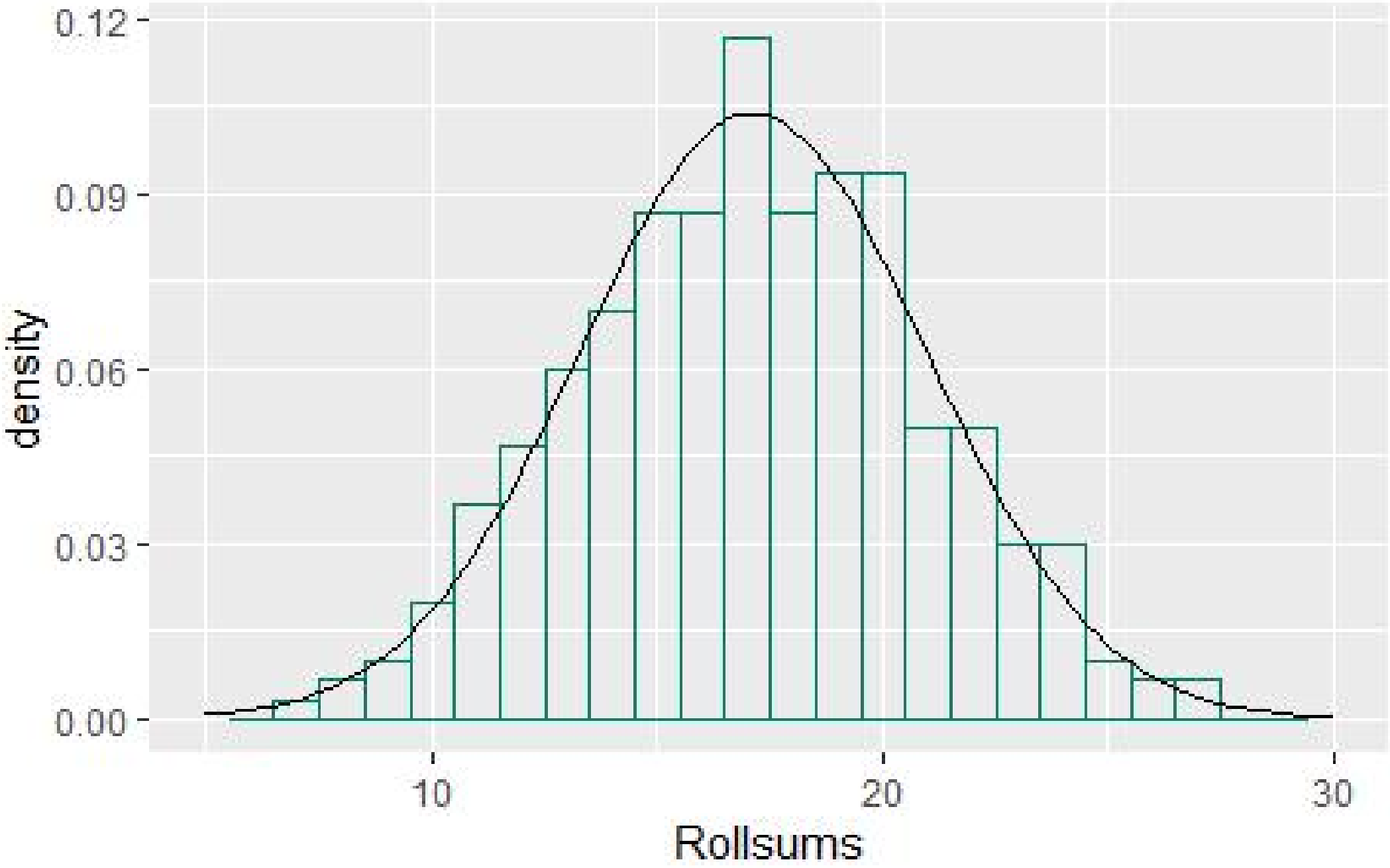
6(b) area under the histogram between 12 and 21 $\approx \int_{a}^{b} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$.

note that, mu = 17.12333

sigma = 3.845006

$x \approx \frac{x - mu}{sigma}$ ⇒ $a = \frac{12 - 17.12333}{3.845006} \approx -1.33$

$b = \frac{21 - 17.12333}{3.845006} \approx 1.01$


0.12
0.09
0.06
0.03
0.00
density
10
20
30
Rollsums