Probability: Random Variables in R

In this document we show how to find probabilities of Binomial and Normal random variables in R.

Binomial Random Variable

We start with this example: *Suppose in one hospital 10 babies are born each day. The probability that a new born baby is a Boy is 50%. We are interested in the distribution of the number of boys born each day in this hospital.*
Let

Y = Number of Boys born on a day

We know that distribution of Y is Binomial with $n = 10$ and $p = 0.50$

$$Y \sim Bin(10, 0.5)$$

Let's find 'Probability Mass Function' for this random variable. First we construct a variable 'x' that stores values $0, 1, 2 \ldots 10$ (since these values random variable Y can take). We can do it conveniently using 'seq' function ('by' argument specifies distance between adjacent values):

```
x=seq(from=0,to=10,by=1)
x
```

```
 [1]  0  1  2  3  4  5  6  7  8  9 10
```

To calculate probabilities of 'Binomial' random variable we use 'dbinom' function ('size' argument determines the number of trials and 'prob' probability of success)

```
prob=dbinom (x,size=10,prob=0.5)
```

We can round these probabilities to 3 decimal places using 'round' function, and put 'x' and 'prob' variables together using 'rbind' to get a nice table

```
rbind(x,round(prob,3))
```

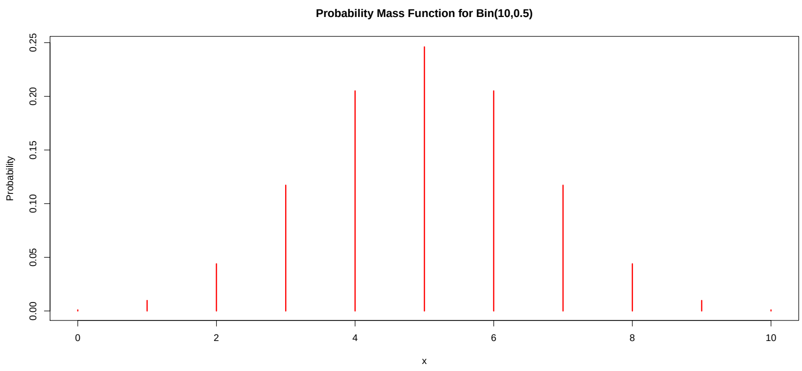
```
   [,1] [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11]
x 0.000 1.00 2.000 3.000 4.000 5.000 6.000 7.000 8.000  9.00 1e+01
  0.001 0.01 0.044 0.117 0.205 0.246 0.205 0.117 0.044  0.01 1e-03
```

From this table we can say for example that probability that 7 boys are born on a given day is 0.117:

$$P(Y = 7) = 0.117$$

Next we want to plot 'Probability Mass Function', which is just 'prob' variable versus 'x':

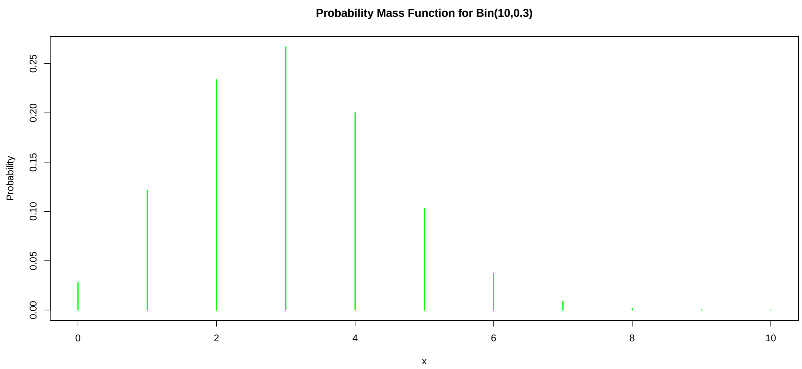
```
plot(x,prob,ylab='Probability',type='h',col='red',lwd=2,
  main='Probability Mass Function for Bin(10,0.5)')
```



Here 'type='h" produces vertical bars and 'lwd' specifies the width of these bars.

If the random variable Y had $Bin(10, 0.3)$ then 'Probability Mass Function' would look like that:

```
x=seq(from=0,to=10,by=1)
prob=dbinom (x,size=10,prob=0.3)
plot(x,prob,ylab='Probability',type='h',col='green',lwd=2,
  main='Probability Mass Function for Bin(10,0.3)')
```



Observe that distribution is no longer symmetric but shifted to the left.

Suppose for the $Y \sim Bin(10, 0.3)$ we want to find probability of being less than or equal to 2. Then we can just add probabilities:

$$P(Y \leq 2) = P(Y = 0) + P(Y = 1) + P(Y = 2)$$

```
dbinom(0,size=10,prob=0.3) + dbinom(1,size=10,prob=0.3) +
  dbinom(2,size=10,prob=0.3)
```

```
[1] 0.3827828
```

So $P(Y \leq 2) \approx 0.3828$.
Instead of finding this probability by adding, it is much more convenient to use 'pbinom' function that calculates probability of being less than or equal of any value:

```
pbinom(2,size=10,prob=0.3)
```

```
[1] 0.3827828
```

Now we immediately get that $P(Y \leq 2) \approx 0.3828$ without adding probabilities.

Normal Random Variable

Consider the next problem:

Suppose the birth weights of the infants born at a hospital follow a Normal distribution with mean of 3700 g and a standard deviation of 350 g.

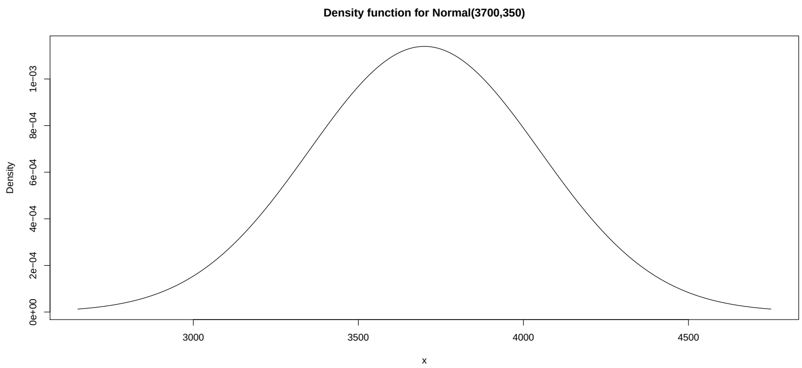
We will work out the answers to these questions:
1. What percent of infants born at the hospital weigh less than 3000 grams?
2. What percent of infants weigh between 4000 and 4500 grams?
3. What birth weight is the first quartile?

Let X be a birth weight of an infant then we know:

$$X \sim N(3700, 350)$$

Which means X has a normal distribution with mean 3700g and standard deviation 350g.
Let's first plot the density of this normal using 'dnorm' function. First we create a variable 'x' that stores values from 2650 (mean − 3 standard deviations) to 4750 (mean + 3 standard deviations), then find density of $N(3700, 350)$ evaluated at 'x' and finally make a plot:

```
x=seq(from=2650,to=4750,by=1)
y=dnorm(x,mean=3700,sd=350)
plot(x,y,ylab='Density',type='l',
  main='Density function for Normal(3700,350)')
```



This has a familiar bell shaped curve centered at 3700. Now let's look at the questions:

1. What percent of infants born at the hospital weigh less than 3000 grams?
In this question we need to find $P(X < 3000)$. We can easily find this quantity using 'pnorm' function.

```
pnorm(q=3000, mean=3700, sd=350)
```

```
[1] 0.02275013
```

So we have the answer:

$$P(X < 3000) \approx 0.0228$$

Equivalently we can first standardize 3000 by subtracting the mean and divide by standard deviation, and then find probability for **Standard** Normal random variable to be less than this standardized number:

```
z=(x-3700)/350
pnorm(z, mean=0, sd=1)
```

```
[1] 0.02275013
```

Exactly the same probability!

2. What percent of infants weigh between 4000 and 4500 grams?
Here we need $P(4000 \leq X < 4500)$. Observe that we can find this probability by calculating $P(X < 4000)$ and $P(X < 4500)$ since

$$P(4000 \leq X < 4500) = P(X < 4500) - P(X < 4000)$$

We can find these two probabilities as in the first question.

```
pnorm(4500, mean=3700, sd=350) - pnorm(4000, mean=3700, sd=350)
```

```
[1] 0.1845475
```

Hence:

$$P(4000 \leq X < 4500) \approx 0.1846$$

3. What birth weight is the first quartile?
Here we have an opposite problem, we need x such that $P(X < x) = 0.25$. This value is found from 'qnorm' function.

```
qnorm(p=0.25, mean=3700, sd=350)
```

```
[1] 3463.929
```

So we get

$$P(X < 3463.929) = 0.25$$

And hence the first quartile is about 3464g.
Equivalently we can first use 'qnorm' for standard normal and then multiply the result by standard deviation and add the mean:

```
qnorm(p=0.25, mean=0, sd=1) * 350 +3700
```

```
[1] 3463.929
```

Summary of R Functions

We give a short summary of all new and/or important R functions [and arguments] that we used in this Module:

Probabilities

```
dbinom() [x,size,prob]
pbinom() [q,size,prob]
dnorm() [x,mean,sd]
pnorm() [q,mean,sd]
qnorm() [p,mean,sd]
```

Miscellaneous

```
rbind()
round()
seq() [from,to,by]
```