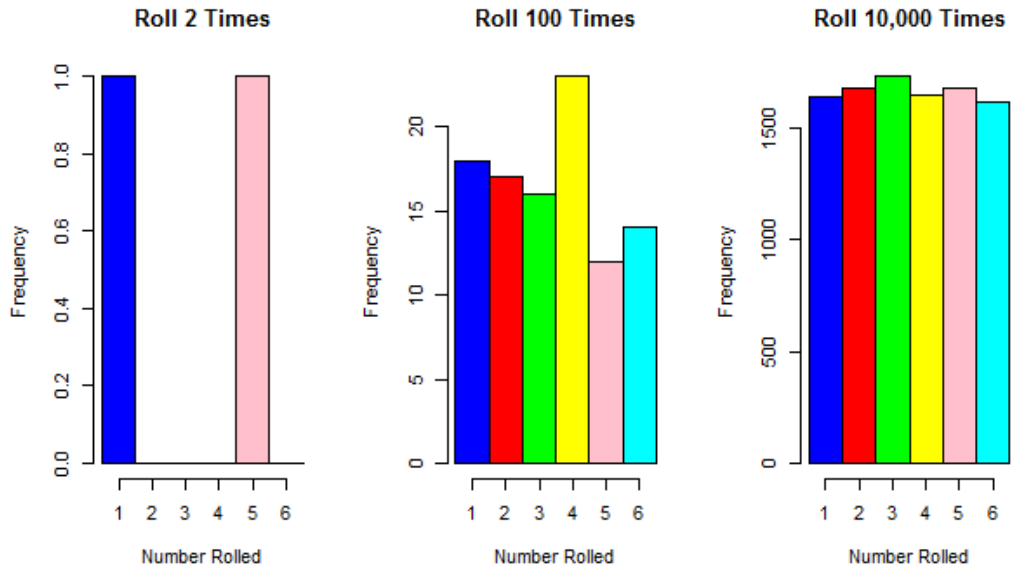
1. Show (with the use of histograms) the distribution of sample values generated using uniform distribution on integers that model rolling a dice multiple times. Is there a difference in the sample's distribution if we roll 2 times, 100 times or 10000 times? Try to explain the results.]

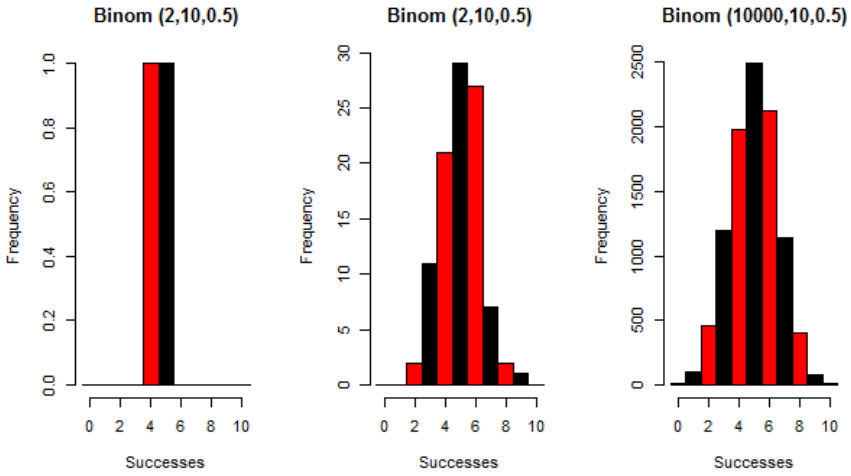
```
> cols = c("blue", "red", "green", "yellow", "pink", "cyan")
> par(mfrow = c(1,3))
>
> hist(as.integer(runif(2,1,7)), col = cols, breaks=seq(0.5,6.5,1),
+      xlab="Number Rolled", main ="Roll 2 Times")
> hist(as.integer(runif(100,1,7)), col = cols, breaks=seq(0.5,6.5,1),
+      xlab="Number Rolled", main ="Roll 100 Times")
> hist(as.integer(runif(10000,1,7)), col = cols, breaks=seq(0.5,6.5,1),
+      xlab="Number Rolled", main ="Roll 10,000 Times")
```



There is a difference in the samples. When we roll only twice at most 2 numbers can have been rolled – as with 1 and 5 above. This does not look anything like the uniform distribution, which resembles a rectangle, that we expect when we roll dice. In the second diagram we have rolled a 100 times are a bit closer to the uniform distribution- with 4 a significant outlier having been rolled much more than expected. Finally in third diagram we have rolled 10,000 times and we are very close to the rectangle we expect for the uniform distribution.

2. Perform a similar analysis for the binomial distribution - check how the histogram of a sample changes if you perform different number of experiments. Let's say we are interested in outcomes from 10 trials, with probability p = 0.5.

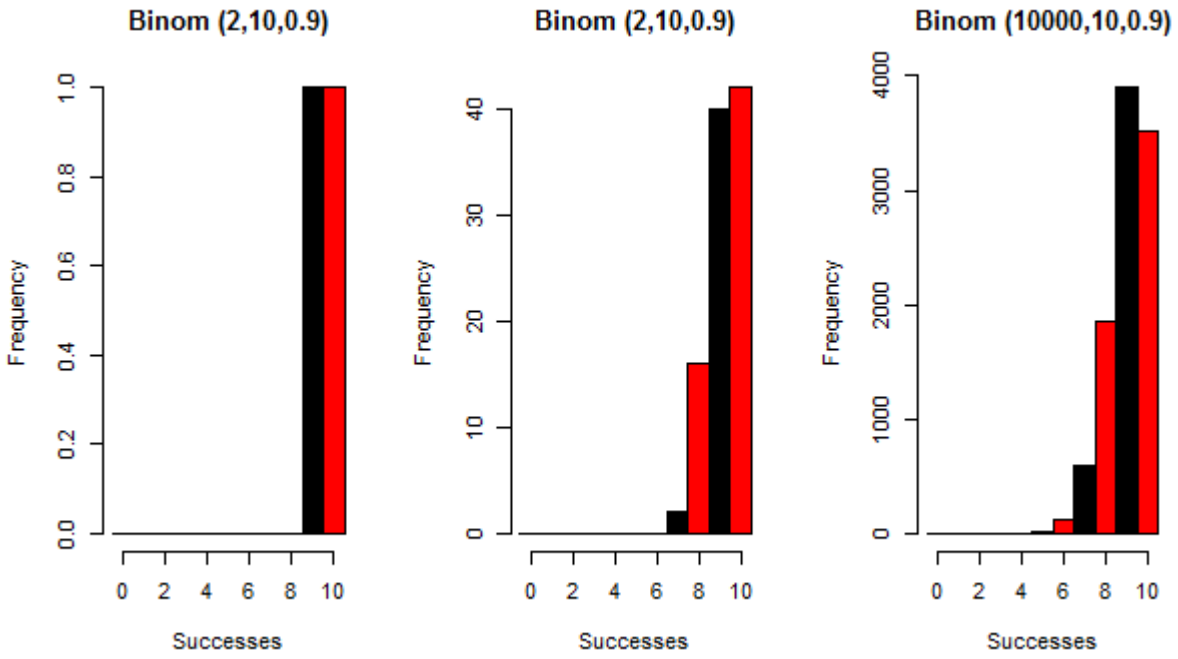
```
> cols = c("red", "black")
> hist(rbinom(2,10,0.5), breaks=seq(-0.5,10.5,1), xlab= "Successes",
+      main= "Binom (2,10,0.5)", col = cols)
> hist(rbinom(100,10,0.5), breaks=seq(-0.5,10.5,1), xlab= "Successes",
+      main= "Binom (2,10,0.5)", col = cols)
> hist(rbinom(10000,10,0.5), breaks=seq(-0.5,10.5,1), xlab= "Successes",
+      main= "Binom (10000,10,0.5)", col = cols)
```



Similar to Q1 with 2 experiments we can only get 1 or two outcomes – here 4 and 5 successes. With 100 experiments we see that 5 is the most frequent outcome - as we would expect since the mean is n*p = 10*0.5= 5 - and we have seen our outcomes expand to a value as low as 1 success and as high as 9 successes. However as we move away from our mean the columns should be symmetric – match red to red and black to black. This is no the case in the second diagram. With 10,000 trials we have a distribution that is more symmetric but interestingly we still have not registered either 0 or 10 successes. If this was a coin toss that would be 0 or 10 successes would be equivalent to tossing all heads or all tails. But the first toss has to come up heads or tails and now you only need 9 more and (1/2)^9 is 1 in 512 times.

3. Check what is the impact of the probability p. Try different values for the same number of trials, e.g. 10 and a high number of experiments. Show different situations and comment on them.

```
> hist(rbinom(2,10,0.9), breaks=seq(-0.5,10.5,1), xlab= "Successes",
+      main= "Binom (2,10,0.9)", col = cols)
> hist(rbinom(100,10,0.9), breaks=seq(-0.5,10.5,1), xlab= "Successes",
+      main= "Binom (2,10,0.9)", col = cols)
> hist(rbinom(10000,10,0.9), breaks=seq(-0.5,10.5,1), xlab= "Successes",
+      main= "Binom (10000,10,0.9)", col = cols)
```

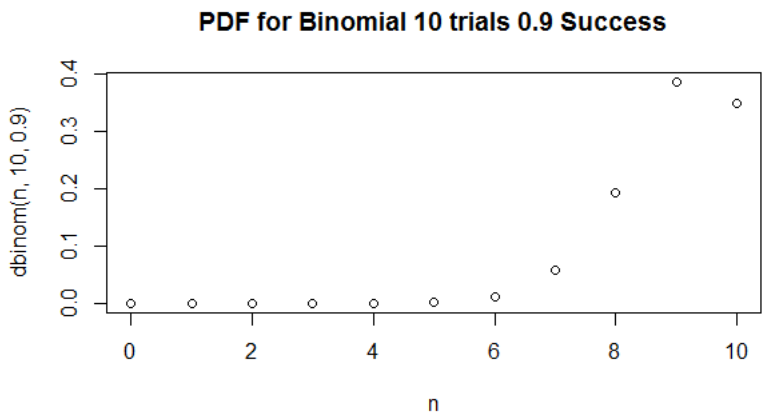


Again we only have two outcomes but because of the high probability of success we expect them to be on the high end. This is an interesting distribution. All success has over 40% of the outcomes with 9 successes having almost another 40%. The variance for binomial is:

$$n*p*(1-p) = 10*0.9*0.1 = 0.9$$

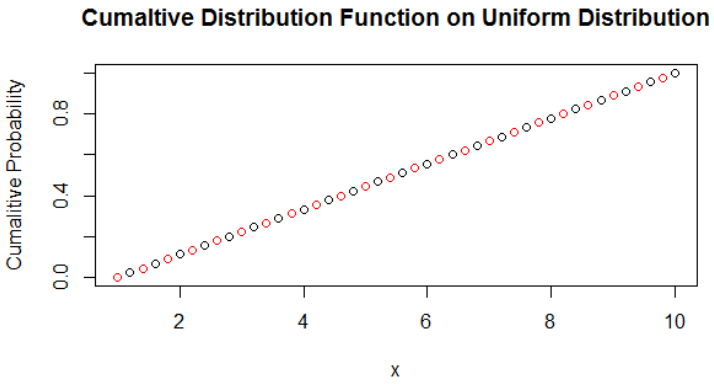
So we have a standard deviation of just 0.3 which explains the small spread in values from 7 to 10 successes.

We can check below what the density function for a binomial looks like. Our 10,000 experiments looks very much like this.



4. punif function returns a value that represents the cumulative probability at a given point. Plot the CDF in the interval from 1 to 10.

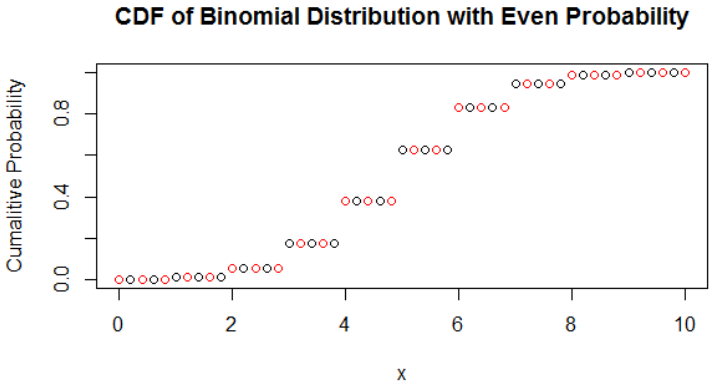
```
x <- seq(1,10,0.2)
plot(x, punif(x, 1, 10), pch=1,
col = c("red", "black"),
ylab="Cumalitive Probability",
xlim=c(1,10), main="Cumaltive
Distribution Function on Uniform
Distribution")
```



This is a straight line which we would expect since PDF is the derivative of the CDF. Since the PDF of the uniform distribution is a constant and the only functions with constant derivatives are lines then the CDF must be a line.

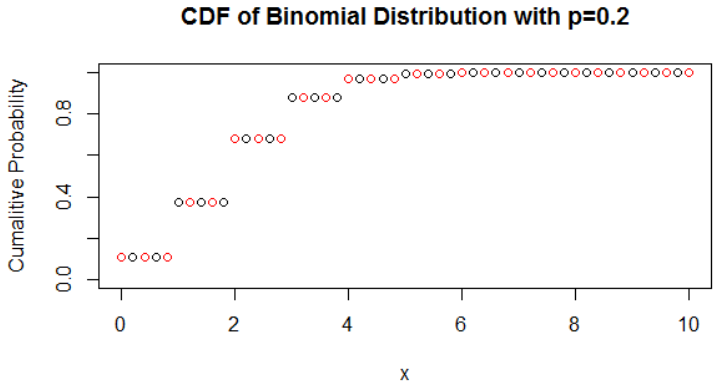
5. Using a similar function for the binomial distribution (use 10 trials and p = 0.5), plot the CDF. How does this function change if you change the probability p?

```
> x <- seq(0,10,0.2)
plot(x, pbinom(x, 10, 0.5), pch=1,
col = c("red", "black"),
     ylab="Cumalitive
Probability", xlim=c(0,10),
     main="CDF of Binomial
Distribution with Even
Probability",
     ylim=c(0,1))
```



Here we see the values move horizontally until a new integer is reached and then it jumps to the next step. With p=0.5 the jumps are correlated with their closeness to 5 successes with barely noticeable jumps from 0-1 and 9-10, which are furthest from 5 in this example.

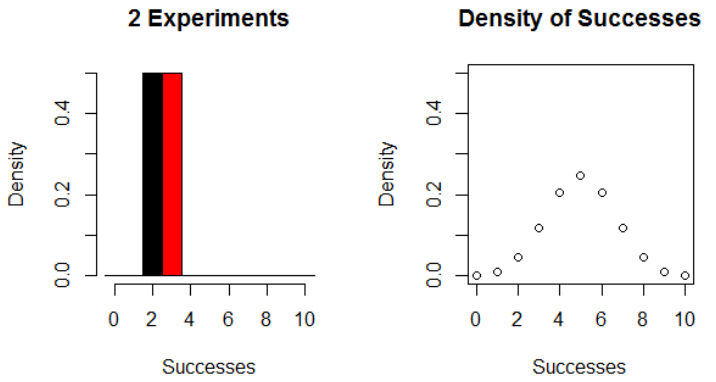
```
> plot(x, pbinom(x, 10, 0.2),
pch=1, col = c("red", "black"),
+      ylab="Cumalitive
Probability", xlim=c(0,10),
+      main="CDF of Binomial
Distribution with p=0.2",
+      ylim=c(0,1))
```



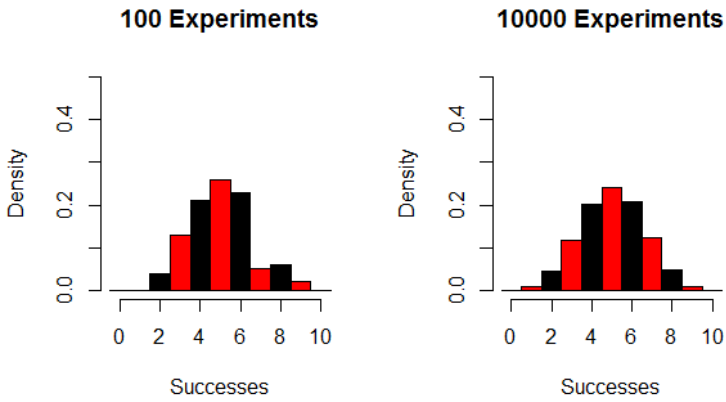
To reshow the last point I present the CDF when p=0.2. The jumps in probability occur around the mean with very little activity in the right half of the plot.

6. Use the dbinom function to check if probabilities of each possible outcome correspond to the histogram that you have got in 2). Do not check absolute values, check the behaviour.

```
> x = 0:10
> par(mfrow=c(1,2))
> hist(rbinom(2,10,0.5)/2, breaks=seq(-0.5,10.5,1), xlab= "Successes",
+      main="2 Experiments", freq=F,
+      col = c("black", "red"))
> plot(x, dbinom(x,10,0.5), ylim=c(0,0.5), main="Density of Successes",
+      xlab="Successes", ylab="Density")
>
> par(mfrow=c(1,2))
> hist(rbinom(100,10,0.5), breaks=seq(-0.5,10.5,1), xlab= "Successes",
+      col = c("black", "red"), ylim=c(0,0.5), freq=F,
+      main = "100 Experiments")
> hist(rbinom(10000,10,0.5), breaks=seq(-0.5,10.5,1), xlab= "Successes",
+      col = c("black", "red"), ylim=c(0,0.5), freq=F,
+      main = "10000 Experiments")
```

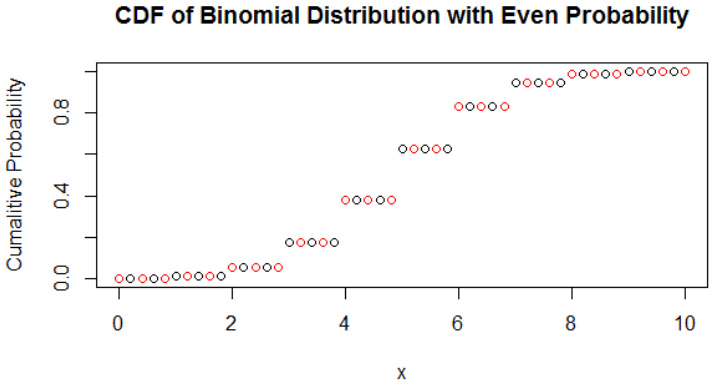


Obviously 2 experiments cannot approximate the distribution as their are insufficient points. 100 experiments looks much closer to the required shape but here we have a very flat right tail. 10,000 experiments does an excellent job of approximating the distribution.



7. We have a binomial distribution with p = 0.5 and 10 trials. Check the values you would get for three quantiles: 0.4, 0.5 and 0.6. Can you explain why all of these return the same value? You can answer this question if you check these values at the CDF.

```
> pbinom(0.4,10,0.5)
[1] 0.0009765625
> pbinom(0.5,10,0.5)
[1] 0.0009765625
> pbinom(0.6,10,0.5)
[1] 0.0009765625
```



The answer can be found in the horizontal lines in the CDF plot. The binomial dustribution is discrete and only has values at natural numbers from 0 to n where n is the number of trials. Numbers other than the natural numbers take the value of the previous natural number. So in the example above the three examples return the value of:

```
> pbinom(0,10,0.5)
[1] 0.0009765625
```