Central Limit Theorem

The Central Limit Theorem

The Central Limit Theorem says that: When drawing with replacement from a box, the probability histogram for the sum (and the average) will follow the normal curve, even if the contents of the box do not. The histogram must be put into standard units, and the number of draws must be reasonably large.

For a sum, the mean and standard deviation of this normal curve are:

$$\text{EV}_{\text{sum}} = n \cdot \text{AV}_{\text{box}}$$

and

$$\text{SE}_{\text{sum}} = \sqrt{n} \cdot \text{SD}_{\text{box}}$$

For an average, the mean and standard deviation of this normal curve are:

$$\text{EV}_{\text{av}} = \text{AV}_{\text{box}}$$

and

$$\text{SE}_{\text{av}} = \frac{\text{SD}_{\text{box}}}{\sqrt{n}}$$

The Central Limit Theorem is one reason that the normal curve plays such a central role in this course.

Function Definitions

If you look in the Rmd file, you will see some functions defined in this section. They are used later in these notes to run simulations and make plots but are not printed here.

Here is a function for computing the *population standard deviation* of a box. It has n in the denominator rather than $n - 1$. The *sample standard deviation* uses $n - 1$.

```
SD <- function(box) {
  n = length(box)
  return (sd(box) * sqrt((n-1)/n))
}
```

Flipping Coins: and the Normal Curve

Flipping coins and counting the number of heads is like taking the sum of a random sample (with replacement) from a box with two tickets 1 (for heads) and 0 (for tails).

Five Coins: Sum of the Number of Heads Follows a Normal Curve

The following simulates flipping five coins 100 times.

```
data = boxSimulation(c(0, 1), 5, 100)
data
```

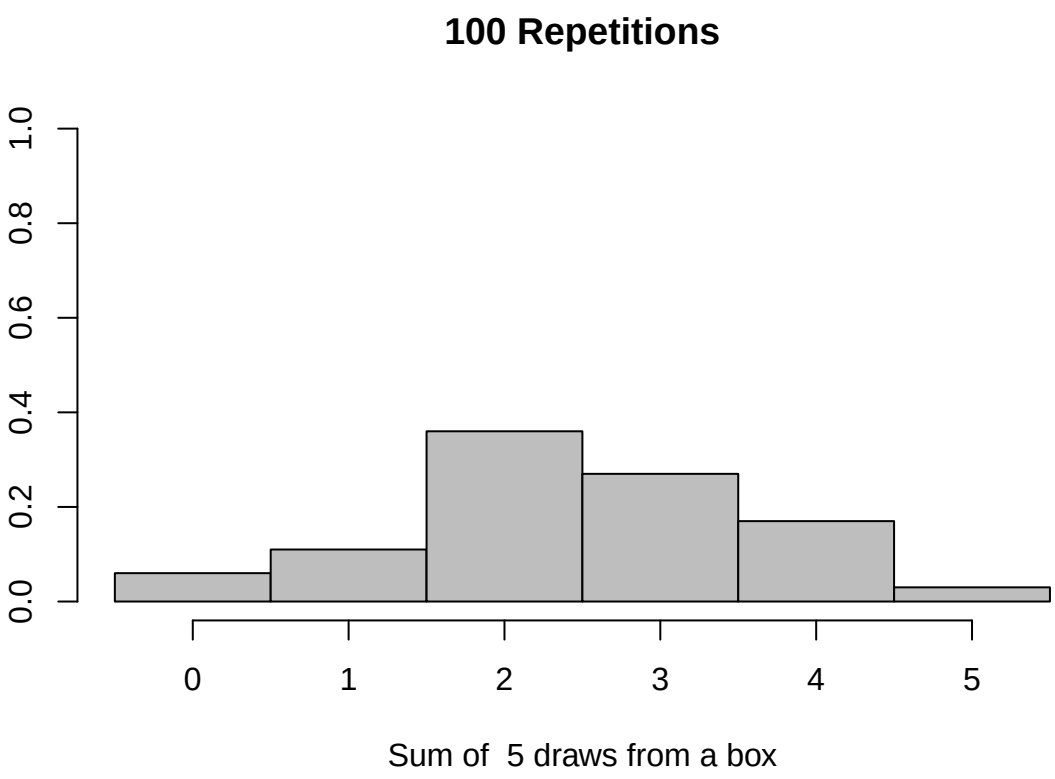
```
##   [1] 2 0 3 3 3 3 3 3 2 1 4 3 3 5 3 3 3 2 3 2 1 2 3 1 3 2 0 2 2 0 2 2 5 3 3
##  [36] 3 3 3 3 4 1 1 1 3 2 4 2 2 3 3 4 3 3 0 4 1 3 1 2 4 2 2 1 1 1 2 2 4 4 3
##  [71] 1 2 4 3 2 4 3 4 2 3 3 4 3 3 2 1 2 2 4 2 3 1 3 4 2 2 3 4 2 5
```

```
table(data)
```

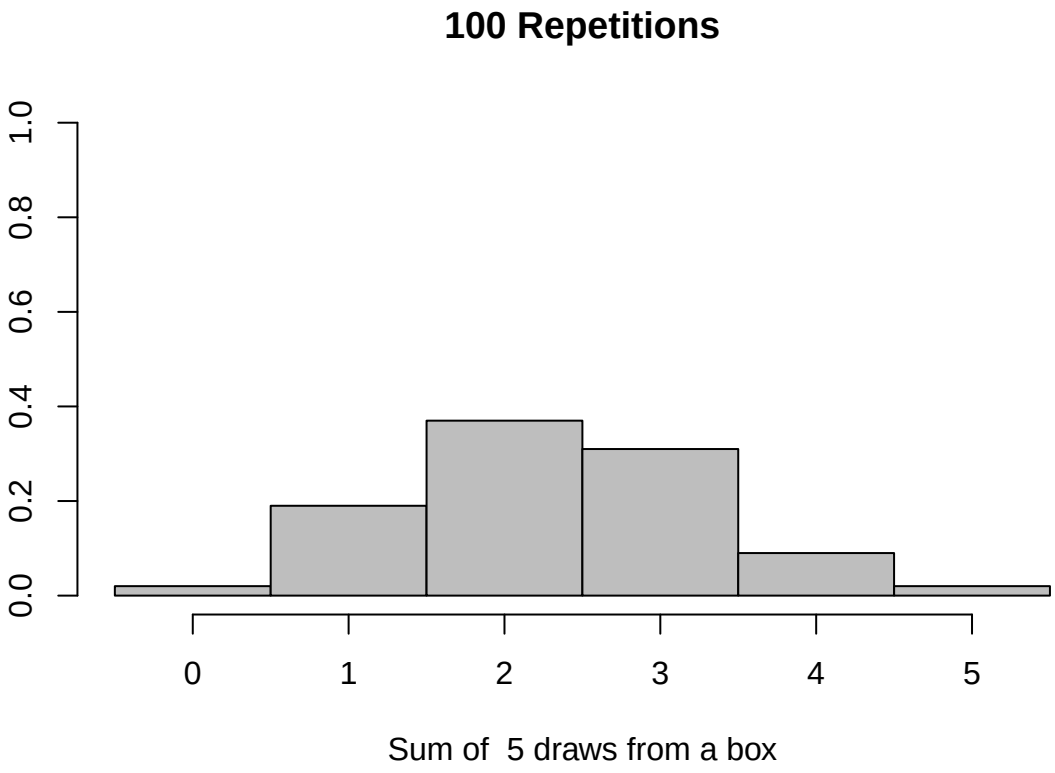
```
## data
##  0  1  2  3  4  5
##  4 14 28 36 15  3
```

Here are histograms of repetitions of this experiment:

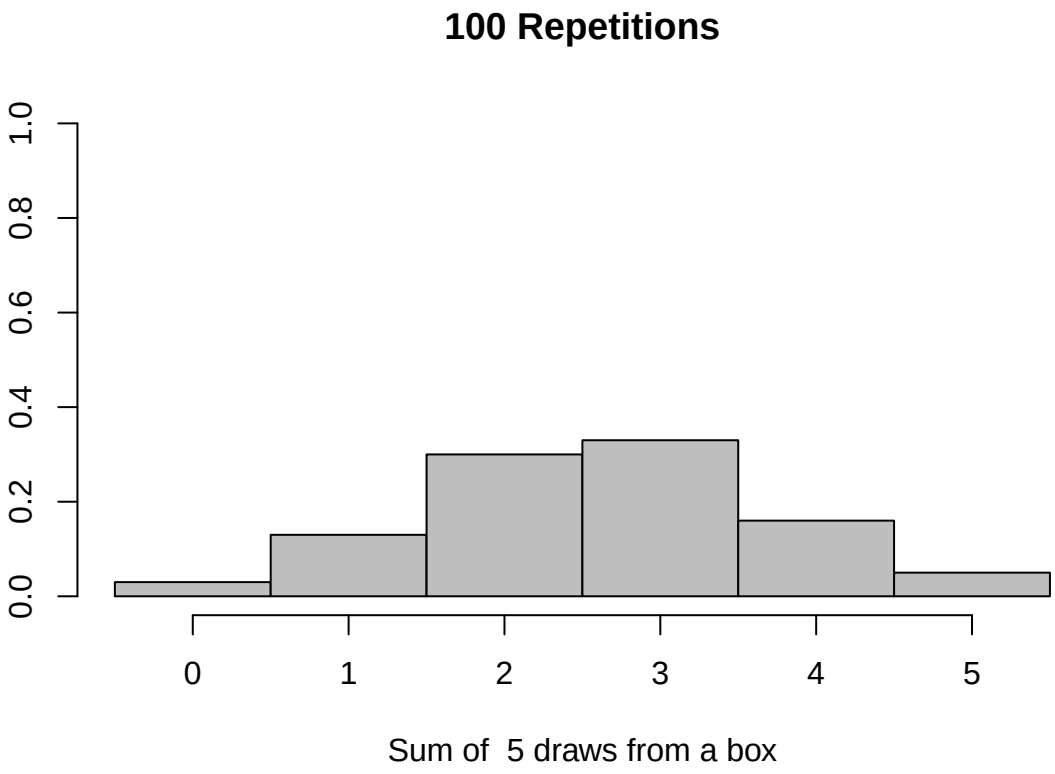
```
boxHistogram(c(0,1), 5, 100, breaks=-0.5+(0:6))
```



```
boxHistogram(c(0,1), 5, 100, breaks=-0.5+(0:6))
```

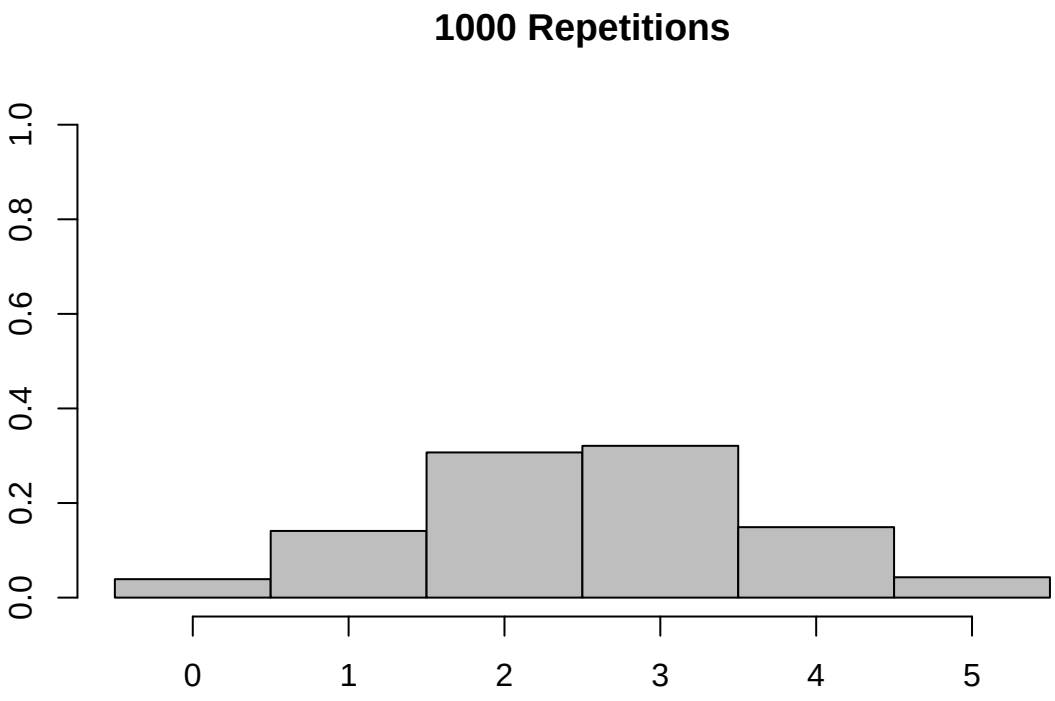



```
boxHistogram(c(0,1), 5, 100, breaks=-0.5+(0:6))
```

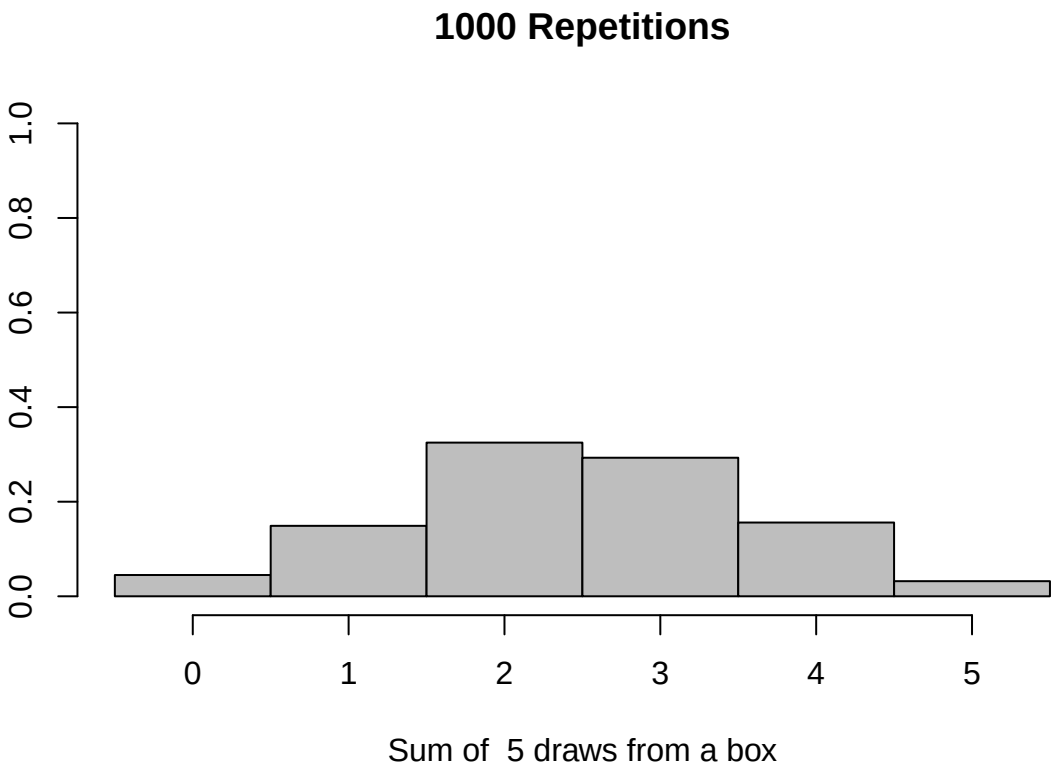


It looks a bit like a normal curve but changes from one sample to another. Let's flip five coins 1000 times.

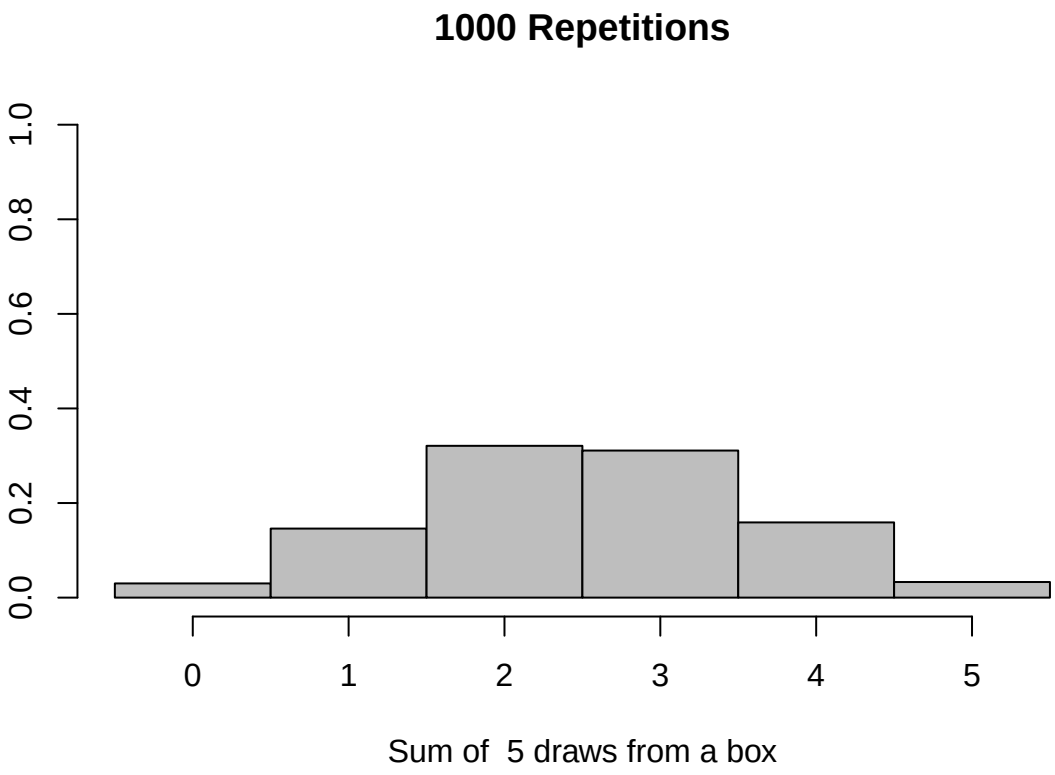
```
boxHistogram(c(0,1), 5, 1000, breaks=-0.5+(0:6))
```




```
boxHistogram(c(0,1), 5, 1000, breaks=-0.5+(0:6))
```



```
boxHistogram(c(0,1), 5, 1000, breaks=-0.5+(0:6))
```



Ten Coins: Sum of the Number of Heads Follows a Normal Curve

The following simulates flipping five coins 100 times.

```
data = boxSimulation(c(0, 1), 10, 100)
data
```

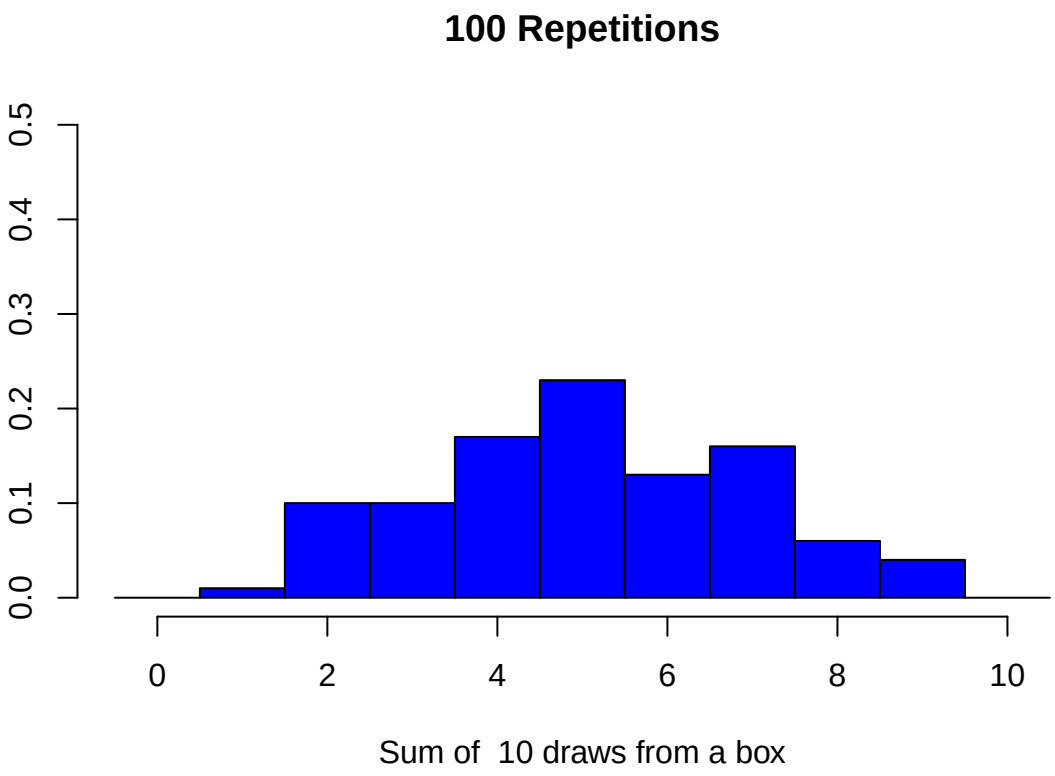
```
##   [1] 4 5 6 6 7 8 5 5 5 4 4 4 4 5 7 4 7 5 5 7 4 4 6 6 8 7 4 4 4 7 4 4 4 5 3
##  [36] 5 6 5 5 4 6 2 4 7 7 7 6 4 4 4 3 4 2 5 5 4 3 2 4 2 6 4 5 5 5 7 5 3 7 6
##  [71] 4 4 5 3 9 7 6 4 4 4 4 5 3 6 4 7 3 6 6 6 7 5 3 5 4 8 2 3 7 4
```

```
table(data)
```

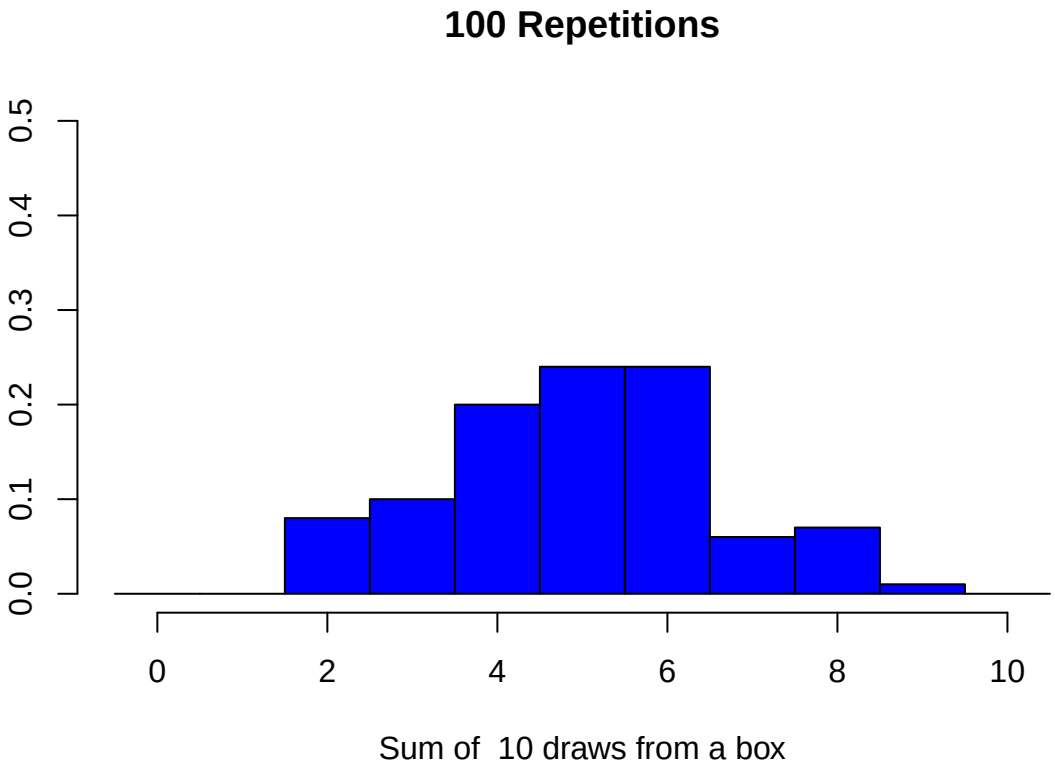
```
## data
##  2  3  4  5  6  7  8  9
##  5  9 32 21 14 15  3  1
```

Here are histograms of repetitions of this experiment:

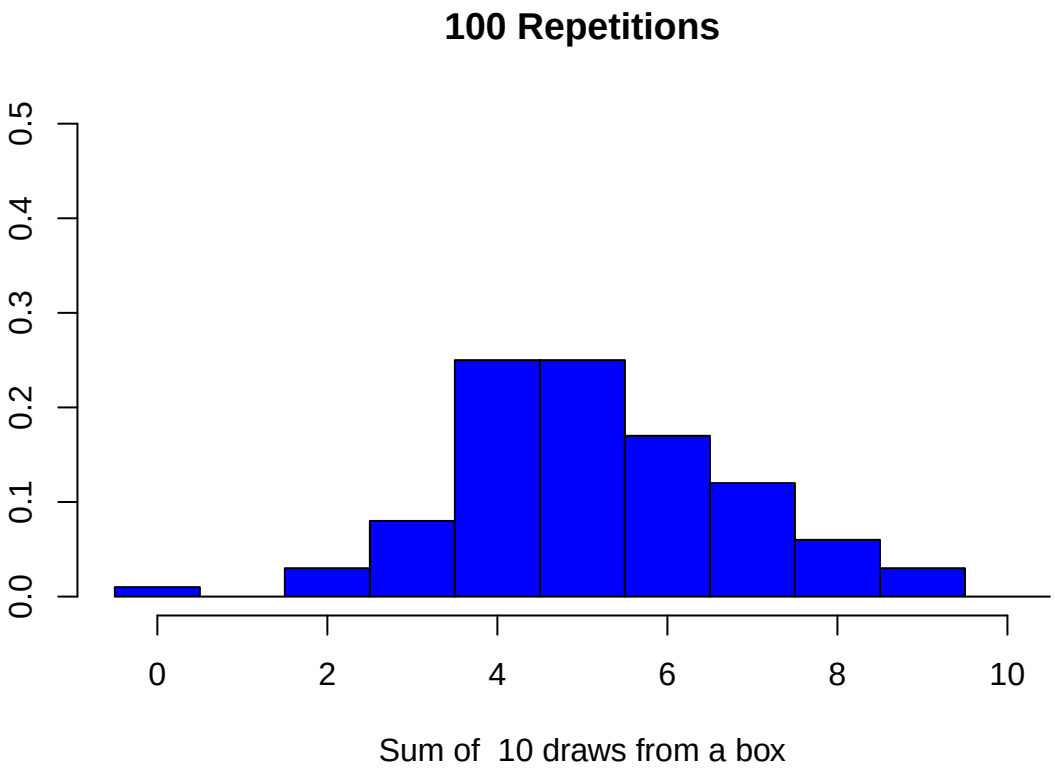
```
ymax=0.5
boxHistogram(c(0,1), 10, 100, breaks=-0.5+(0:11), ymax=ymax, col='blue')
```



```
boxHistogram(c(0,1), 10, 100, breaks=-0.5+(0:11), ymax=ymax, col='blue')
```

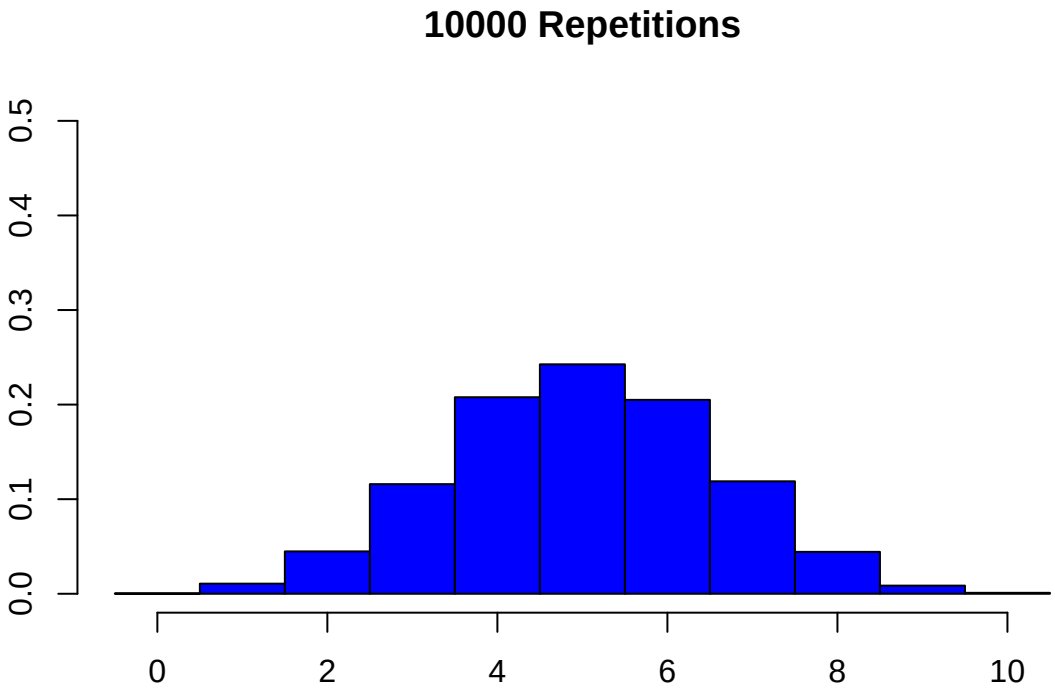



```
boxHistogram(c(0,1), 10, 100, breaks=-0.5+(0:11), ymax=ymax, col='blue')
```

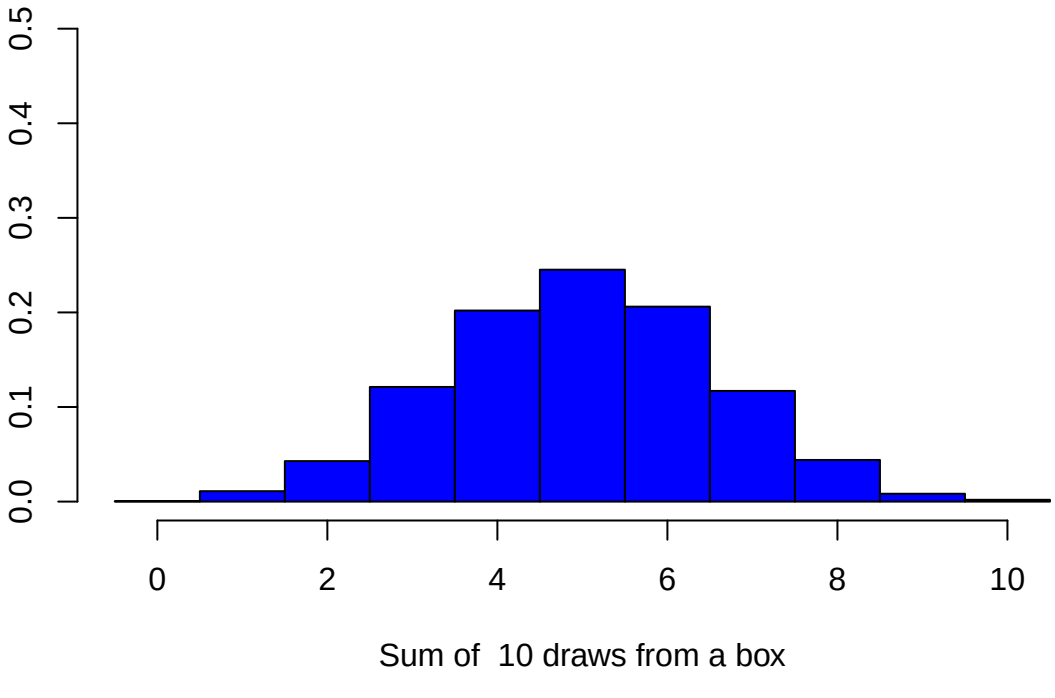


It looks a bit like a normal curve but changes from one sample to another. Let's flip 10 coins 10,000 times.

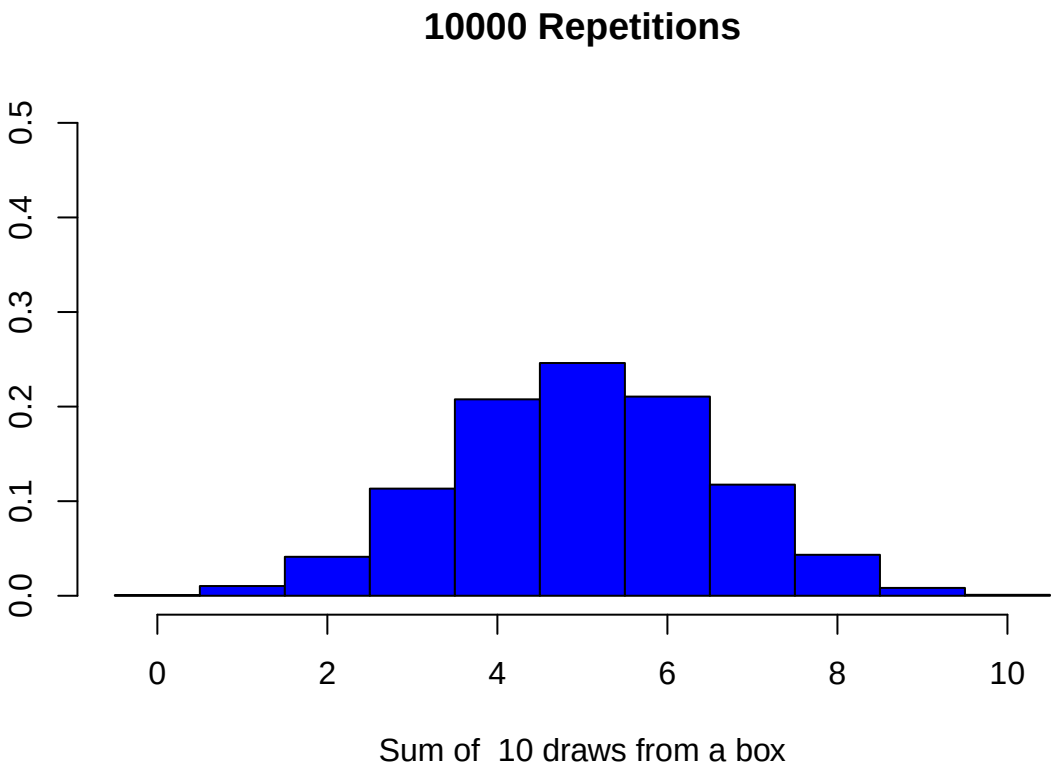
```
boxHistogram(c(0,1), 10, 10000, breaks=-0.5+(0:11), ymax=ymax, col='blue')
```




```
boxHistogram(c(0,1), 10, 10000, breaks=-0.5+(0:11), ymax=ymax, col='blue')
```



```
boxHistogram(c(0,1), 10, 10000, breaks=-0.5+(0:11), ymax=ymax, col='blue')
```



Rolling Dice

We can simulate rolling dice and taking the sum)of the numbers shown.

Two Dice: Sum of the Number of Dots Follows a Normal Curve

We could compute the frequencies using the 6 x 6 table of possible outcomes.

```
data = diceSimulation(100)
data
```

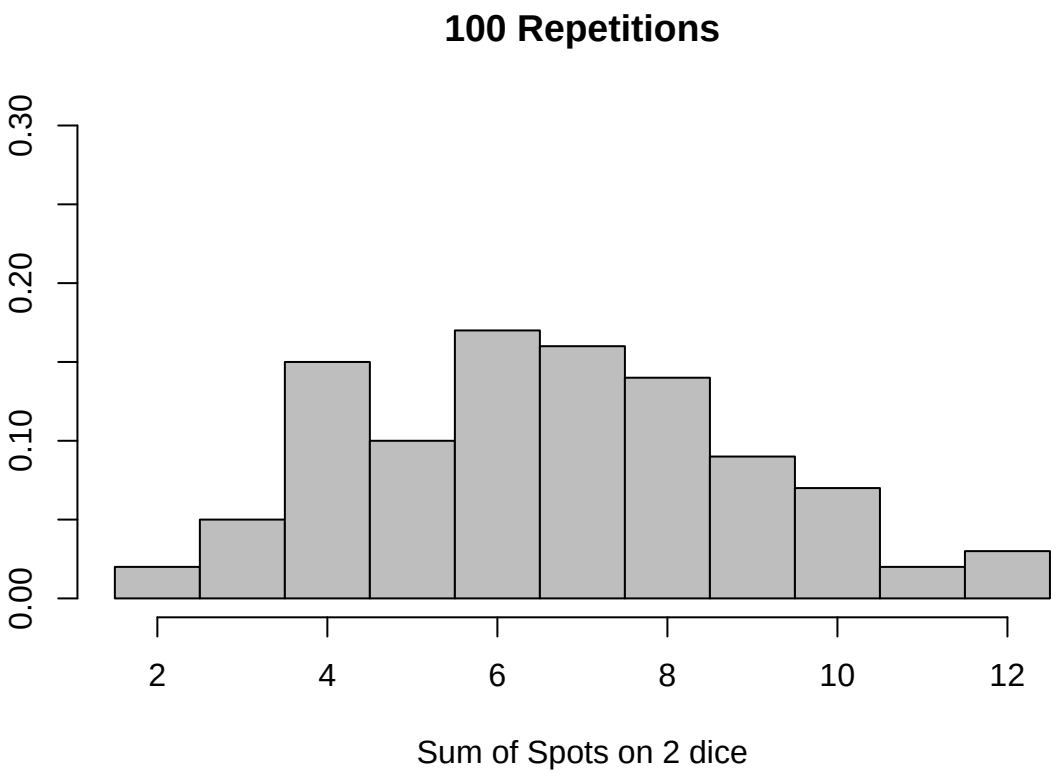
```
##   [1]  7  7 11  6  7 12  3  3  5  7  7  8  2  6  8 10  6  5  4  6  7 10 12
##  [24]  5 10  7  3  5  7  6 10  4  4 11 10  9 10  7 10 11  9 12 10  9 10  6
##  [47]  9  6 10  3  4  7  8  5  8 10 11  2  6  9 11  6  4  7  8 10  8 10 11
##  [70]  7  6  9  4  9  3  3  7  8  5  6  6  7 11 11  7  4  5  3  9 12  7  5
##  [93]  3 10 10  3 10  2  5  8
```

```
table(data)
```

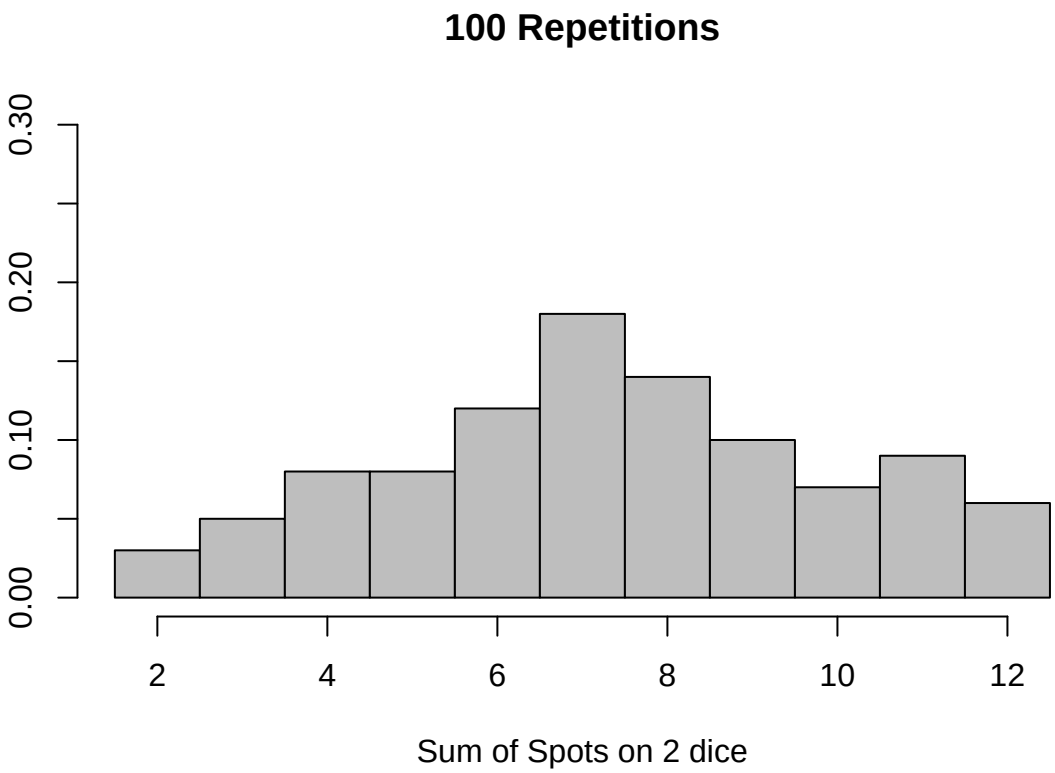
```
## data
##  2  3  4  5  6  7  8  9 10 11 12
##  3  9  7  9 12 16  8  8 16  8  4
```

Let's look at histograms of 100 trials taking the sum of two dice.

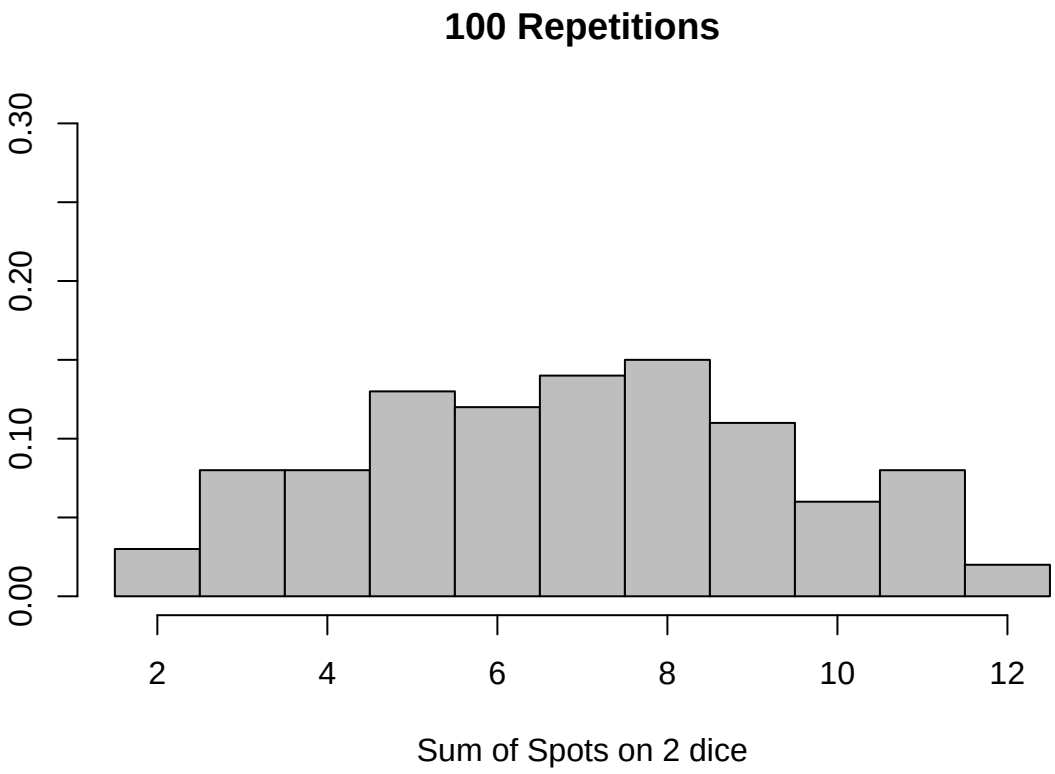
```
diceHistogram(100)
```



```
diceHistogram(100)
```



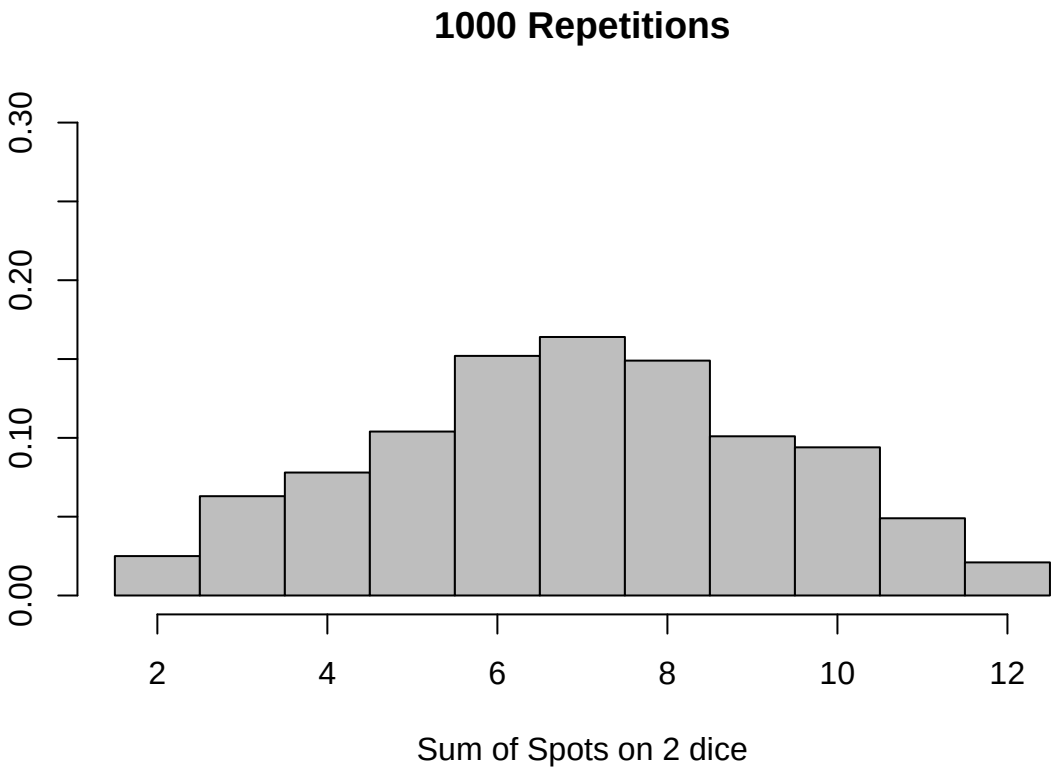

```
diceHistogram(100)
```



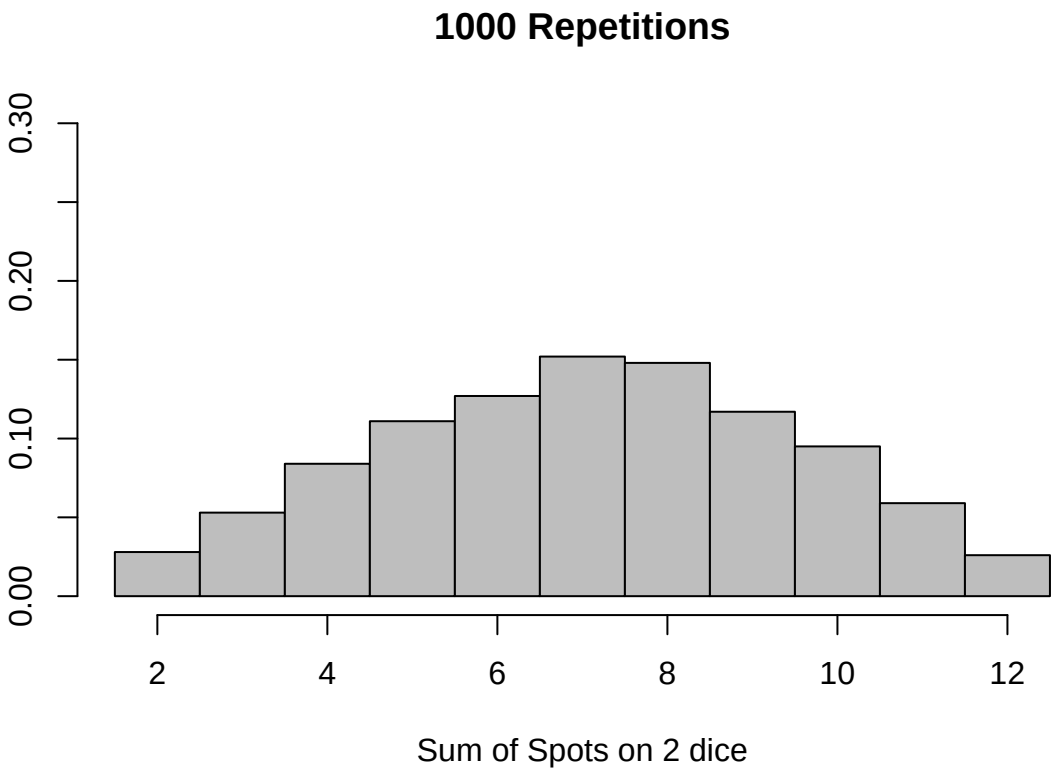
The results vary a bit each time we run 100 trials. If we run 1000 trials, the results will be closer to the

expected values.

```
diceHistogram(1000)
```



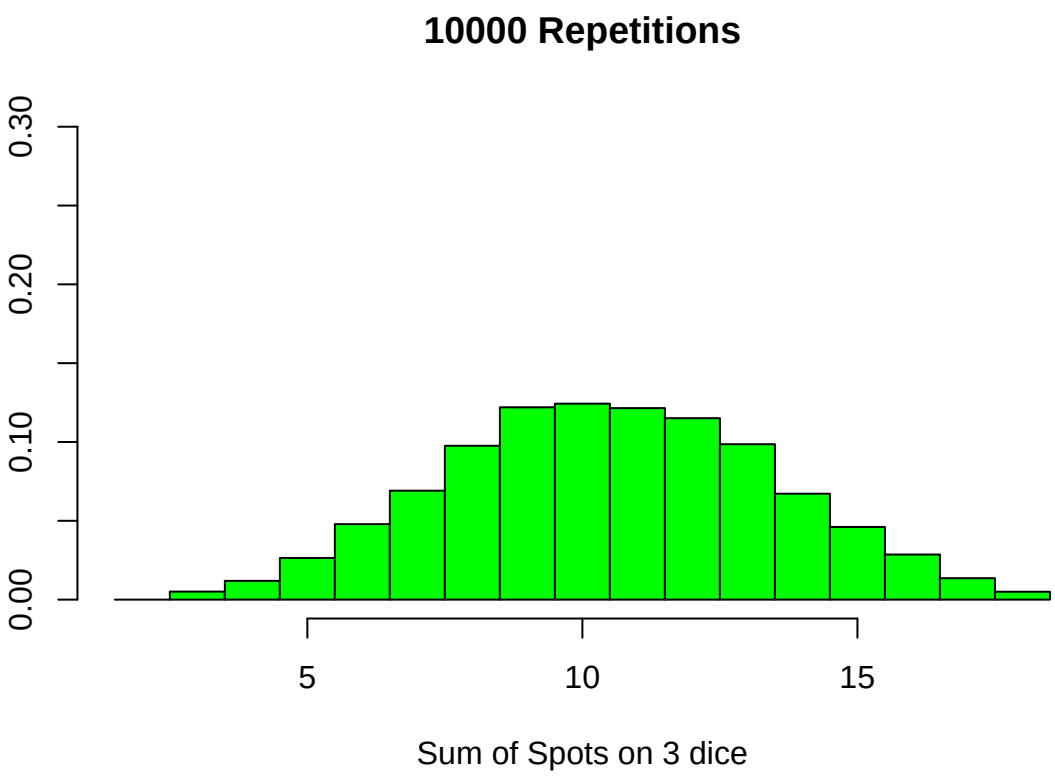

```
diceHistogram(1000)
```



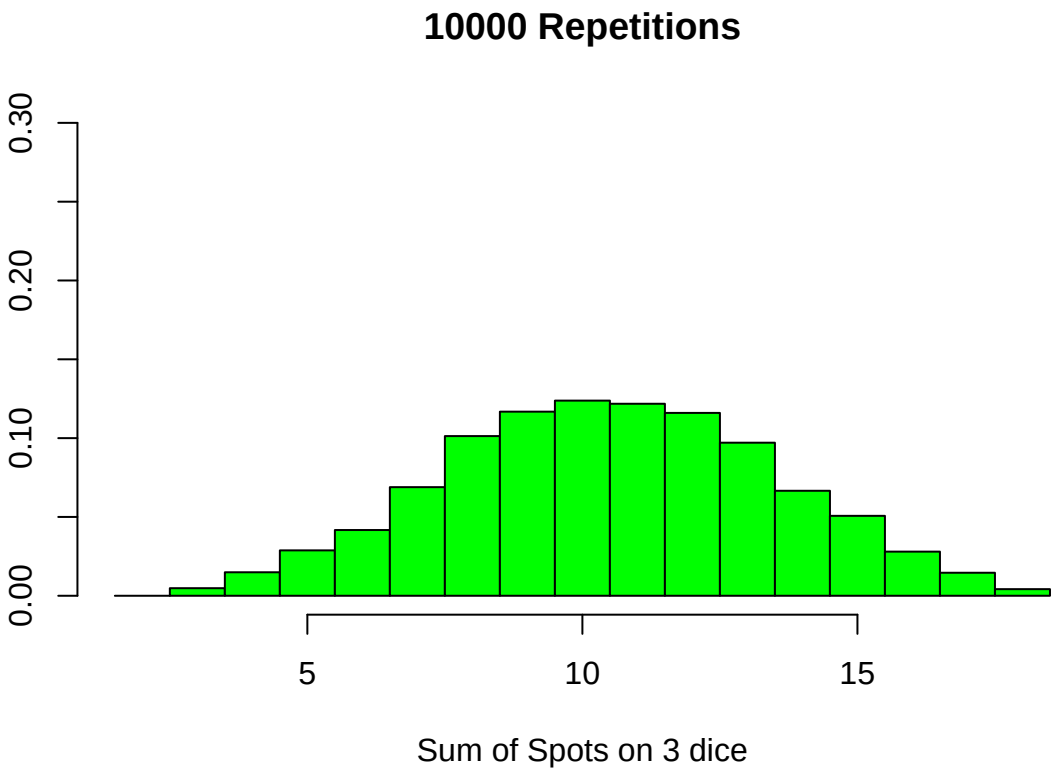
Three Dice: Sum of the Number of Sots Follows a Normal Curve

Let's try the experiment with three dice.

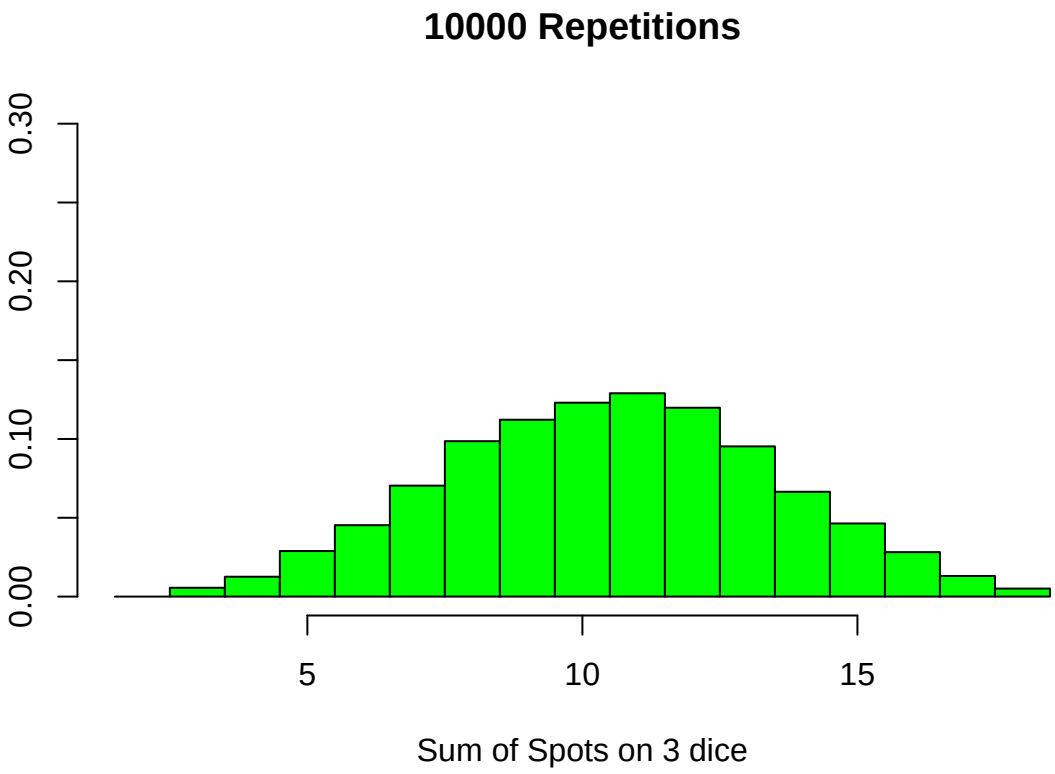
```
diceHistogram(10000, dice=3, color="green")
```



```
diceHistogram(10000, dice=3, color="green")
```




```
diceHistogram(10000, dice=3, color="green")
```

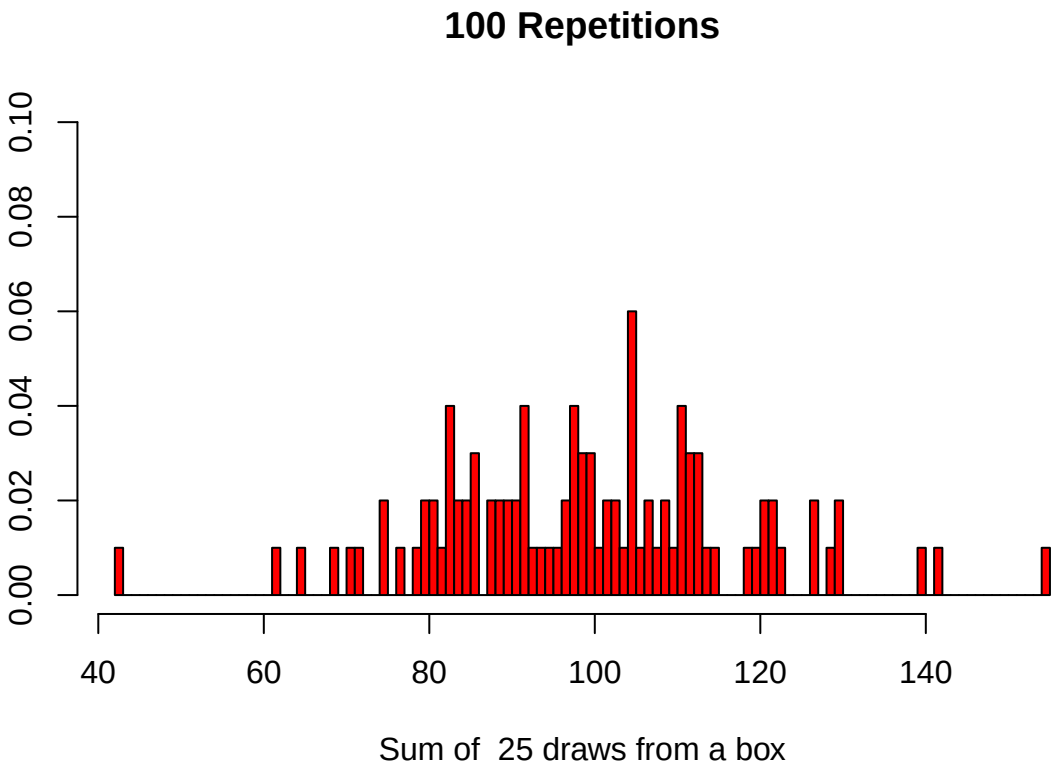


Sampling from a Box.

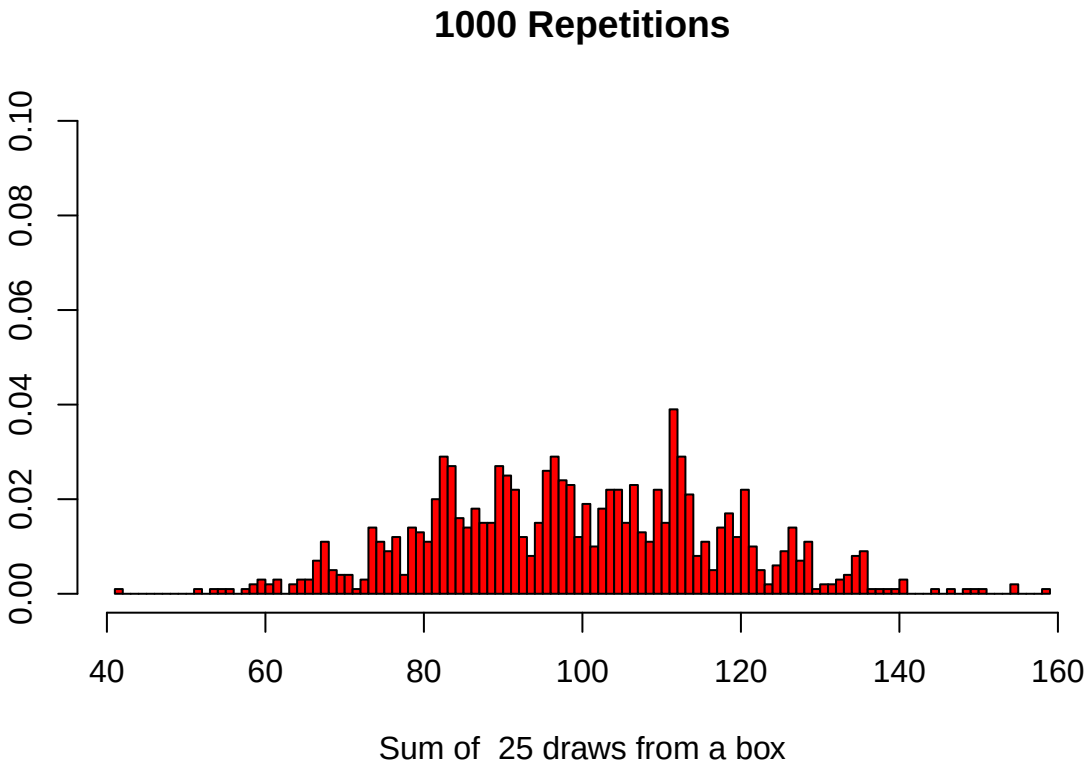
Some times it can take a very large sample size before the probability histogram follows a normal curve very closely.

Sample of Size 25 from {1, 2, 9}

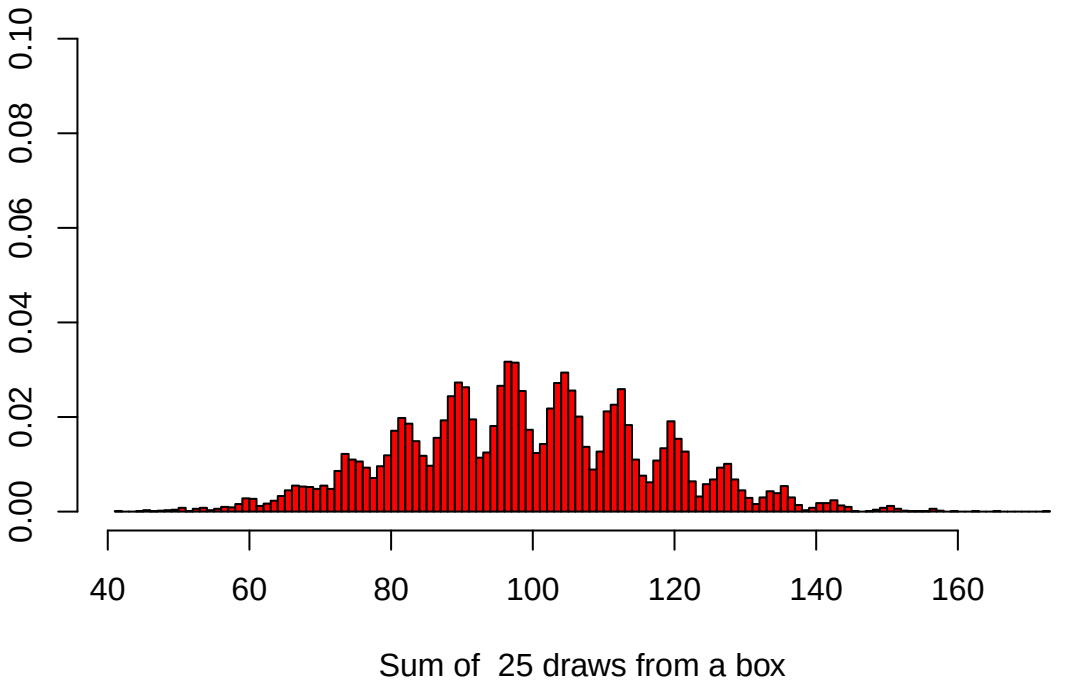
```
boxHistogram(c(1,2,9), 25, 100, ymax=0.1, breaks=160, col='red')
```



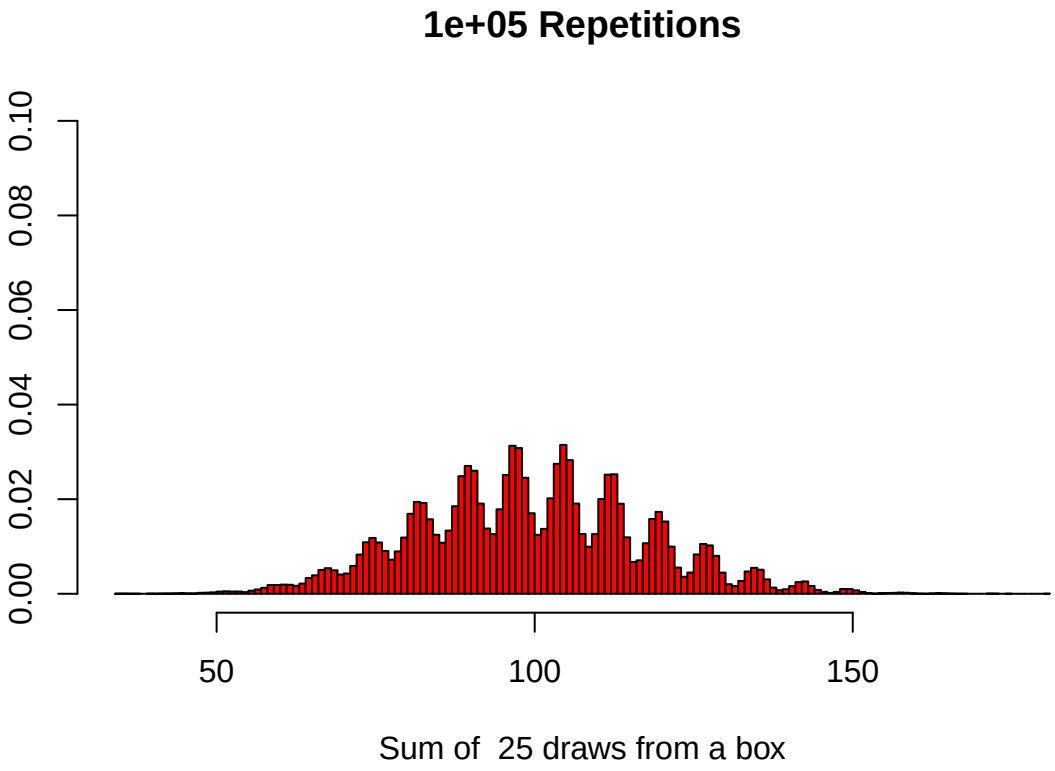
```
boxHistogram(c(1,2,9), 25, 1000, ymax=0.1, breaks=160, col='red')
```




```
boxHistogram(c(1,2,9), 25, 10000, ymax=0.1, breaks=160, col='red')
```

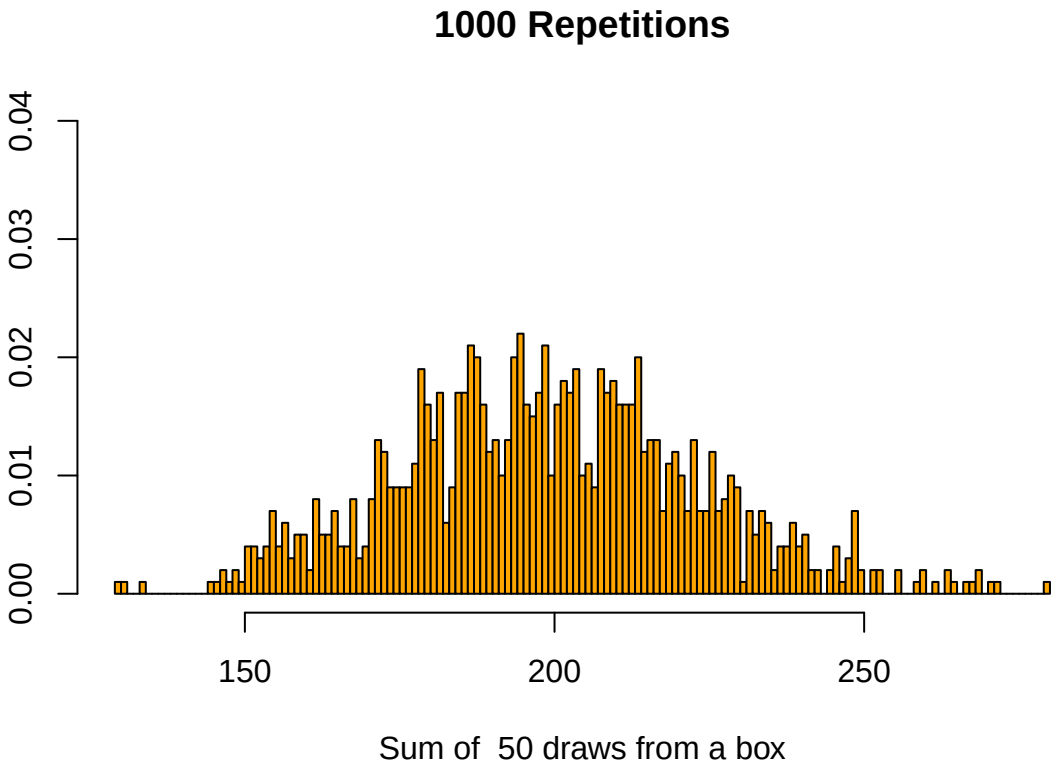


```
boxHistogram(c(1,2,9), 25, 100000, ymax=0.1, breaks=160, col='red')
```

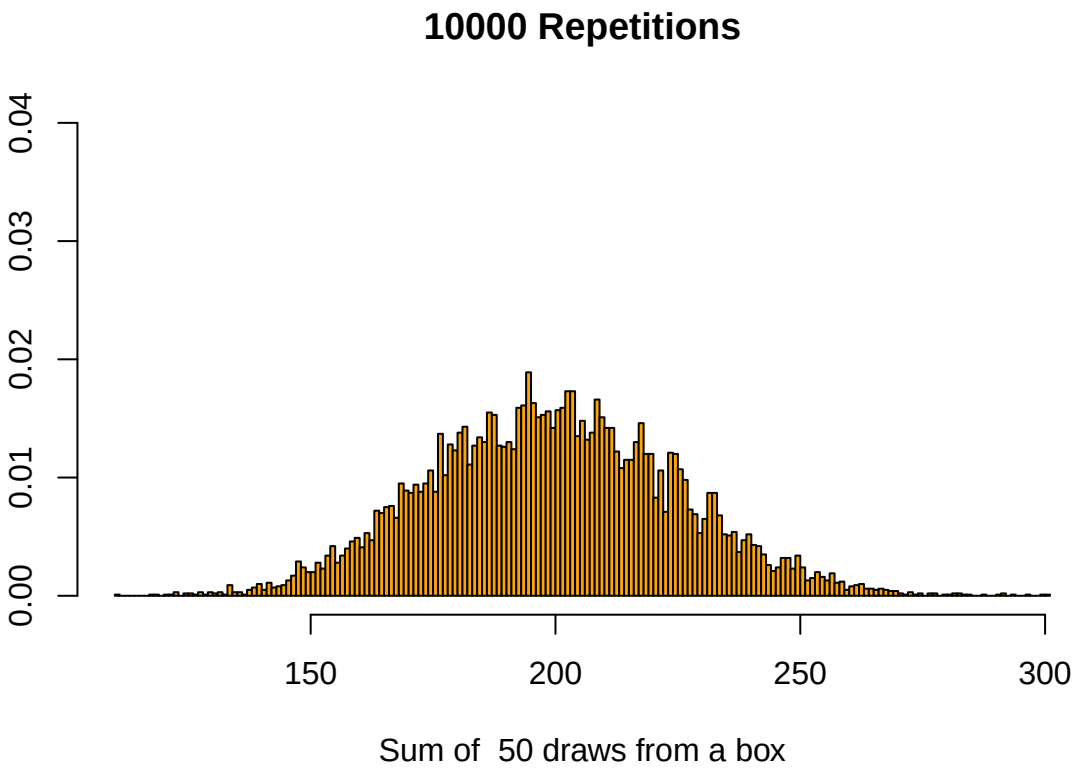


Sample of Size 50 from {1, 2, 9}

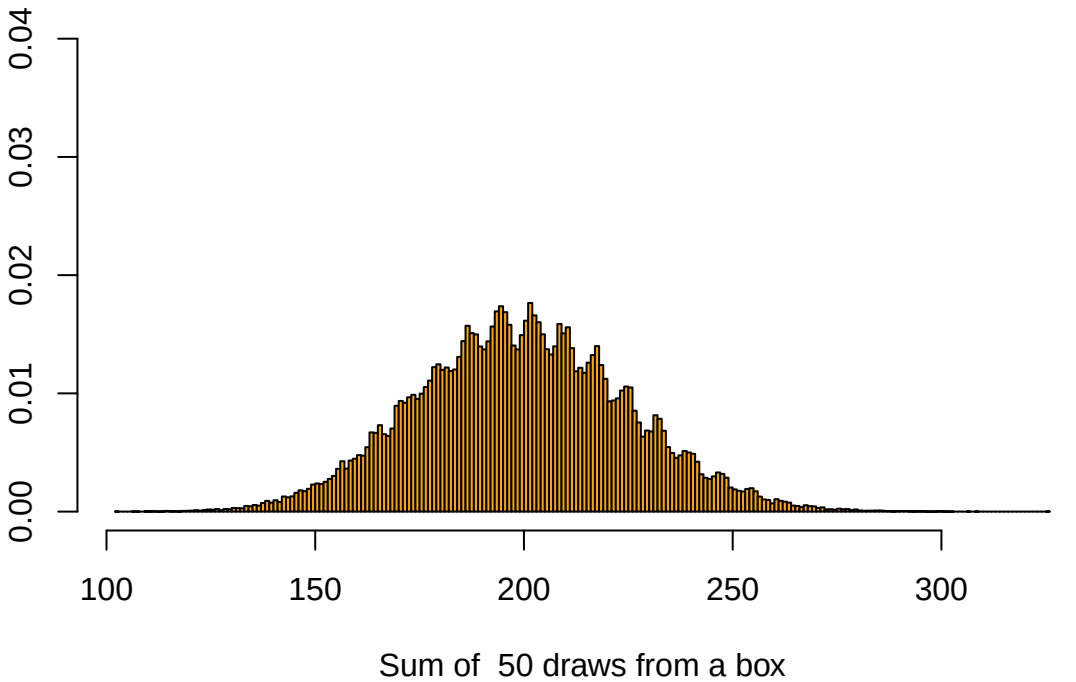
```
boxHistogram(c(1,2,9), 50, 1000, ymax=0.04, breaks=160, col='orange')
```



```
boxHistogram(c(1,2,9), 50, 10000, ymax=0.04, breaks=160, col='orange')
```




```
boxHistogram(c(1,2,9), 50, 100000, ymax=0.04, breaks=160, col='orange')
```



Product

The Central Limit Theorem says that when drawing with replacement from a box, the probability histogram for the sum (and the average) will follow the normal curve, even if the contents of the box do not. It doesn't say anything about the product or some other combination of the sample values.

Product of Dice Rolls

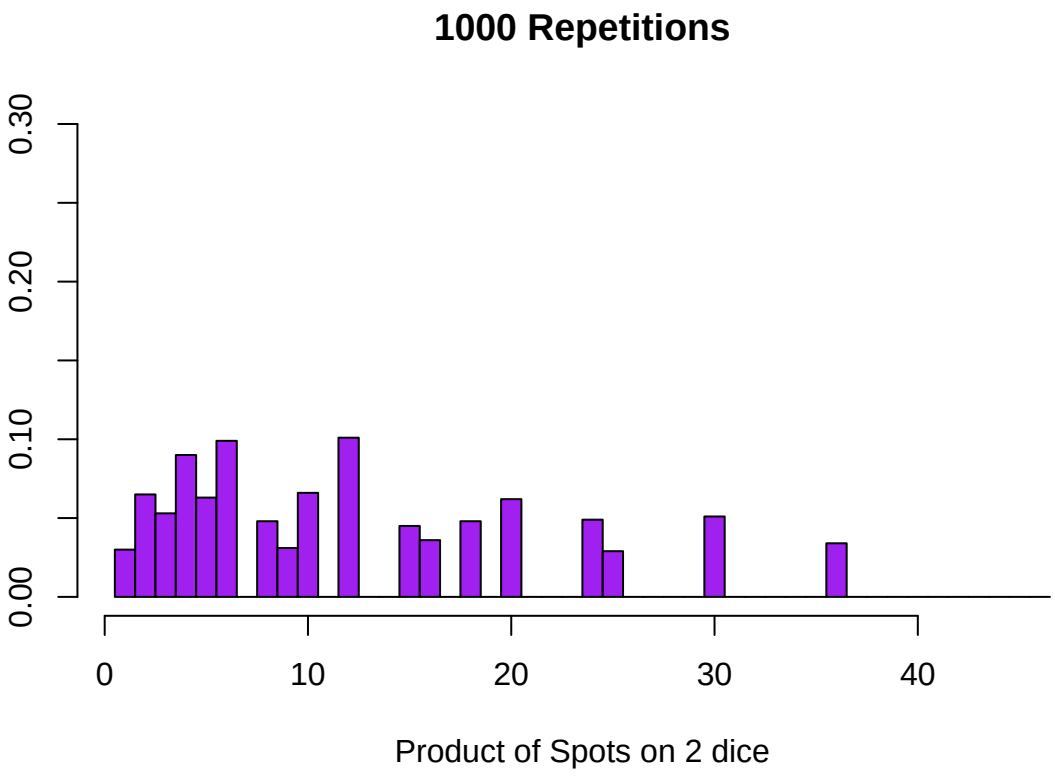
If we roll two dice and take the product, we can get certain values between $1 = 1 \times 1$ and $36 = 6 \times 6$. Here is data generated from 100 trials.

```
diceSimulation(100, diceSum=FALSE)
```

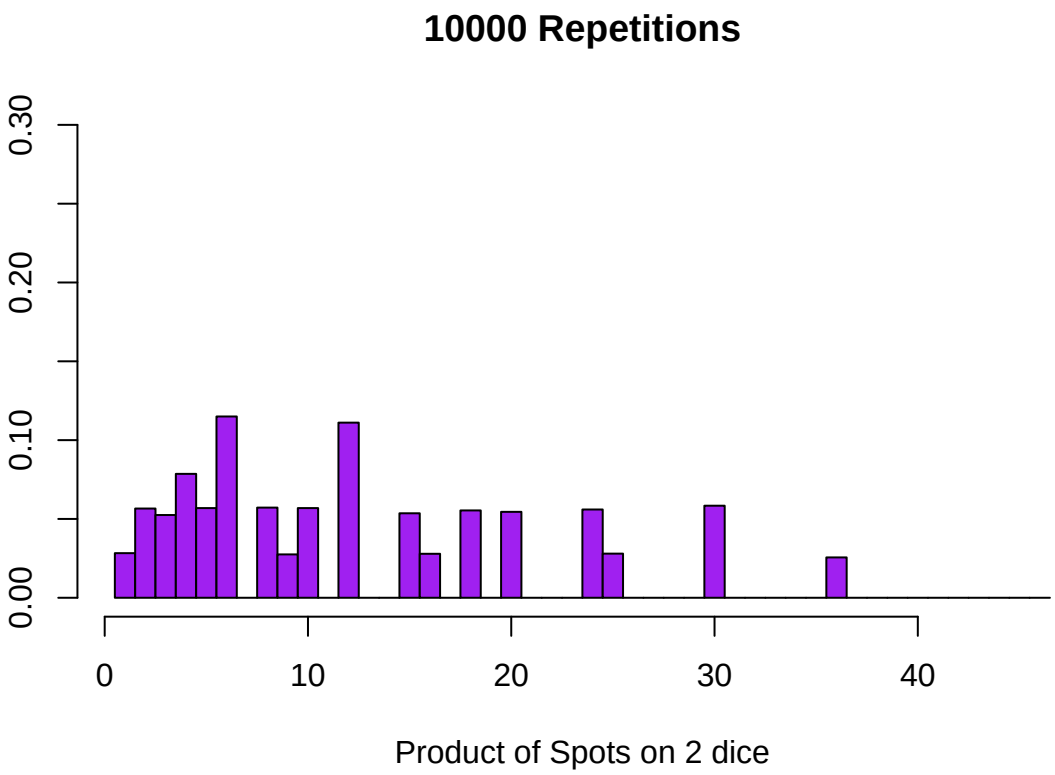
```
##   [1] 12  8  6  8 20  6 24  6 15 20  9  5 10  1 18 36  8 12 20  3 30  6  3
##  [24] 10 12  6 12 30 30 24 10 12  6  1  6 25  4  8  3 25  4 25 25  8 18 18
##  [47] 30 15  3  4 12 18  8  4  3  8  6 25  4 10  1  6  5  4 18 24  5 18 20
##  [70] 10  9  8 10  9 15  5 24 12 18  5 15 16 20 25 25 15  8 18  2  4 12  5
##  [93]  8 10  4 20 12  1 15 36
```

Here are histograms from two such experiments:

```
diceHistogram(1000, diceSum=FALSE, col='purple')
```



```
diceHistogram(10000, diceSum=FALSE, col='purple')
```



These don't look like normal curves at all. Why not?