CLT

Sampling Distribution

Sampling distribution is a term used for probability distribution of any sample quantity (mean, standard deviation, median, etc.)

Law of large numbers

The weak law of large numbers asserts that the sample mean of a large number of independent identically distributed random variables is very close to the true mean, with high probability.

Consider a sequence $X_1, X_2, ...$ of independent identically distributed random variables with mean µ and variance σ^2 and define the sample mean by $M_n = \frac{X_1, X_2, ... X_n}{n}$.

Then E[M_n]=E[$\frac{X_1+...+X_n}{n}$] = $\frac{n\mu}{n} = \mu$ and var(M_n)=$\frac{var(X_1+X_2+...+X_n)}{n^2}$=$\frac{var(X_1)+...+var(X_n)}{n^2}$=$\frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$.

According to the weak law of large numbers, the distribution of the sample mean M_n is increasingly concentrated in the near vicinity of the true mean μ. In particular, its variance tends to zero. The sum, $S_n = X_1 + ... + X_n = nM_n$, on the other hand, increases to infinity and the distribution of S_n does not converge.

CLT for sample mean

Let $X_1, X_2, ..., X_n$ be a sequence of independent identically distributed random variables with common mean µ and variance σ^2 and define $Z_n = \frac{X_1, X_2, ..., X_n - n\mu}{\sigma\sqrt{n}}$.

Then, the CDF of Z_n converges to the standard normal CDF

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{x^2}{2}} dx$$

in the sense that $\lim_{n\to\infty} P(Z_n \leq z) = \Phi(z)$ for every z.

Examples

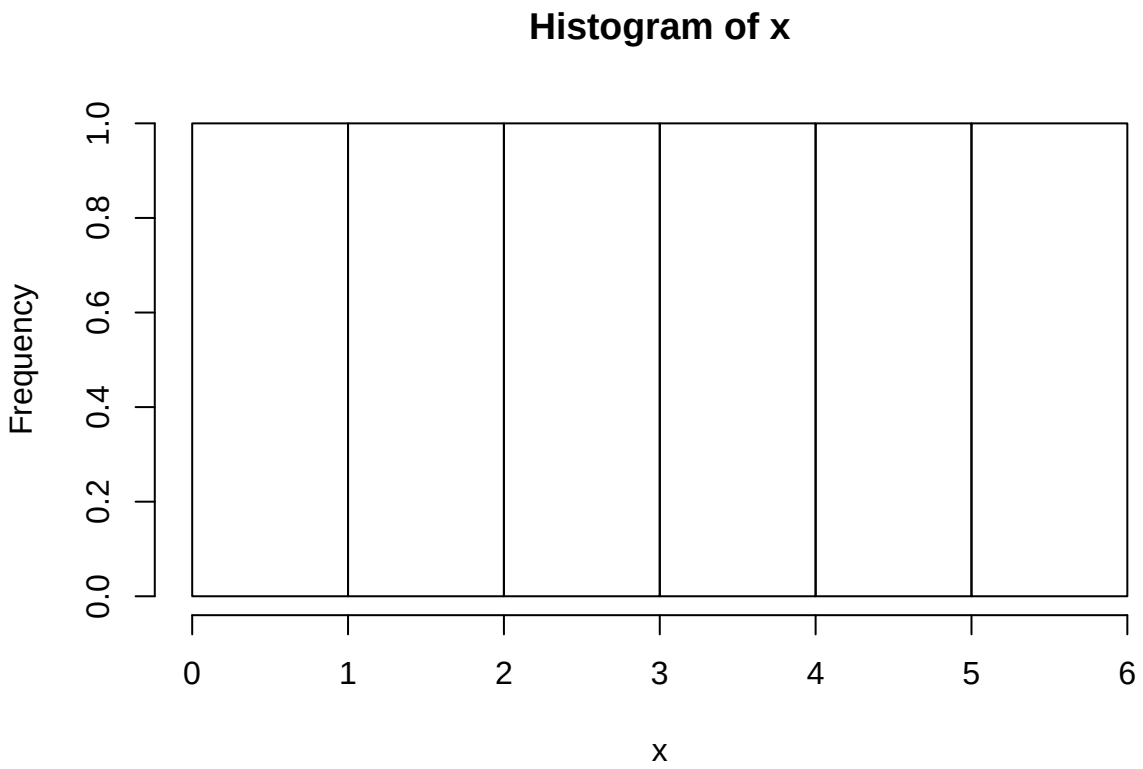
Fair Die

Single Sample

Suppose you have a fair die. Each time you toss the die, any of the 6 sides have a equal probability of landing face up.

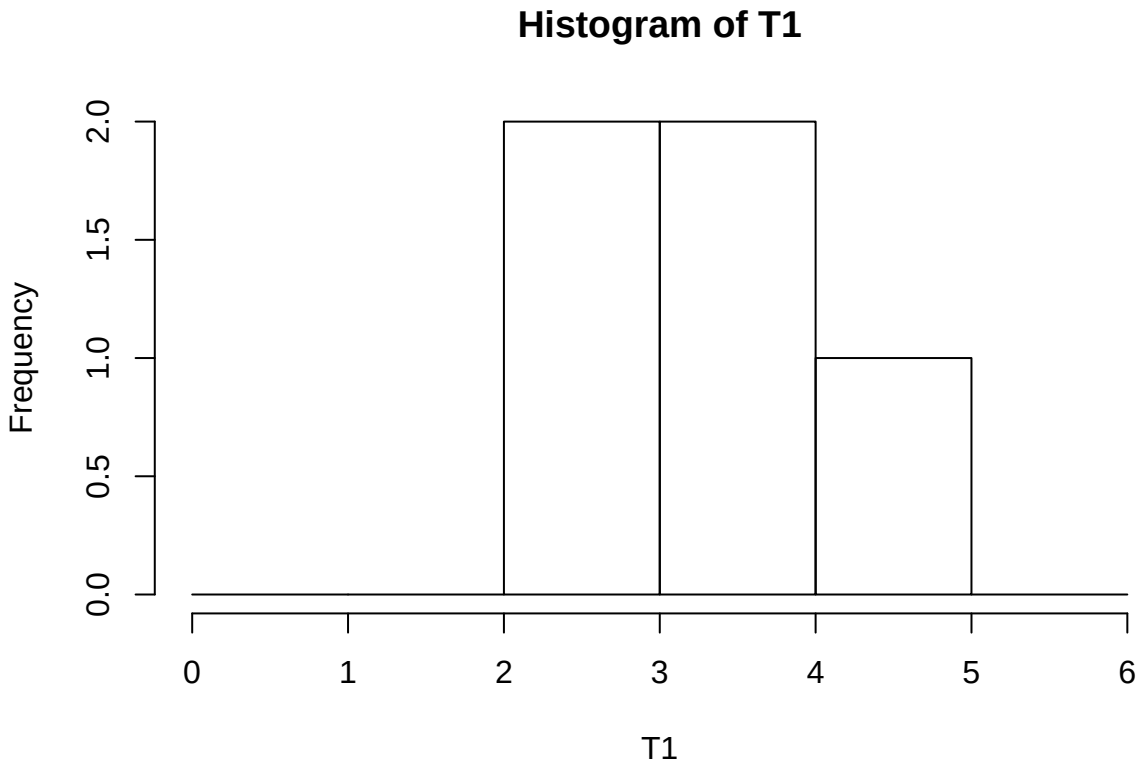
```
set.seed(111)

p<-c(1/6,1/6,1/6,1/6,1/6,1/6)
x <- c(1,2,3,4,5,6)
hist(x,breaks=0:6)
```



Now suppose you decide to toss the die 5 times. For the first toss, we get 4, 5, 3, 4, and 3 with mean 3.8.

```
set.seed(111)
T1 <- sample(x, 5,replace=TRUE)
hist(T1,breaks=0:6)
```



```
T1
```

```
## [1] 4 5 3 4 3
```

```
mean(T1)
```

```
## [1] 3.8
```

Now suppose you decide to toss the die 25 times. For the first toss, we get the sequence {4, 5, 3, 4, 3, 3, 1, 4, 3, 1, 4, 4, 1, 1, 1, 3, 2, 6, 2, 4, 3, 2, 3, 3, 6} with mean 3.04.

```
set.seed(111)
T2 <- sample(x, 25, replace=TRUE)
hist(T2,breaks=0:6)
```


```
T2
```

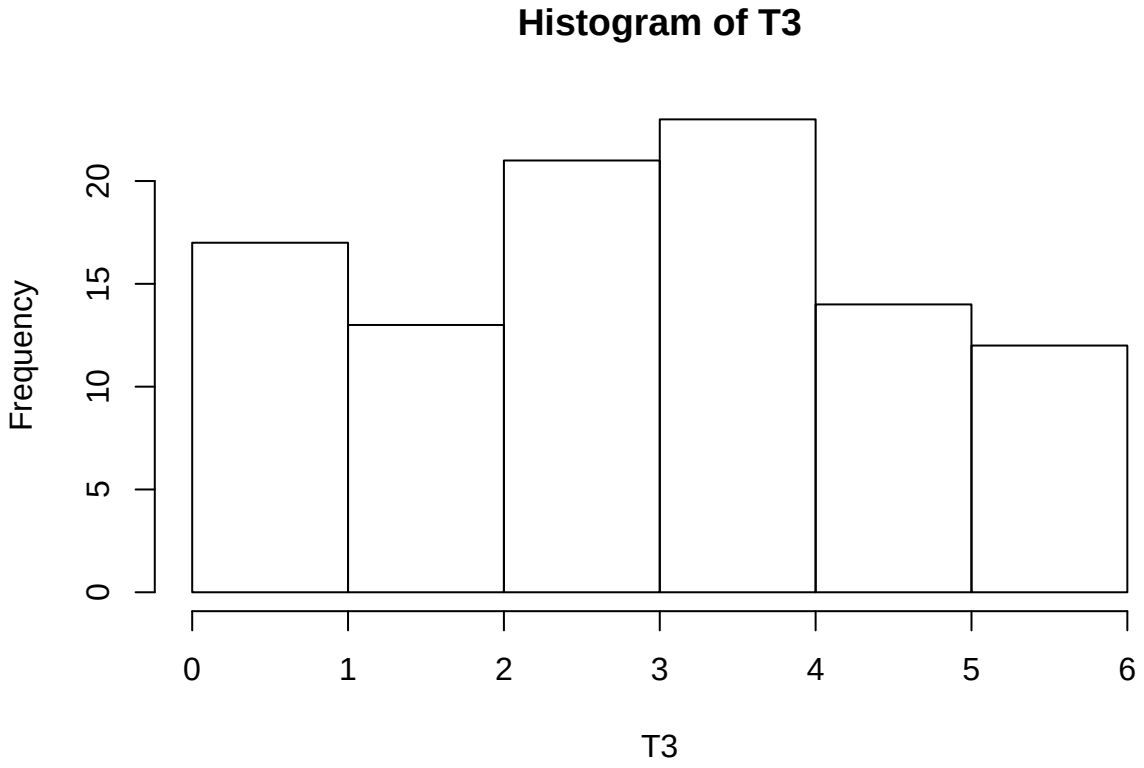
```
##  [1] 4 5 3 4 3 3 1 4 3 1 4 4 1 1 1 3 2 6 2 4 3 2 3 3 6
```

```
mean(T2)
```

```
## [1] 3.04
```

Now suppose you decide to toss the die 100 times. For the first toss, we get the sequence {4, 5, 3, 4, 3, 3, 1, 4, 3, 1, 4, 4, 1, 1, 1, 3, 2, 6, 2, 4, 3, 2, 3, 3, 6} with mean 3.04.

```
set.seed(111)
T3 <- sample(x, 100, replace=TRUE)
hist(T3,breaks=0:6)
```




```
T3
```

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

```
mean(T3)
```

```
## [1] 3.4
```

Multiple Samples

Now lets see what distribution looks like for the sample mean as we increase the number of samples.

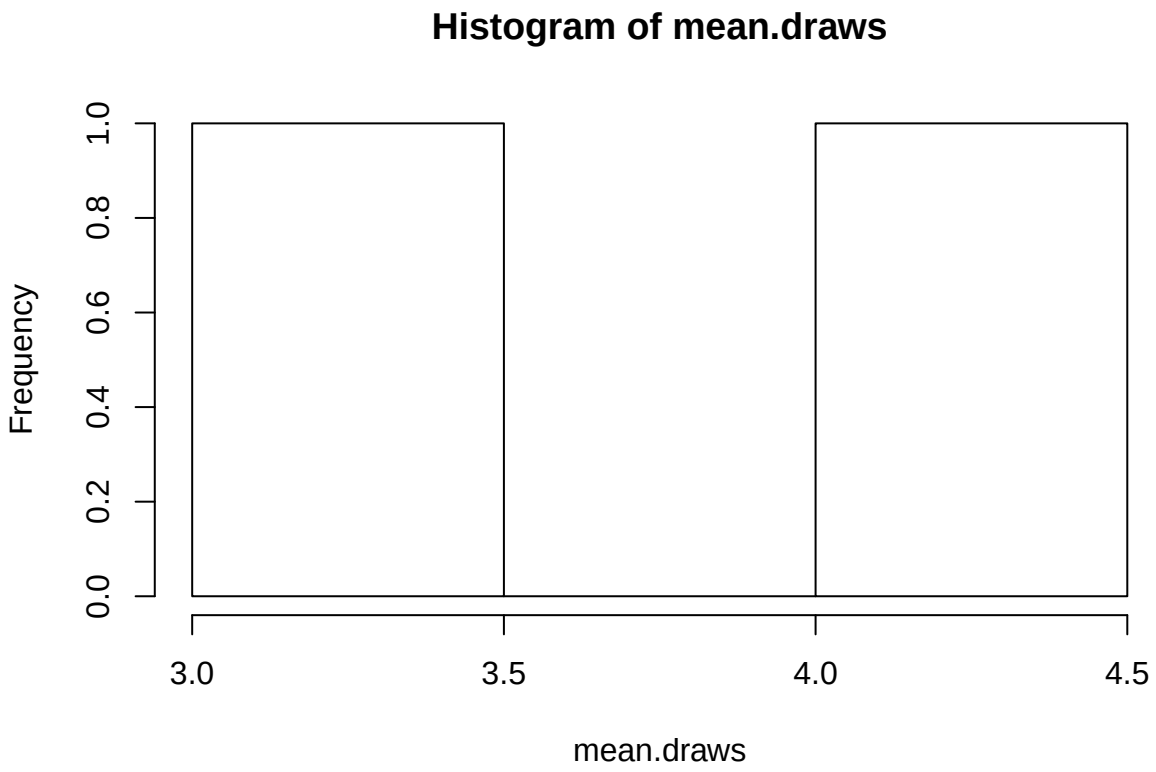
Suppose we have 2 people who each record 10 rolls of the dice. We get 2 means, 3.0 and 4.3.

```
set.seed(1)
N=2
n=10
p = c(1/6,1/6,1/6,1/6,1/6,1/6)
mean.draws <- replicate(N, mean(sample(c(1,2,3,4,5,6), n, prob = p, replace = TRUE)))

mean.draws
```

```
## [1] 3.0 4.3
```

```
hist(mean.draws)
```



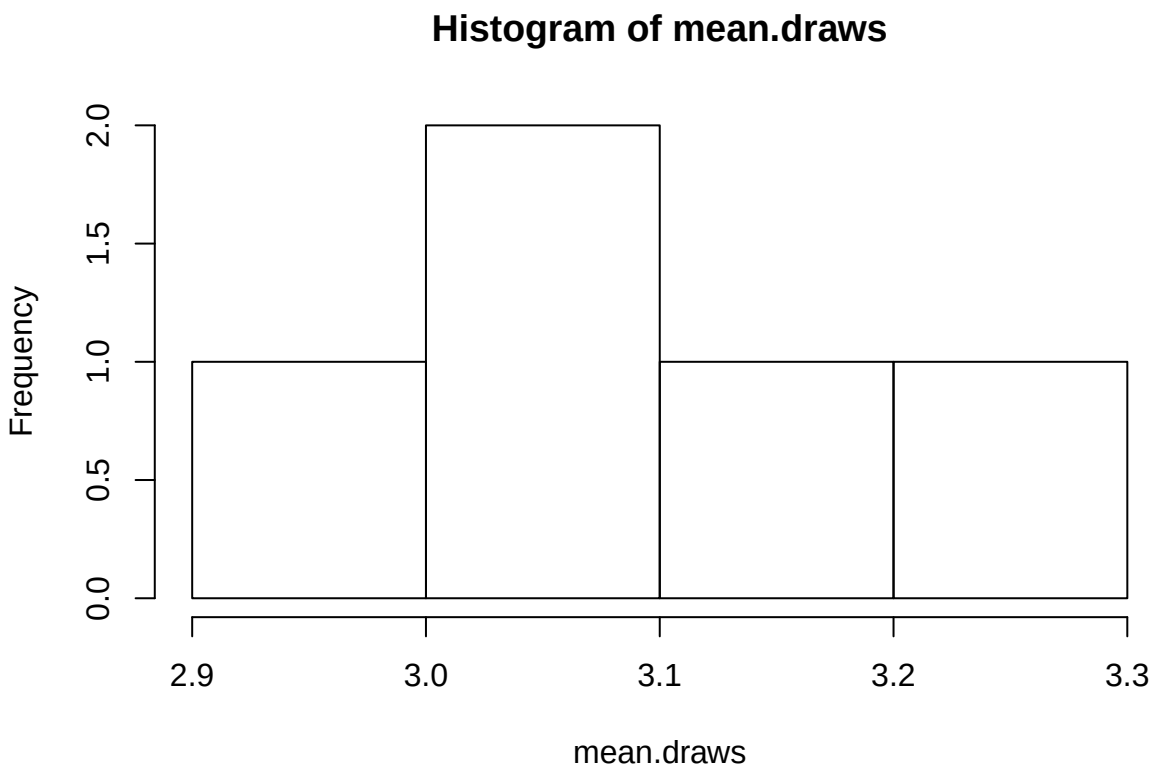
Now suppose we have 5 people who each record 10 rolls of the dice. We get means 3.1, 3.2, 2.9, 3.3, 3.1.

```
set.seed(2)
N=5
n=10
p = c(1/6,1/6,1/6,1/6,1/6,1/6)
mean.draws <- replicate(N, mean(sample(c(1,2,3,4,5,6), n, prob = p, replace = TRUE)))

mean.draws
```

```
## [1] 3.1 3.2 2.9 3.3 3.1
```

```
hist(mean.draws)
```



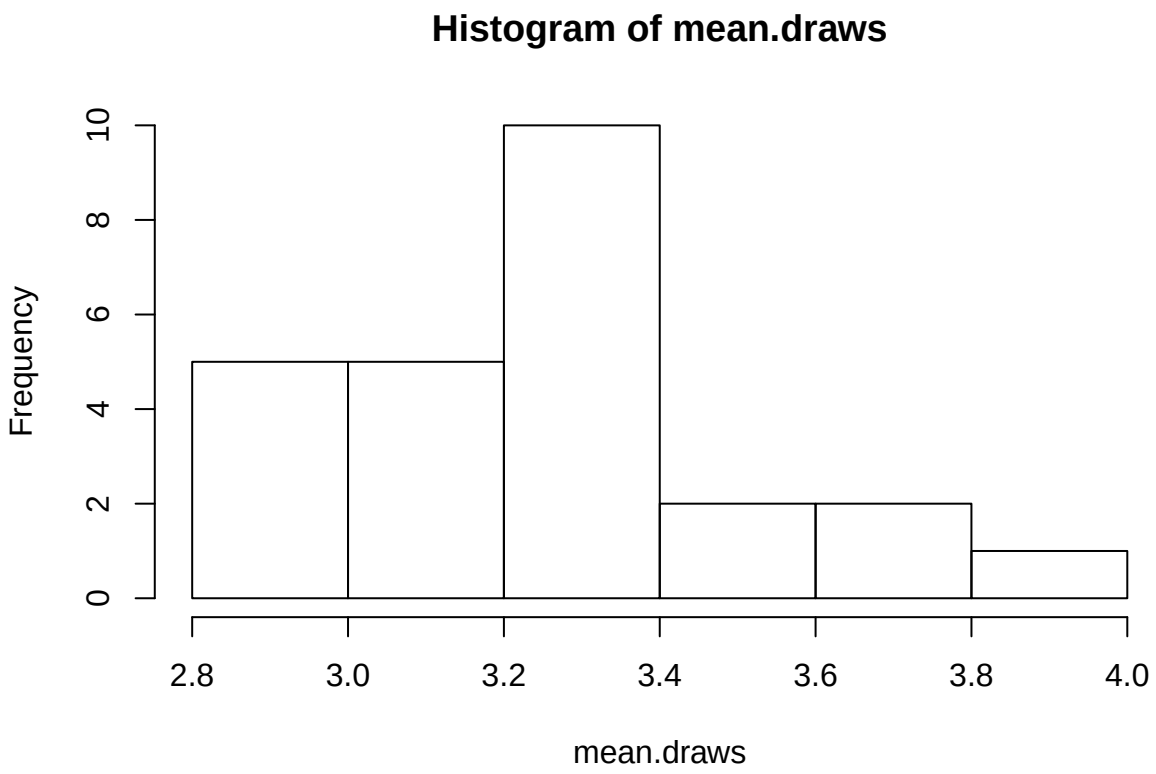
Now suppose we have 25 people who each record 10 rolls of the dice.

```
set.seed(2)
N=25
n=10
p = c(1/6,1/6,1/6,1/6,1/6,1/6)
mean.draws <- replicate(N, mean(sample(c(1,2,3,4,5,6), n, prob = p, replace = TRUE)))

mean.draws
```

```
##  [1] 3.1 3.2 2.9 3.3 3.1 3.8 3.3 3.3 3.4 3.3 3.1 3.8 3.4 3.3 4.0 3.0 2.8
## [18] 3.3 2.8 2.8 3.4 3.4 3.5 3.5 3.2
```

```
hist(mean.draws)
```



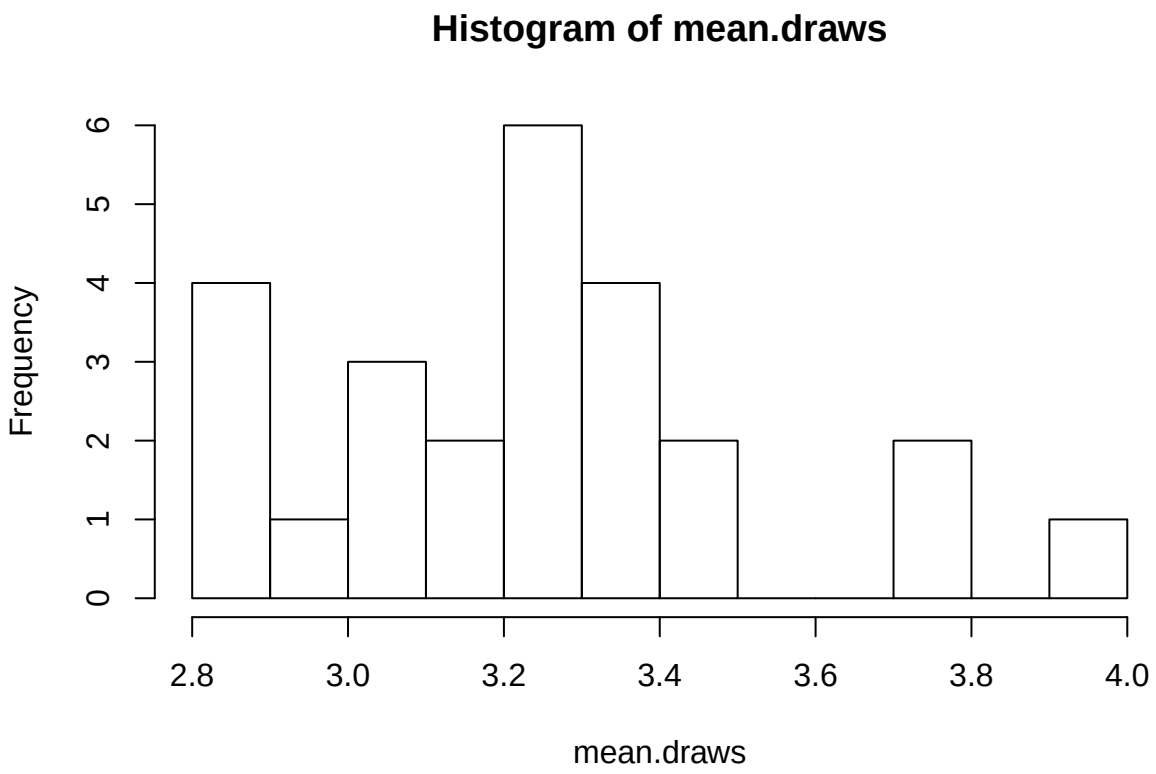
Now suppose we have 25 people who each record 10 rolls of the dice.

```
set.seed(2)
N=25
n=10
p = c(1/6,1/6,1/6,1/6,1/6,1/6)
mean.draws <- replicate(N, mean(sample(c(1,2,3,4,5,6), n, prob = p, replace = TRUE)))

mean.draws
```

```
##  [1] 3.1 3.2 2.9 3.3 3.1 3.8 3.3 3.3 3.4 3.3 3.1 3.8 3.4 3.3 4.0 3.0 2.8
## [18] 3.3 2.8 2.8 3.4 3.4 3.5 3.5 3.2
```

```
hist(mean.draws,breaks=10)
```



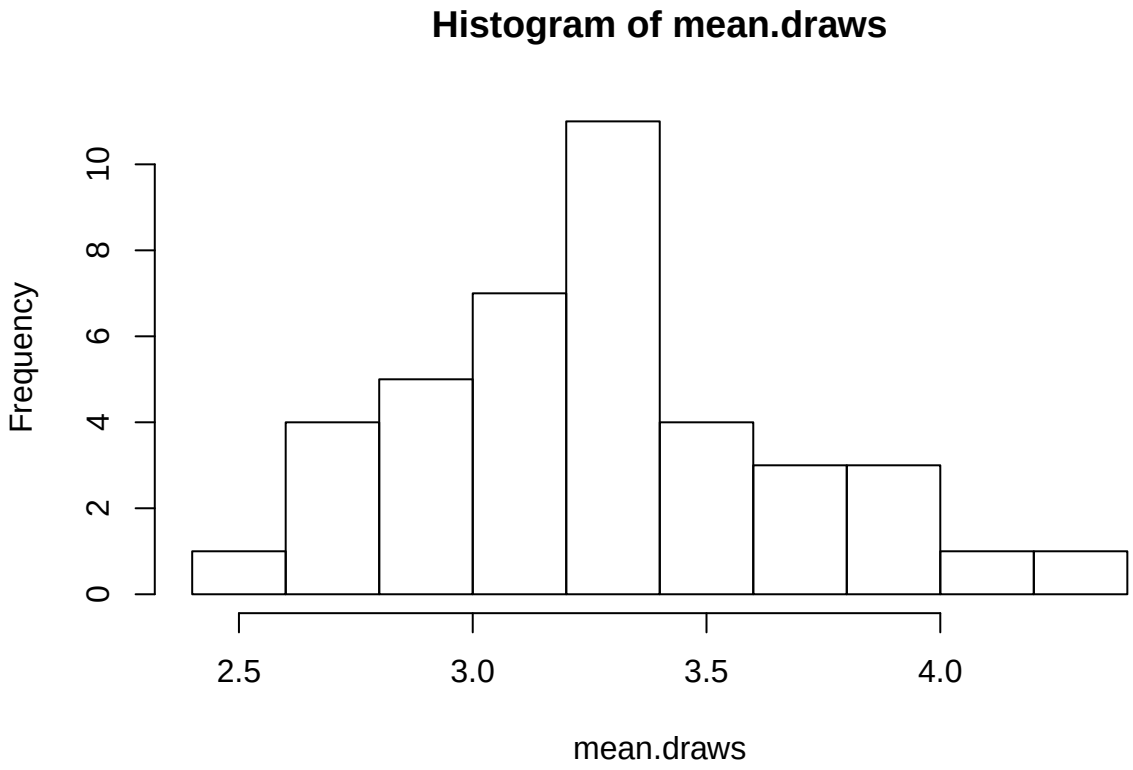
Now suppose we have 40 people who each record 10 rolls of the dice.

```
set.seed(2)
N=40
n=10
p = c(1/6,1/6,1/6,1/6,1/6,1/6)
mean.draws <- replicate(N, mean(sample(c(1,2,3,4,5,6), n, prob = p, replace = TRUE)))

mean.draws
```

```
##  [1] 3.1 3.2 2.9 3.3 3.1 3.8 3.3 3.3 3.4 3.3 3.1 3.8 3.4 3.3 4.0 3.0 2.8
## [18] 3.3 2.8 2.8 3.4 3.4 3.5 3.5 3.2 3.3 4.3 2.5 3.9 3.8 3.0 4.0 3.1 2.7
## [35] 3.2 4.1 2.9 3.6 3.6 3.0
```

```
hist(mean.draws, breaks=10)
```



Now suppose we have 100 people who each record 10 rolls of the dice.

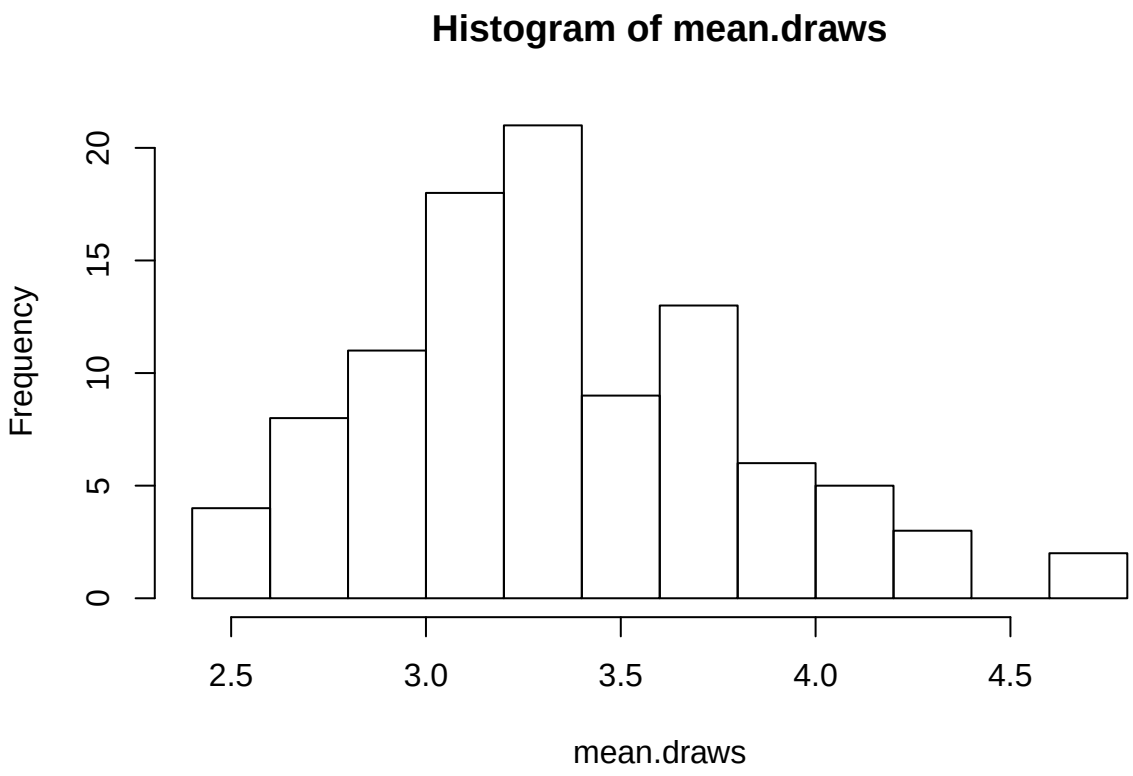
```
set.seed(2)
N=100
n=10
p = c(1/6,1/6,1/6,1/6,1/6,1/6)

mean.draws <- replicate(N, mean(sample(c(1,2,3,4,5,6), n, prob = p, replace = TRUE)))

mean.draws
```

```
##   [1] 3.1 3.2 2.9 3.3 3.1 3.8 3.3 3.3 3.4 3.3 3.1 3.8 3.4 3.3 4.0 3.0 2.8
##  [18] 3.3 2.8 2.8 3.4 3.4 3.5 3.5 3.2 3.3 4.3 2.5 3.9 3.8 3.0 4.0 3.1 2.7
##  [35] 3.2 4.1 2.9 3.6 3.6 3.0 3.4 4.7 2.9 2.7 4.2 4.1 3.4 3.4 3.9 3.4 4.8
##  [52] 3.6 3.1 3.2 3.4 3.5 2.8 2.9 3.9 3.3 3.7 3.7 3.8 4.2 3.8 3.1 3.3 3.2
##  [69] 4.3 3.7 2.7 3.7 2.5 3.2 3.5 3.1 3.5 3.7 3.3 4.2 3.4 3.0 3.2 2.9 3.0
##  [86] 3.7 2.5 3.8 3.1 3.2 3.3 3.6 4.4 3.2 3.1 2.8 2.6 3.8 3.9 3.0
```

```
hist(mean.draws, breaks=10)
```

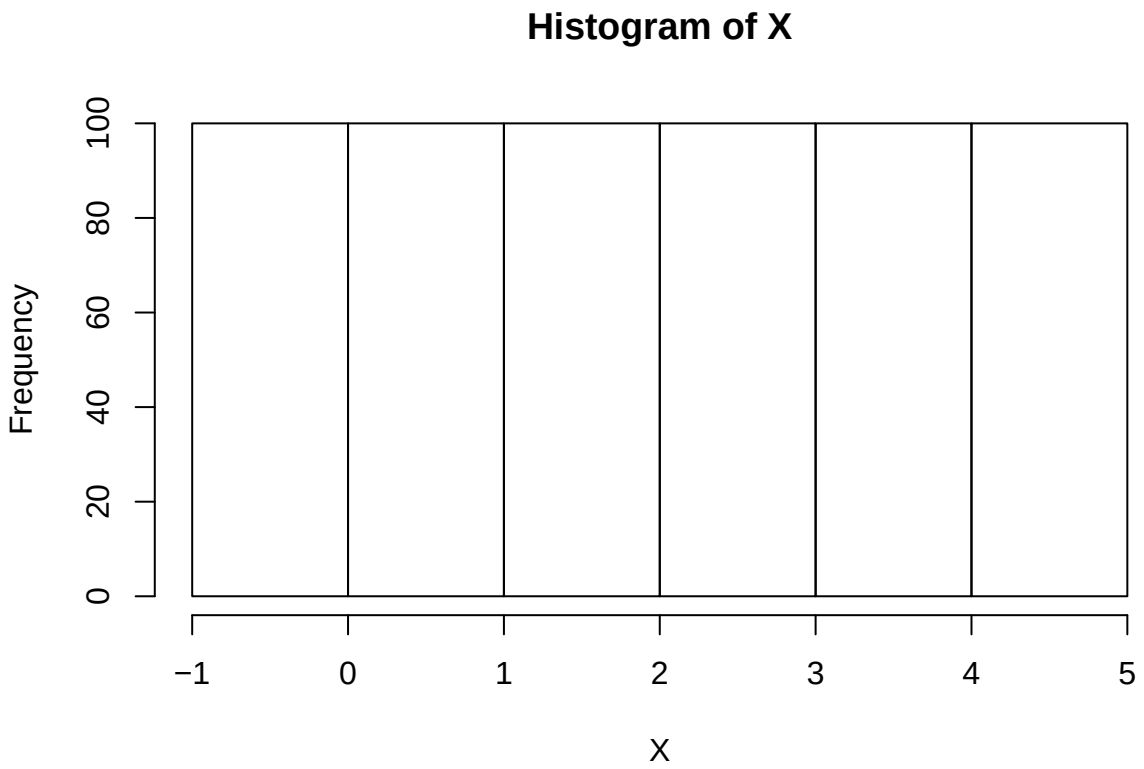


Question

Suppose out of a group of 500 people, 100 have never watched Star Wars, 100 have watched Star Wars once, 100 have watched Star Wars twice, 100 have watched Star Wars three times, 100 have watched Star Wars 4 times, and 100 have watched Star Wars 5 times.

The mean is 2 and the distribution is uniform as shown below:

```
N<-100
X <- c(rep(0,N),rep(1,N),rep(2,N),rep(3,N),rep(4,N),rep(5,N))
hist(X,breaks=-1:5)
```




```
set.seed(111)
mean(X)
```

```
## [1] 2.5
```

When we take one sample of 10 people, we get a mean of 2.8. From sample of 30 people, we get a mean of 2.46. From a sample of 100 people, we get a mean of 2.48. From a sample of 200 people, we get a mean of 2.62. The mean of the sample means is 2.8.

```
set.seed(1)
N=1
mean.draws1 <- replicate(N, mean(sample(X, 10, replace = FALSE)))
paste("mean of draw 1: ", mean.draws1)
```

```
## [1] "mean of draw 1:  2.8"
```

```
mean.draws2 <- replicate(N, mean(sample(X, 30, replace = FALSE)))
paste0("mean of draw 2: ", round(mean.draws2,2))
```

```
## [1] "mean of draw 2: 2.47"
```

```
mean.draws3 <- replicate(N, mean(sample(X, 100, replace = FALSE)))
paste0("mean of draw 3: ",mean.draws3)
```

```
## [1] "mean of draw 3: 2.48"
```

```
mean.draws4 <- replicate(N, mean(sample(X, 300, replace = FALSE)))
paste0("mean of draw 4: ",mean.draws4)
```

```
## [1] "mean of draw 4: 2.62"
```

```
paste0("mean of mean from draw 1 and 2: ", round((mean.draws1+mean.draws2)/2,2))
```

```
## [1] "mean of mean from draw 1 and 2: 2.63"
```

```
paste0("mean of mean from draw 1,2, and 3: ", round((mean.draws1+mean.draws2+mean.draws3)/3,3))
```

```
## [1] "mean of mean from draw 1,2, and 3: 2.582"
```

```
paste0("mean of mean from draw 1,2, 3 and 4: ", round((mean.draws1+mean.draws2+mean.draws3+mean.draws4),
```

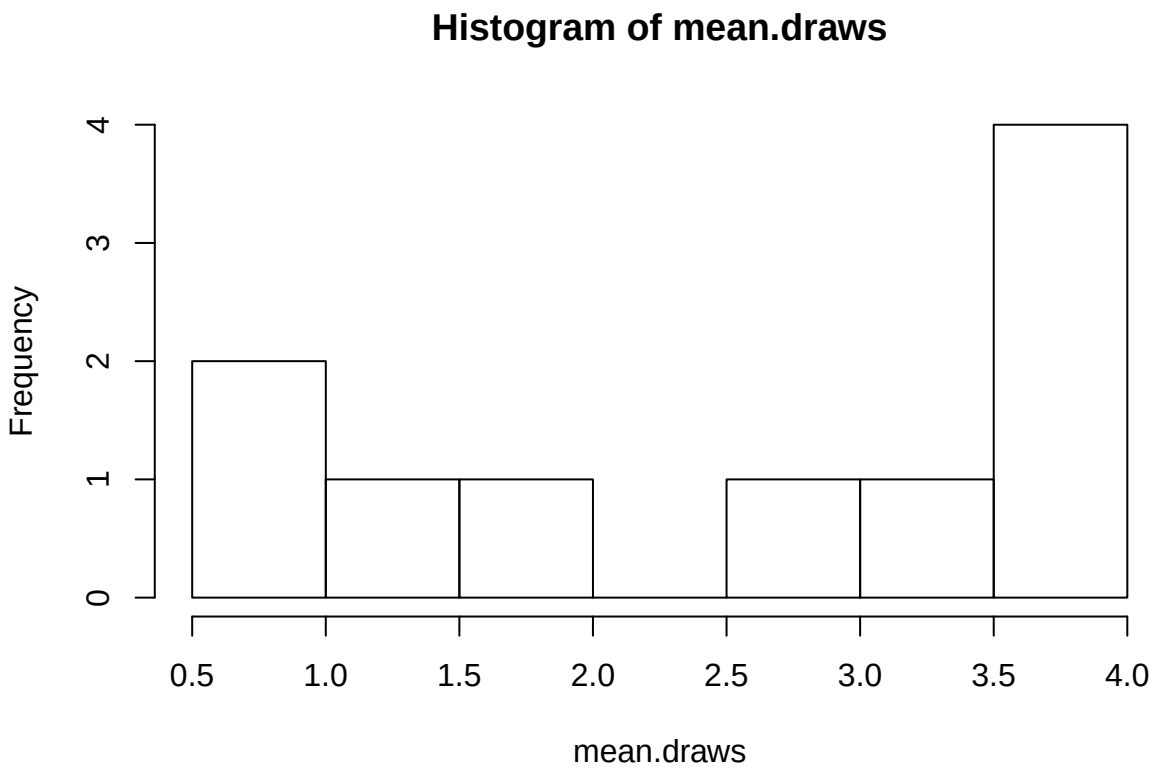
```
## [1] "mean of mean from draw 1,2, 3 and 4: 2.592"
```

Now let's suppose we take a sample of 3 people and repeat the sampling process 10 times.

```
set.seed(1)
N=10
n=3
p = c(1/6,1/6,1/6,1/6,1/6,1/6)
mean.draws <- replicate(N, mean(sample(X, n, replace = FALSE)))
mean.draws
```

```
##  [1] 2.0000000 3.6666667 3.6666667 0.6666667 3.3333333 3.6666667 3.6666667
##  [8] 1.3333333 1.0000000 3.0000000
```

```
hist(mean.draws,breaks=5)
```



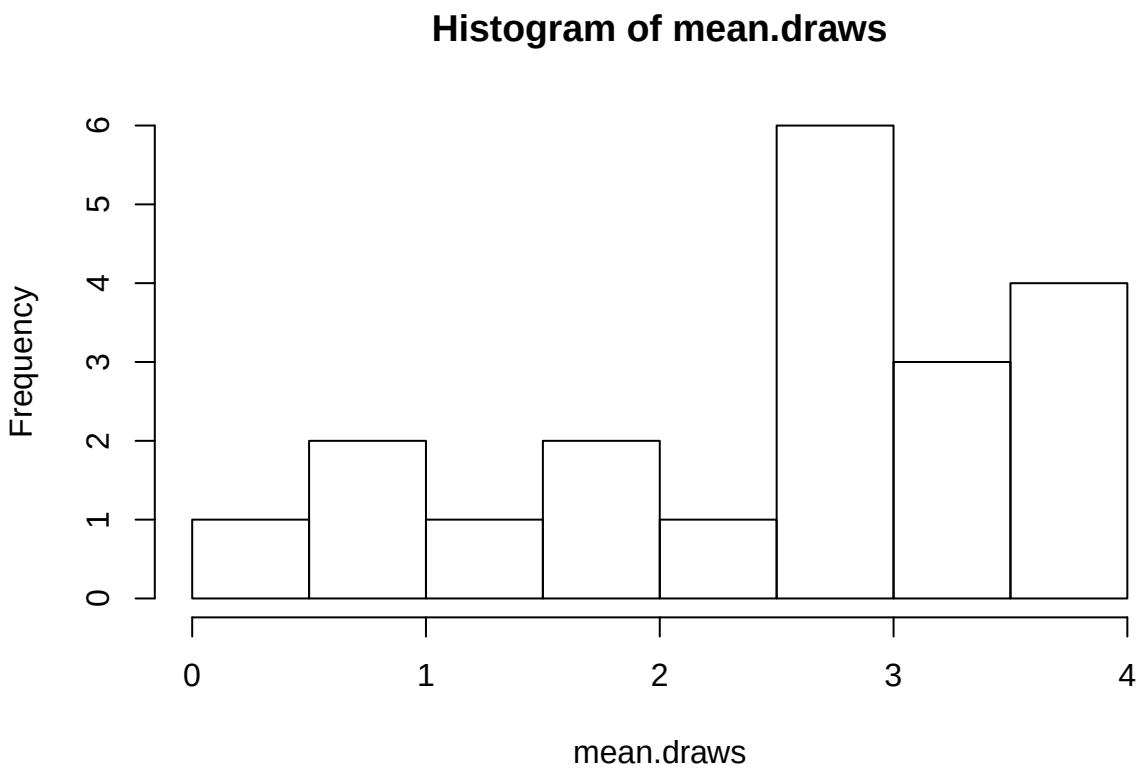
Now let's suppose we take a sample of 3 people and repeat the sampling process 20 times.

```
set.seed(1)
N=20
n=3
p = c(1/6,1/6,1/6,1/6,1/6,1/6)
mean.draws <- replicate(N, mean(sample(X, n, replace = FALSE)))
mean.draws
```

```
##  [1] 2.0000000 3.6666667 3.6666667 0.6666667 3.3333333 3.6666667 3.6666667
##  [8] 1.3333333 1.0000000 3.0000000 2.3333333 2.6666667 2.6666667 3.0000000
```

```
## [15] 3.3333333 2.0000000 3.3333333 2.6666667 0.3333333 2.6666667
```

```
hist(mean.draws,breaks=5)
```

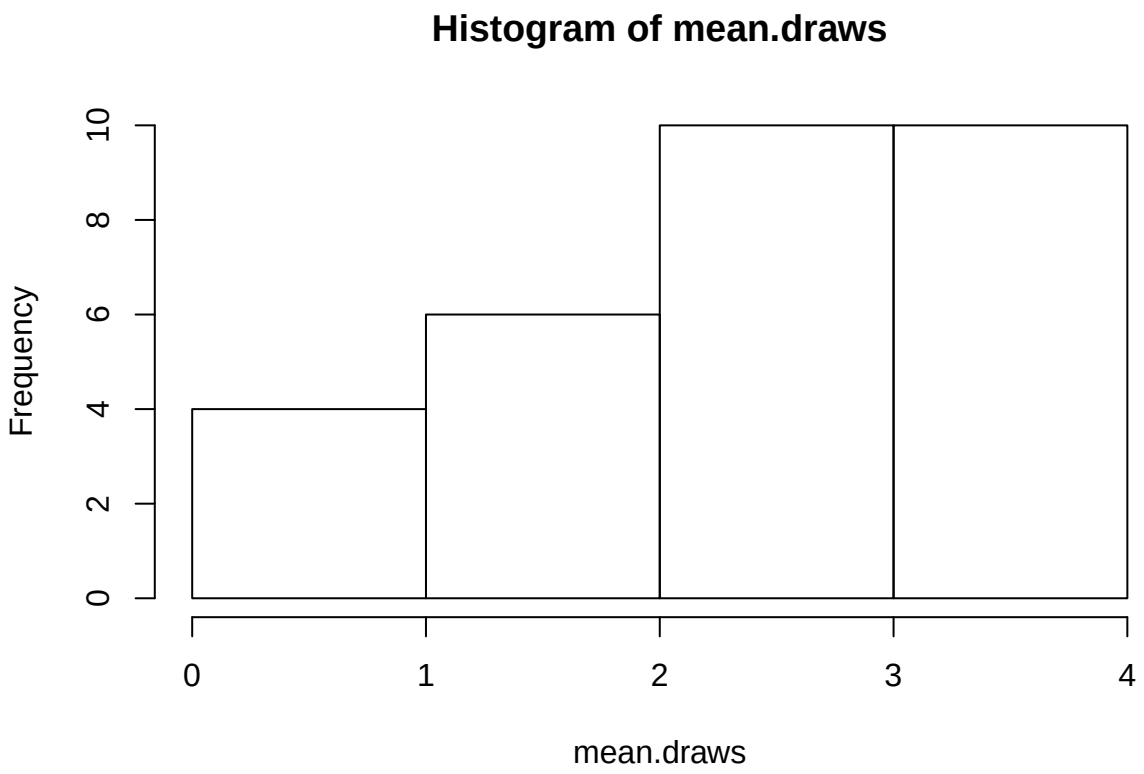


Now let's suppose we take a sample of 3 people and repeat the sampling process 30 times.

```
set.seed(1)
N=30
n=3
p = c(1/6,1/6,1/6,1/6,1/6,1/6)
mean.draws <- replicate(N, mean(sample(X, n, replace = FALSE)))
mean.draws
```

```
##  [1] 2.0000000 3.6666667 3.6666667 0.6666667 3.3333333 3.6666667 3.6666667
##  [8] 1.3333333 1.0000000 3.0000000 2.3333333 2.6666667 2.6666667 3.0000000
## [15] 3.3333333 2.0000000 3.3333333 2.6666667 0.3333333 2.6666667 2.6666667
## [22] 1.6666667 2.0000000 4.0000000 1.6666667 4.0000000 3.6666667 2.3333333
## [29] 3.0000000 0.3333333
```

```
hist(mean.draws,breaks=5)
```

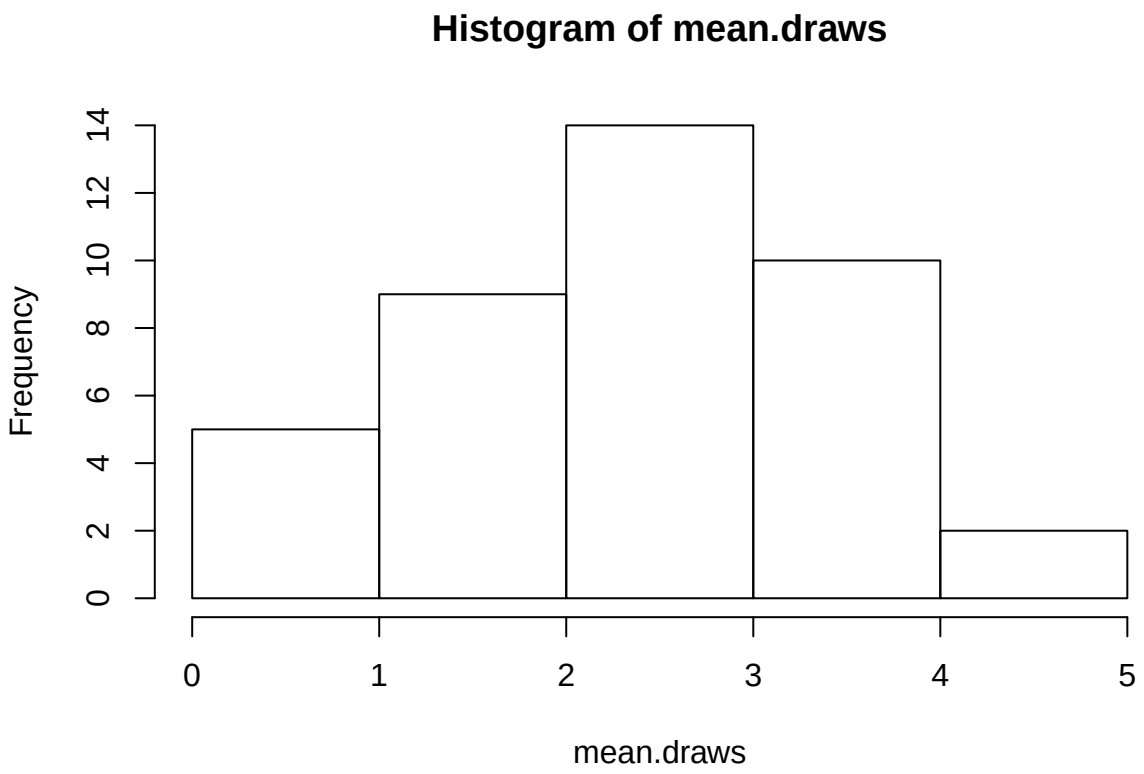


Now let's suppose we take a sample of 3 people and repeat the sampling process 40 times.

```
set.seed(1)
N=40
n=3
p = c(1/6,1/6,1/6,1/6,1/6,1/6)
mean.draws <- replicate(N, mean(sample(X, n, replace = FALSE)))
mean.draws
```

```
##  [1] 2.0000000 3.6666667 3.6666667 0.6666667 3.3333333 3.6666667 3.6666667
##  [8] 1.3333333 1.0000000 3.0000000 2.3333333 2.6666667 2.6666667 3.0000000
## [15] 3.3333333 2.0000000 3.3333333 2.6666667 0.3333333 2.6666667 2.6666667
## [22] 1.6666667 2.0000000 4.0000000 1.6666667 4.0000000 3.6666667 2.3333333
## [29] 3.0000000 0.3333333 1.3333333 4.3333333 2.6666667 2.6666667 3.0000000
## [36] 1.0000000 4.3333333 2.6666667 1.3333333 1.6666667
```

```
hist(mean.draws,breaks=5)
```

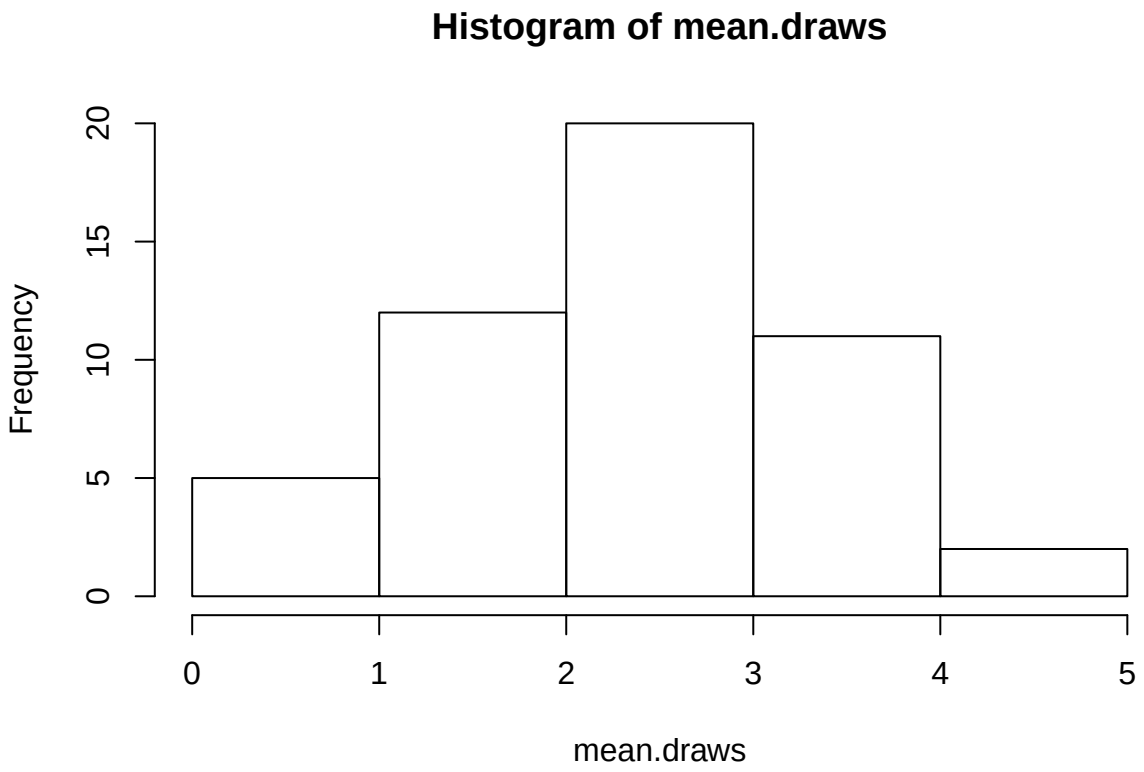


Now let's suppose we take a sample of 3 people and repeat the sampling process 50 times.

```
set.seed(1)
N=50
n=3
p = c(1/6,1/6,1/6,1/6,1/6,1/6)
mean.draws <- replicate(N, mean(sample(X, n, replace = FALSE)))
mean.draws
```

```
##  [1] 2.0000000 3.6666667 3.6666667 0.6666667 3.3333333 3.6666667 3.6666667
##  [8] 1.3333333 1.0000000 3.0000000 2.3333333 2.6666667 2.6666667 3.0000000
## [15] 3.3333333 2.0000000 3.3333333 2.6666667 0.3333333 2.6666667 2.6666667
## [22] 1.6666667 2.0000000 4.0000000 1.6666667 4.0000000 3.6666667 2.3333333
## [29] 3.0000000 0.3333333 1.3333333 4.3333333 2.6666667 2.6666667 3.0000000
## [36] 1.0000000 4.3333333 2.6666667 1.3333333 1.6666667 3.0000000 2.3333333
## [43] 1.6666667 2.0000000 2.6666667 3.0000000 4.0000000 1.6666667 2.3333333
## [50] 3.0000000
```

```
hist(mean.draws,breaks=5)
```


Histogram of mean.draws
Frequency
0
5
10
15
20
0
1
2
3
4
5
mean.draws