# STA 371G: Statistics and Modeling

## Introduction to Monte Carlo Simulation

Mingyuan Zhou
McCombs School of Business
The University of Texas at Austin

http://mingyuanzhou.github.io/STA371G

# Why Simulation?

- Extremely useful to model uncertainty
- Can be incorporated into spreadsheet models
- Allow the input variables to be random
- Provide an entire distribution of results, including not only the most likely, but also the worst and the best
- Help analyze a system's sensitivity to changes in operation conditions
- Can be implemented in Excel with the RAND() function or some add-in package (such as @Risk)
- Extremely convenient to implement in R (also in Matlab, Python, ...)

# Example: Simulation of Market Return

- Incorporate uncertainty into spreadsheet models

| | A | B | C | D |
|---|---|---|---|---|
| 1 | Economy | Probability | Market Return | (Market Return - Mean)^2 |
| 2 | Rapid Expanding | 12.00% | 23% | 0.005929 |
| 3 | Moderate Expanding | 40.00% | 18% | 0.000729 |
| 4 | No Growth | 25.00% | 15% | 9E-06 |
| 5 | Moderate Contraction | 15.00% | 9% | 0.003969 |
| 6 | Serious Contraction | 8.00% | 3% | 0.015129 |
| 7 | | | | |
| 8 | Expected Return | 0.153 | Simulated Return | 0.152489879 |
| 9 | Variance of Return | 0.002811 | Simulated variance | 0.002769478 |
| 10 | Standard Deviation of Return | 0.053018865 | Simulated sd | 0.052625831 |
| 11 | | | | |
| 12 | Random # | Simulated Return | Lookup Probability Table | |
| 13 | 0.39065967 | 18.00% | Cumulative probability | Market Return |
| 14 | 0.951676654 | 3.00% | 0 | 23% |
| 15 | 0.902555852 | 9.00% | 12% | 18% |
| 16 | 0.572895871 | 15.00% | 52% | 15% |
| 17 | 0.929102096 | 3.00% | 77% | 9% |
| 18 | 0.428496674 | 18.00% | 92% | 3% |
| 19 | 0.683789167 | 15.00% | | |
| 20 | 0.582831026 | 15.00% | | |
| 21 | 0.362137678 | 18.00% | | |

Let $X$ denote the market return, then $X$ is a random variable whose distribution can be described as $P(X = 23\%) = 0.12$, $P(X = 18\%) = 0.40$, $P(X = 15\%) = 0.25$, $P(X = 9\%) = 0.15$ and $P(X = 3\%) = 0.08$

- Simulate 1000 random numbers in R

```
Prob = c(0.12,0.40,0.25,0.15,0.08)
Return = c(0.23,0.18,0.15,0.09,0.03)
SimulatedReturn = sample(Return,size=1000,replace=TRUE,prob=Prob)
```

- The mean $\mu$ and standard deviation $\sigma$ of this random variable:

```
Mu = sum(Prob*Return)
Sigma = sqrt(sum(Prob*(Return-Mu)^2))
```

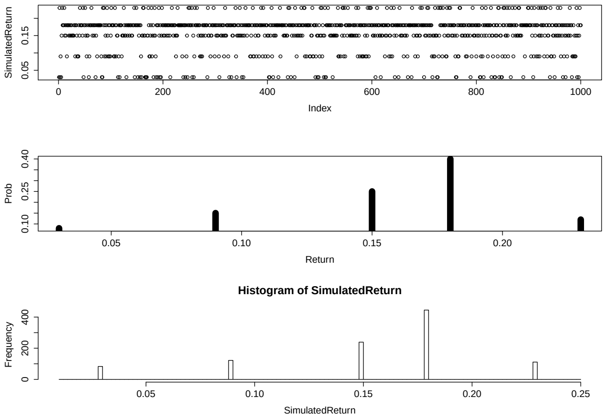
- Find the sample mean $\bar{X}$ and sample standard deviation $s_X$:

```
mean(SimulatedReturn)
sd(SimulatedReturn)
```

- Repeat the same experiment over and over, what will you find?

- Plot the results:

```
par(mfrow=c(3,1))
plot(SimulatedReturn)
plot(Return,Prob,type="h",lwd=10)
hist(SimulatedReturn,breaks=seq(0.01,0.25,0.002))
```



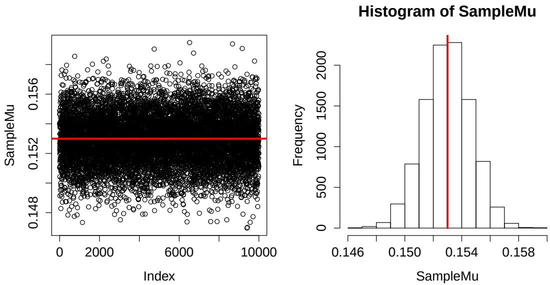
Let's repeat the same experiment 10,000 times, each time we record the sample mean

- What would be the distribution of the sample mean?
- Below is the R code:

```
#Sampling distribution of the sample mean
SampleMu = matrix(0,10000,1)
for (i in 1:10000){
  SimulatedReturn = sample(Return,size=1000,replace=TRUE,prob=Prob)
  SampleMu[i]=mean(SimulatedReturn)
}
par(mfrow=c(1,2))
plot(SampleMu)
abline(Mu,0,col='Red',lwd=3)
hist(SampleMu)
lines(c(Mu,Mu),c(0,4000),col='Red',lwd=3)
```

Let's repeat the same experiment 10,000 times, each time we record the sample mean

- Does this looks like a normal distribution?
- What's the mean of this random variable?
- What's the variance of this random variable?
- Sigma/sd(SampleMu)≈?
- Do we have a better understanding about the sampling distribution of the sample mean now?



Suppose the CEO gets a bonus of $Y = 80 + 10000X$

- What's the mean of $Y$?
- What's the variance of $Y$?
- Use simulation to estimate the mean and variance of $Y$.

# Uniform Random Variable

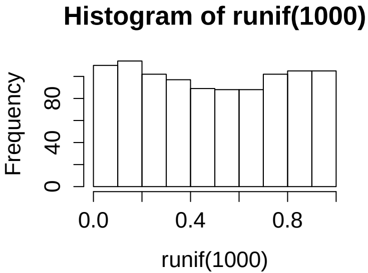
- A random number uniformly distributed between 0 and 1
- Can be used to generate any other random numbers
- In Excel: =RAND()
  - Flip a coin: =IF(RAND()>0.5,"Head","Tail")
  - Roll a die:
    - =RANDBETWEEN(1,6)
    - or use =RAND() and a lookup table
- In R: runif
  - Flip a coin: runif(1)>0.5 or sample(0:1,1)
  - Roll a die: =ceiling(runif(1)*6) or sample(1:6,1)

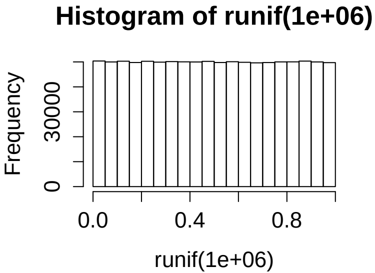
# Uniform Random Variable

Histogram of uniformly distributed random numbers

In Excel: copy RAND() to n cells , where $n$ is the number of experiments. What if $n = 10^9$?

In R: hist(runif(n))



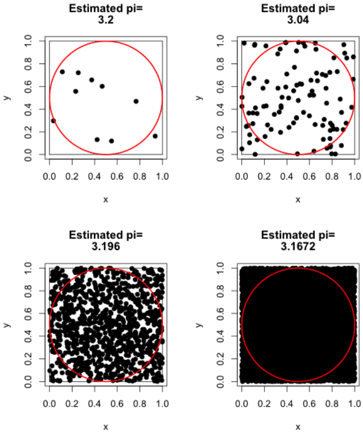



# Example: Estimate $\pi$ with Monte Carlo Simulation

- The area of a circle with the radius of $1/2$ is $\pi/4$
- The area of a unit square is 1
- A uniform random point in the unit square falls within the circle with probability $\pi/4$
- Estimate $\pi$ with 10, 100, 1000, or 10000 uniform random numbers
- R code:

```
for (n in c(10,100,1000,10000))
{
  x=runif(n)
  y=runif(n)
  Pi = sum(sqrt((x-0.5)^2+(y-0.5)^2)<0.5)/n*4
  plot(x,y,xlim=c(0,1),ylim=c(0,1),pch=19)
  lines(c(0,0),c(0,1))
  lines(c(1,0),c(0,0))
  lines(c(1,0),c(1,1))
  lines(c(1,1),c(1,0))
  r = 1/2
  theta = seq(-pi,pi,0.001)
  lines(r*cos(theta)+0.5,r*sin(theta)+0.5,type="s",lwd=2,col='red')
  title(c('Estimated pi=',Pi))
}
```

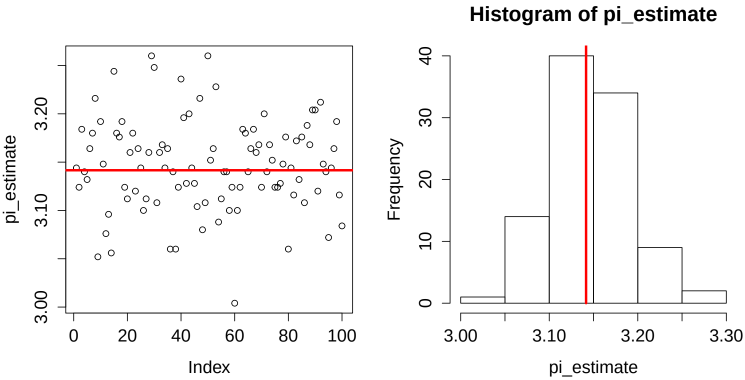
- Below are what we get:



- Let's simulate $\pi$ 100 times, each time using 1000 uniform random numbers
- R code:

```
count=matrix(0,100,1)
for (iter in (1:100)){
  x = runif(1000)
  y = runif(1000)
  for (i in (1:length(x))){
    if (sqrt((x[i]-0.5)^2+(y[i]-0.5)^2)<=0.5)
    {
      count[iter]=count[iter]+1
    }
  }
}
count/length(x)*4
pi_estimate = count/length(x)*4
par(mfrow=c(1,2))
plot(pi_estimate)
abline(pi,0,col='Red',lwd=3)
hist(pi_estimate)
lines(c(pi,pi),c(0,800),col='Red',lwd=3)
```

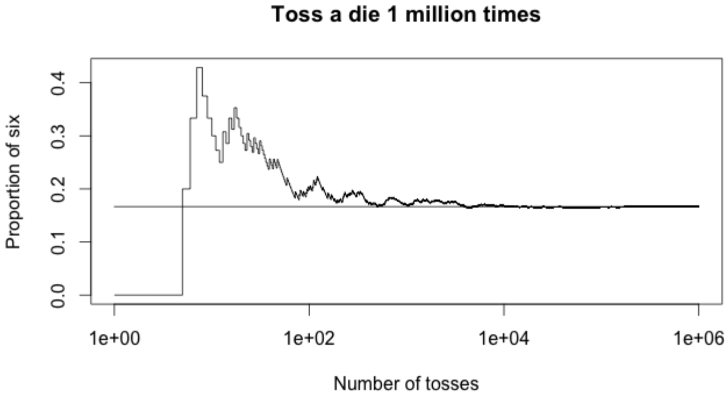
- Below are what we get:



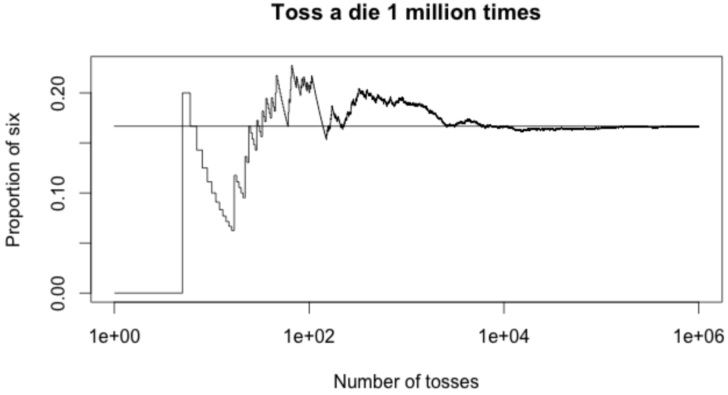

- How to improve the accuracy of estimation?

# Law of Large Numbers

- Toss a die one million times
- Record the proportion of die tosses that are 6 after each toss
- The proportion tends to get closer to the probability $1/6$ as the number of tosses increases



Repeat the same experiment: toss a die one million times



# Law of Large Numbers

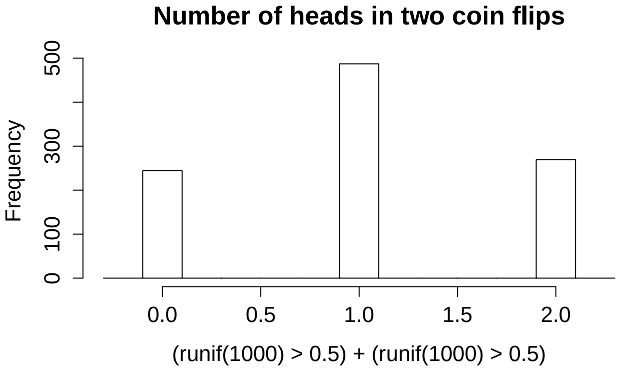
- Probability: the probability of a random outcome is the proportion of the times that it would occur if the experiment is repeated an infinite number of times.
- Law of Large Numbers: As the number of experiments $n$ increases towards $\infty$, the proportion of a particular random outcome $\hat{p}_n$ converges to the true probability $p$ of that outcome.

# Example: Flip Two Coins

In Excel: copy =(RAND()>0.5)+(RAND()>0.5) to *n* cells, use scatter plot to visualize the results

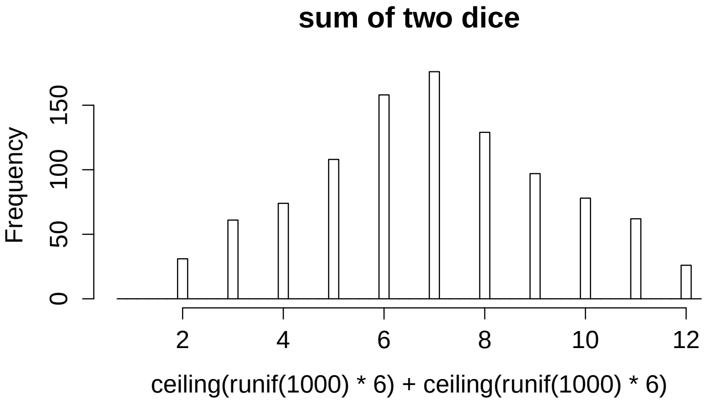
In R: x = (runif(n)>0.5) + (runif(n)>0.5)

use hist(x) and plot(x) to visualize the results



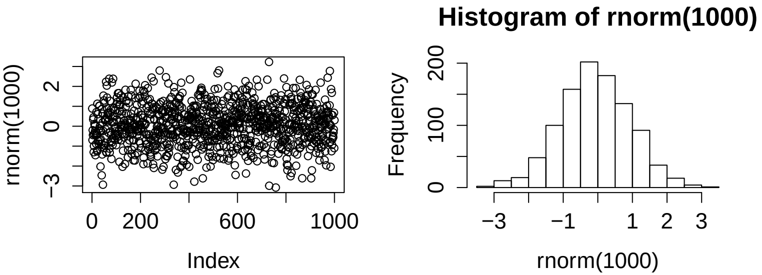
# Example: Toss Two Dice

In R: hist(ceiling(runif(n)*6) + ceiling(runif(n)*6) ) or
hist(sample(1:6,n,replace=TRUE) + sample(1:6,n,replace=TRUE))



# Normal Random Variable

- $X \sim \mathcal{N}(\mu, \sigma^2)$
- In Excel: $=$ NORMINV(RAND(), $\mu$, $\sigma$)
- In R: rnorm(n,$\mu$,$\sigma$) , where *n* is the number of normal random numbers that we ask R to generate
- Let's simulate 1000 standard normal random numbers and visualize the results:



# Normal Random Variable

- $X \sim \mathcal{N}(\mu, \sigma^2)$
- Let's simulate 10000 standard normal random numbers

```
x = rnorm(10000,0,1)
```

- Can you guess what are the following values:

```
x= rnorm(10000,0,1)
sum(x< 0)/10000
sum(x> -1 & x< 1)/10000
sum(x> -2 & x< 2)/10000
sum(x< -2)/10000
sum(x> -3 & x< 3)/10000
y=sort(x)
y[0.025*10000]
```

## Normal Distribution – Example

- Assume the annual returns on the SP500 are normally distributed with mean 6% and standard deviation 15%. SP500 $\sim \mathcal{N}(0.06, (0.15)^2)$.

- $(i)$ $Pr(SP500 < 0) =?$

```
n=10000
x= rnorm(n,0.06,0.15)
sum(x< 0)/n
```

- $(ii)$ $Pr(SP500 <?) = 0.02$

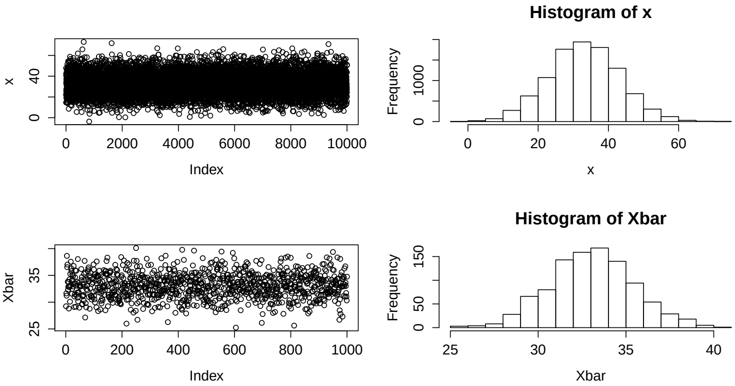
```
n=10000
x= rnorm(n,0.06,0.15)
y=sort(x)
y[round(0.02*n)]
```

# Sampling Distribution of the Sample Mean

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then the sample mean $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ follows a normal distribution: $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

```
par(mfrow=c(2,2))
n=10000
x= rnorm(n,33,10)
plot(x)
hist(x)
Xbar=matrix(0,1000,1)
for (i in 1:1000){
  Xbar[i]=mean(sample(x,size=16,replace=FALSE))
}
plot(Xbar)
hist(Xbar)
```

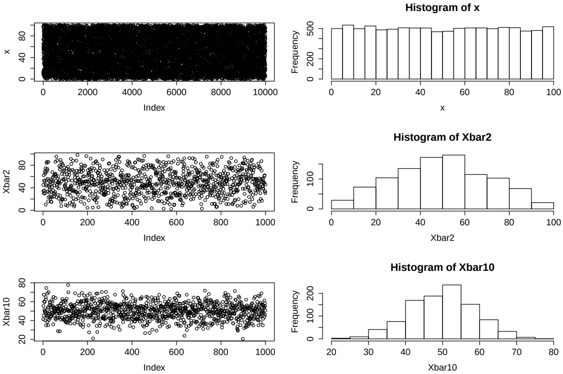
- Take 1000 samples of size 16 from the population, plot the means of these samples:



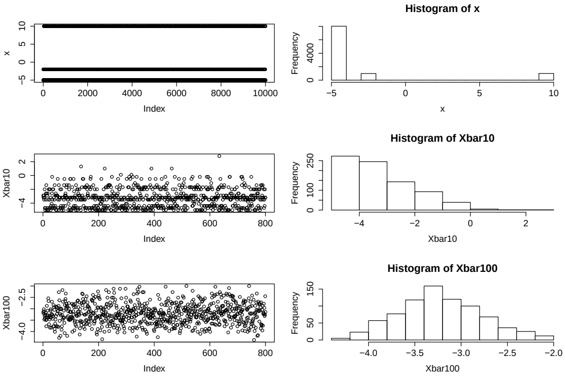
# Central Limit Theorem

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then the sample mean $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ follows a normal distribution: $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$
- What if $X$ does not follow a normal distribution?
- Let $X_1, \cdots, X_n$ be a random sample from a distribution (any distribution!) with a finite mean $\mu$ and a finite variance $\sigma^2$, the Central Limit Theorem tells us that as the sample size $n$ increases, the distribution of the sample mean $\bar{X}$ approaches the normal distribution $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
- We will demonstrate this using Simulation with R.

- Sampling distribution of the sample mean for a uniform distribution
- Uniform random numbers between 0 and 100, the distribution of the mean of two random points, and the distribution of the mean of 10 random points



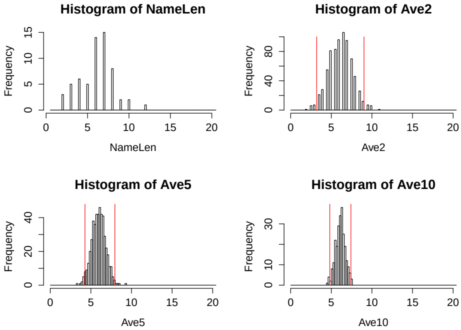
- Sampling distribution of the sample mean for an asymmetric distribution
- 10000 random numbers from that distribution, the distribution of the mean of 10 random points, and the distribution of the mean of 100 random points




- As long as $n$ is large enough, then the sample mean...

# Example: STA 371G Student Last Name Length

- Randomly sample *n* students from the class
- Record the average length of the students' last names.
- Repeat 1000 times and plot the histogram of the averages




- What can you observe from the figure?

# Simulation

To make predictions and decisions we build probability models that we think describe what is going on in the real world.

Sometimes the calculations we need to make are too hard!!

We can often do the calculations by *simulation*!!

*We can ask the computer for independent, and identically distributed (IID) draws from a given distribution!!!*

Let's suppose that demand every day is just another IID draw from:

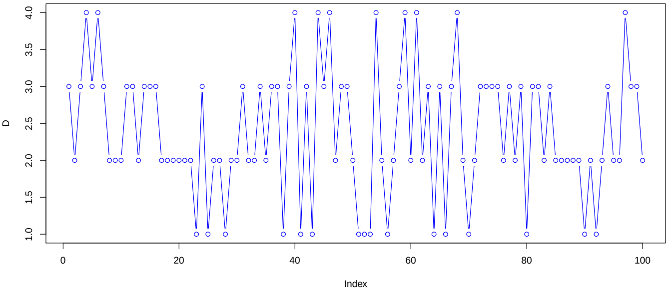
```
     P(D=d)    d
[1,]    0.2    1
[2,]    0.4    2
[3,]    0.3    3
[4,]    0.1    4
```

We can ask the computer to simulate what will happen over the next 100 days!

Now let $D_i$ be demand on day $i$.

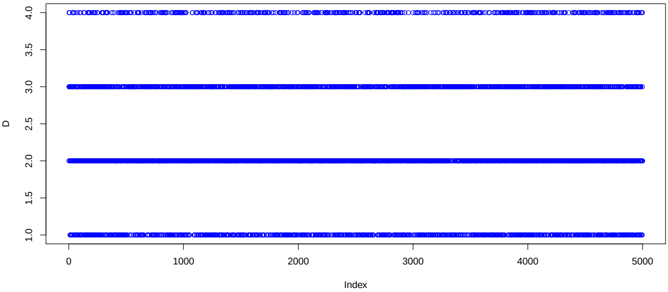
We assume the $D_i$ are IID and the common distribution is the one above.

Here are 100 IID draws from the computer:




How would you predict the next one?

Here are 5,000 IID draws from the computer:




About how many of the draws are equal to 3?

Here are the counts of the different outcomes
($D$ is 1,2,3, or 4).

```
D
   1    2    3    4 
 965 2107 1394  534 
```

If we divide each count by 5,000 we get

```
D
     1      2      3      4 
0.1930 0.4214 0.2788 0.1068 
```

which is pretty close to $P(D = i),\ i = 1, 2, 3, 4.$

## Monte Carlo Estimation of a Probability:

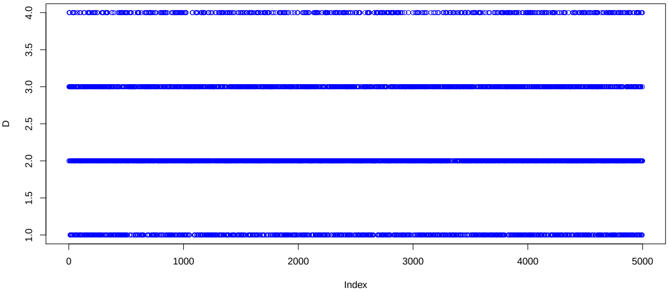
*If I want the probability of an event given a distribution, I can count how often the event happens with IID draws from that distribution!!!*

## Example:

$$P(D > 2) \approx \frac{1394 + 534}{5000} = \frac{1929}{5000} = .3856.$$

If I need a more accurate answer, I need to make more draws.

Here are 5,000 IID draws from the computer:




What is the average of the $D_i$ draws above?

Let $n_1$ be the number of times a $D$ is 1.

$n_1$ is 965.

Same for $n_2$, $n_3$, and $n_4$.

$n = 5000$.

$$
\begin{aligned}
\bar{D} &= \frac{D_1 + D_2 + \ldots + D_n}{n}. \\
&= \frac{n_1(1) + n_2(2) + n_3(3) + n_4(4)}{n} \\
&= \frac{n_1}{n}(1) + \frac{n_2}{n}(2) + \frac{n_3}{n}(3) + \frac{n_4}{n}(4) \\
&\approx P(D = 1)(1) + P(D = 2)(2) + P(D = 3)(3) + P(D = 4)(4)
\end{aligned}
$$

## Monte Carlo Estimation of an Expected Value:

*The average of a large number of IID draws from a distribution will be close to the expected value of the distribution you are drawing from.*

## Example:

For our 5,000 draws:

$\bar{D} = 2.229$

$$E(D) = .2(1) + .4(2) + .3(3) + .1(4) = 2.3.$$

If I need a more accurate answer, I need to make more draws.

Note:

We can also use the idea to *interpret* the expected value of a random variable.

If we "do it over and over" *independently* the expected value is the *long run average*.

Example:

If daily demand is IID, then $E(D) = 2.3$ is long run average daily demand.

# Example, Simulating Weekly Demand

We have a distribution describing our uncertain demand for a day.

What about demand for next week!!

We can do this by simulation.

We will assume that the demand is IID on the 5 week days.

Let $D_i$ be the demand on week-day $i$, $i = 1, 2, 3, 4, 5$.

We draw the $D_i$ and sum them to get the demand for the week.

Let $W$ denote the demand for the whole week.

$$W = D_1 + D_2 + D_3 + D_4 + D_5.$$

I got, $D_1 = 2$, $D_2 = 3$, $D_3 = 2$, $D_4 = 2$, and $D_5 = 3$.

So, for $W$, I got 12.

The 12, is one draw from the distribution of $W$!!!

What use is that???!!

If I do it over and over, I can get Monte Carlo estimates of anything I want about the distribution of $W$!!!

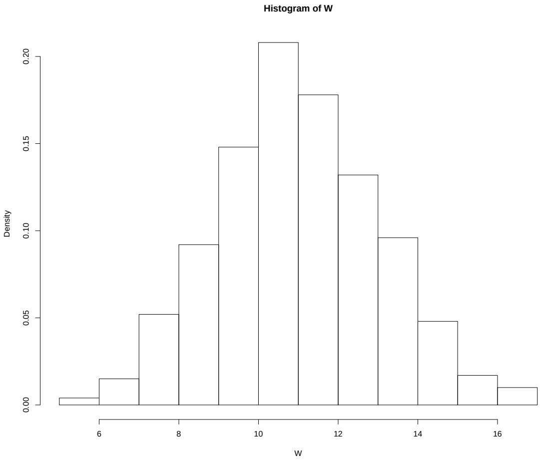
I draw an array of $D$ values.

Each row, corresponds to the draws for a week.

The sum of values in each row is the draws of total demand for that week.

```
     D1 D2 D3 D4 D5  W
[1,]  2  3  2  2  3 12
[2,]  3  3  3  3  4 16
[3,]  4  2  2  2  1 11
[4,]  3  3  3  3  2 14
[5,]  4  4  3  2  2 15
...
[997,]  2  3  2  2  2 11
[998,]  2  2  2  1  3 10
[999,]  2  2  4  2  4 14
[1000,]  2  2  1  4  2 11
```

Here, is the histogram of the $W$ draws:




$E(W) \approx \bar{W} = 11.5.$

Suppose my inventory is 12.

What is the chance I run out by the end of week?

I can use Monte Carlo to estimate this by simply counting the fraction of $W$ draws which are bigger than 12!!

30% of the $W$ draws are bigger than 12!!

$$P(W > 12) \approx .3.$$

# Monte Carlo Simulation

Monte Carlo:

1. Make up a bunch of random variables describing what you think might happen.
2. Draw the variables and compute something of interest.
3. Repeat the draws "many" times.
4. Look at all the draws of the "quantity of interest" to see what the likely outcomes for it are.