STA 371G: Statistics and Modeling

Introduction to Monte Carlo Simulation: Examples

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

http://mingyuanzhou.github.io/STA371G

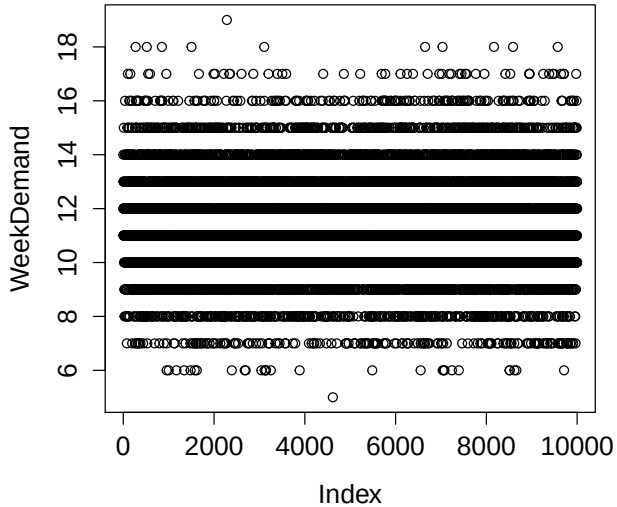
Simulation and Decision

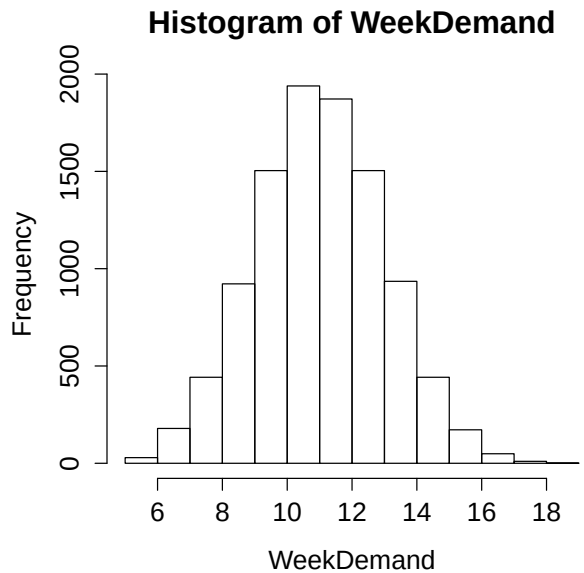
With 12 units in our inventory, we run a pretty high chance ($\approx .3$)

or running out before the end of the week!

We have the option of ordering 4 more units so that our inventory goes up to 16.

Should we do it?





What are our costs and benefits?

In this case, rather than looking at payoffs we will use *losses*.

For each possible W outcome we will determine our loss if we order (inventory $= 16$) or if we don't order (inventory $= 12$).

Our loss table would have 16 rows (W from 5 to 20) and two columns (no order:inventory 12, order:inventory 16).

Rather than give the table, let's plot the losses:
The loss for

every possible w

and two possible

actions:

(no order/order).

For small w

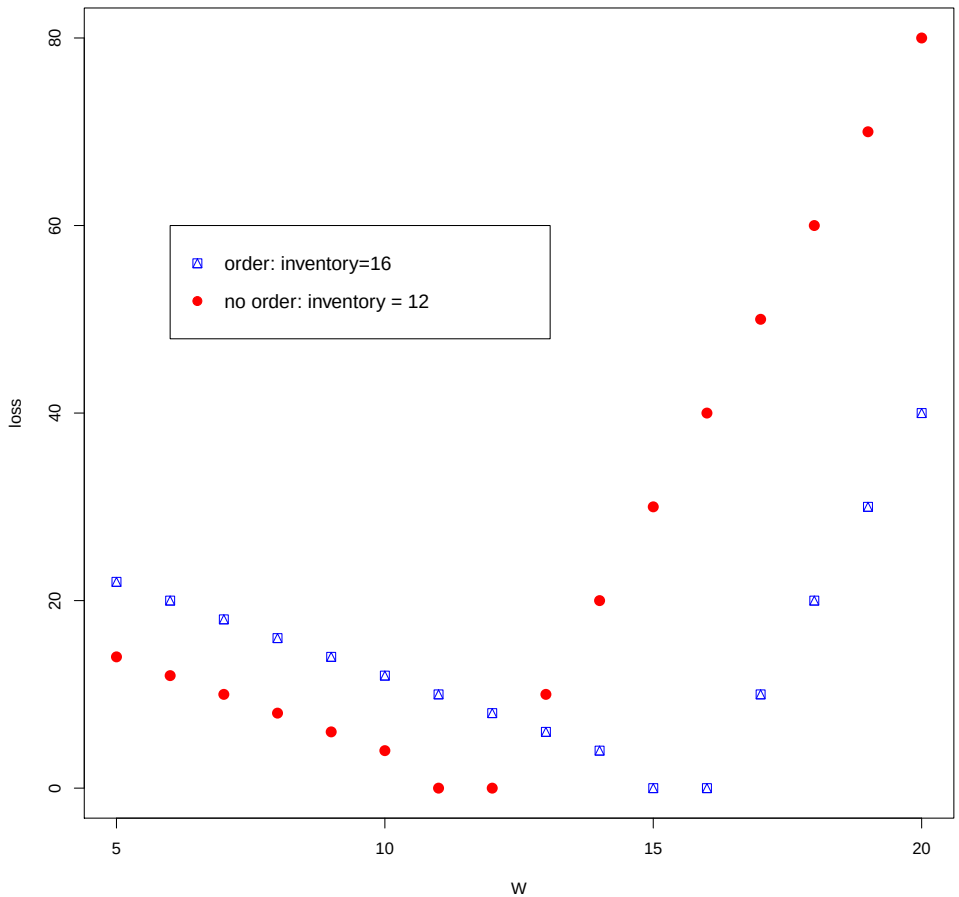
you are

overstocked.

For big w

you are

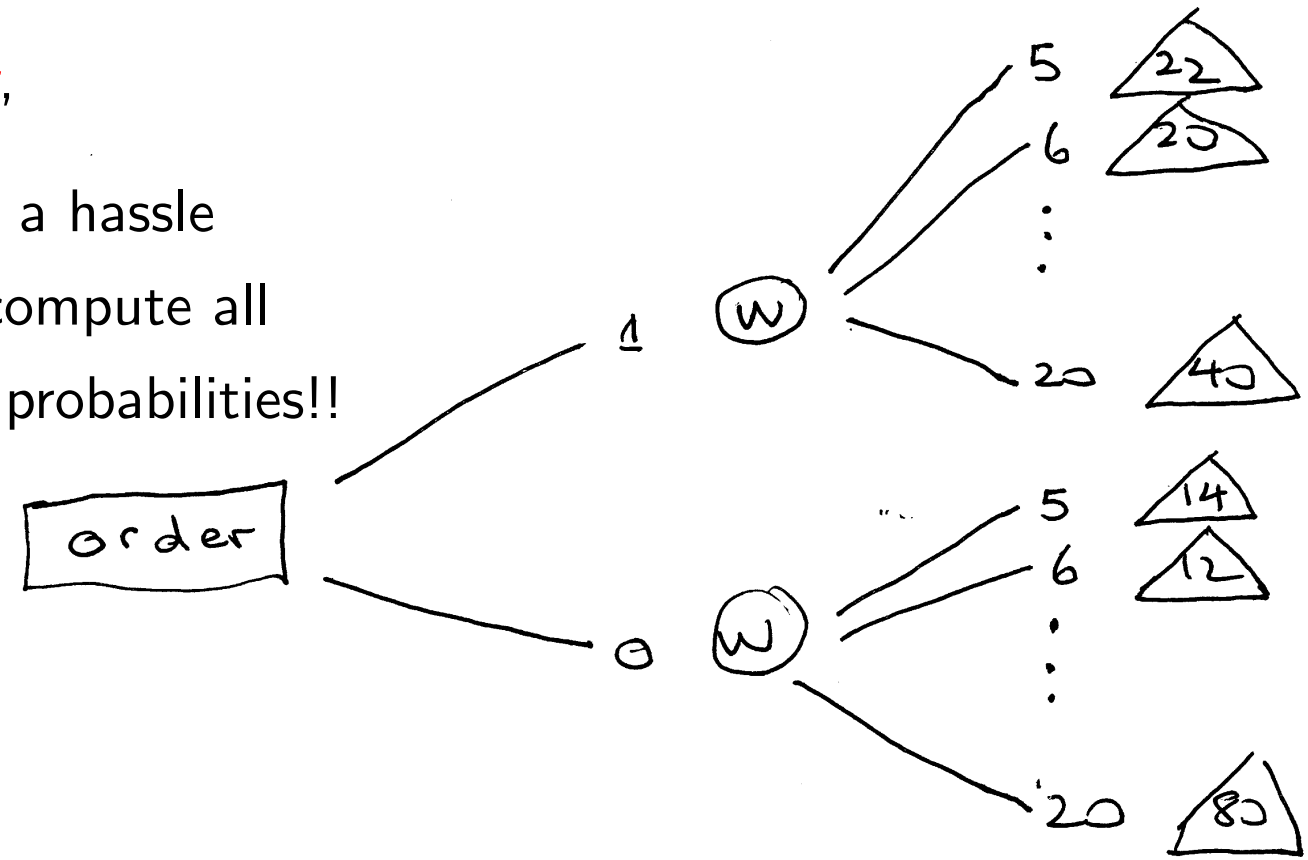
understocked.



In principle, we could do
what we did before.

But,

it is a hassle
to compute all
the probabilities!!



We can easily do all
the computation
by simulation!!

For each simulation of W compute the loss under order and no-order.

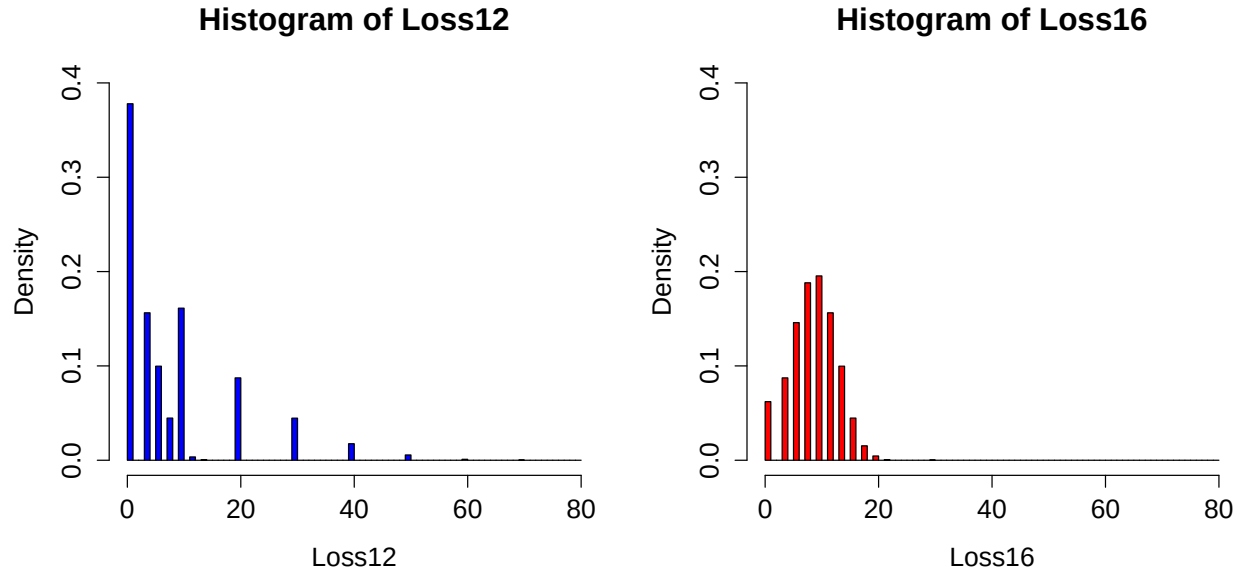
Here are the results for the first 8 draws:

D1	D2	D3	D4	D5	W	l12	l16
2	3	2	4	2	13	10	6
3	3	2	4	2	14	20	4
4	3	2	2	2	13	10	6
3	3	2	4	1	13	10	6
4	2	2	2	3	13	10	6
3	3	4	2	2	14	20	4
4	3	3	3	1	14	20	4
3	1	2	2	3	11	0	10

The `l12` column is the loss for each W draw if we go have inventory 12 (no order).

The `l16` column is the loss for each W draw if we go have inventory 16 (order).

Here are the histograms of the L12 (on left) and L16 (on right) draws (based on 10,000 iid samples).



The average of the 10,000 L12 draws is **7.4**.

It is our Monte Carlo simulation estimate of the expected loss with inventory 12.

The average of the 10,000 L16 draws is **9.1**.

It is our Monte Carlo simulation estimate of the expected loss with inventory 16

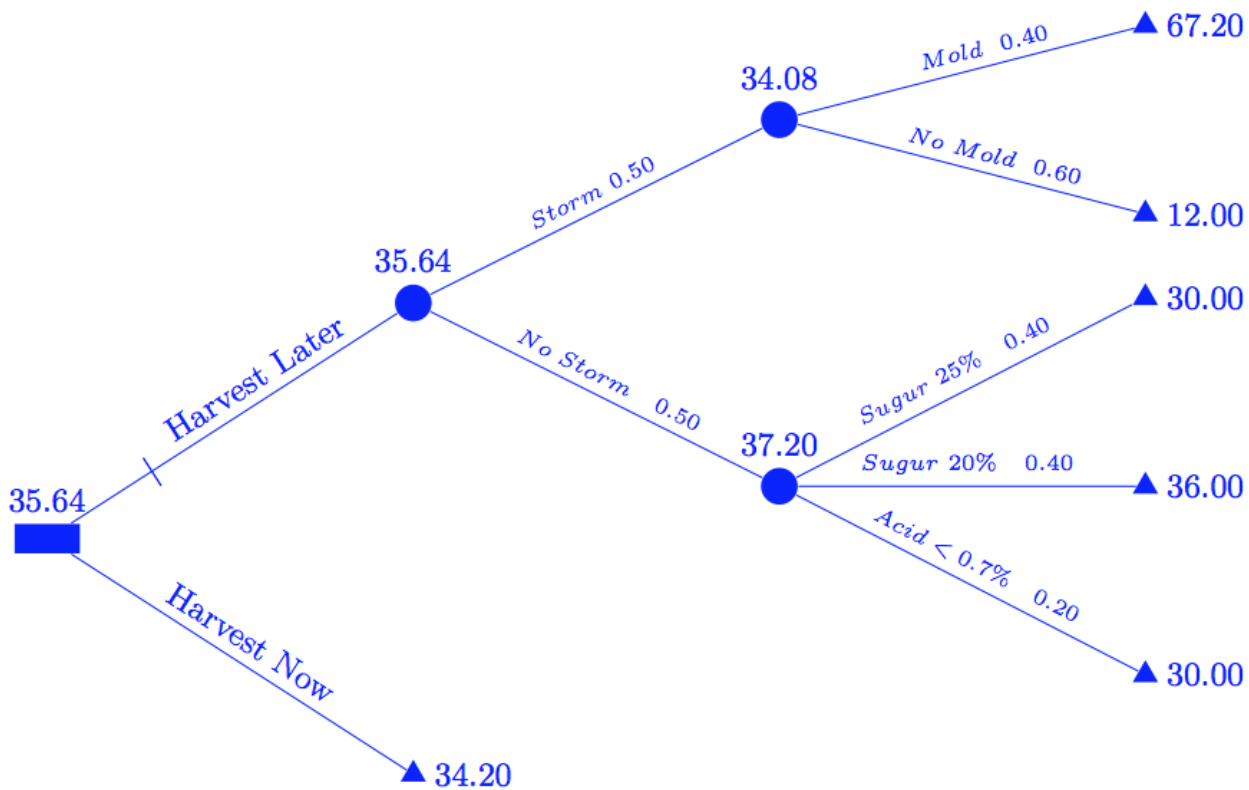
What should you do?

If you want to minimize the expected value of your loss then you do not order.

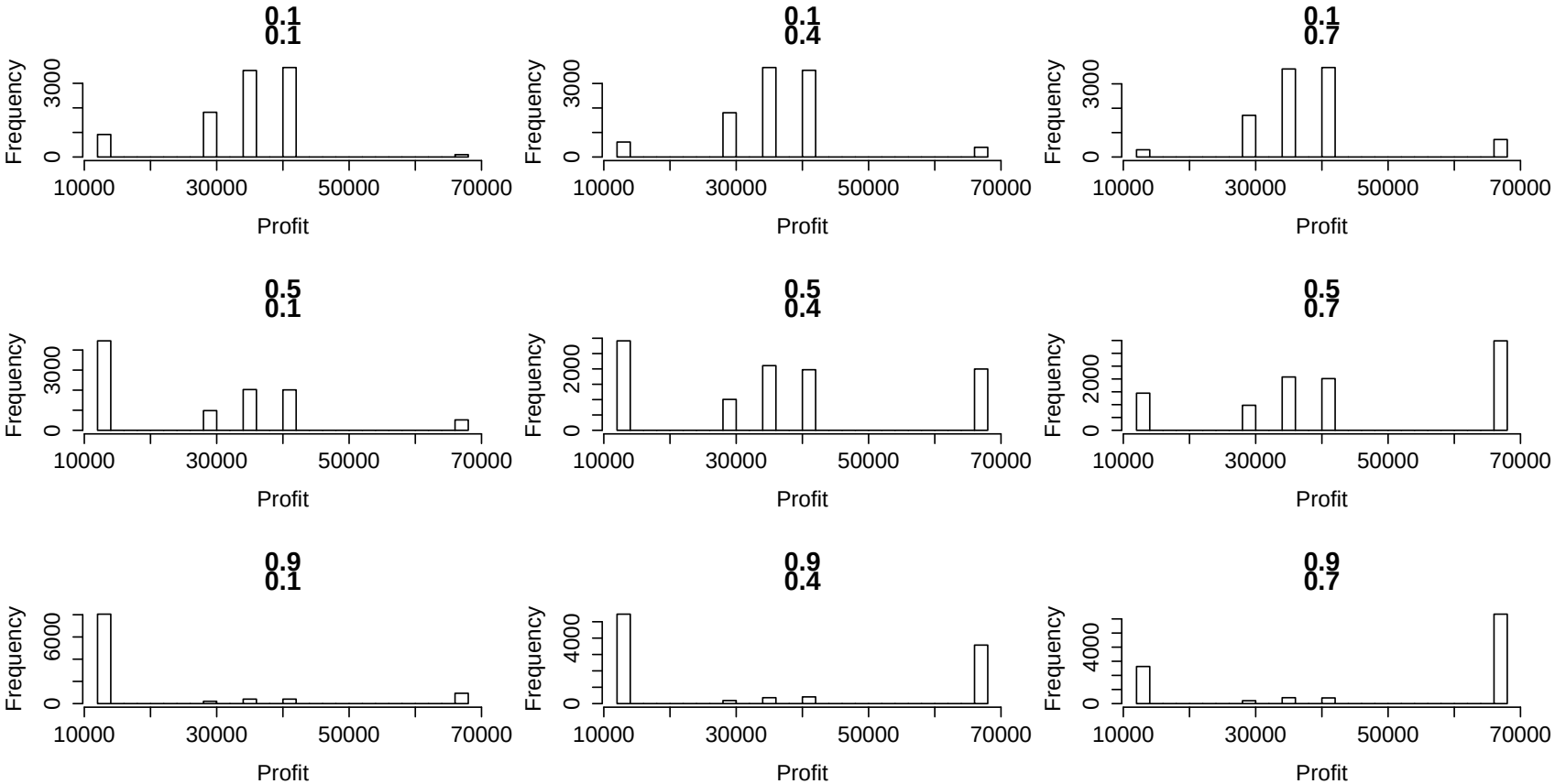
However, if you really want to avoid a big loss, maybe you should order!!

Freemark Abbey Winery

Let's consider Harvest Later. How to find the probability distributions of different outcomes when $P(\text{Storm}) = 0.1$, 0.5, or 0.9 and $P(\text{Mold}) = 0.1$, 0.4 or 0.7?



Below are the histograms of 10,000 iid draws for $P(\text{Storm}) = 0.1$, 0.5, or 0.9 and $P(\text{Mold}) = 0.1$, 0.4 or 0.7.



Example: Portfolio Analysis

- $100,000 is to be invested in a portfolio of 8 stocks
- What would be the expected return if the portfolio is constructed according to the table shown below?

	Stock 1	Stock 2	Stock 3	Stock 4	Stock 5	Stock 6	Stock 7	Stock 8
Weights	10500	16300	9600	9300	9500	15400	14300	15100
Means	0.101	0.073	0.118	0.099	0.118	0.091	0.096	0.123
Stdevs	0.143	0.139	0.152	0.158	0.1731	0.174	0.133	0.188

- What would be the variance of the return?
- What's the probability that the return is negative?
- What's the probability that the return is more than $10,000?

Sum of Independent Random Variables

- if $Y = aX$, then $E(Y) = aE(X)$ and $Var(Y) = a^2 Var(X)$
- If $Y = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$, then

$$E(Y) = a_0 + a_1 E(X_1) + a_2 E(X_2) + \cdots + a_n E(X_n)$$

- If X_i and X_j are independent for $i \neq j$, then we further have

$$Var(Y) = a_1^2 Var(X_1) + a_2^2 Var(X_2) + \cdots + a_n^2 Var(X_n)$$

Correlated Random Variables

- Covariance between X and Y

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \sum_i (x_i - \mathbb{E}[X])(y_i - \mathbb{E}[Y])P(X = x_i, Y = y_i)\end{aligned}$$

- Correlation between X and Y

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Sum of Correlated Random Variables

If X_j and X_j are not independent:

- If $Y = a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n$, then

$$E(Y) = a_0 + a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n)$$

and

$$Var(Y) = \sum_{i=1}^{n} a_i^2 Var(X_i) + \sum_{i<j} 2a_i a_j Cov(X_i, X_j)$$

Example: Portfolio Analysis

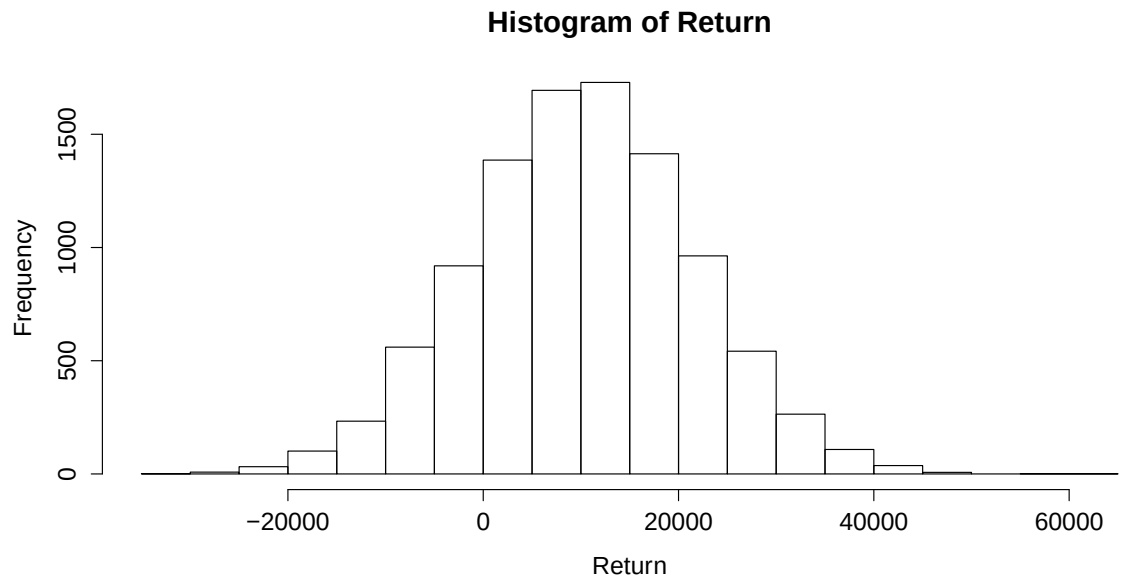
- The returns of these stocks are correlated
- The covariance matrix:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]
[1,]	0.0204	0.0047	0.0061	0.0107	0.0157	0.0120	0.0077	0.0121
[2,]	0.0047	0.0192	0.0065	0.0131	0.0081	0.0151	0.0128	0.0099
[3,]	0.0061	0.0065	0.0230	0.0062	0.0182	0.0081	0.0051	0.0163
[4,]	0.0107	0.0131	0.0062	0.0249	0.0151	0.0112	0.0080	0.0164
[5,]	0.0157	0.0081	0.0182	0.0151	0.0300	0.0146	0.0075	0.0148
[6,]	0.0120	0.0151	0.0081	0.0112	0.0146	0.0303	0.0165	0.0094
[7,]	0.0077	0.0128	0.0051	0.0080	0.0075	0.0165	0.0178	0.0128
[8,]	0.0121	0.0099	0.0163	0.0164	0.0148	0.0094	0.0128	0.0353

- What would be the variance of the return?
- What's the probability that the return is negative?
- What's the probability that the return is more than $30,000?

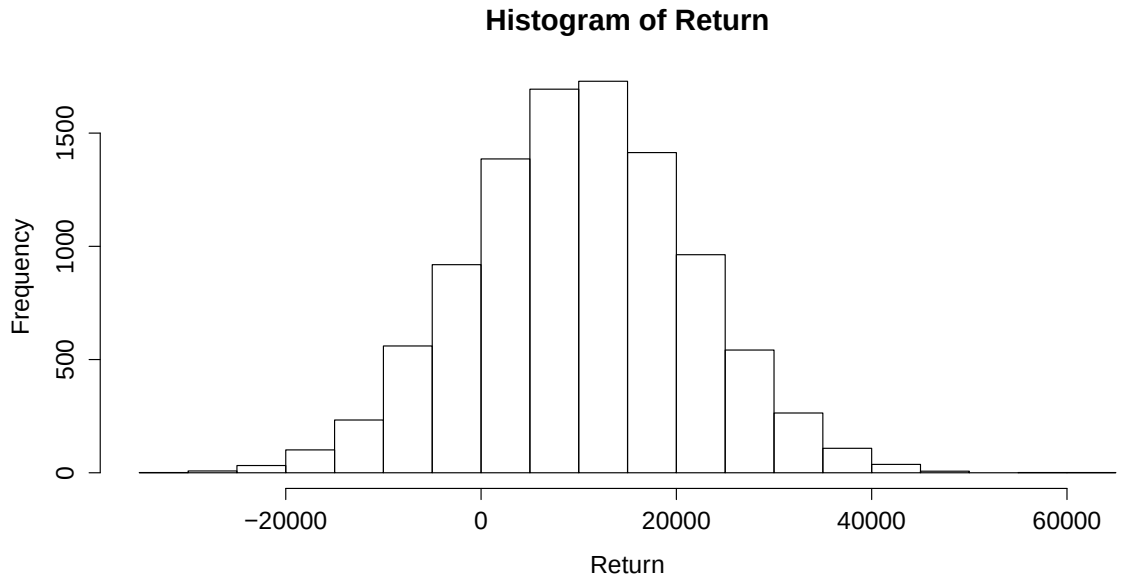
Example: Portfolio Analysis

- We can simulate iid multivariate normal random variables to find the distribution of the portfolio return



- With calculation, we find that $\mu = 10056$ and $\sigma = 11468$
- With 10,000 iid draws, our estimates of the mean and standard deviation are 10153 and 11405, respectively.

Example: Portfolio Analysis



- With simulation, we find that $P(Return < 0) \approx 0.185$
- With simulation, we find that $P(Return > 30,000) \approx 0.042$

Example: Portfolio Analysis

```
######Portfolio analysis
Weights =  c(10500, 16300,9600,9300,9500,15400,14300,15100)
Mu =  c(10.1,7.3,11.8,9.9,11.8,9.1,9.6,12.3)/100
Sigma=matrix(
  +c(0.0154,  0.0047,  0.0061,0.0107,0.0157,0.0120,0.0077,0.0121,
     +0.0047,0.0142,0.0065,0.0131,0.0081,0.0151,0.0128,0.0099,
     +0.0061,0.0065,0.0180,0.0062,0.0182,0.0081,0.0051,0.0163,
     +0.0107,0.0131,0.0062,0.0199,0.0151,0.0112,0.0080,0.0164,
     +0.0157,0.0081,0.0182,0.0151,0.0250,0.0146,0.0075,0.0148,
     +0.0120,0.0151,0.0081,0.0112,0.0146,0.0253,0.0165,0.0094,
     +0.0077,0.0128,0.0051,0.0080,0.0075,0.0165,0.0128,0.0128,
     +0.0121,0.0099,0.0163,0.0164,0.0148,0.0094,0.0128,0.0303),
  +8,8)+diag(0.005,8,8)
library(MASS)
Return=matrix(0,1,10000)
for (iter in (1:10000)){
  Return[iter] = sum(Weights*mvrnorm(n=1,Mu,Sigma))
}
hist(Return)
#Truth
c(sum(Weights*Mu), sqrt(Weights%*%Sigma%*%Weights))
#Simulated
c(mean(Return), sd(Return))
sum(Return<0)/10000
sum(Return>30000)/10000
```

Simulate the number of non-Texas Residents

Suppose 87% of the 4000 McCombs students are Texas residents, find the distribution of non-Texas residents from a sample (without replacement) of 60 McCombs students.

The number of success in n draws without replacement from a finite population of size N follows a hypergeometric distribution...

Although we have never discussed this distribution in this class, we can use simulation to get our answers!

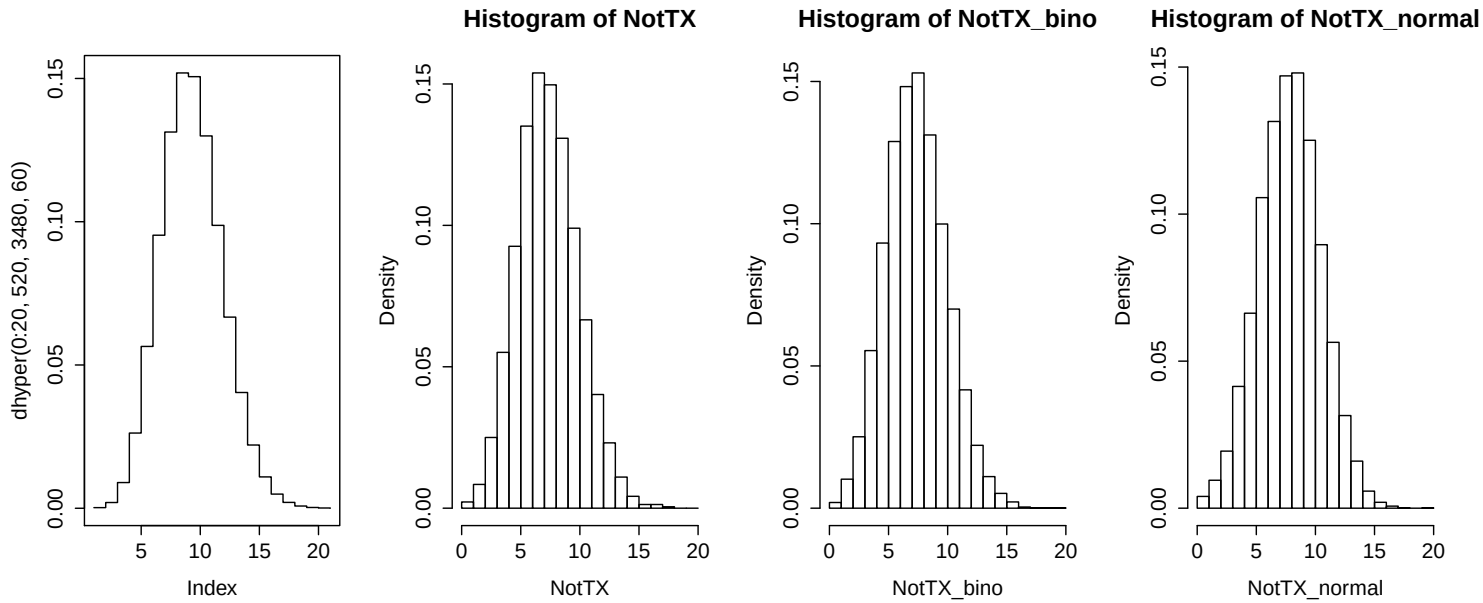
- Exact: sample 60 students without replacement from a population with 3480 TX residents and 520 non-TX residents.
- Approximate: draw $X \sim \text{Binomial}(n = 60, p = 0.13)$
- Approximate: draw $X \sim \mathcal{N}(\mu, \sigma^2)$, where $\mu = np = 60 * 0.13$ and $\sigma = \sqrt{np(1-p)} = \sqrt{60 * 0.13 * 0.87}$.

Simulate the number of non-Texas Residents

```
par(mfrow=c(1,4))
plot(dhyper(0:20,520,3480,60),type='s')
sum(dhyper(11:60,520,3480,60))

McCombs=c(matrix(1,1,870*3),matrix(0,1,130*3)) #4000 students, 87% of which are from Texas
#Find the distribution of the number of non-Texas residents
NotTX = matrix(0,10000,1)
for (iter in 1:10000){
  NotTX[iter]=60-sum(sample(x=McCombs,size=60,replace=FALSE))
}
hist(NotTX,freq=FALSE,breaks=0:20)
sum(NotTX>10)/10000

NotTX_bino = rbinom(n=10000,size=60,prob=1-0.87)
hist(NotTX_bino,freq=FALSE,breaks=0:20)
sum(NotTX_bino>10)/10000

NotTX_normal = rnorm(10000,60*(1-0.87),sqrt(60*0.87*(1-0.87)))
NotTX_normal[NotTX_normal<0]=0
hist(NotTX_normal,freq=FALSE,breaks=0:20)
sum(NotTX_normal>10.5)/10000
```

Simulate the number of non-Texas Residents

- Let's do 10000 IID draws and visualize the results:



- They look very similar!

Simulate the number of non-Texas Residents

- The binomial distribution is accurate enough to describe the number of Yes answers in a random sample of size $n = 60$ from a finite population of 4000.
- The binomial distribution can be approximated with a normal distribution, since $n = 60$ is not too small and $p = 0.13$ is not too close to 0 or 1.
- $P(\text{non-Texas Residents} \geq 11)$
  - Exact: 0.1483
  - Exact Simulate: 0.1482
  - Binomial Simulate: 0.1529
  - Normal Simulate: 0.1538

How many consecutive successful free throws?

Suppose you go to the Gregory Gym and make free throws. Each time you shoot the basketball at the free throw line, you succeeds with probability p. What would be the distribution of the number of consecutive successful free throws?

If each trail succeeds with probability p, then the random number of successes in independent trials before observing a failure follows a geometric distribution...

Although we have never discussed this distribution before, we can use simulation to get our answers!

Your job is to describe a method to simulate this random number repeatedly.

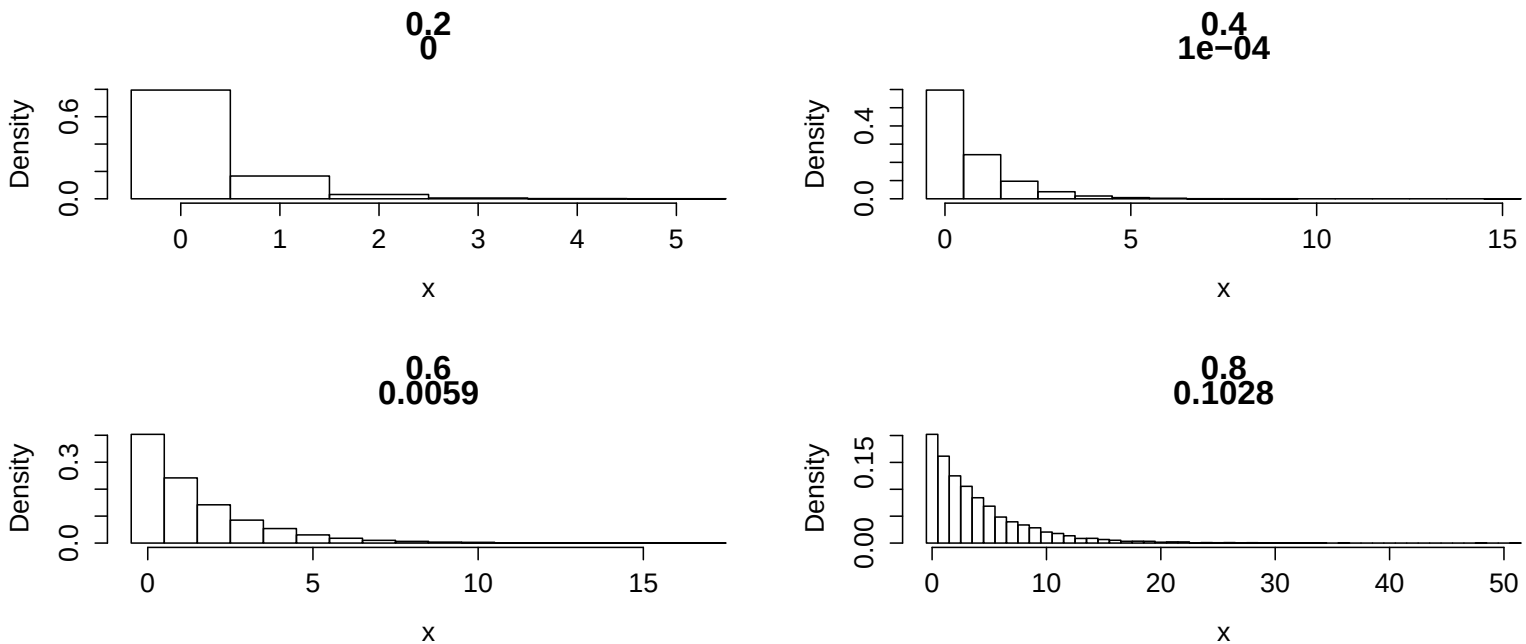
How many consecutive successful free throws?

```
###Free throws###
#p is the probability to sucessfully make a free throw
par(mfrow=c(2,2))
for (p in c(0.2,0.4,0.6,0.8))
{
  x = matrix(0,10000,1)
  for (iter in 1:10000){
    count=0
    while (runif(1)<p){
      count = count+1
    }
    x[iter]=count
  }
  hist(x,freq=FALSE,breaks=seq(-0.5,max(x)+0.5,1),main=c(p,sum(x >= 10)/10000))
}
```

How many consecutive successful free throws?

With simulation we find that

- If $p = 0.2$, then $P(X \geq 10) \approx 0$ (Truth: $1.02 * 10^{-7}$)
- If $p = 0.4$, then $P(X \geq 10) \approx 10^{-4}$ (Truth: $1.05 * 10^{-4}$)
- If $p = 0.6$, then $P(X \geq 10) \approx 0.0059$ (Truth: $6.05 * 10^{-3}$)
- If $p = 0.8$, then $P(X \geq 10) \approx 0.103$ (Truth: 0.107)



Student t distribution

Recall that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

- If we know the population variance, then we estimate the confidence intervals using $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$
- If we do not know the population variance, then we estimate the confidence intervals using $\frac{\bar{X}-\mu}{s/\sqrt{n}} = \sim t_{(n-1)}$

As n increases, $t_{(n-1)}$ (t distribution with $n-1$ degrees of freedom) approaches the standard normal distribution. But when n is small, t distribution can be quite different from the standard normal distribution.

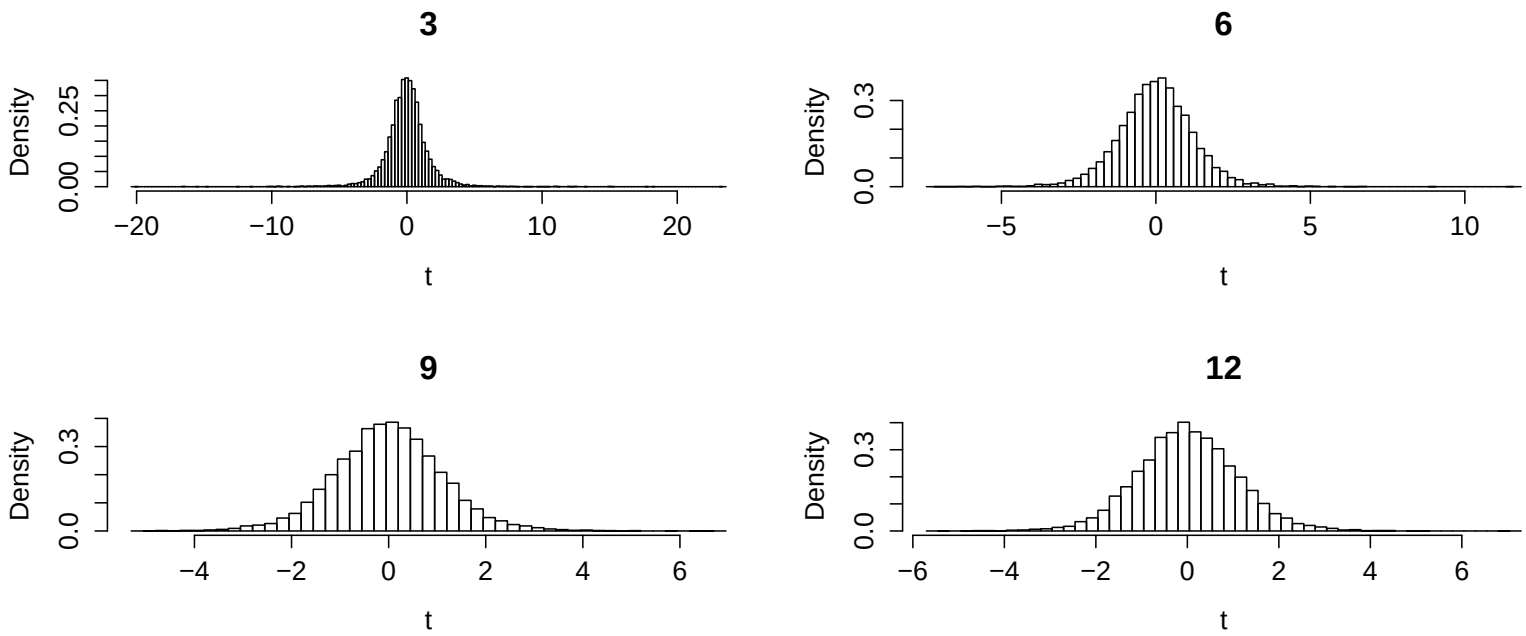
Your job is to describe a procedure to simulate $t_{(n-1)}$ random variables.

Student t distribution

```
###Simulate t distribution####
par(mfrow=c(2,2))
for (n in c(4,7,10,13)){
  t=matrix(0,10000,1)
  for (iter in 1:10000){
    mu=runif(1)*100
    sigma=runif(1)*100
    x=rnorm(n,mu,sigma)
    t[iter]=(mean(x)-mu)/(sd(x)/sqrt(n))
  }
  hist(t,freq=FALSE,breaks=seq(-0.5+min(t),max(t)+0.5,0.25),main=n-1)
}
```

Student t distribution

Simulate t random variables with 3, 6, 9 or 12 degrees of freedom.



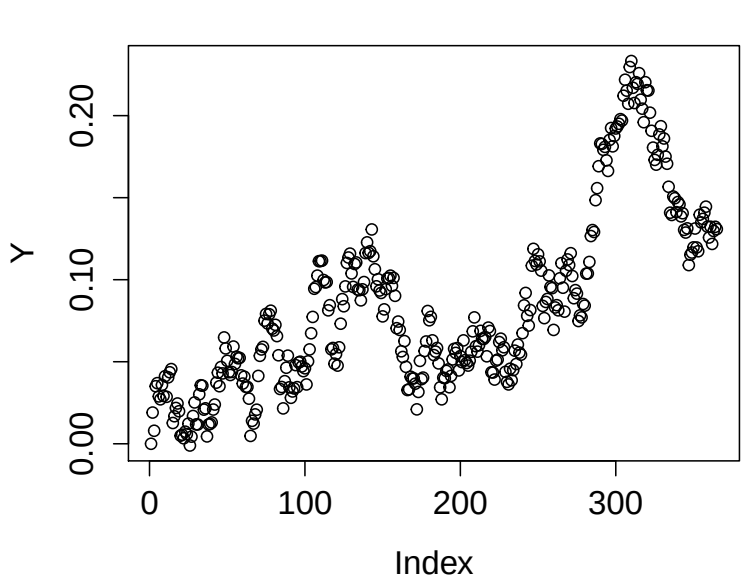
Simulate a Random Walk Model

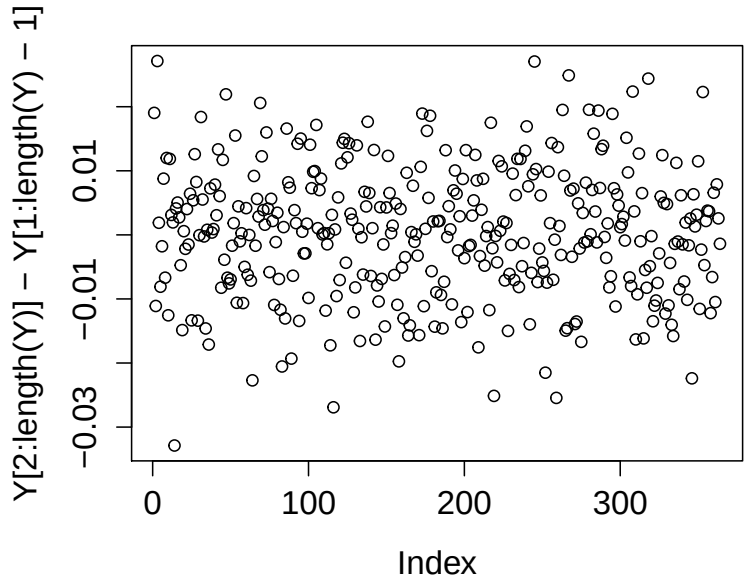
- $Y_t = \mu + Y_{t-1} + \epsilon_t,\ \epsilon_t \sim \mathcal{N}(0, \sigma^2)$
- In Excel: enter $= \mu +$ B1+NORMINV(RAND(), 0, σ) at cell B2 and copy B2 all the way down
- In R:

```
Y=matrix(0,365,1)
for (i in 1:364){
      Y[i+1]=Y[i]+rnorm(1,mu,sigma)
}
plot(Y)
plot(Y[2:length(Y)]-Y[1:length(Y)-1])
```

Simulate a Random Walk Model

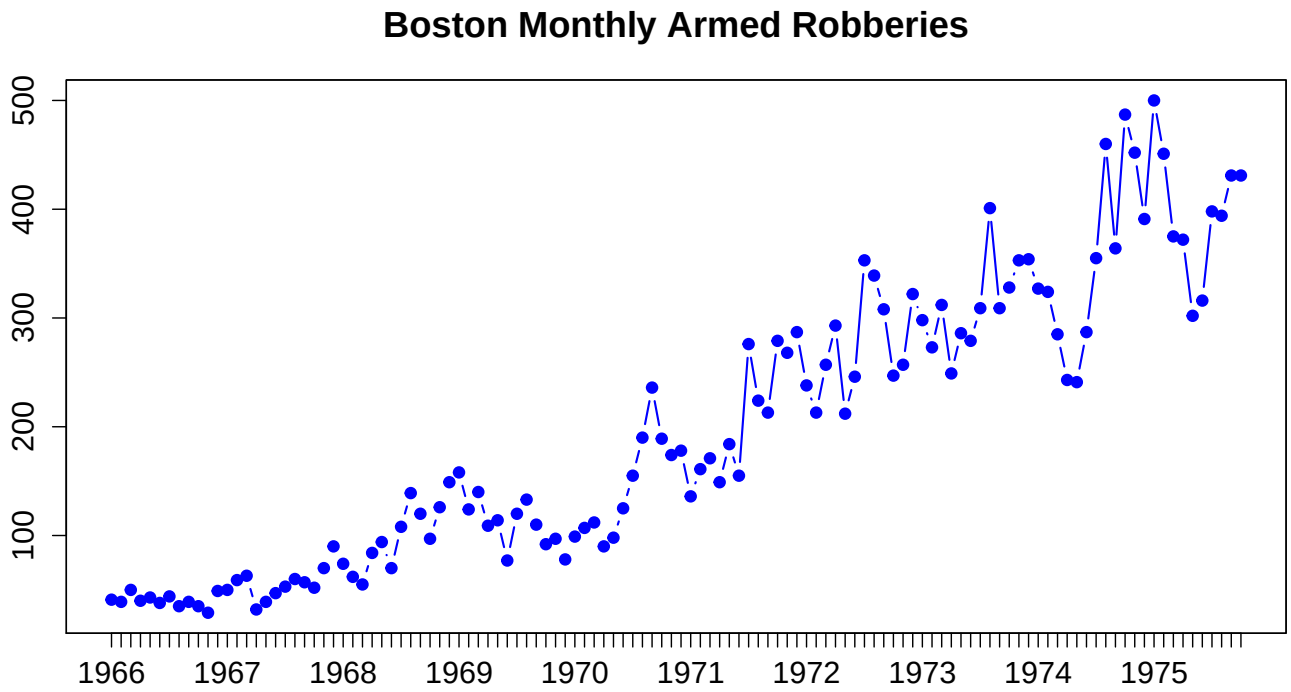
- $Y_t = Y_{t-1} + \epsilon_t,\ \epsilon_t \sim \mathcal{N}(0, 0.01^2)$





Example: Monthly Boston Armed Robberies

Jan.1966-Oct.1975



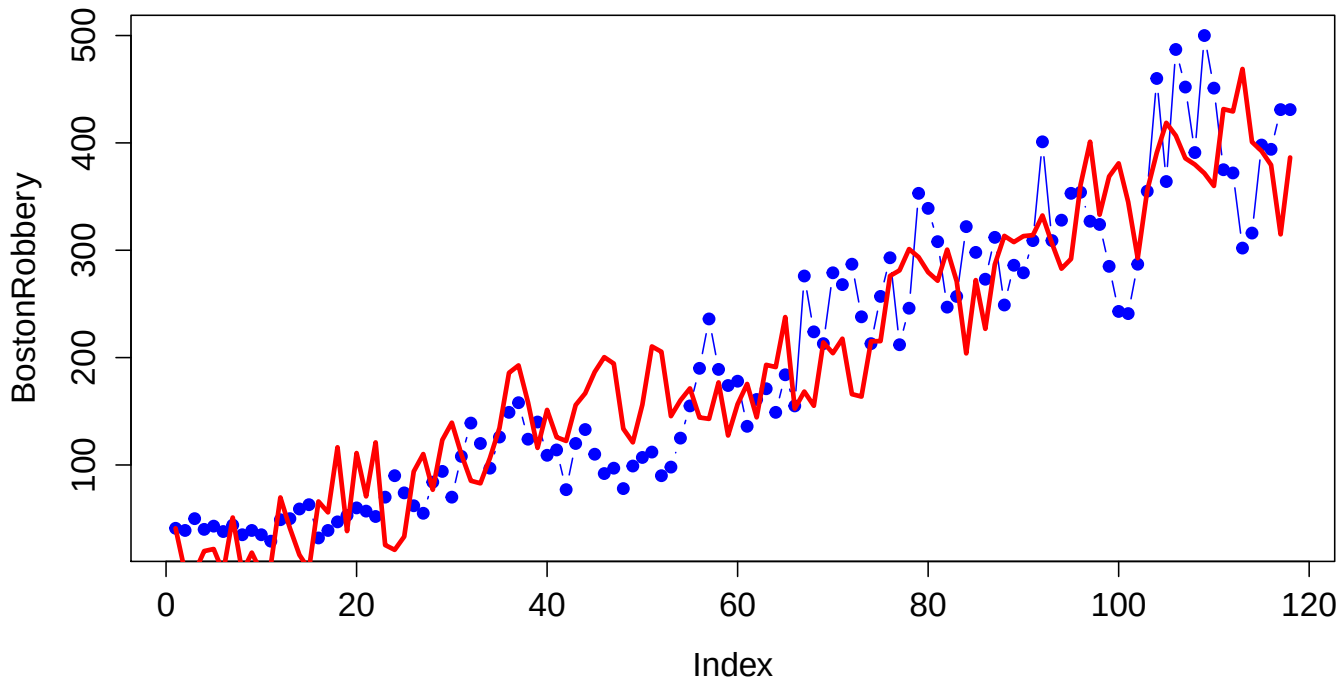
Example: Boston Armed Robberies

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 t + \epsilon_t$$

```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               -4.06758    6.89469  -0.590    0.556
BostonRobbery[1:length(BostonRobbery) - 1]  0.56397    0.07691   7.333 3.52e-11 ***
Time                                        1.56124    0.28733   5.434 3.17e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.7 on 114 degrees of freedom
Multiple R-squared:  0.9189,    Adjusted R-squared:  0.9175
F-statistic: 645.9 on 2 and 114 DF,  p-value: < 2.2e-16
```

Simulated Time Series

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 t + \epsilon_t$$



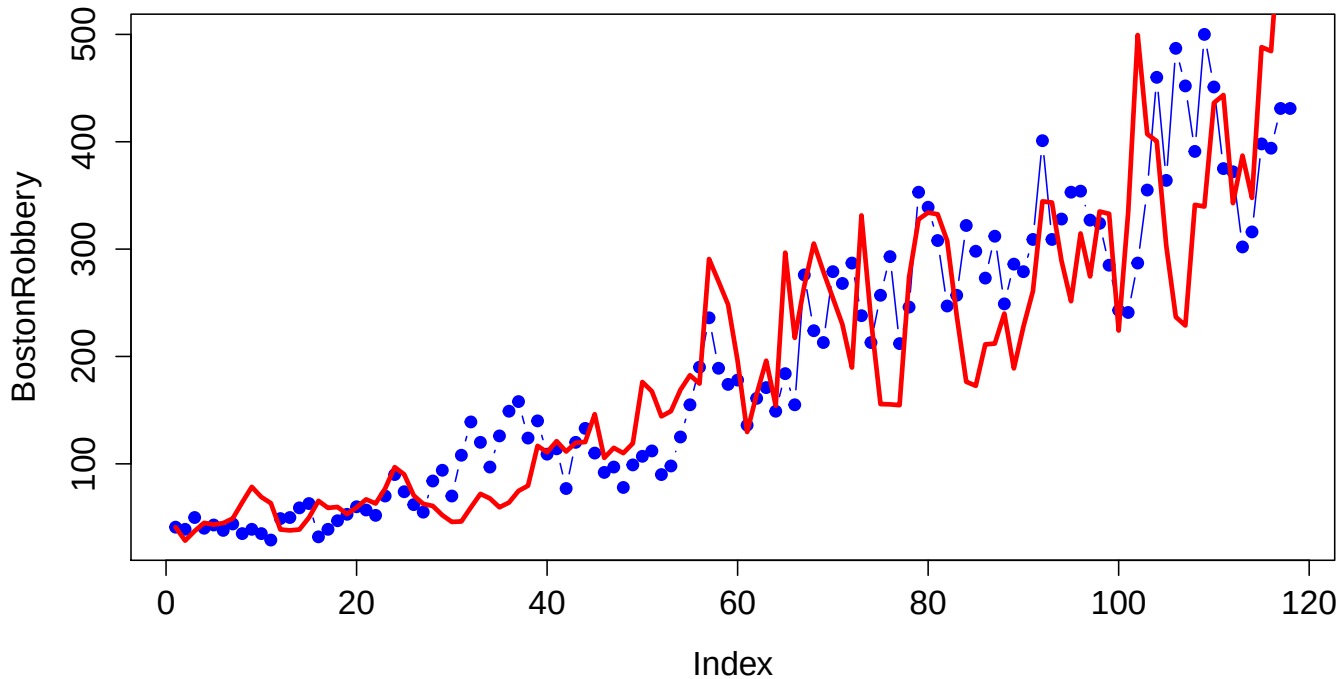
Example: Boston Armed Robberies

$$Log(Y_t) = \beta_0 + \beta_1 Log(Y_{t-1}) + \beta_2 t + \epsilon_t$$

```
Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     1.329217   0.268444   4.952 2.57e-06 ***
log(BostonRobbery[1:length(BostonRobbery) - 1]) 0.648915   0.071723   9.047 4.57e-15 ***
Time                                            0.007598   0.001658   4.582 1.19e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1898 on 114 degrees of freedom
Multiple R-squared:  0.9419,    Adjusted R-squared:  0.9409
F-statistic: 924.5 on 2 and 114 DF,  p-value: < 2.2e-16
```

Simulated Time Series

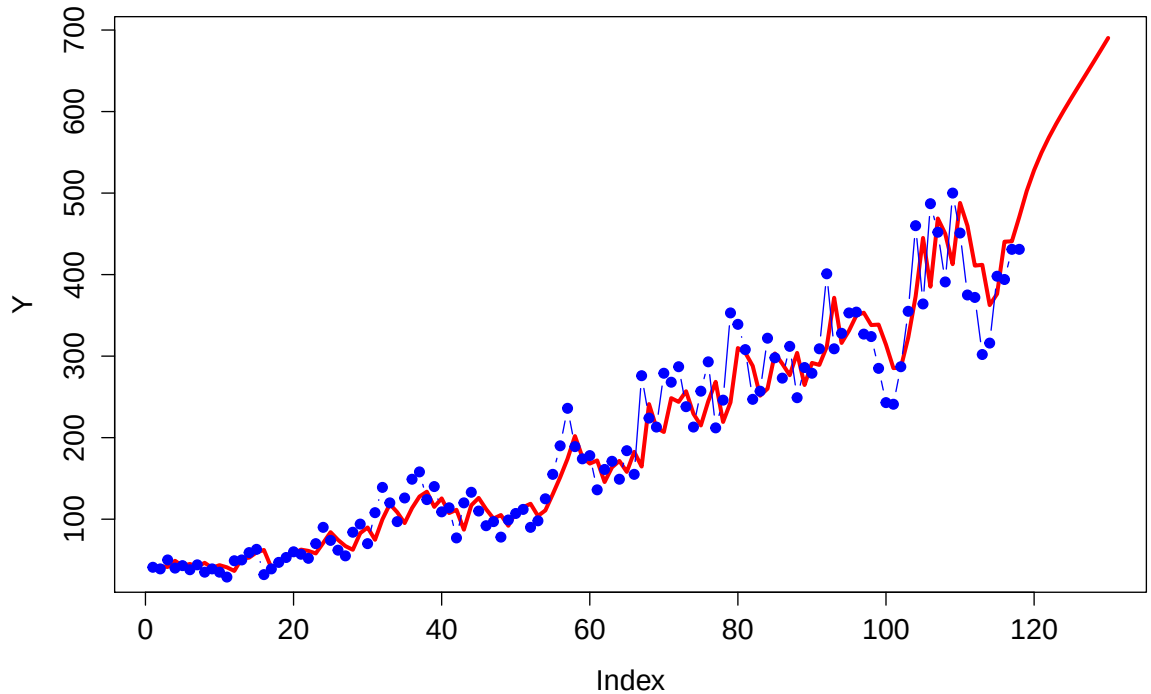
$$Log(Y_t) = \beta_0 + \beta_1 Log(Y_{t-1}) + \beta_2 t + \epsilon_t$$



Prediction

Calculating the predicted values is easy. The difficult part is to describe the uncertainty about these predictions.

Your job is to use simulation to find the 95% prediction intervals for the number of armed robberies for the next 12 months.



Prediction

$$Log(Y_t) = \beta_0 + \beta_1 Log(Y_{t-1}) + \beta_2 t + \epsilon_t$$