Intro to Simulation in R

ZERO TO HERO

THE DEFINITIVE INTRO SERIES

Corey Chivers
Department of Biology, McGill University

Assess your prior knowledge...

1) What is a random number?

2) What is (numerical) simulation, and why do we use it?

3) Name some common probability distributions, and in what type(s) of data or processes might you find them.

4) What two *parameters* describe the normal distribution?

Learning Objectives

- *The participant will:*

1) Describe why we use **simulation**
2) Draw **random samples** from a set
3) Draw random samples from a **probability distribution**
4) Describe a model in terms of its **deterministic** and **stochastic** parts
5) **Simulate data** from a **model**

Script Format

https://gist.github.com/cjbayesian/5220711

- The script is divided into sections:

```
##@  x.x  @##
…
...some commands...
…
###  --  ###
```

##@ 0.1 @##

- Keep your house in order:

```
##@  0.1  @##

rm(list=ls())                          # Housekeeping
install.packages("RCurl")
library(RCurl)

###  --  ###
```

Challenges

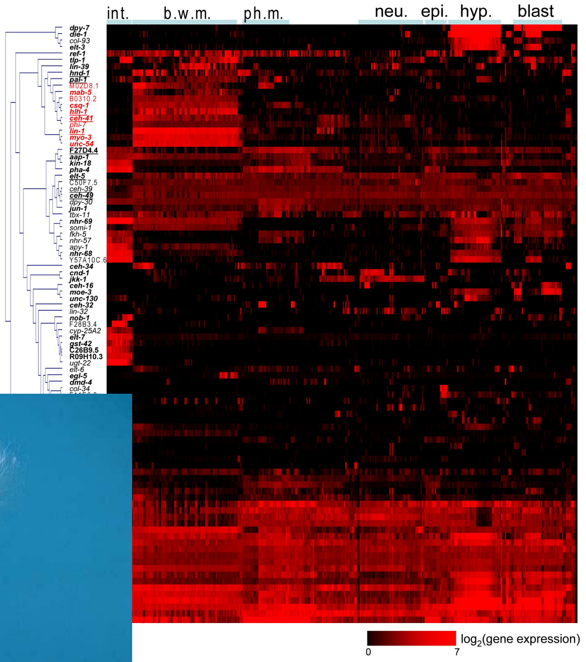
- Like before, these will be items for you to work on yourself and with your neighbour.

Flip a coin ten times, and record the number of heads.

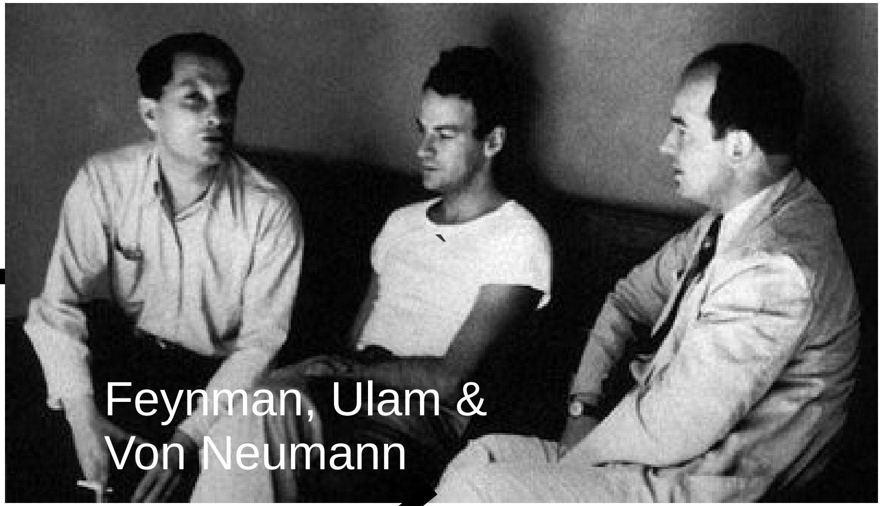
Save this number for later.

That's like, so *random!*

- Random events have outcomes which are not known with <u>certainty</u>.
  - Coin tosses
  - Dice rolls
  - Seed location
  - Number of offspring
  - Gene expression level
  - Oncogenesis



Computerized Randomization



What is (numerical) Simulation?

- Using random numbers to generate data with known characteristics and which follow hypothesized processes.
- We can collect vast amounts of virtual data to test out hypotheses before we collect any 'wet' data.
- Computers (and R) make this easy!

What do we do simulations for?

- Figuring out what to expect
  - Proposals/Grant applications. Conducting a test on 'dummy' data.
- Testing hypotheses about detectability
  - If I measure X and the effect is D, will I be able to detect it?
- Experimenting with model structure
  - In simulation we **know** the processes and parameters
- Analyzing complex systems
  - We can manipulate complex systems in ways which may not be possible in the real world

Drawing random samples

- Recall that R has a built in variable called letters
- We can ask R to give us a random* letter from the alphabet using:

```
> sample(letters,1)
```

*Note that by *random*, we mean that each letter has the same probability. The outcome is not known, but the probability is.

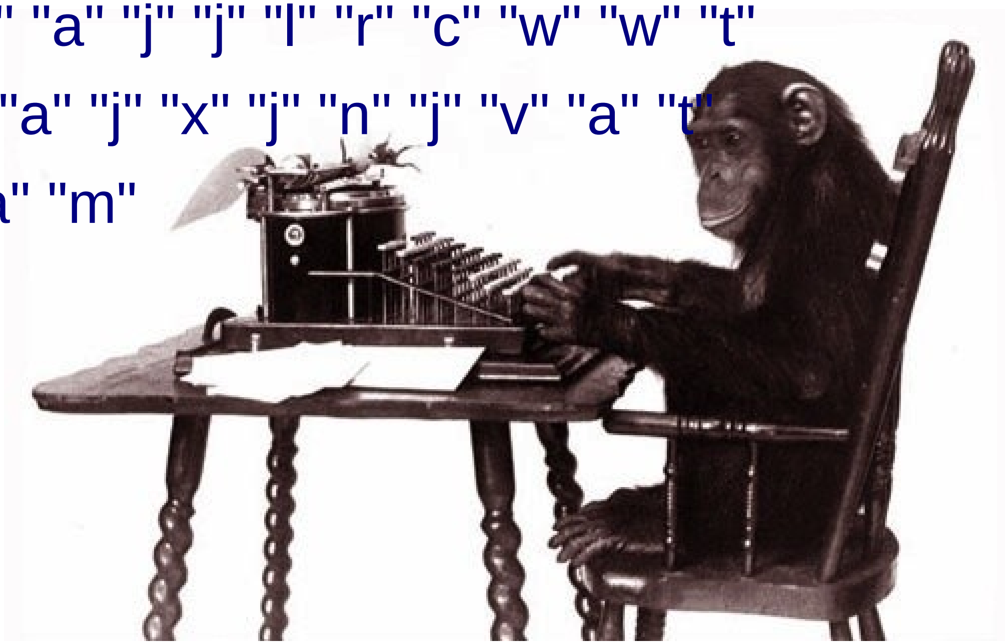
Challenge

- Use ?sample for help with this challenge.

Generate a vector of 100 random letters

Drawing random samples

```
> sample(letters,100,replace=TRUE)

  [1] "a" "o" "w" "p" "u" "r" "c" "e" "c" "h" "r" "i" "k" "m" "i" "e" "i" "z"
 [19] "q" "n" "q" "r" "g" "c" "m" "l" "y" "t" "c" "o" "t" "y" "v" "h" "z" "k"
 [37] "p" "w" "y" "v" "u" "m" "p" "m" "p" "d" "k" "z" "c" "w" "j" "r" "q" "o"
 [55] "e" "b" "m" "h" "n" "l" "z" "y" "d" "a" "j" "j" "l" "r" "c" "w" "w" "t"
 [73] "v" "f" "a" "n" "i" "m" "j" "g" "o" "a" "j" "x" "j" "n" "j" "v" "a" "t"
 [91] "x" "z" "u" "m" "d" "r" "w" "o" "a" "m"
```



The Ecologist's Quarter

- Lands tails (caribou up) 60% of the time

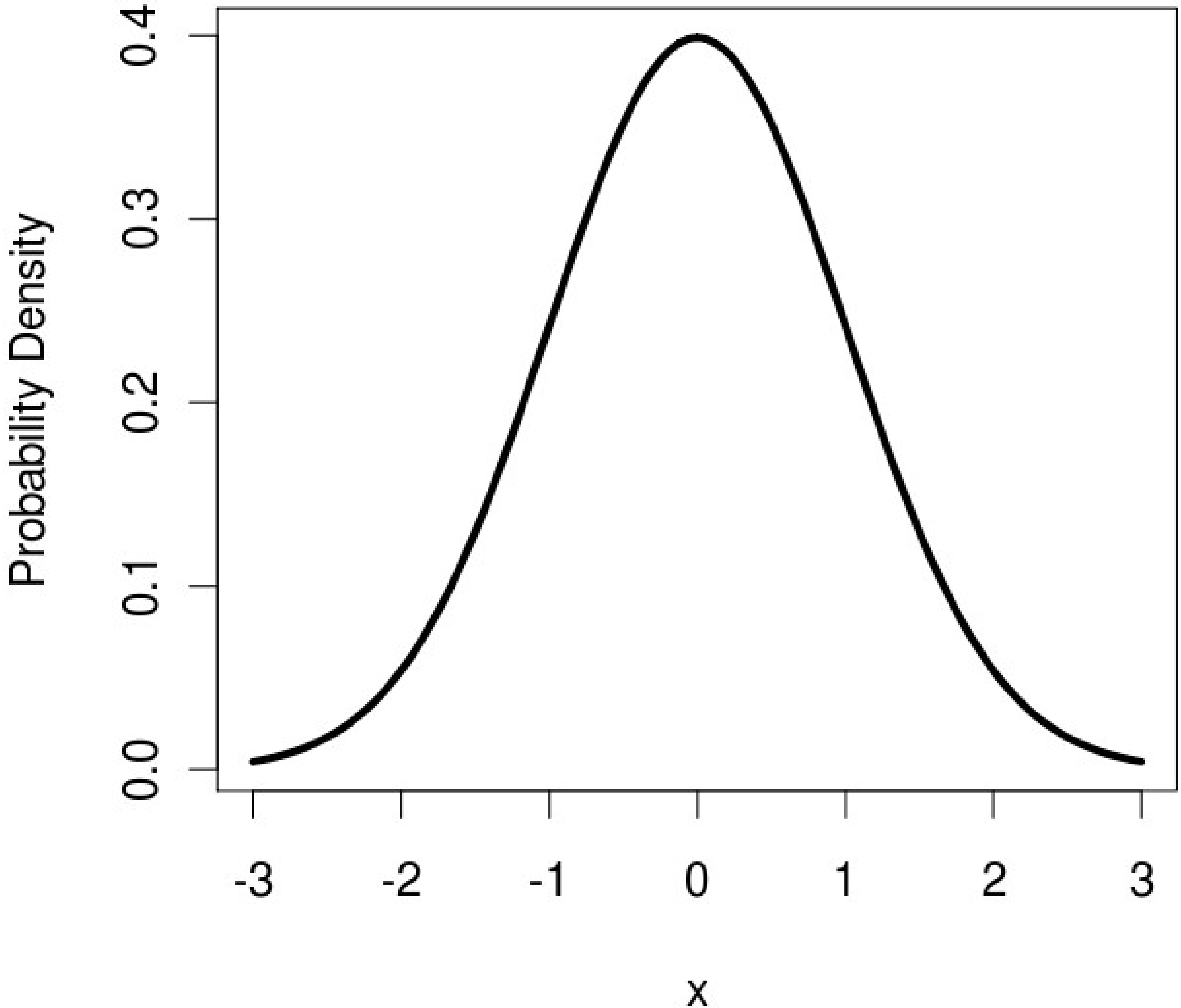
```
> EQ<-c('heads','tails')
> sample(EQ,20,replace=TRUE,p=c(0.4,0.6))

 [1] "heads" "tails" "heads" "tails" "tails" "tails" "heads" "heads" "heads"
[10] "heads" "tails" "heads" "tails" "tails" "heads" "heads" "heads" "tails"
[19] "tails" "tails"
```

Challenge

Repeat the physical coin flipping experiment you did before, this time <u>in silico.</u>

Discrete vs Continuous

- We can use sample() to draw items from a discrete (countable) set.
- What about Continuous values?



Probability Density
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
x

##@ 1.3 @##

- Simulate a group of 30 participants using rnorm()

rnorm()

- Stands for random variate from the normal distribution.

- Usage:

rnorm(n=100,mean=0,sd=1)

Number of points

Parameters

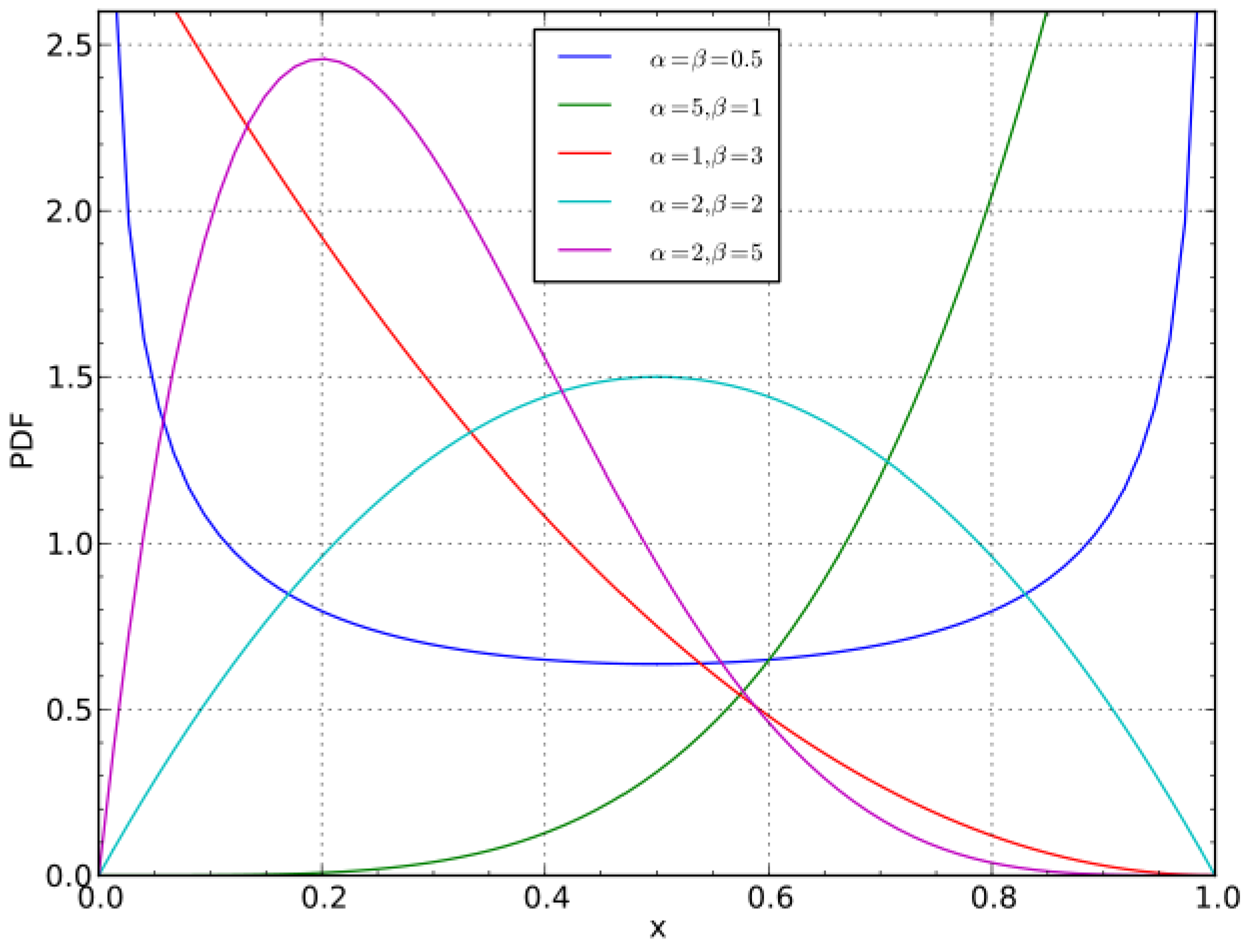
Other continuous distributions

- While the normal distribution is the most common, there are many other continuous distributions.
- R can simulate from these distributions using `r<dist>()`.
- Section ##@ **1.4** @## has commands to simulate from and plot several of them.

Beta

- Parameters:
  α,β

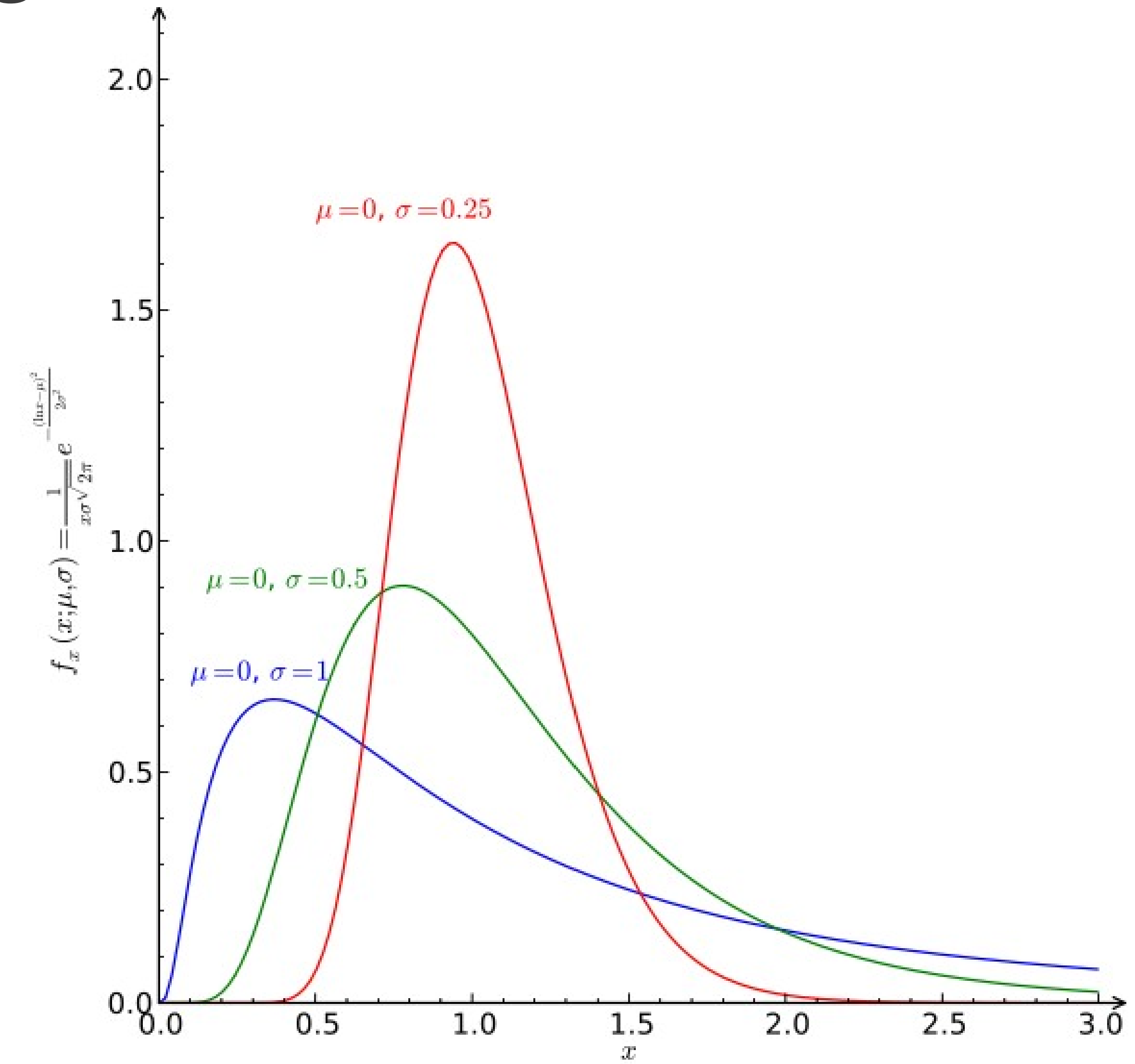
- *Uncertain proportions (what <u>kind</u> of coin is it?)*



Log-Normal

- Parameters:
  μ, σ^2

- *Multiplicative errors*

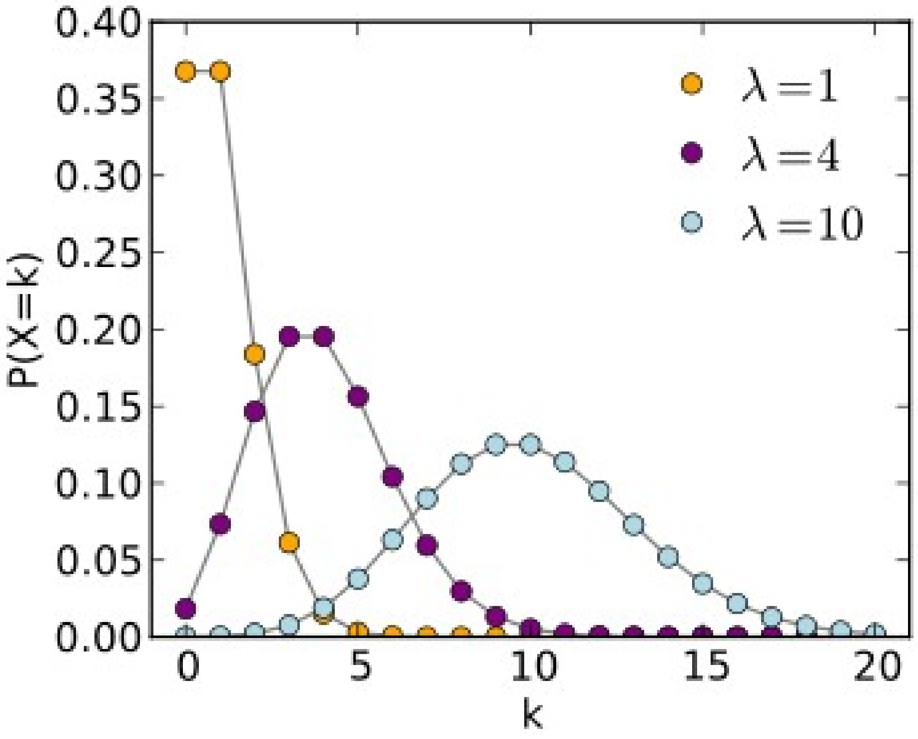


Poisson

- Parameters
  λ

- *Number of events in a fixed amount of time.*

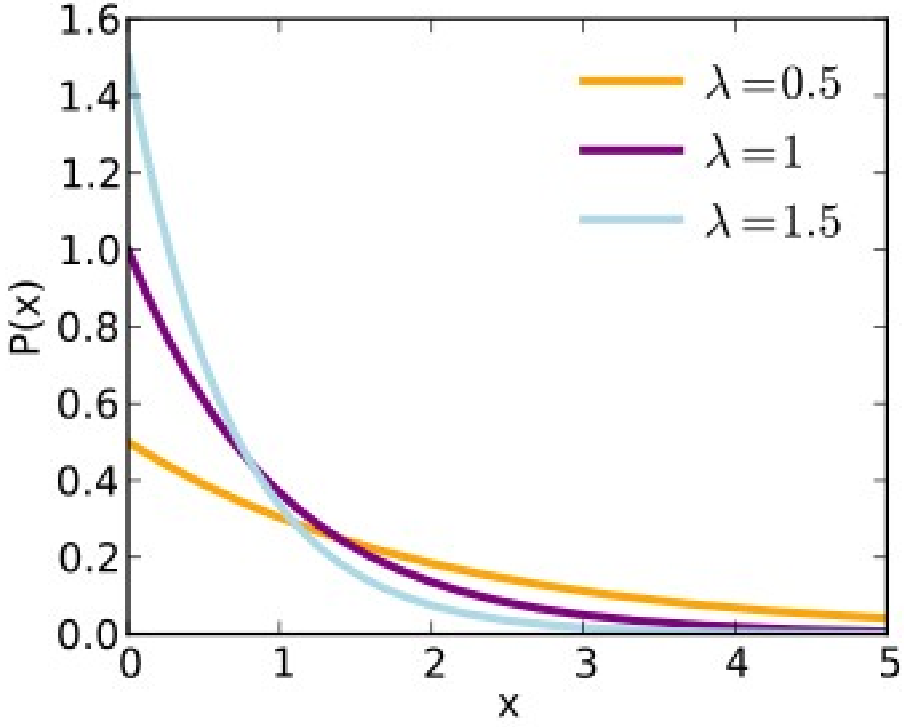
- *Discrete!*



Exponential

- Parameters:
  λ

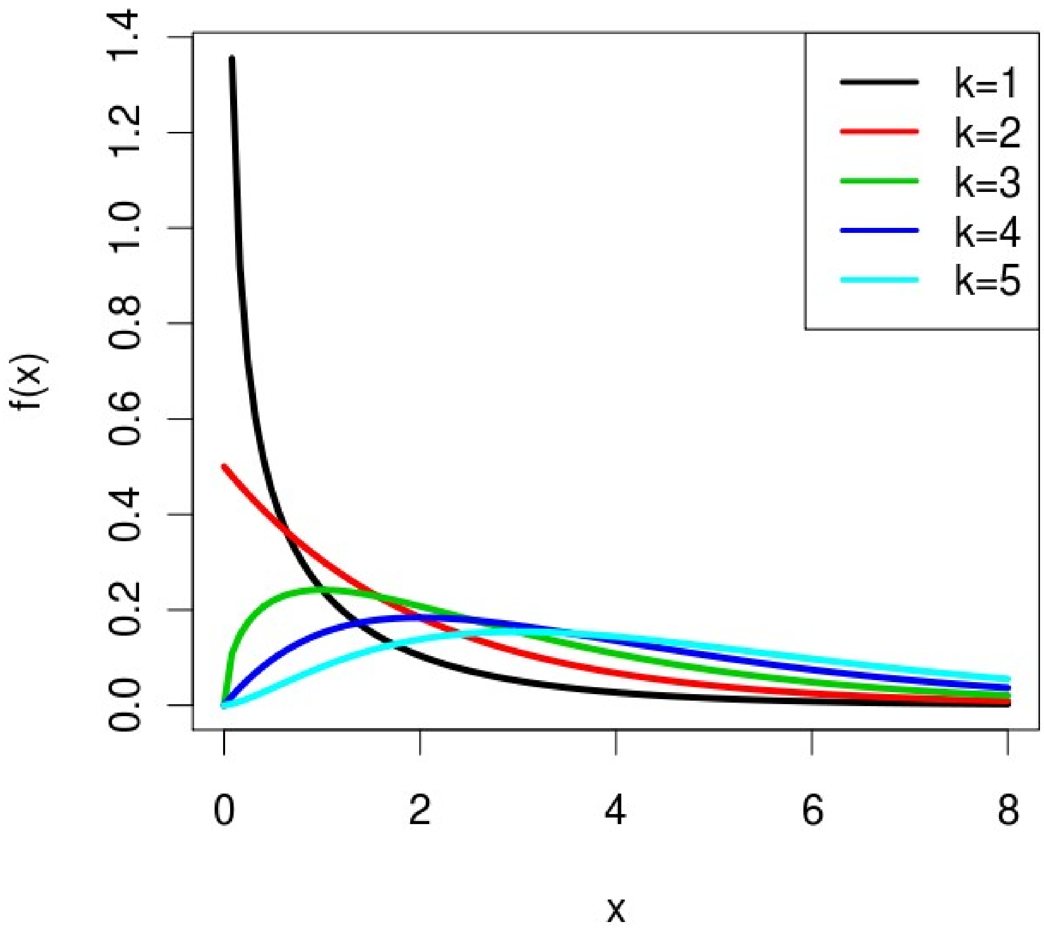
- Time between Poisson events.



Chi-Squared

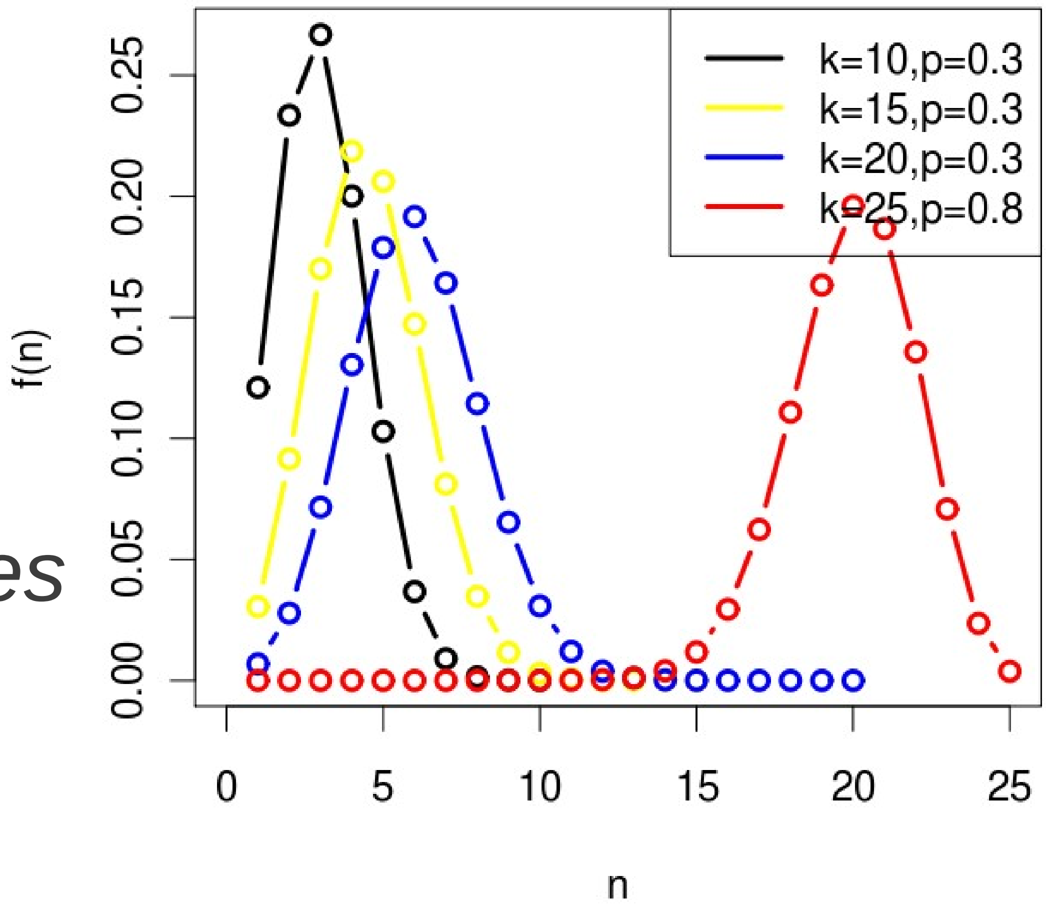
- Parameters:
  K (df)

- *Distribution of sums of squares of a normal random variate*



Binomial

- Parameters:
  k (trials)
  p (probability)



- *Number of successes in a series of binary trials.*

- *Discrete!*

Distribution	R name	additional arguments
beta	beta	shape1, shape2, ncp
binomial	binom	size, prob
Cauchy	cauchy	location, scale
chi-squared	chisq	df, ncp
exponential	exp	rate
F	f	df1, df2, ncp
gamma	gamma	shape, scale
geometric	geom	prob
hypergeometric	hyper	m, n, k
log-normal	lnorm	meanlog, sdlog
logistic	logis	location, scale
negative binomial	nbinom	size, prob
normal	norm	mean, sd
Poisson	pois	lambda
signed rank	signrank	n
Student's t	t	df, ncp
uniform	unif	min, max
Weibull	weibull	shape, scale
Wilcoxon	wilcox	m, n

Simulating from models

- A *Model* is just a mathematical representation of a process, often including two components:

  1) Deterministic
     - The structural part
  2) Stochastic
     - The unexplained variation, error, uncertainty

Simulating from models

- 1) Formulate the model
- 2) Simulate the independent variable(s)
  - In the range which you expect to observe
  - runif() is handy for this step
- 3) Simulate the dependent variables by feeding the independent variables through the deterministic and stochastic parts of the model

Simulating from models I: Linear models

- Large fish swim faster than small fish:

$$Y = \beta_0 + \beta_1 X$$

Deterministic

- There are additional, unknown factors which determine how fast fish swim:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

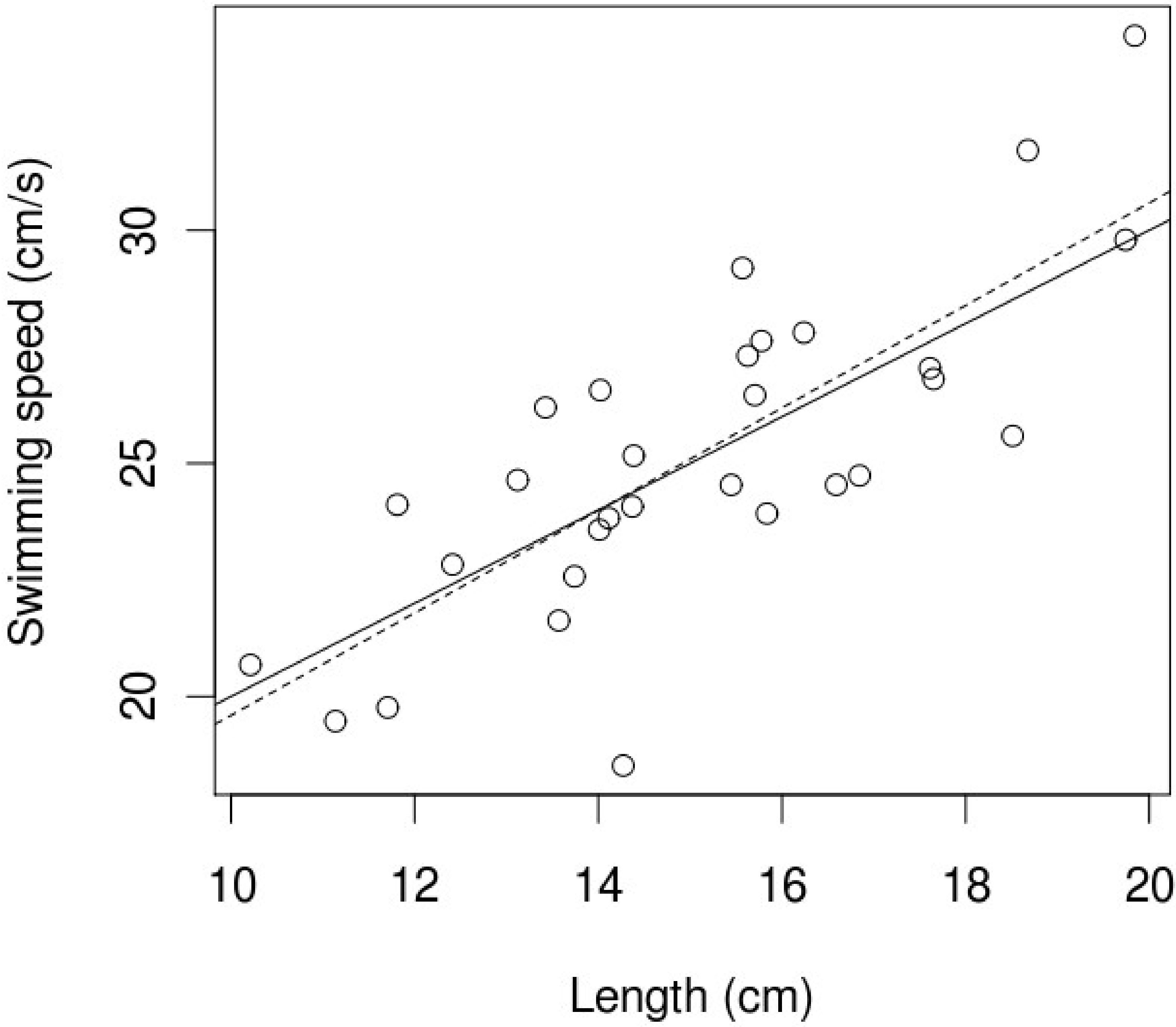
Stochastic

##@ 2.1.1 @##

```
#Model parameters
intercept<-10     #B_0
slope<-1          #B_1
error_sd<-2       #sigma

n<-30                   #number of data points
x<-runif(n,min=10,max=20)    #Simulate x values
```

##@ 2.1.2 @##

```
#Simulate from the model
y<- intercept + slope*x          #Deterministic
y<-y + rnorm(n,0,error_sd)   #Stochastic


plot(x,y,
    xlab='Length (cm)',
    ylab='Swimming speed (cm/s)')
```



Swimming speed (cm/s)
20
25
30
10
12
14
16
18
20
Length (cm)

Challenge

What might the data look like if you collected only 10 individuals, from a population where the true mean swimming speed was 20cm/s and there was no real relationship between length and swimming speed?

Simulating from models II: Tadpole Predation

- Suppose tadpole predators have a Holling type-II functional response:

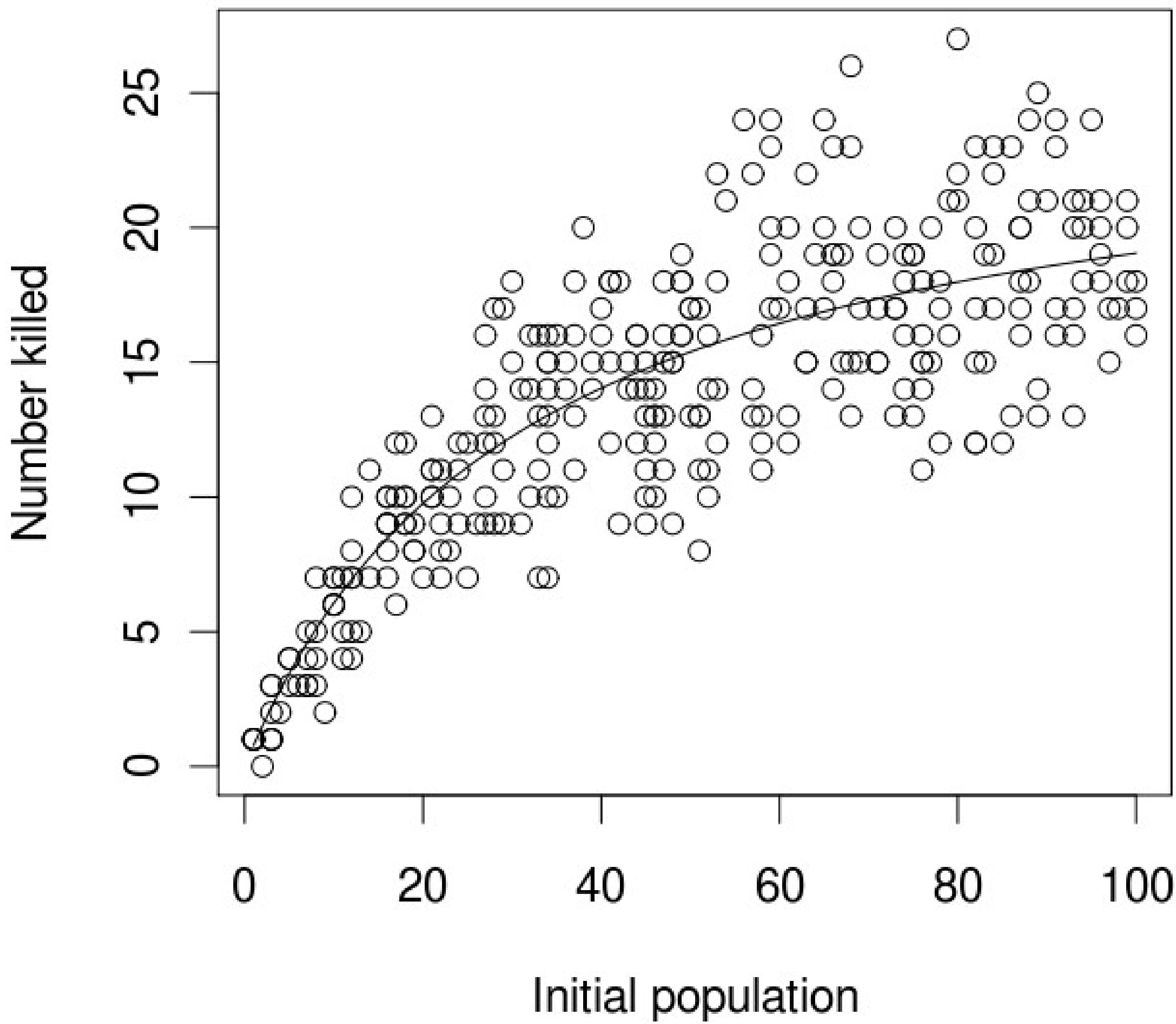
$$p = \frac{a}{1+ahN}$$ ← ***Deterministic***

- The realized number eaten is binomial with probability p:

$$k \sim \mathrm{Binom}(p, N)$$ ← ***Stochastic***

##@ 2.2 @##

```
#Model Parameters
a<-0.5
h<-0.012
n<-300
N<-sample(1:100,n,replace=TRUE)
#Simulate from the model
predprob<- a/(1+a*h*N) #Deterministic part
killed<- rbinom(n,prob=predprob,size=N) #Stochastic part
plot(N,killed,
     xlab='Initial population',
     ylab='Number killed')
```



Number killed
0
5
10
15
20
25
0
20
40
60
80
100
Initial population

Challenge

Simulate data from your research system.

- ***How would you start?***
- ***What are the hypothesized processes(ie model)?***
- ***Are there parameter estimates in the literature?***

Summary

- Simulation is the process of using computer generated random numbers to create data.
- We can simulate data from discrete and continuous probability distributions.
- We can combine random processes with deterministic ones to simulate data from models.