

Simulations: Unit 2's Kickoff

Grinnell College

October 2025



Motivation






It's 3:30am on a Sunday in a seedy, smoke-filled, underground bar in Port Arthur TX

You've drank a bottle and some change of rye since the night started

You are playing a texas hold 'em against some biker who looks like he did time and you got a lot of money riding on this hand

Cards dealt are:

▶ You:  

▶ Community:     

▶ Biker: ? ?

What is the probability you have gas money back to Iowa?

Motivation

My old DnD character could get a “critical hit” if I rolled a 19 or 20 on a 20 sided dice roll.

What is the probability I would 'crit when I rolled with advantage (roll twice and take the higher number)?

What is the probability I would 'crit when I rolled with disadvantage (roll twice and take the lower number)?

Analytical Solutions

We can solve these via math

- ▶ Could solve this using their distributional forms
 - ▶ eg advantage dice rolls would be the maximum of two independent, discrete uniform variables with values $1, 2, \dots, 20$
- ▶ Could write out all possible combinations of cards dealt or dice rolls (counting arguments)
 - ▶ Count the proportion where (EVENT) happens
 - ▶ eg number of times we win our hand at poker
- ▶ Kind of annoying to do in larger, more complex situations

So instead of annoying math let's....SIMULATE!!

The general strategy is as follows:

1. We have some question we are interested in
 - ▶ How often do I crit with disadvantage?
2. We use a computer to simulate that process
 - ▶ We roll some dice
3. We then use those simulations to state something of interest
 - ▶ Find the proportion of simulated rolls that were crits
4. Profit

Starting Point

We start with the belief that we can approximate the scientific mechanism being studied

- ▶ Simulation is only as valid as it's assumptions:
 - ▶ Simulated model reflects the data generating process
 - ▶ Simplifications do not unreasonably change the model
 - ★ eg Using Newtonian physics to estimate where a trebuchet will toss a rock is valid enough, we can ignore(?) general relativity
- ▶ Creates a simulated distribution about how the statistic of interest behaves

Safety Check

If we randomly grab a dot in the unit square, what will be the probability it lies within a circle with radius 1?

Safety Check

If we randomly grab a dot in the unit square, what will be the probability it lies within a circle with radius 1?

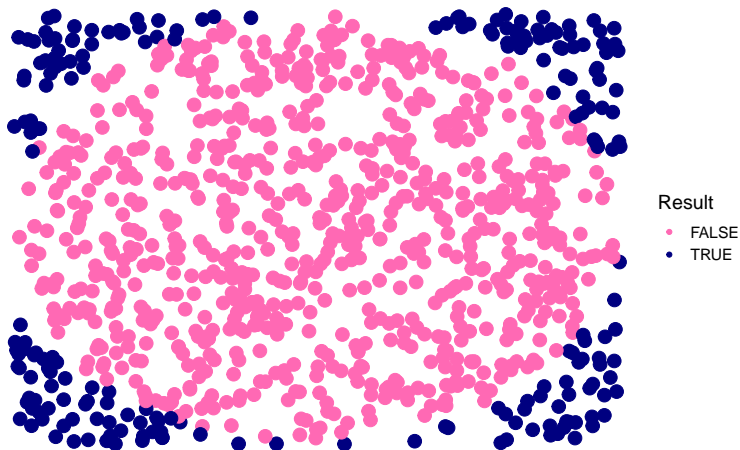
Geometry

- ▶ Area of a circle is $\pi * r^2$
- ▶ Area of a unit square is 1×1
- ▶ Probability is the ratio so $\pi * .25/1 = .7853$

Or we can simulate it...

Safety Check

.7860 via simulation with 100 random draws



Why are we off?

Why are we off?

Differing names but the idea is **simulation error**, that is our simulation uses randomness to generate samples such that we won't be exactly right for non-infinite sample sizes

Generally, as the simulation count increases we get closer

Simulation Size	Result
10	.8
50	.74
100	.72
250	.74
1,000	.788
10,000	.7813
100,000	.7826
1,000,000	.7859
Theoretical	.7853

A General Example

For a circle with a radius of 1, what is the average distance between two points?

A General Example

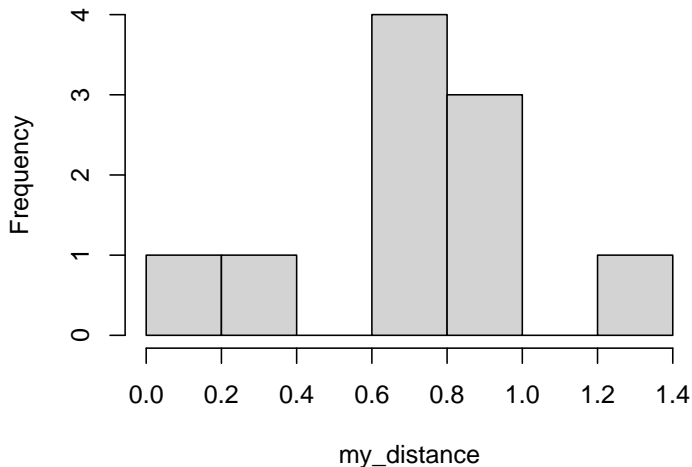
For a circle with a radius of 1, what is the average distance between two points?

Strategy:

1. We simulate two points uniformly from the unit square
 - ▶ Kick out the pair if either are outside our circle
 - ▶ This is a simplification for coding
 - ▶ “uniformly sampled” means no area of the circle is sampled more heavily than another (outside of random chance)
2. We calculate the two point's distance from each other
 - ▶ I.e. we calculate a *statistic* using those sampled points
3. Repeat steps 1 and 2 a lot...like a lot a lot
4. Profit

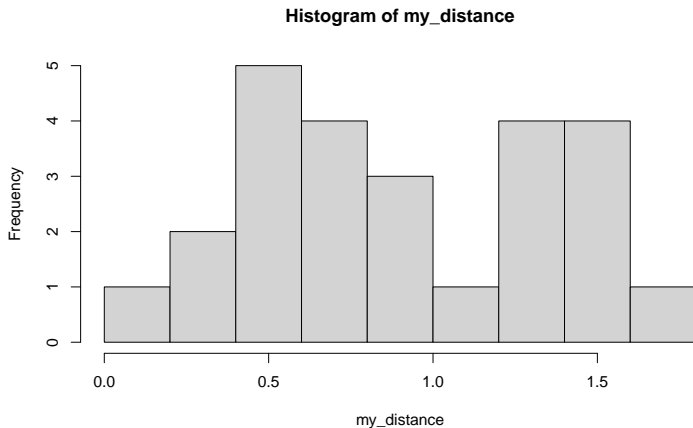
Simulated Results, Number of Sim.s = 10

Histogram of my_distance

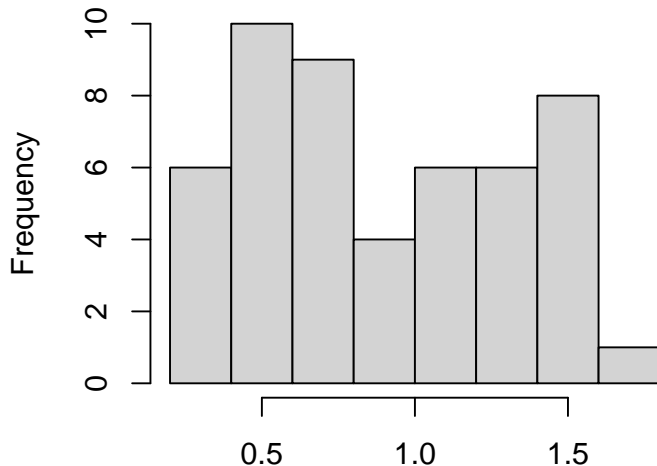


Simulated Results, Number of Sim.s = 25

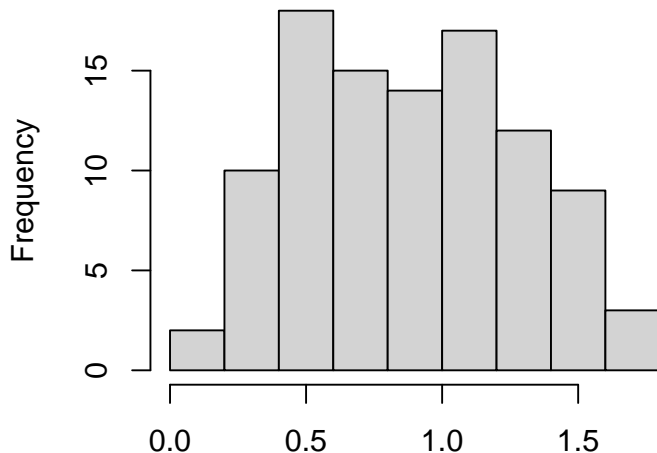
mean = .9040



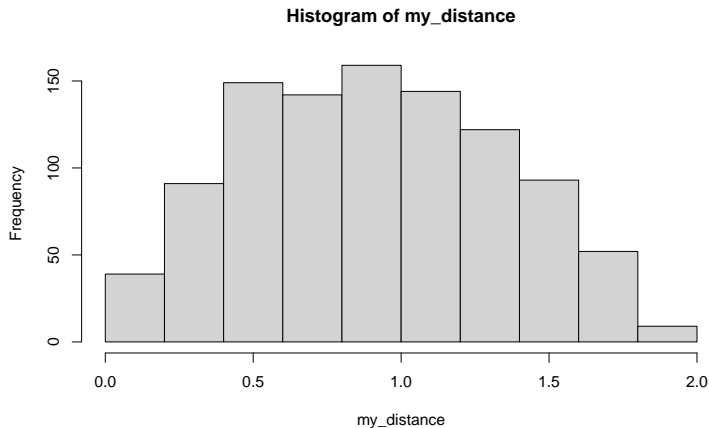
Histogram of my_distance



Histogram of my_distance

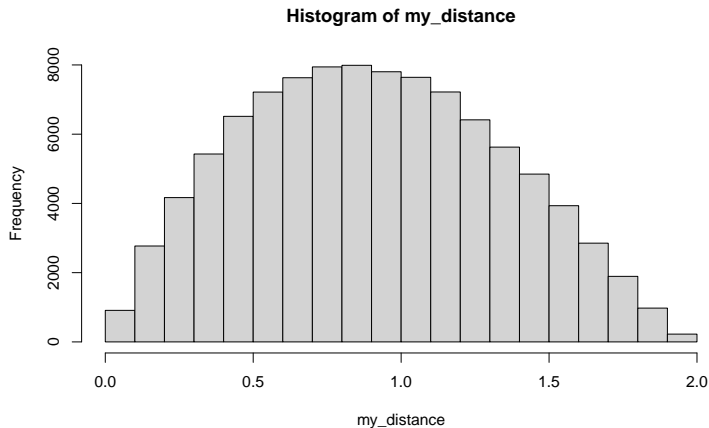


Simulated Results, Number of Sim.s = 1,000



.9112

Simulated Results, Number of Sim.s = 100,000



.9076

What are we noticing?

- ▶ A pattern/distribution seems to be forming
- ▶ More simulations seems to be make our estimate “stable”
 - ▶ Simulations take a lot of simulations to become believable
- ▶ Slowly our simulated distribution is starting to match the true distribution of distances between points
 - ▶ The real distribution is difficult to calculate, even if an analytical solution exists

Leveraging these, we can use the simulated distribution to make a statement about the actual distribution

Simulated Results, Uses

Questions that these simulated distributions can help us answer....

- ▶ What is the center (mean) of the population?
 - ▶ $\text{mean} = .9076$
- ▶ What is the spread of the data?
 - ▶ $\text{st. dev.} = .4247$
- ▶ Where is the middle 50% of the data located?
 - ▶ $Q_1 = .5765$ to $Q_3 = 1.225$
- ▶ Where is the middle 90% of the data located?

Simulated Results, Uses

Questions that these simulated distributions can help us answer....

- ▶ What is the center (mean, median) of the population?
 - ▶ mean = .9076
 - ▶ median = .8934
- ▶ What is the spread of the data?
 - ▶ st. dev. = .4247
- ▶ Where is the middle 50% of the data located?
 - ▶ $Q_1 = .5765$ to $Q_3 = 1.225$
- ▶ Where is the middle 90% of the data located?
 - ▶ Need the 5 and 95 percentiles
 - ▶ 5% = .2340 and 95% = 1.621

Foreshadowing confidence intervals.....

Drawbacks

Realistically analytical solutions are better when possible (no error)

Generally we don't know the data generating process or the population distribution's parameters

- ▶ eg the above examples were mathematically defined
- ▶ Easy to simulate, say, dice rolls
 - ▶ Not even! Most dice are imperfectly weighted
- ▶ How would I begin simulate students' weight, height, major, and gpa?
 - ▶ I don't know any of the distributions those are coming from

We want to build a simulation that let's us approximate a long-term behavior, which we don't know.....right?

Looking Back

So far....

- ▶ A sample hopefully looks like the population
 - ▶ ie our data was produced using the real data generating process
- ▶ We calculate a statistics from our sample
- ▶ which hopefully is close to the population's parameter.
- ▶ What can we say about the behavior of that statistic?
 - ▶ eg if I grabbed a new sample where would I expect my mean to be?
 - ▶ Grabbing new samples is expensive....

Pull yourself up by your own bootstraps

The **bootstrap method** was developed by Bradley Efron tries to do this

1. We collect real data (with sample size Q)
2. We estimate a statistic
3. We then randomly sample from our existing sample
 - ▶ We allow ourselves to draw an observation twice or more
4. Calculate a new statistic
5. Do steps 3 and 4 a lot
6. We end up with a distribution where questions from earlier can be discussed