Section 1 : Introduction to the Potential Outcomes Framework

Andrew Bertoli

4 September 2013

Roadmap

1. Preview
2. Helpful Tips
3. Potential Outcomes Framework
4. Experiments vs. Observational Studies
5. Monte Carlo Simulation

Preview

1. Causality and Experiments
2. Regression
3. Matching
4. Natural Experiments
5. Regression Discontinuity
6. Instrumental Variables

Preview

Regression Example

Question: Do poor houses reduce poverty rates?

Design: Explain poverty as well as you can using age and population (with a linear model), and then see how much of the residual is explained by poor houses.

Preview

Matching Example

Question: Does coffee have a positive effect on academic performance?

Design: Compare a group of students who drink coffee to a very similar group of students who do not drink coffee.

Preview

Natural Experiment Example

Question: How did going to Vietnam affect future earnings for individual soldiers?

Design: Compare the wealth of men who were and were not drafted to go to the war.

Preview

Regression Discontinuity

Question: High school students took a test and everyone who scored above 50% got a college scholarship. Did the scholarship increase the likelihood that students would go to college.

Design: Compare students who scored 50% to students that got 49%.

Preview

Instrumental Variables Example

Question: To what extent do smart phones cause car accidents?

Design: Take advantage of the fact that many people bought smart phones when Verizon was offered for the iPhone. The sudden increase in smart phones may correspond to an increase in traffic accidents.

Helpful Tips

Advice for R

1. If you are new to R, do some online tutorials

http://tryr.codeschool.com/levels/1/challenges/1

http://www.cyclismo.org/tutorial/R/

2. Look things up online

3. E-mail me any questions

Helpful Tips

Advice for the Homework

1. Work in groups (3-4 people)
2. Use the readings and the Internet
3. Use R for problems where it's not required
4. Come to section with questions

Potential Outcomes Framework

Basic Definition of Causality:

X is a cause of Y if Y would not have occurred without X.

There are other versions of causality, but this is the one we will use.

Two Reasons

1. Significant scientific progress has been made using this understanding of causality

2. It is consistent with what we often mean when we speak in causal terms

Potential Outcomes Framework

Potential Outcomes Framework

Potential Outcomes Framework

Potential Outcomes Framework

Potential Outcomes Framework

Potential Outcomes Framework

Key Points

1. We need a clear sense of the counterfactual world where X is not present.

2. No causation without manipulation (Holland 1986)

3. Our job is to determine what Y would have been in the absence of X, which can be very hard.

Potential Outcomes Framework

Potential Outcomes Framework

Another Meaning of "Causality"

Basic Idea: To explain something is to link it to a larger group

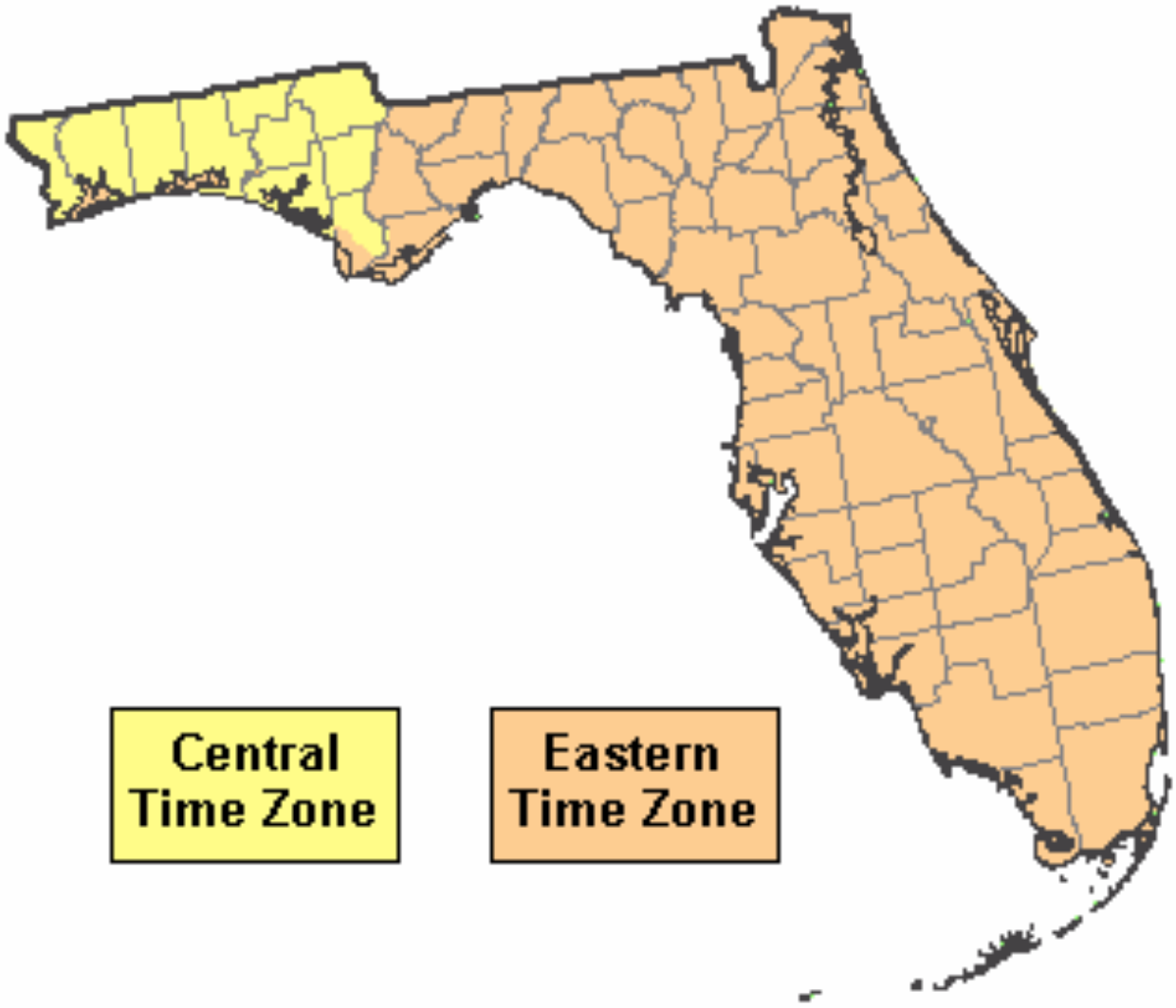
Question: Why is your dog chasing a squirrel?

Response: Because dogs like to chase squirrels.

Very different than trying to estimate the effect of being a dog on the desire to chase squirrels.

Yet some papers try to estimate the effect of being a democracy on GDP growth.

Potential Outcomes Framework



Potential Outcomes Framework

Henry Brady made a compelling case that the finding was wrong.

In this example, looking at the case was helpful.

Because there was a clear counterfactual and straightforward evidence, we could determine the causal effect.

Potential Outcomes Framework

Many studies in social science do not have a clearly defined counterfactual.

One Example: Democratic Peace Theory

Several problems arise when there is not a clear counterfactual

1. No idea what the world look like in the absence of X
2. Difficult to know what to control for
3. All possible control variables can become post-treatment

Potential Outcomes Framework

Example: Democratic Peace Theory

Question 1: Did the shared democracy between Britain and France prevent war in the early 1900's?

Problems

1. What would the world look like if either Britain or France was not a democracy?

2. Should you consider economic interdependence an alternative explanation or part of the mechanism?

Potential Outcomes Framework

Example: Democratic Peace Theory

Question 2: Is there a statistically significant causal relationship between shared democracy and peace?

Problems:

1. Do you control for economic interdependence or participation in international organizations, or are they part of the mechanism?

2. Almost anything you might want to control for is post-treatment for countries that have been democracies for a long time.

Potential Outcomes Framework

Final Remarks

1. To estimate a treatment effect, we need to have some understanding of the counterfactual world where X is not present.

2. When there is no clear counterfactual, the problem of what to control for can leave major debates in the social sciences unresolved.

3. This does not mean that attributes like democracy do not play some role in the causal process. It just means that the independent treatment effect of these factors is not well-defined.

4. In social science, it is often very difficult to determine the effect of some factor in a single case.

Observational Studies vs. Experiments

Set-up

1. Instead of trying to determine the effect of X in a single case, we're going to estimate its average effect across a large number of cases. Assume we have n units in our study.

2. If Unit i receives the treatment, then we observe Y_{it}. If not, we observe Y_{ic}.

3. The causal effect for Unit i is

$$Y_{it} - Y_{ic}$$

which we can never compute directly, but can be estimated.

4. The average treatment effect (ATE) for the sample is

$$\frac{1}{n} \sum_i (Y_{it} - Y_{ic})$$

which also cannot be computed directly, but can be estimated.

Observational Studies vs. Experiments

Two reasons to prefer experiments

1. More realistic assumptions
2. Less opportunities for dishonest research

Observational Studies vs. Experiments

Assumptions for an Experiment

1. $\{Y_{it}, Y_{ic}\} \perp\!\!\!\perp T_i$
2. $\{Y_{it}^I, Y_{ic}^I\} = \{Y_{it}^L, Y_{ic}^L\}$ (SUTVA or non-interference)

These assumptions can be rewritten as

$\{Y_{it}, Y_{ic}\} \perp\!\!\!\perp T$

where T is the vector of treatment assignments.

Observational Studies vs. Experiments

Running an experiment is like randomly sampling from all the Y_{it}'s and Y_{ic}'s.

Treatment
(Y_{it}'s)

Control
(Y_{ic}'s)

Observational Studies vs. Experiments

Assumptions for Regression

1. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$

2. All independent and control variables are fixed (no measurement error)

3. There is no deterministic linear relationship between the X variables (no collinearity)

4. $E[\epsilon_i] = 0$ for all i

5. $\epsilon_i \sim N(0, \sigma^2)$ for all i

Assumptions 1-4 are necessary for $\hat{\beta}$ to be unbiased. Assumption 5 is required for the standard errors, p-values, and confidence intervals to be correct.

Observational Studies vs. Experiments

Assumptions for Matching

1. $\{Y_{it}, Y_{ic}\} \perp\!\!\!\perp T_i | X$
2. $\{Y_{it}^I, Y_{ic}^I\} = \{Y_{it}^L, Y_{ic}^L\}$ (SUTVA or non-interference)

Observational Studies vs. Experiments

Dishonest Research

1. It is often unclear which observable factors you should control for

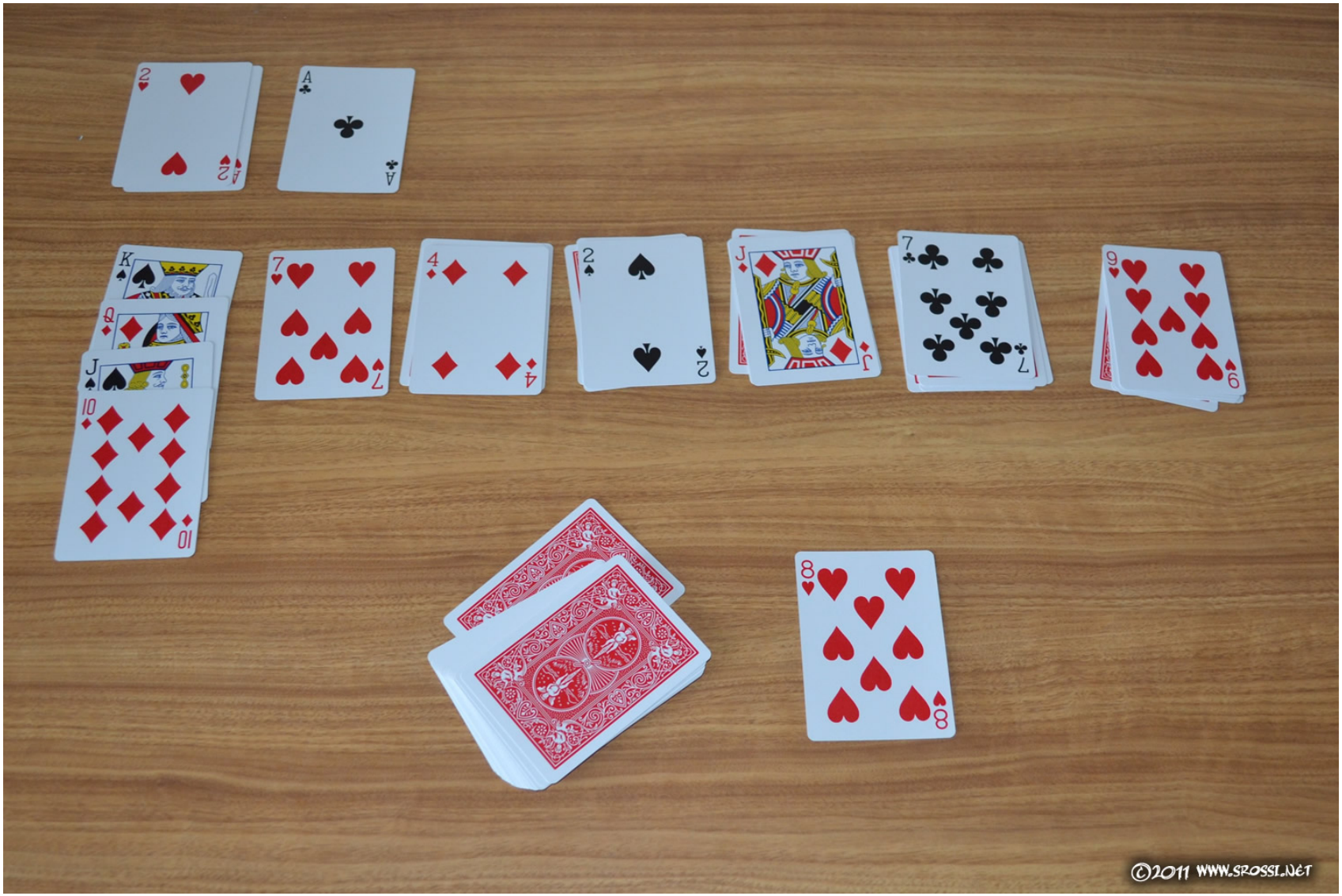
2. It can be tempting to choose the control variables that lead to the lowest p-value

3. In experiments, the main concern is that the researcher looked at many outcomes and only reported the significant ones.

4. In observational studies, researchers can look at many outcomes and many control variables for each outcome, drastically increasing the number of tests they can run.

Monte Carlo Simulation

Monte Carlo Simulation



Monte Carlo Simulation

Problem: There are $8.06581751709 \cdot 10^{67}$ possible hands.

Solution: Deal 1000 hands and count the number that are winnable. The probability of getting a winnable hand can be estimated by

$P(\text{Winnable hand}) \approx \text{Number of winnable hands}/1000$

Monte Carlo Simulation

Example 1: If you go through a deck of cards one at a time, what is the probability that you will see your phone number?

Phone Number: 793-3229

Monte Carlo Simulation

```
Success.Vector=rep(0,100000000)

for(i in 1:100000000){

deck=sample(c(rep(2:14,4), 52, replace=FALSE)

seven=which(deck==7)
five=which(deck[seven+1]==5)
three=which(deck[five+1]==3)
four=which(deck[three+1]==4)
two=which(deck[four+1]==2)
one=which(deck[two+1]==1)
nine=which(deck[one+1]==9)

if(length(nine)==0)Success.Vector[i]=0
if(length(nine)>0)Success.Vector[i]=1
}

sum(Success.Vector)
```

Monte Carlo Simulation

Example 1: If you go through a deck of cards one at a time, what is the probability that you will see your phone number?

Phone Number: 793-3229

Answer: $P(\text{Seeing Phone Number}) \approx 5/100,000,000$

Monte Carlo Simulation

Example 2: The Monte Hall Problem

Monte Carlo Simulation

```
for(i in 1:1000){
Doors=1:3
Prize.Door=sample(Doors,1)
First.Choice=sample(Doors,1)
if(First.Choice != Prize.Door){Open.Door=Doors[-c(Prize.Door,First.Choice)]}
if(First.Choice == Prize.Door){Open.Door=sample(Doors[-First.Choice],1)}
Doors.Remaining=Doors[-Open.Door]
isTRUE(Prize.Door==First.Choice)
isTRUE(Prize.Door!=First.Choice)
if(Prize.Door==First.Choice)Success.For.Switching.Doors[i]=0
if(Prize.Door!=First.Choice)Success.For.Switching.Doors[i]=1
}
sum(Success.For.Switching.Doors)/length(Success.For.Switching.Doors)
```

0.664

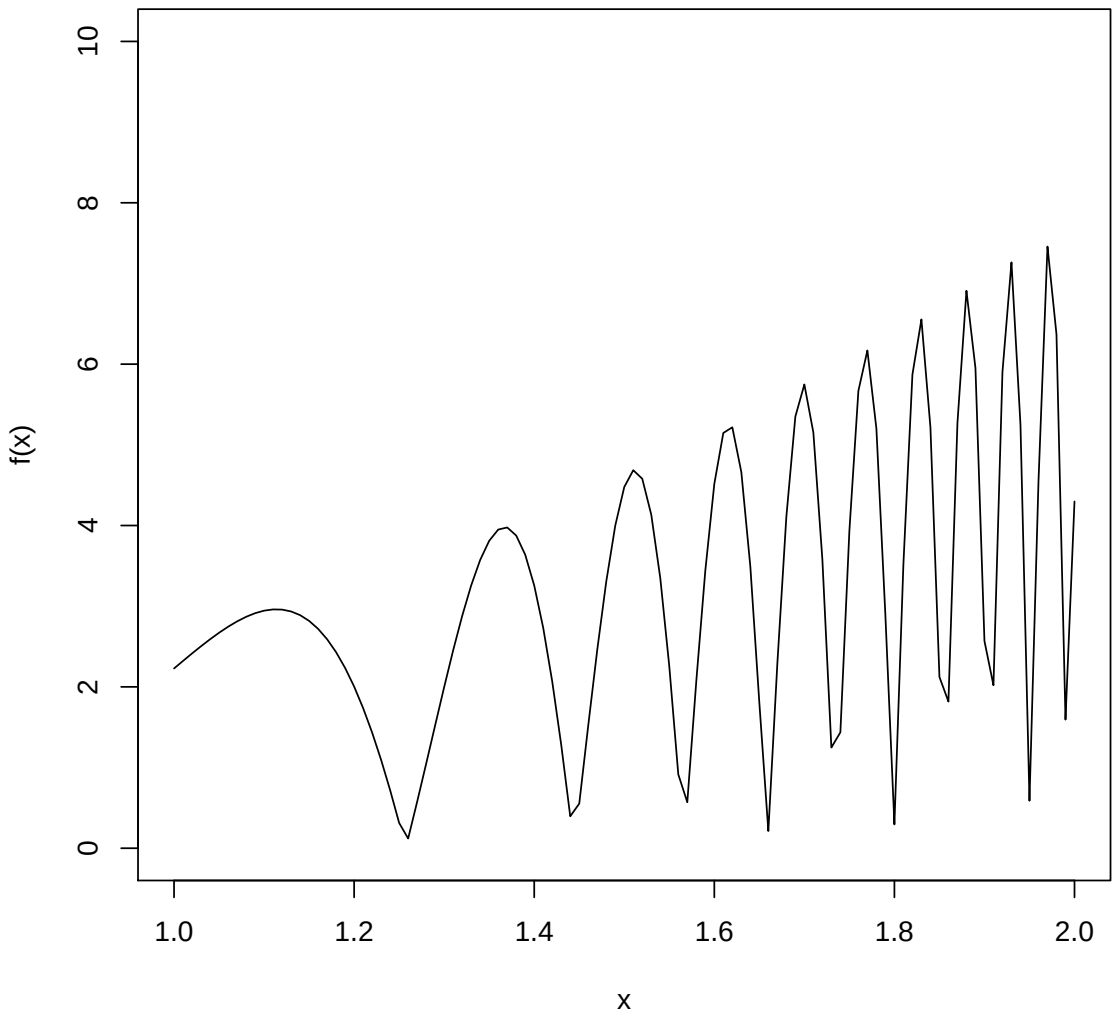
Monte Carlo Simulation

Example 3: Evaluating a Hard Integral

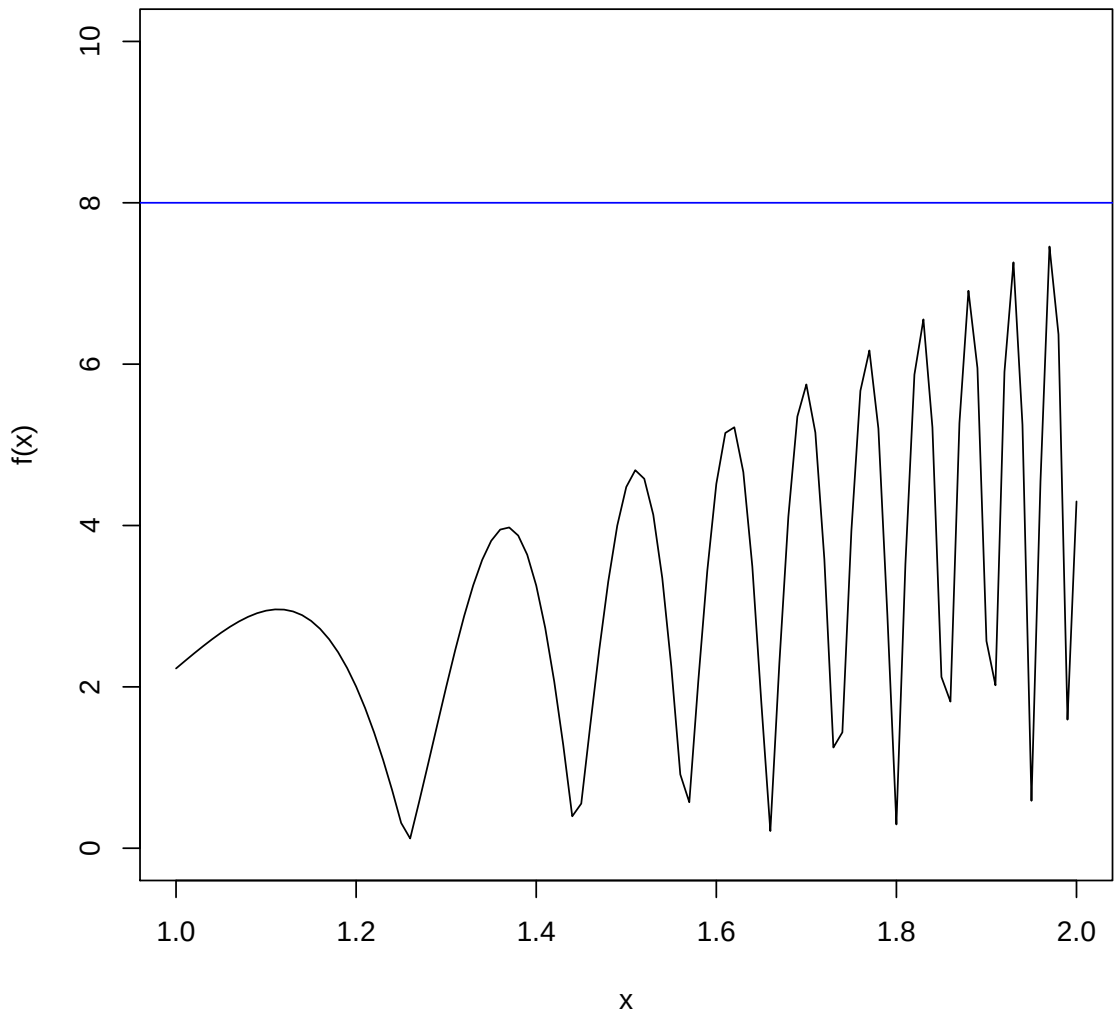
$$\int_1^2 \left| \frac{log(x^2/(x+1)^3) * sin(x^5)}{arctan(1/x^2)} \right| dx$$

Solution: Hit-or-Miss Monte Carlo

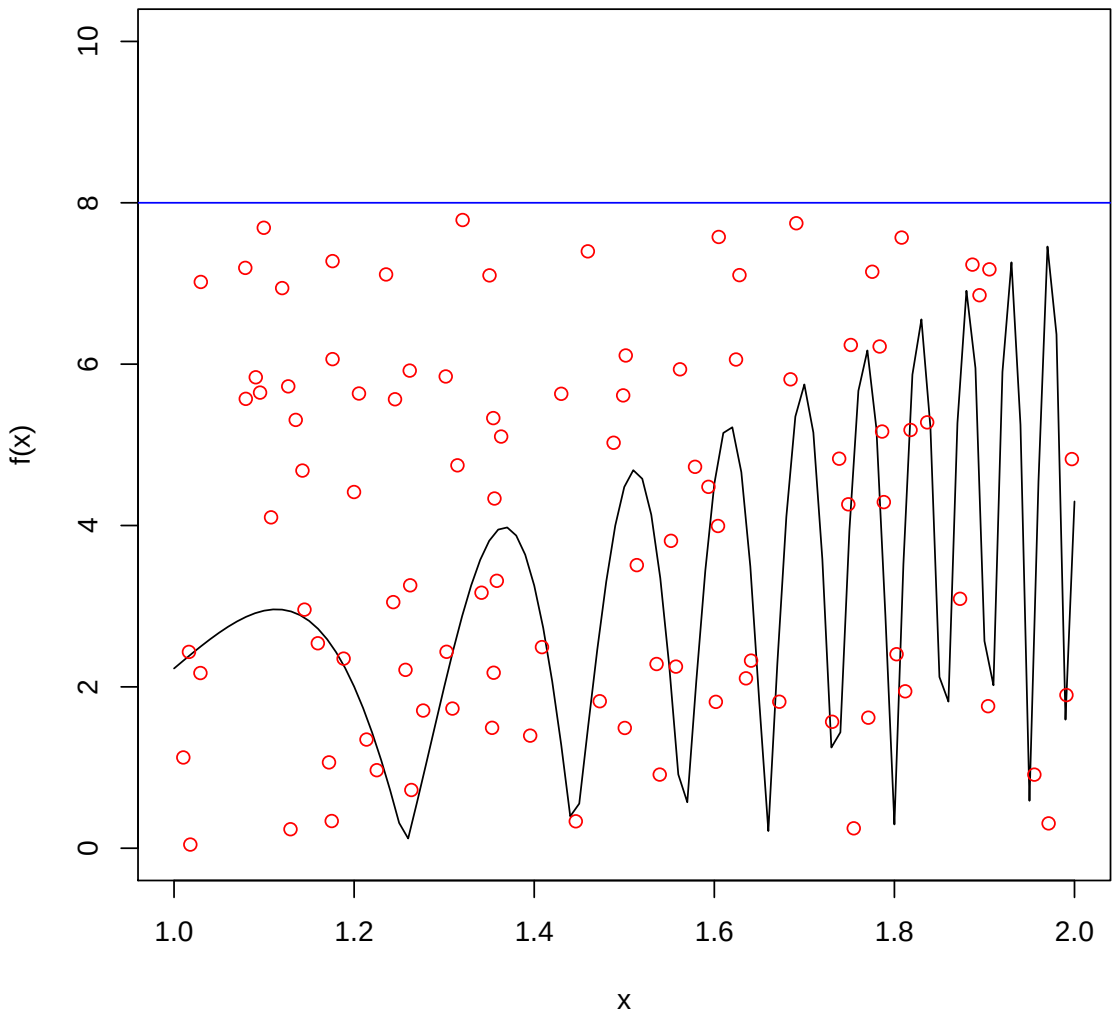
Monte Carlo Simulation



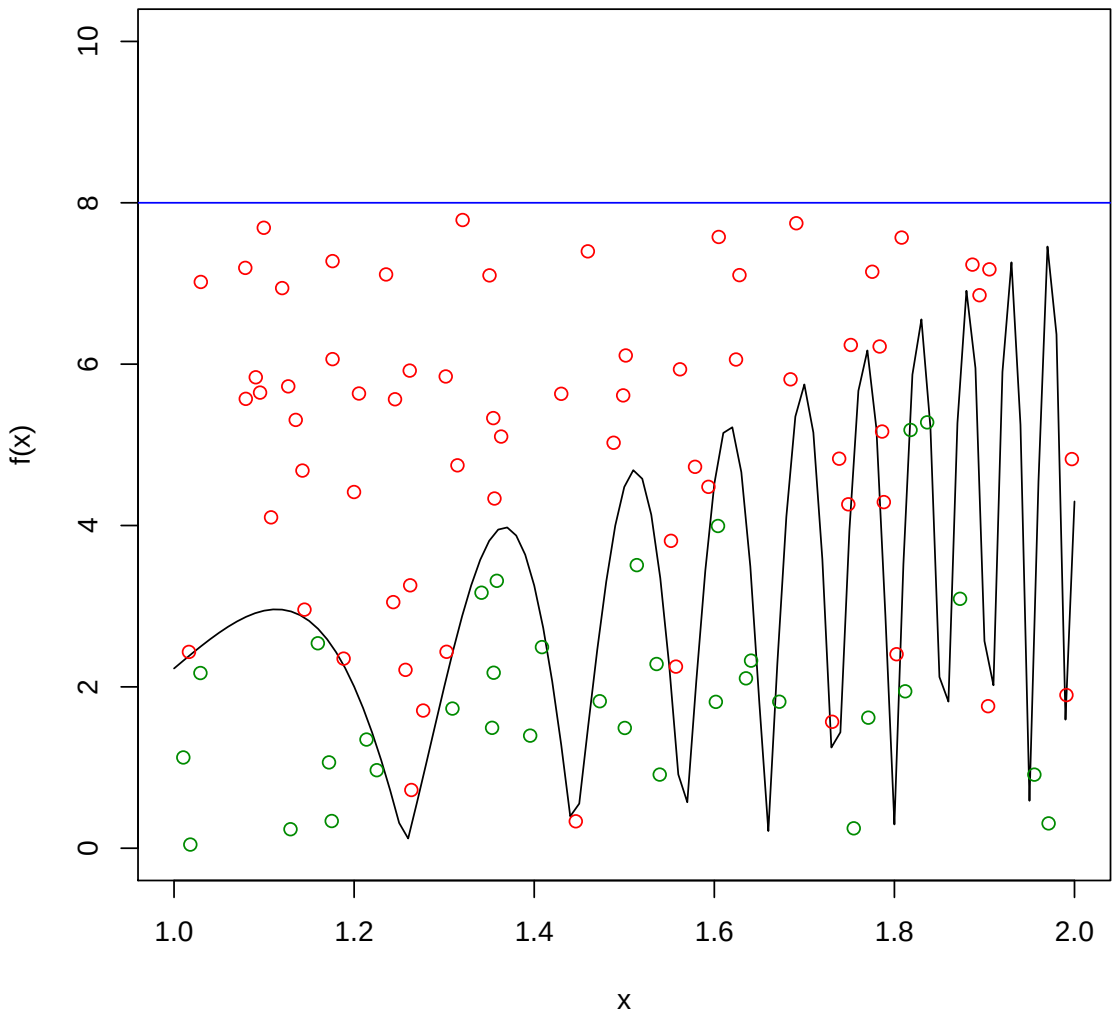
Monte Carlo Simulation



Monte Carlo Simulation



Monte Carlo Simulation



Monte Carlo Simulation

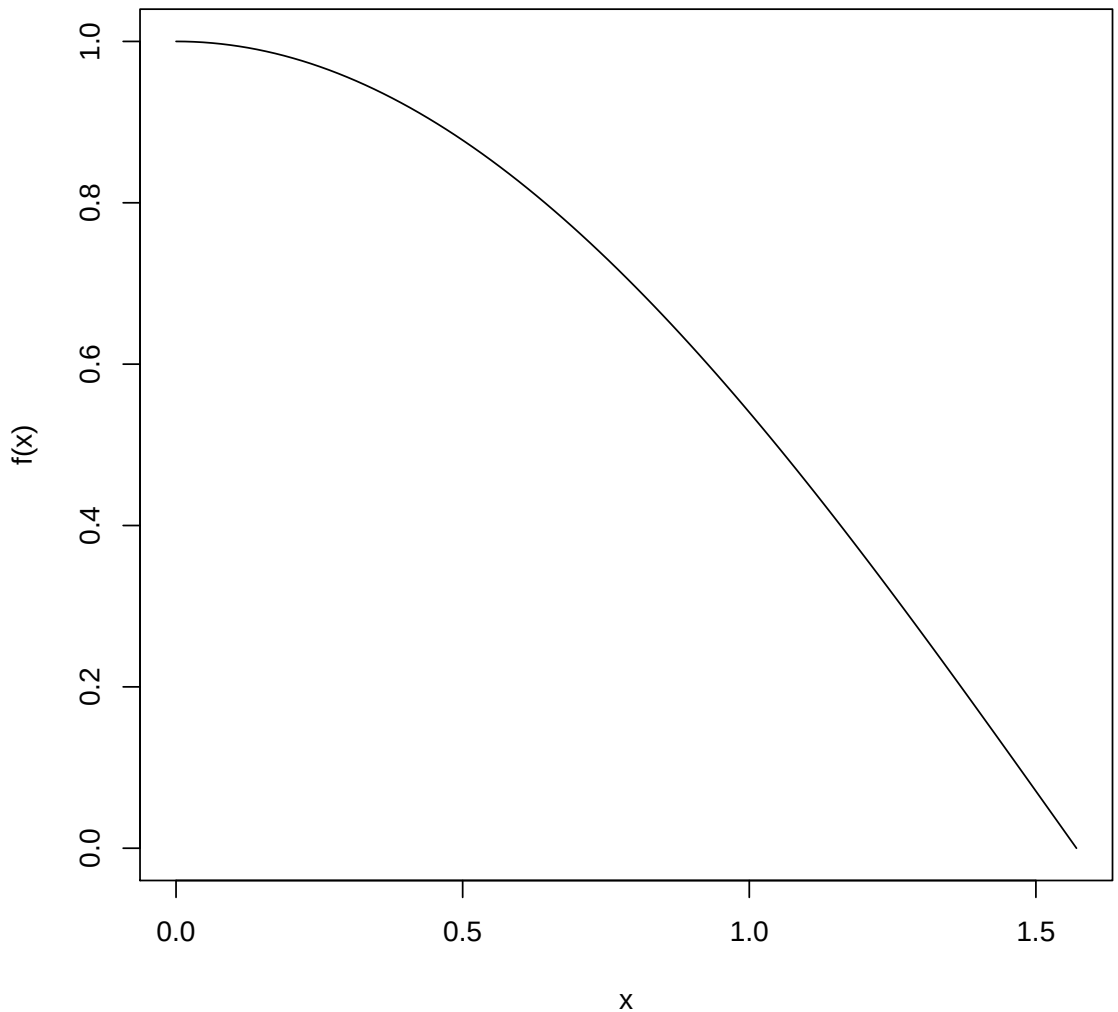
```
Xs=runif(10000,1,2)
Ys=runif(10000,0,10)

Area.Rectangle=1*10

Function.Values=abs(log(Xs^2/(Xs+1)^3)*sin(Xs^5)/atan(1/Xs^2))

hit.rate=length(which(Ys<Function.Values))/length(Function.Values)

integral=hit.rate*Area.Rectangle

integral

3.109
```

Monte Carlo Simulation

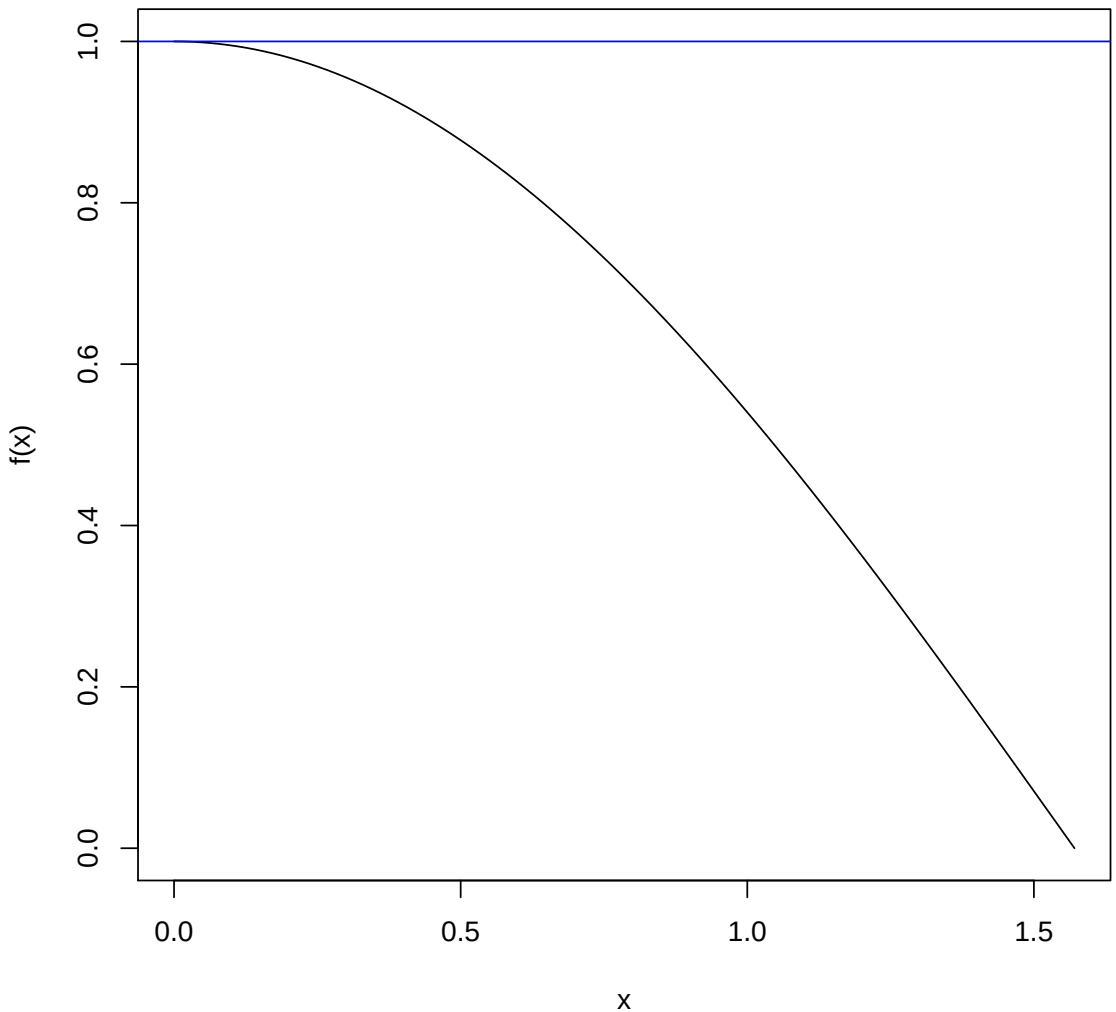
Example 4: Taking draws of an unusual random variable

$f(x) = cos(x)$ for $0 \leq x \leq \frac{\pi}{2}$, 0 otherwise

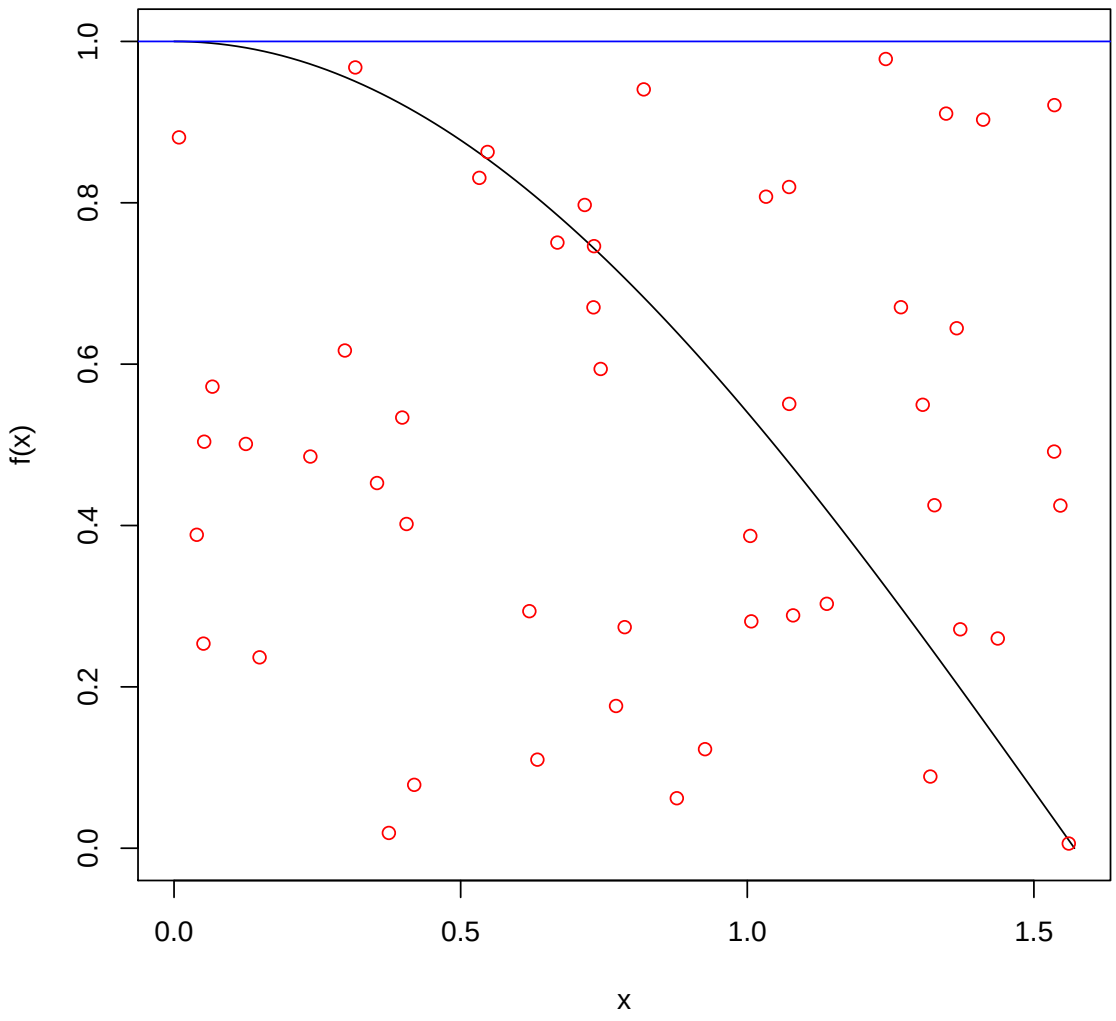
Monte Carlo Simulation



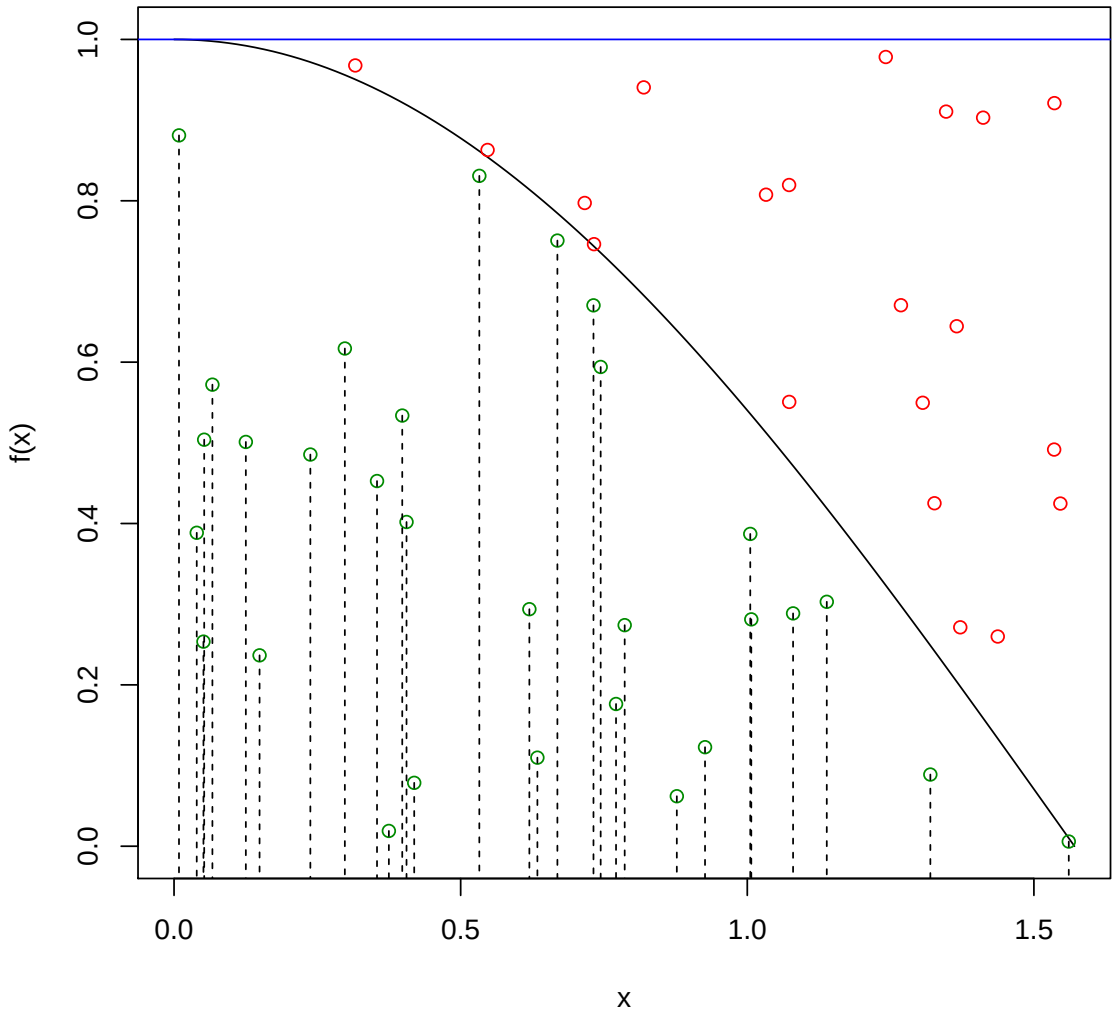
Monte Carlo Simulation



Monte Carlo Simulation



Monte Carlo Simulation



Monte Carlo Simulation

```
Xs=runif(50,0,pi/2)
Ys=runif(50,0,1)

Function.Values=cos(Xs)

hits=which(Ys<Function.Values)

draws=Xs[hits]

draws
```

Monte Carlo Simulation

Example 5: Analyzing an Experimental with Permutation Inference

Question: How unlikely is this data under the assumption of no treatment effect?

Monte Carlo Simulation

Permutation Inference

Basic Idea: Under the assumption of no treatment effect, it should not matter what units got treated. In other words, the order of 1's and 0's in T should not affect the outcome.

Steps

1. Calculate the test statistic for the real T used in the experiment. Usually, the test statistic is the estimated ATE $\hat{\tau}$.

2. Generate a large number of possible treatment assignments (1,0,1,..,1)

3. Calculate the test statistic for each treatment assignment

4. Determine how rare the actual test statistic is compared to the other test statistics

Monte Carlo Simulation

```
t=rnorm(50,1,3)
c=rnorm(50,0,3)

real.t.stat=mean(t)-mean(c)

fake.t.stats=rep(0,1000)

for(i in 1:1000){

treatmentassignment=sample(c(rep(0,50),rep(1,50)),100,
replace=FALSE)
outcomes=c(t,c)
fake.t=outcomes[treatmentassignment==1]
fake.c=outcomes[treatmentassignment==0]
fake.t.stats[i]=mean(fake.t)-mean(fake.c)

}

pvalue=length(which(abs(fake.t.stats)>=real.t.stat))/length(fake.t.stats)

pvalue
[1] 0.014
```

Monte Carlo Simulation

Permutation Inference p-value

$p = 0.014$

t-test p-value

$p = 0.009519$