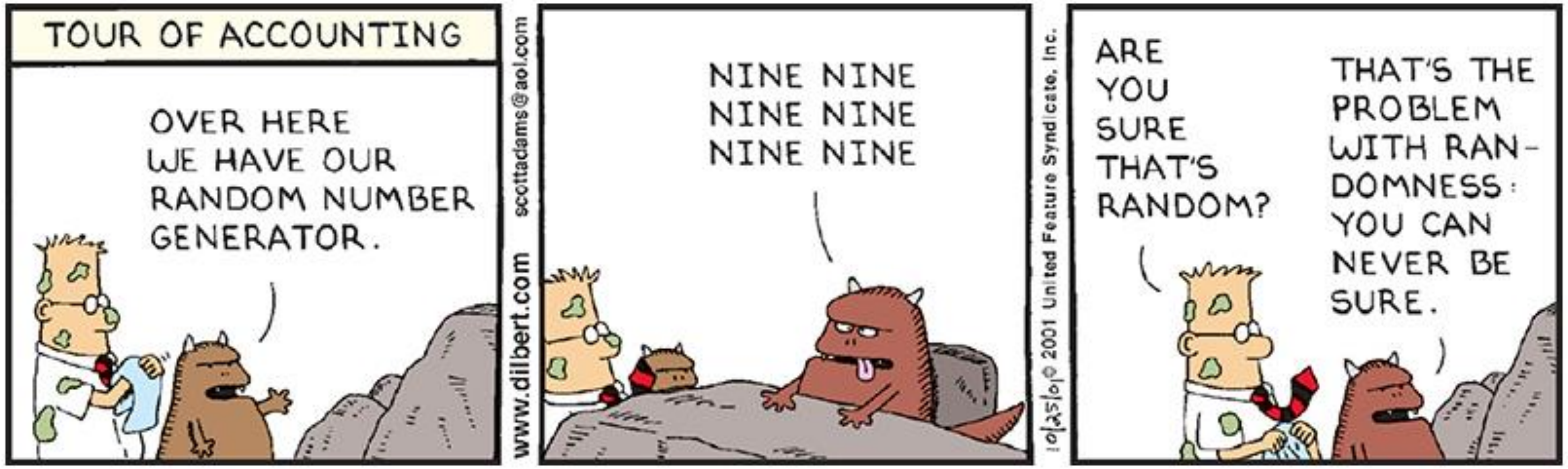
Probability, simulations, sampling, & distributions

Introduction to data analysis: Lecture 5

Ori Plonsky

Spring 2023



Slides partially based on Berkeley data8

Re: Data science: What?

Extracting useful insights from data using computation, statistics, and visualization

- Exploration
  - Play with the data, let it guide the questions
  - Identify patterns in information
  - Uses visualizations
- Inference
  - Start with a formal question
  - Quantify whether patterns are reliable
  - Uses randomization and statistical decision theory
- Prediction
  - Make informed guesses
  - Uses machine learning

Domain knowledge!

Goal

Reminder: Data typically includes a sample from a larger population

- Population includes all the elements (individuals) of interest
- Sample is a subset if observations from a population

Our goal (inference):

Answer questions that concern the population using a sample taken from it

Reminder

- In an experiment, we compare a treatment group with the control group
- Establishing causality: If the two groups are similar apart from the treatment, differences between outcomes can be ascribed to the treatment.
- If assignment of individuals to treatment and control is done at random, then the two groups are *likely* to be similar apart from the treatment

Reminder

- In an experiment, we compare a treatment group with the control group
- Establishing causality: If the two groups are similar apart from the treatment, differences between outcomes can be ascribed to the treatment.
- If assignment of individuals to treatment and control is done at random, then the two groups are *likely* to be similar apart from the treatment
- We need a randomization device!
    - Jupyter notebook

Probability: Basic definitions

- Random trial – a trial (experiment) the exact outcome(s) of which cannot be known in advance
  - Rolling a die
- Sample space – a set of possible outcomes of an experiment
  - {1, 2, 3, 4, 5, 6}

Probability: Basic definitions

- Random trial – a trial (experiment) the exact outcome(s) of which cannot be known in advance
  - Rolling a die
- Sample space – a set of possible outcomes of an experiment
  - {1, 2, 3, 4, 5, 6}
- Event – subset of the sample space
  - A = {5}: “a five” (elementary event)
  - B = {1, 3, 5}: “an uneven number”
  - C = {1, 2}: “a number under 3”
  - D = {1, 2, 3, 4, 5, 6}

Probability: Basic definitions

- Random trial – a trial (experiment) the exact outcome(s) of which cannot be known in advance
  - Rolling a die
- Sample space – a set of possible outcomes of an experiment
  - {1, 2, 3, 4, 5, 6}
- Event – subset of the sample space
  - A = {5}: “a five” (elementary event)
  - B = {1, 3, 5}: “an uneven number”
  - C = {1, 2}: “a number under 3”
  - D = {1, 2, 3, 4, 5, 6}
- Complementary event – the complementary set of an event
  - what are the complementary events A, B, C, D?

Probability: Basic definitions

- The chance for an event occurring is its *probability of occurrence*

Probability: Basic definitions

- The chance for an event occurring is its *probability of occurrence*
- If an event is **impossible** its probability is 0
  - Lowest probability value

Probability: Basic definitions

- The chance for an event occurring is its *probability of occurrence*
- If an event is **impossible** its probability is 0
    - Lowest probability value
- If an event is **certain** its probability is 1
    - Highest probability value

Probability: Basic definitions

- The chance for an event occurring is its *probability of occurrence*
- If an event is **impossible** its probability is 0
  - Lowest probability value
- If an event is **certain** its probability is 1
  - Highest probability value
- If an event has probability p,
  the probability of its complementary event is $1-p$

Equally likely outcomes

- If each possible outcome is equally likely (all elementary events in the sample space have the same chance of happening), then the probability of event A is:

$$P\ \ A\ = \frac{Number\,of\,outcomes\,that\,make\,A\,occur}{Total\,number\,of\,possible\,outcomes}$$

Equally likely outcomes

- If each possible outcome is equally likely (all elementary events in the sample space have the same chance of happening), then the probability of event A is:

$$P \quad A = \frac{Number\, of\, outcomes\, that\, make\, A\, occur}{Total\, number\, of\, possible\, outcomes}$$

- What are the probabilities of A, B, C, D? Of their complementary events?

- A = {5}
- B = {1, 3, 5}
- C = {1, 2}
- D = {1, 2, 3, 4, 5, 6}

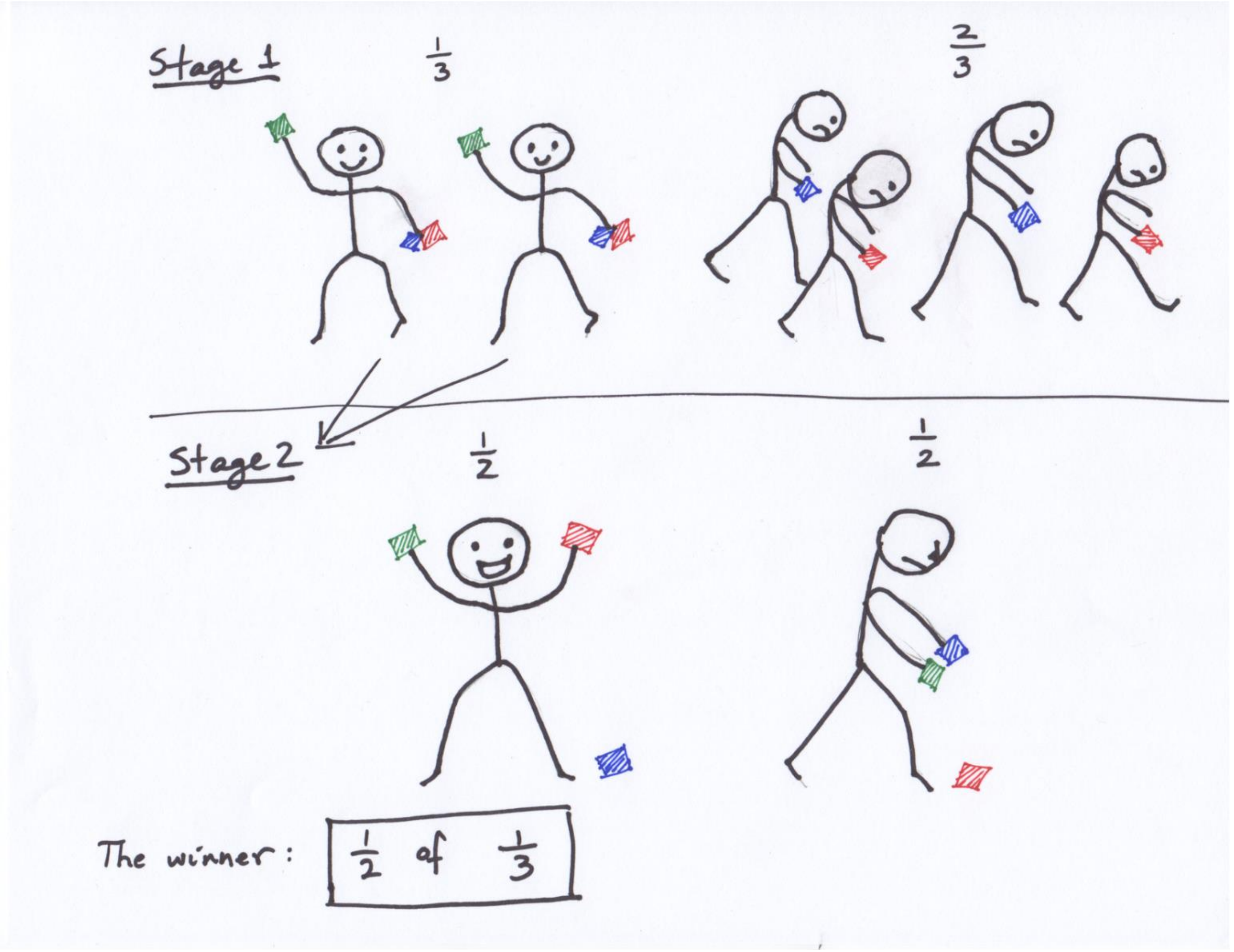
Conditional probability

- Think of the following game-of-chance:
  - There are three tickets in a hat: Red, Blue, and Green
  - Stage 1: You draw a ticket from the hat.
    - If it is Green, you move to Stage 2;
    - otherwise, you lose
  - Stage 2: without putting the Green ticket back in the hat, you draw another ticket
    - If it is Red, you win the game
    - otherwise, you lose.

Conditional probability

- Think of the following game-of-chance:
  - There are three tickets in a hat: Red, Blue, and Green
  - Stage 1: You draw a ticket from the hat.
    - If it is Green, you move to Stage 2;
    - otherwise, you lose
  - Stage 2: without putting the Green ticket back in the hat, you draw another ticket
    - If it is Red, you win the game
    - otherwise, you lose.

- What are the chances of winning the game?



Stage 1
1/3
2/3
Stage 2
1/2
1/2
The winner:
1/2 of 1/3

Multiplication rule

Chance that two events *A* and *B* both happen

= P(*A* happens) x P(*B* happens given that *A* has happened)

Multiplication rule

Chance that two events *A* and *B* both happen

= P(*A* happens) x P(*B* happens given that *A* has happened)

- The answer is **less than or equal** *to* each of the two chances multiplied

Multiplication rule

Chance that two events *A* and *B* both happen

= P(*A* happens) x P(*B* happens given that *A* has happened)

- The answer is **less than or equal** *to* each of the two chances multiplied
- The more conditions you have to satisfy, the less likely you are to satisfy them all

Yoram Lass is at **Neve Avivim**.

2 hrs • Tel Aviv, Israel •

לא נפטרים ממנה

בתגובה על "מעקב קורונה – ישראל, כל הנתונים",
("הארץ", 25.4)

תוחלת החיים בישראל היא 82 שנה. זה הגיל הממוצע שבו נפטרים בישראל. הגיל ממוצע של הנפטרים בגלל מחלת הקורונה הוא גם כן כ-82 שנה. מכאן שבממוצע, בכלל לא מתים ממחלת הקורונה (לידיעת כל מי שמדבר על ימי הביניים ועל השואה).

יורם לס, תל אביב

376 162 Comments • 53 Shares

Problem 1

Please choose ‘A’ or ‘B’:

A:
300 with certainty

B:
400 with chance of 4/5
0 with chance of 1/5

A

B

Problem 2

Please choose ‘A’ or ‘B’:

A:
300 with chance of 1/4
0 with chance of 3/4

B:
400 with chance of 1/5
0 with chance of 4/5

A

B

Problem 2

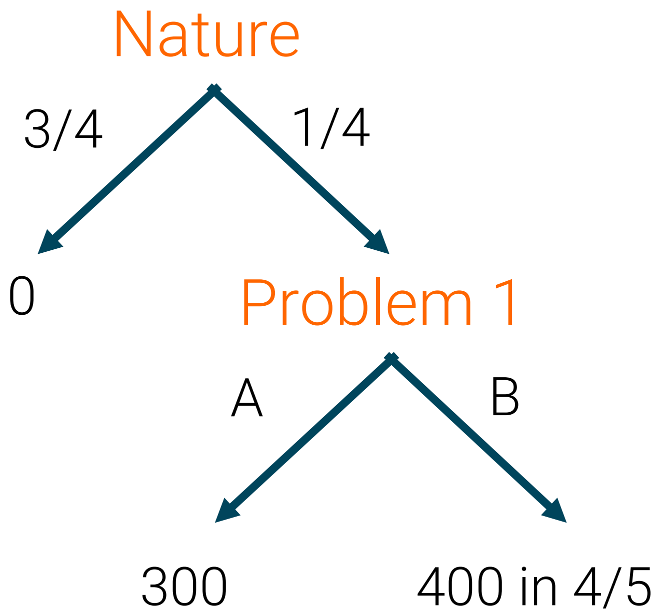
Please choose ‘A’ or ‘B’:

A:
300 with chance of 1/4
0 with chance of 3/4

B:
400 with chance of 1/5
0 with chance of 4/5

A

B



Probability of disjoint events

- Think of the following game-of-chance:
  - There are three tickets in a hat: Red, Blue, and Green
  - You draw 2 tickets one after the other
    - If one of them is Green and the other Red, you win
    - Otherwise, you lose

- Would you prefer playing this game or the previous game?

Probability of disjoint events

- Think of the following game-of-chance:
  - There are three tickets in a hat: Red, Blue, and Green
  - You draw 2 tickets one after the other
    - If one of them is Green and the other Red, you win
    - Otherwise, you lose

- Would you prefer playing this game or the previous game?
- What are the chances of winning the game?

Addition rule

If event A can happen in exactly one of two ways, and these two ways cannot happen at the same time (disjoint):

$$P(A) = P(\text{first way}) + P(\text{second way})$$

Addition rule

If event A can happen in exactly one of two ways, and these two ways cannot happen at the same time (disjoint):

$$P(A) = P(\text{first way}) + P(\text{second way})$$

- The answer is **greater than or equal** *to* the chance of each individual way

Examples: coin tosses

Say you have 2 coin tosses.

- What is the probability of getting exactly one H?

Examples: coin tosses

Say you have 2 coin tosses.

- What is the probability of getting exactly one H?
- What is the probability of getting at least one H?

Examples: coin tosses

Say you have 2 coin tosses.

- What is the probability of getting <u>exactly</u> one H?
- What is the probability of getting <u>at least</u> one H?

- Possible outcomes: HH, HT, TH, TT

At least one success

- Data scientists often work with random samples from populations.
- What is the probability of a random sample to include a specific event?
    - To have at least one “success”

At least one success

- Data scientists often work with random samples from populations.
- What is the probability of a random sample to include a specific event?
  - To have at least one “success”

- What is the probability to get at least one H in 3 tosses?

At least one success

- Data scientists often work with random samples from populations.
- What is the probability of a random sample to include a specific event?
    - To have at least one “success”

- What is the probability to get at least one H in 3 tosses?
    - Any outcome except TTT
    - $P(TTT) = (\frac{1}{2}) \times (\frac{1}{2}) \times (\frac{1}{2}) = \frac{1}{8}$
    - $P(\text{at least one head}) = 1 - P(TTT) = 1 - (\frac{1}{2})^3 = \frac{7}{8} = 87.5\%$

At least one success

- Data scientists often work with random samples from populations.
- What is the probability of a random sample to include a specific event?
  - To have at least one “success”

- What is the probability to get at least one H in 3 tosses?
  - Any outcome except TTT
  - P(TTT) = $(\frac{1}{2}) \times (\frac{1}{2}) \times (\frac{1}{2}) = \frac{1}{8}$
  - P(at least one head) = 1 − P(TTT) = $1 - (\frac{1}{2})^3 = \frac{7}{8}$ = 87.5%

- In 10 tosses?

At least one success

- Data scientists often work with random samples from populations.
- What is the probability of a random sample to include a specific event?
  - To have at least one “success”

- What is the probability to get at least one H in 3 tosses?
  - Any outcome except TTT
  - P(TTT) = $(\frac{1}{2}) \times (\frac{1}{2}) \times (\frac{1}{2}) = \frac{1}{8}$
  - P(at least one head) = 1 − P(TTT) = $1 - (\frac{1}{2})^3 = \frac{7}{8}$ = 87.5%

- In 10 tosses?
  - $1 - (\frac{1}{2})^{10}$ = 99.9%

(Notebook: Probabilities)

Simulating experiments

- Using Probability Theory, we can compute probabilities and other statistical quantities mathematically
  - Like we just did for “at least one success”
- In this course, we focus on simulating such quantities
- We will tell Python to simulate a random trial that gives us what we want and repeat it many times
- We can then summarize the results of all simulations to see the value of the quantity we want

Steps of simulation

Simulation is the process of using a computer to mimic a physical experiment. We will need the following steps:

Steps of simulation

Simulation is the process of using a computer to mimic a physical experiment. We will need the following steps:

1. Specify the quantity you want to simulate.

Steps of simulation

Simulation is the process of using a computer to mimic a physical experiment. We will need the following steps:

1. Specify the quantity you want to simulate.
2. Figure out how to simulate the quantity you specified in Step 1 once.

Steps of simulation

Simulation is the process of using a computer to mimic a physical experiment. We will need the following steps:

1. Specify the quantity you want to simulate.
2. Figure out how to simulate the quantity you specified in Step 1 once.
3. Choose how many times you want to simulate the quantity

Steps of simulation

Simulation is the process of using a computer to mimic a physical experiment. We will need the following steps:

1. Specify the quantity you want to simulate.
2. Figure out how to simulate the quantity you specified in Step 1 once.
3. Choose how many times you want to simulate the quantity
4. Code the simulation

Steps of simulation

Simulation is the process of using a computer to mimic a physical experiment. We will need the following steps:

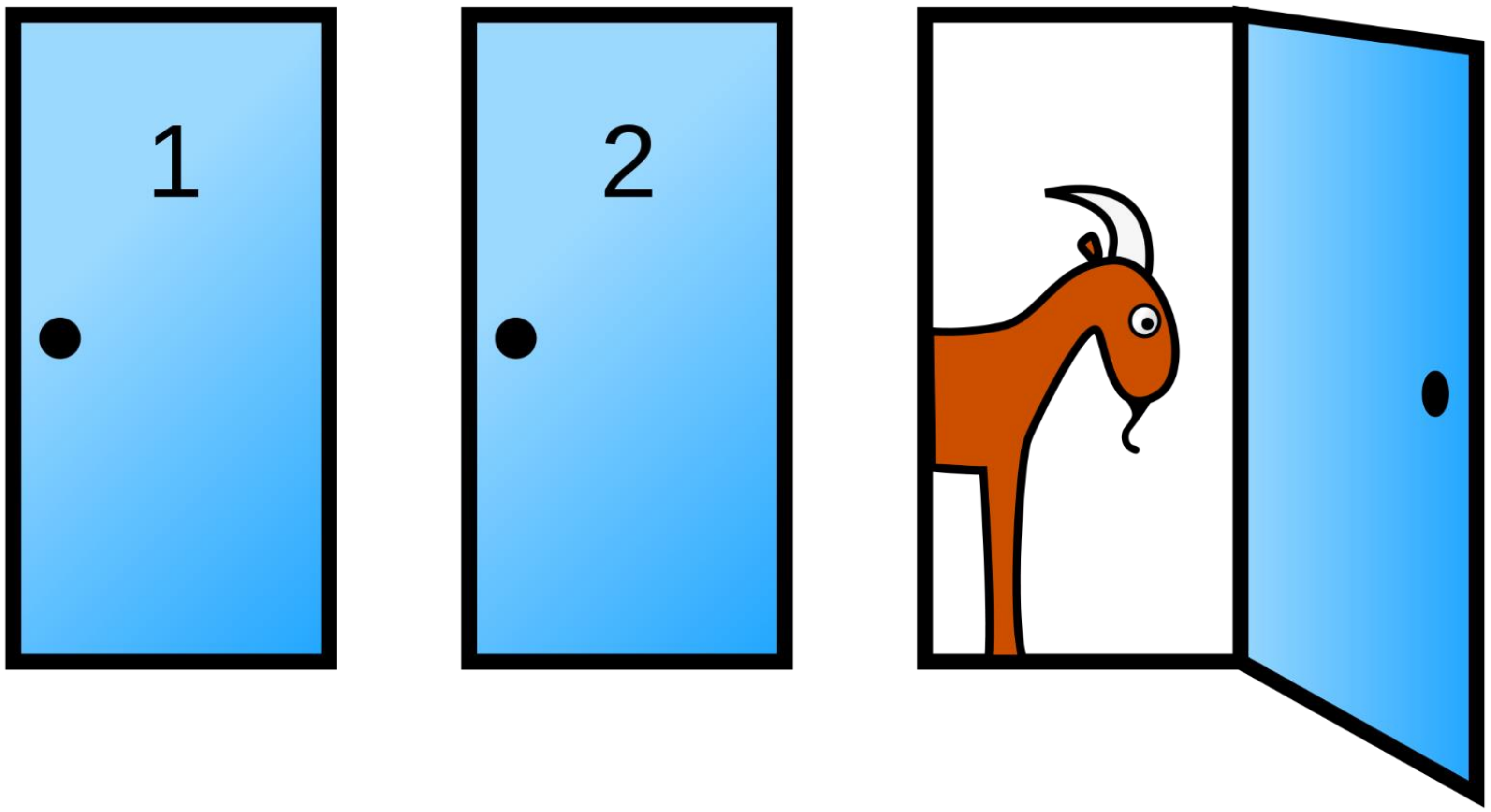
1. Specify the quantity you want to simulate.
2. Figure out how to simulate the quantity you specified in Step 1 once.
3. Choose how many times you want to simulate the quantity
4. Code the simulation
   - Create a “collection array”: an empty array in which we’ll collect the simulated values
   - Create a *for* loop with as many iterations as you defined in Step 3.
   - In each iteration, simulate one value based on the code from Step 2, and append the simulated value to the collection array
   - At the end of the loop, the collection array will hold all simulated values, and you can summarize the results

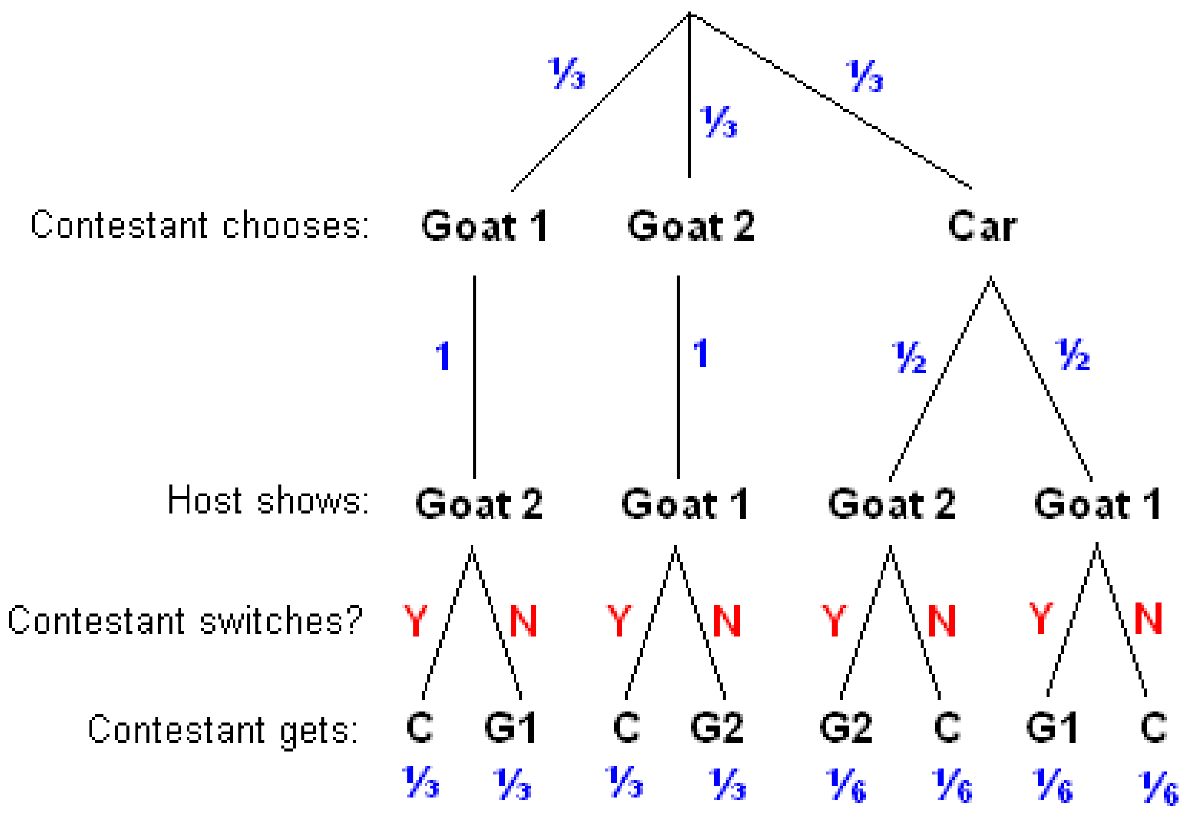
Steps of simulation

Simulation is the process of using a computer to mimic a physical experiment. We will need the following steps:

1. Specify the quantity you want to simulate.
2. Figure out how to simulate the quantity you specified in Step 1 once.
3. Choose how many times you want to simulate the quantity
4. Code the simulation
   - Create a “collection array”: an empty array in which we’ll collect the simulated values
   - Create a *for* loop with as many iterations as you defined in Step 3.
   - In each iteration, simulate one value based on the code from Step 2, and append the simulated value to the collection array
   - At the end of the loop, the collection array will hold all simulated values, and you can summarize the results

(Notebook: Simulation)


1
2



Switch
- Car: ⅔
- Goat 1: ⅙
- Goat 2: ⅙

No switch
- Car: ⅓
- Goat 1: ⅓
- Goat 2: ⅓

Sampling

Deterministic sample:

- Sampling does not involve chance

Probability sample:

- Each individual has a non-zero probability of being selected into the sample
- Before sample is drawn, we need to know the probability of selecting each group of individuals from the population
    - It is not required that each individual will have the same probability of being sampled
    - But it can help if they do
        - If in addition the selection of each individual is independent on the selection of all other individuals, such sampling is called *simple random sampling*

(notebook)

Convenience sampling

- Example: sample consists of whoever walks by
- Just because you think you’re sampling “at random”, doesn’t mean you are

Convenience sampling

- Example: sample consists of whoever walks by
- Just because you think you're sampling "at random", doesn't mean you are
- If you can't figure out ahead of time
    1. what's the population
    2. what's the chance of selection, for each group in the population

then you don't have a random sample

Probability Distribution

- Random quantity with various possible values
- "Probability distribution":
    - All the possible values of the quantity
    - The probability of each of those values

Probability Distribution

- Random quantity with various possible values
- “Probability distribution”:
  - All the possible values of the quantity
  - The probability of each of those values
- If you can do the math, you can work out the probability distribution without ever simulating the random quantity

Empirical Distribution

- Based on observations
- Observations can be from repetitions of an experiment
- "Empirical Distribution":
  - All observed values
  - The proportion of counts of each value

Law of Averages

If a chance experiment is repeated many times, independently, and under the same conditions,
then the proportion of times that an event occurs gets closer to the theoretical probability of the event

Law of Averages

If a chance experiment is repeated many times, independently, and under the same conditions,
then the proportion of times that an event occurs gets closer to the theoretical probability of the event

- Example: As we increase the number of rolls of a die, the proportion of times you see the face with five spots gets closer to 1/6

Law of Averages

If a chance experiment is repeated many times, independently, and under the same conditions,
then the proportion of times that an event occurs gets closer to the theoretical probability of the event

- Example: As we increase the number of rolls of a die, the proportion of times you see the face with five spots gets closer to 1/6
- Example: As we increase the number of times we roll a die four times, the proportion of times we get at least one 6 gets closer to 1-(5/6)^4

Law of Averages

If a chance experiment is repeated many times, independently, and under the same conditions,
then the proportion of times that an event occurs gets closer to the theoretical probability of the event

- Example: As we increase the number of rolls of a die, the proportion of times you see the face with five spots gets closer to 1/6
- Example: As we increase the number of times we roll a die four times, the proportion of times we get at least one 6 gets closer to 1-(5/6)^4

As a result:

If the sample size is large, then the empirical distribution of a **simple random sample** resembles the distribution of the population, with high probability