# YData: An Introduction to Data Science

## Lecture 15: Sampling

Elena Khusainova & John Lafferty
Statistics & Data Science, Yale University
Spring 2021

Credit: data8.org

# For today

- Elementary probability redux
- The Monty Hall puzzle
- Sampling

# Probability

# Basics

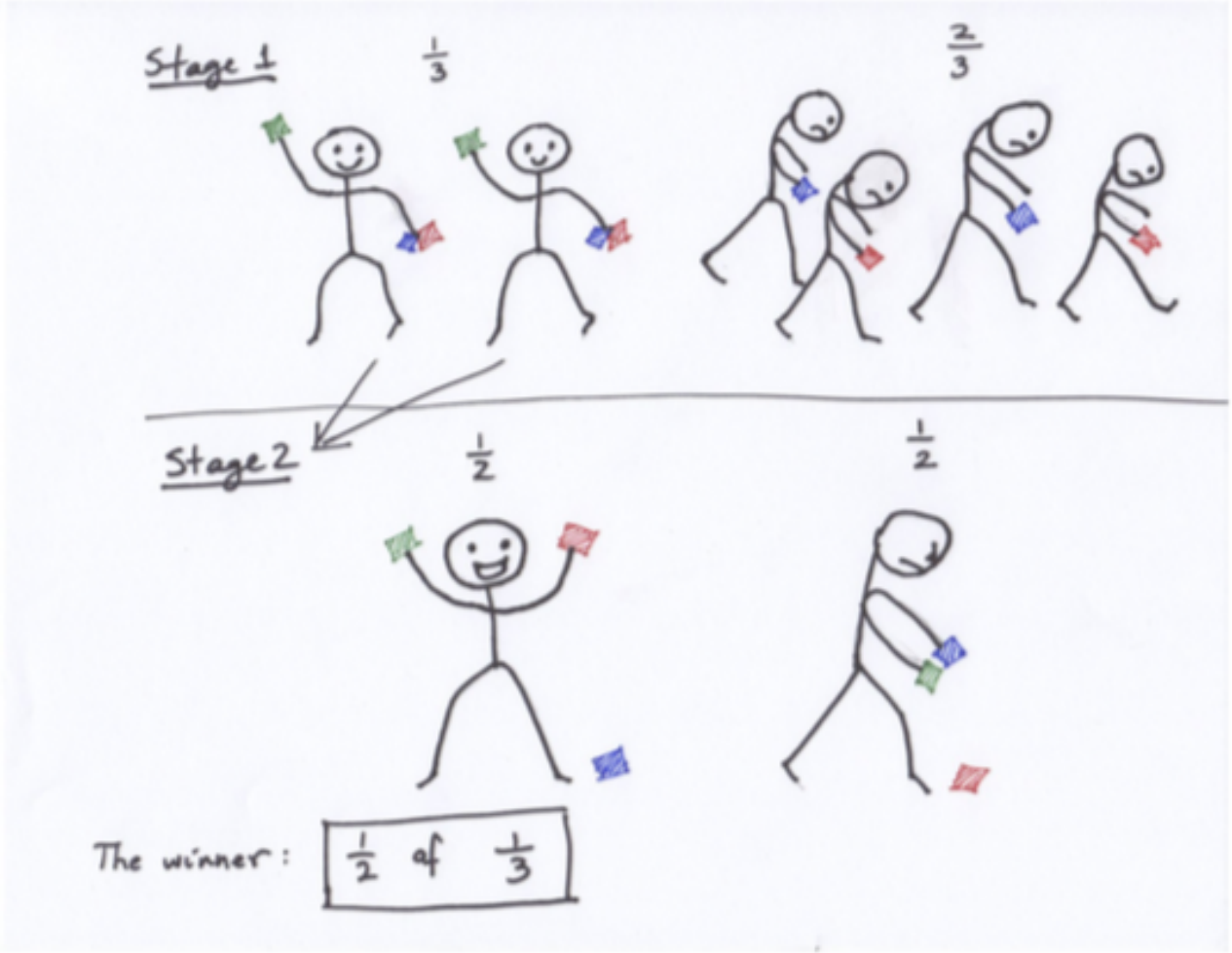
- Lowest value: 0
  - Chance of event that is impossible
- Highest value: 1 (or 100%)
  - Chance of event that is certain
- If an event has chance 70%, then the chance that it doesn't happen is
  - 100% - 70% = 30%
  - 1 - 0.7 = 0.3

# Equally Likely Outcomes

*Assuming all outcomes are equally likely*, chance of an event A is:

$$P(A) = \frac{\text{number of outcomes that make A happen}}{\text{total number of outcomes}}$$

# Fraction of a Fraction



# Multiplication Rule

Chance that two events A and B both happen

$$= P(\text{A happens}) \times P(\text{B happens given that A has happened})$$

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

# Addition Rule

If event A can happen in *exactly* one of two ways, then

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

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

# Example: At Least One Head

- In 3 tosses:
  - Any outcome except TTT
  - P(TTT) $= (1/2) \times (1/2) \times (1/2) = 1/8$
  - P(at least one head) = 1 - P(TTT) $= 7/8 = 87.5\%$
- In 10 tosses:
  - 1 - (1/2)**10
  - 99.9%

# Discussion Question

A population has 100 people, including Mo and Jo.
We sample two people at random without replacement.

(a) P(both Mo and Jo are in the sample)

# Discussion Question

A population has 100 people, including Mo and Jo.
We sample two people at random without replacement.

(a) P(both Mo and Jo are in the sample)
= P(first Mo, then Jo) + P(first Jo, then Mo)

# Discussion Question

A population has 100 people, including Mo and Jo.
We sample two people at random without replacement.

(a) P(both Mo and Jo are in the sample)
= P(first Mo, then Jo) + P(first Jo, then Mo)
= (1/100) * (1/99) + (1/100) * (1/99) = 0.0002

# Discussion Question

A population has 100 people, including Mo and Jo.
We sample two people at random without replacement.

(a) P(both Mo and Jo are in the sample)
= P(first Mo, then Jo) + P(first Jo, then Mo)
= (1/100) * (1/99) + (1/100) * (1/99) = 0.0002

(b) P(neither Mo nor Jo is in the sample)

# Discussion Question

A population has 100 people, including Mo and Jo.
We sample two people at random without replacement.

(a) P(both Mo and Jo are in the sample)
= P(first Mo, then Jo) + P(first Jo, then Mo)
= (1/100) * (1/99) + (1/100) * (1/99) = 0.0002

(b) P(neither Mo nor Jo is in the sample)
= (98/100) * (97/99)

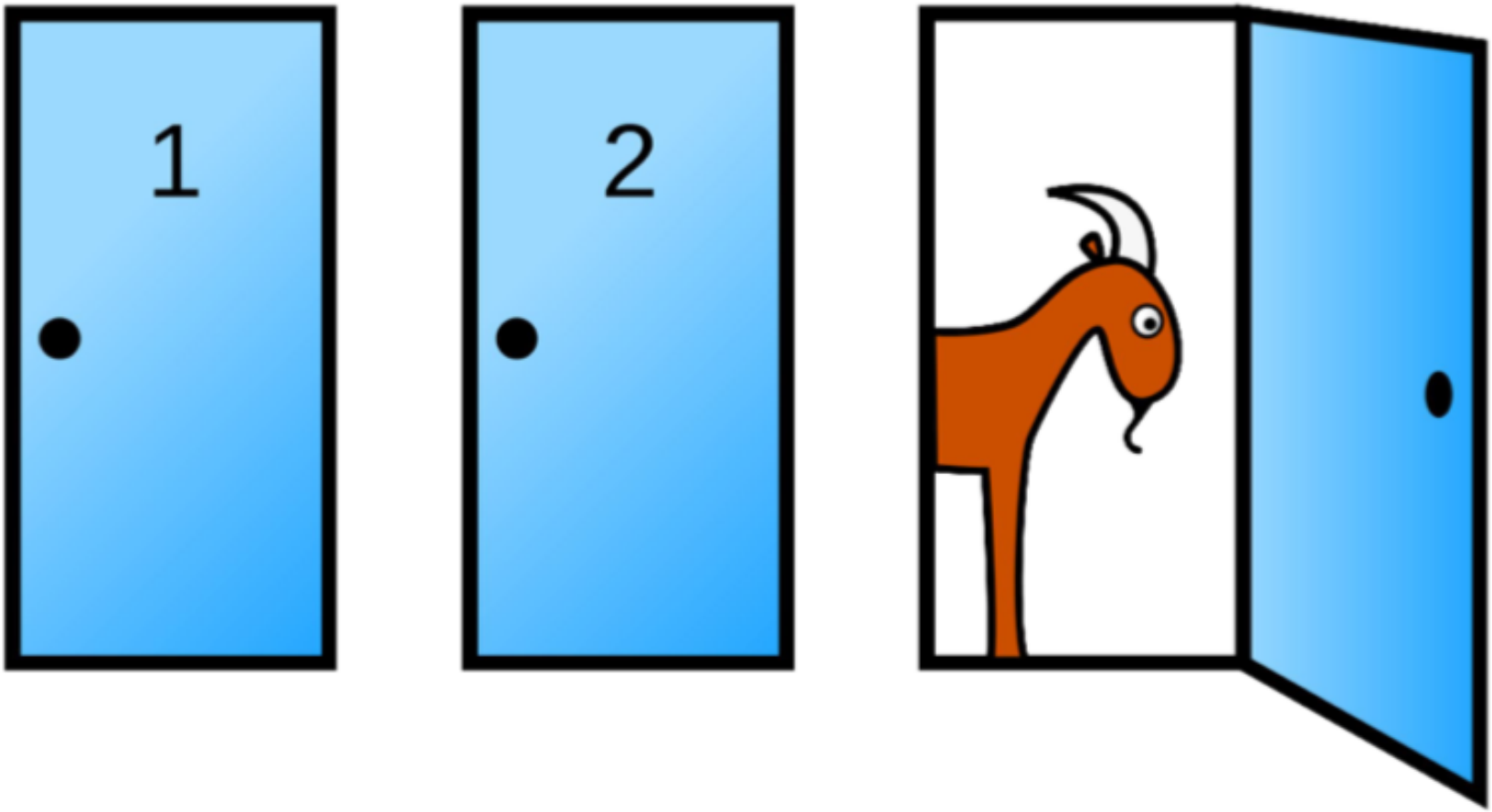
# Discussion Question

A population has 100 people, including Mo and Jo.
We sample two people at random without replacement.

(a) P(both Mo and Jo are in the sample)
= P(first Mo, then Jo) + P(first Jo, then Mo)
= (1/100) * (1/99) + (1/100) * (1/99) = 0.0002

(b) P(neither Mo nor Jo is in the sample)
= (98/100) * (97/99)
= 0.9602

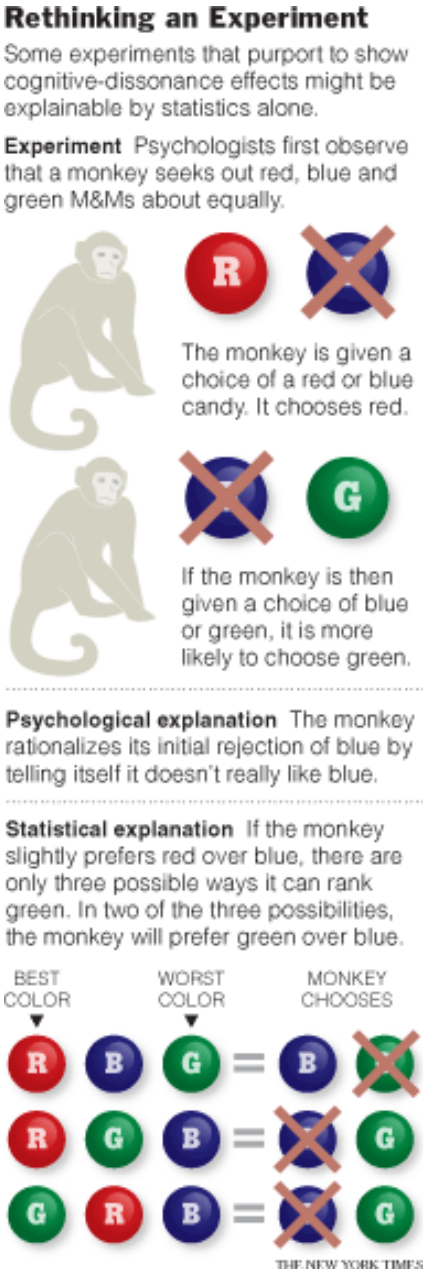
# Monty Hall Problem




(DEMO)

# Food for Thought

https://www.nytimes.com/2008/04/08/science/08tier.html



# Sampling

# Sampling

- Deterministic sample:
  - Sampling scheme doesn't involve chance
- Probability sample:
  - Before the sample is drawn, you have to know the selection probability of every group of people in the population
  - Not all individuals have to have equal chance of being selected

# Sample of Convenience

- Example: sample consists of whomever 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
  - what's the population
  - what's the chance of selection, for each group in the population

  then you don't have a random sample

# Distributions

# 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

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

(DEMO)

# Large Random Samples

# 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

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

# Empirical Distribution of a Sample

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

(DEMO)

# A Statistic

# Inference

- **Statistical Inference:**
  Making conclusions based on data in random samples

- **Example:**
  Use the data to guess the value of a fixed, unknown number

  Create an **estimate** [depends on the random sample] of the unknown quantity

# Terminology

- **Parameter**
  - A number associated with the population
- **Statistic**
  - A number calculated from the sample

A statistic can be used as an estimate of a parameter

(DEMO)

# Probability Distribution of a Statistic

- Values of a statistic vary because random samples vary
- "Sampling distribution" or "probability distribution" of the statistic:
  - All possible values of the statistic,
  - and all the corresponding probabilities
- Can be hard to calculate
  - Either have to do the math
  - Or have to generate all possible samples and calculate the statistic based on each sample

# Empirical Distribution of a Statistic

- Empirical distribution of the statistic:
  - Based on simulated values of the statistic
  - Consists of all the observed values of the statistic,
  - and the proportion of times each value appeared

- Good approximation to the probability distribution of the statistic
  - if the number of repetitions in the simulation is large