QATAR UNIVERSITY

CMPS 360

Data Science Fundamentals

Spring 2025

Session 6

Randomness & Sampling

Ch. 9, 10

Many slides are created by John DeNero (denero@berkeley.edu) and Ani Adhikari (adhikari@berkeley.edu)

Rows,
Applying a Function to a Row

Rows

A row of a table has items and can be aggregated.

```
r = t.row(0) # r is the row at index 0

r.item(1)     # item can take an index
r.item('name') # item can take a label

sum(r)        # Also: np.average, min, max,
etc.
```

Apply with One Argument

`t.apply(f)` for a table `t` and function `f` creates an array of the results of applying `f` to each *row* of `t`.

e.g., `t.apply(sum)` would return the sum of each row as an array.

(Demos)

RANDOM SELECTION & SIMULATION

Random Selection

Random Selection

`np.random.choice`

- Selects uniformly at random
- with replacement
- from an array,
- a specified number of times

```
np.random.choice(some_array, sample_size)
```

(Demos -- Random Selection, A Game)

Appending Arrays

A Longer Array

- **`np.append(array_1, value)`**
  - new array with `value` appended to `array_1`
  - `value` has to be of the same type as elements of `array_1`
- **`np.append(array_1, array_2)`**
  - new array with `array_2` appended to `array_1`
  - `array_2` elements must have the same type as `array_1` elements

(Demo)

Simulation

(Demo)

CHANCE

Revision of
Basics of Probability

Basics

- **Lowest value**: 0
  - Chance of event that is impossible
- **Highest value**: 1 (or 100%)
  - Chance of event that is certain

- **Complement**: 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, the chance of an event A is:

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

A Question

- I have three cards: **ace of hearts**, **king of diamonds**, and **queen of spades**.
- I shuffle them and draw two cards *at random without replacement.*

- What is the chance that I get the Queen followed by the King?

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 being multiplied
- The more conditions you have to satisfy, the less likely you are to satisfy them all

Another Question

- I have three cards: **ace of hearts**, **king of diamonds**, and **queen of spades**.
- I shuffle them and draw two cards *at random without replacement.*

- What is the chance that one of the cards I draw is a King and the other is Queen?

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

Complement: At Least One Head

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

Problem-Solving Method

Here’s a method that works widely:

Ask yourself what event must happen on the first trial.

- If there’s a clear answer (e.g. “not a six”) whose probability you know, you can most likely use the **multiplication rule**.
- If there’s no clear answer (e.g. “could be K or Q, but then the next one would have to be Q or K …”), list all the **distinct ways** your event could occur and **add up their chances**.
- If the list above is long and complicated, look at the **complement**. If the complement is simpler (e.g. the complement of “at least one” is “none”), you can find its chance and subtract that from 1.

Discussion Question

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

(a) P(both Rick and Morty are in the sample)

= P(first Rick, then Morty) + P(first Morty, then Rick)

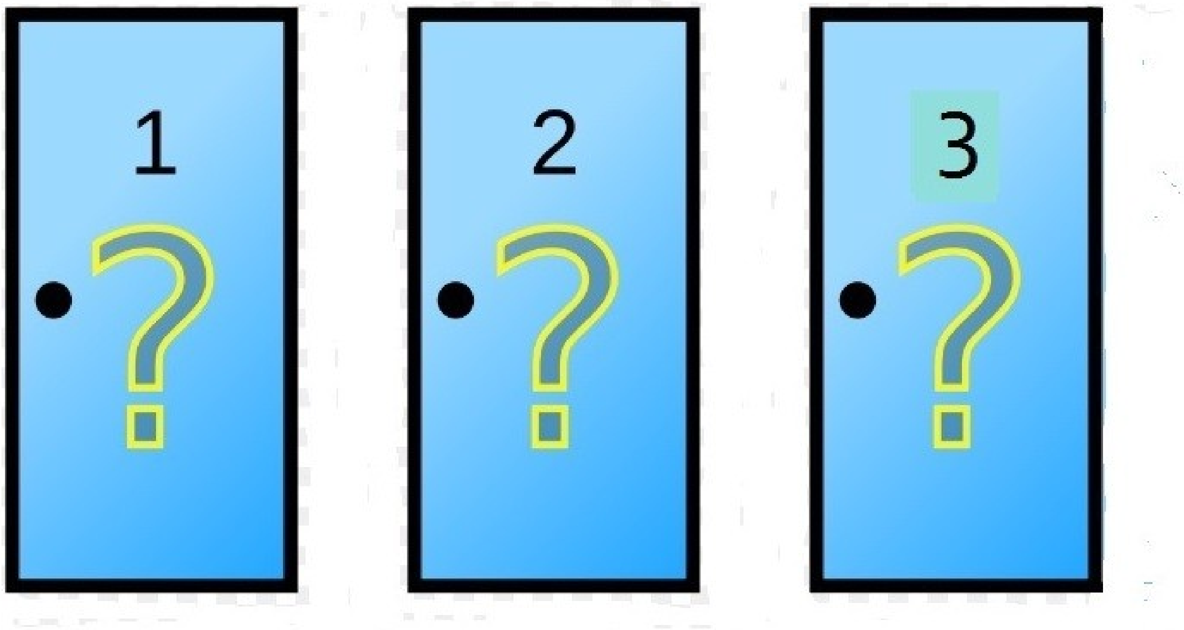
= (1/100) * (1/99) + (1/100) * (1/99) = 0.0002

(b) P(neither Rick nor Morty is in the sample)

= (98/100) * (97/99) = 0.9602

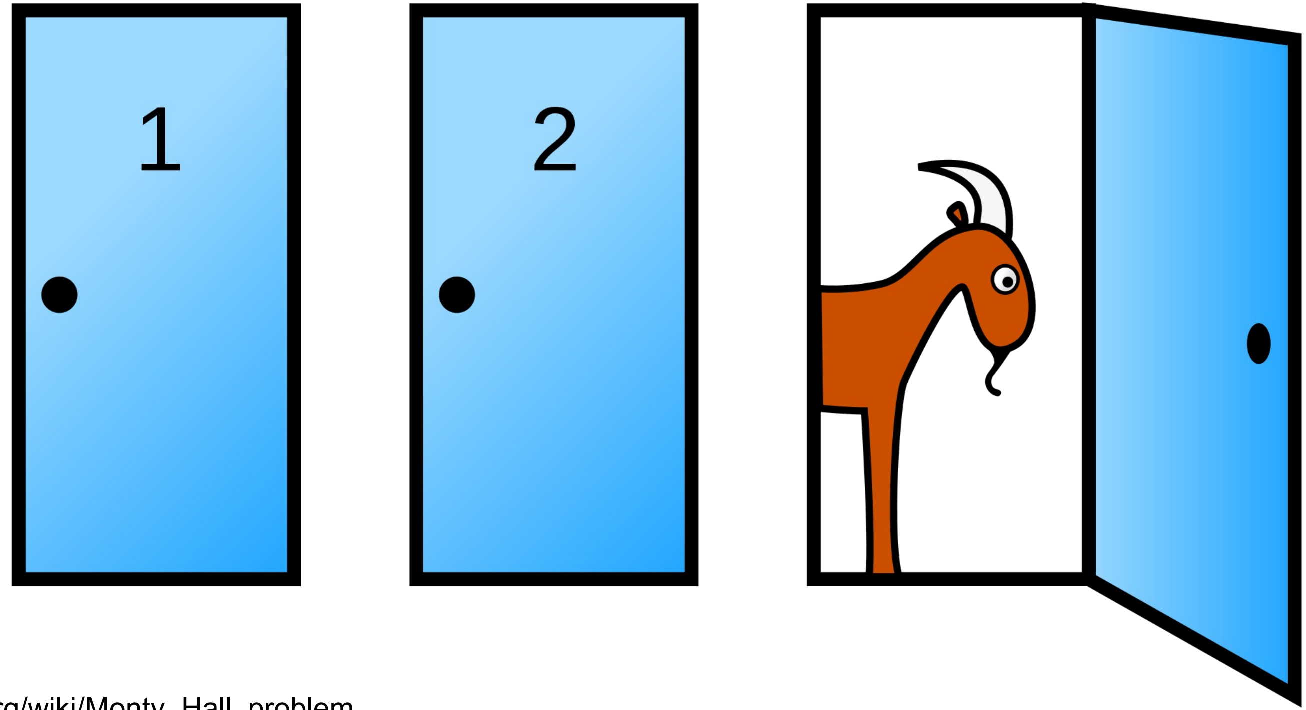
The Monty Hall Problem

Monty Hall Problem



https://probabilityandstats.files.wordpress.com/2017/05/monty-hall-pic-1.jpg

Stay or Switch?

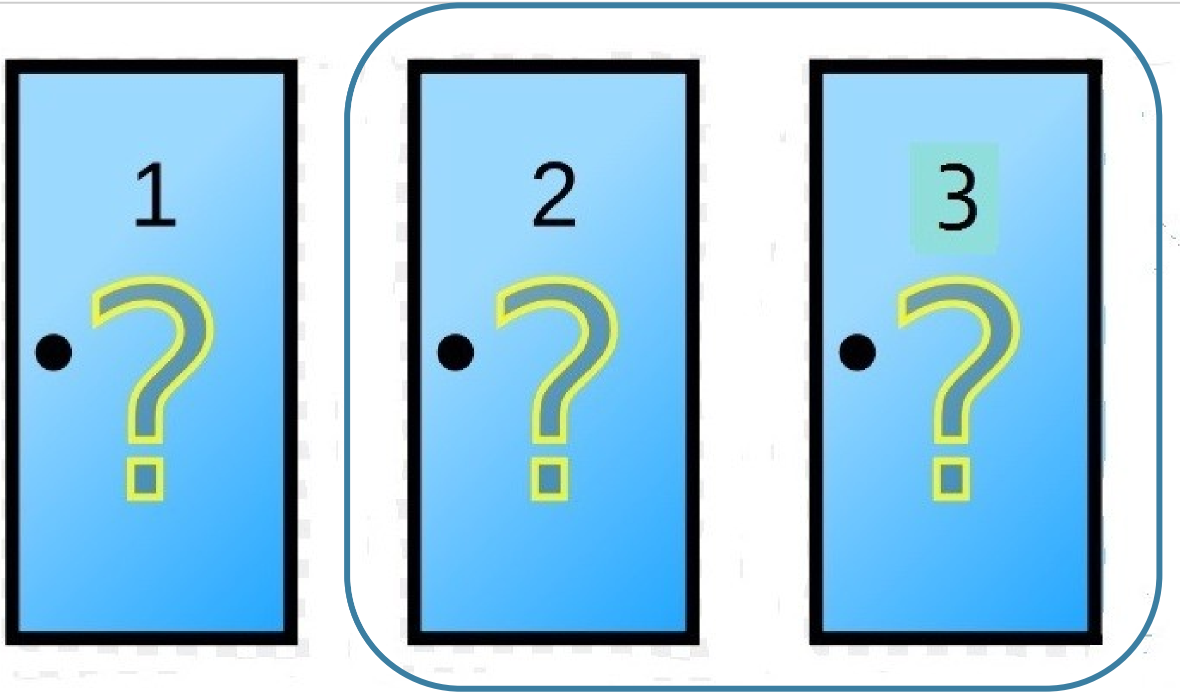


https://en.wikipedia.org/wiki/Monty_Hall_problem

Equally Likely Outcomes

My Choice	Monty's Reveal	Other Choice
🐐 1️⃣	🐐 2️⃣	🚗
🐐 2️⃣	🐐 1️⃣	🚗
🚗	🐐 1️⃣ or 🐐 2️⃣	🐐 2️⃣ or 🐐 1️⃣

Stay or Switch?



https://probabilityandstats.files.wordpress.com/2017/05/monty-hall-pic-1.jpg

Simulating the Monty Hall Problem

(Demo)

SAMPLING

Random Samples

- Deterministic sample:
  - Sampling scheme doesn't involve chance

- Random sample:
  - Before the sample is drawn, you have to know the selection probability of every group of people in the population.
  - Not all individuals / groups have to have equal chance of being selected.

(Demo)

Sample of Convenience

- Example: sample consists of whoever walks by
- Just because you think you're sampling "randomly", doesn't mean you have a random sample.
- 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 variable: 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 it.
- But... simulation is often easier!

Empirical Distribution

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

(Demo)

Large Random Samples

Law of Averages (or Large Numbers)

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 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

Statistical Inference

Making conclusions based on data in random samples.

- **Example**:

fixed

Use the data to guess the value of an unknown number

depends on the random sample

Create an **estimate** 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

(Demo)