Lecture 5

Iteration, Chance, Sampling

Slides adopted from UC Berkeley, data8 course

Announcement

The nba table contains 8 columns.
The first few rows are shown here ->

player	prefix	position	age	salary	games	minutes	points
Al Horford	BOS	C	30	2.65401e+07	68	2193	952
Amir Johnson	BOS	PF	29	1.2e+07	80	1608	520
Avery Bradley	BOS	SG	26	8.26966e+06	55	1835	894
Demetrius Jackson	BOS	PG	22	1.45e+06	5	17	10
Gerald Green	BOS	SF	31	1.4106e+06	47	538	262
Isaiah Thomas	BOS	PG	27	6.58713e+06	76	2569	2199
Jae Crowder	BOS	SF	26	6.28641e+06	72	2335	999
James Young	BOS	SG	21	1.8252e+06	29	220	68
Jaylen Brown	BOS	SF	20	4.743e+06	78	1341	515
Jonas Jerebko	BOS	PF	29	5e+06	78	1232	299

The age of the oldest NBA player.

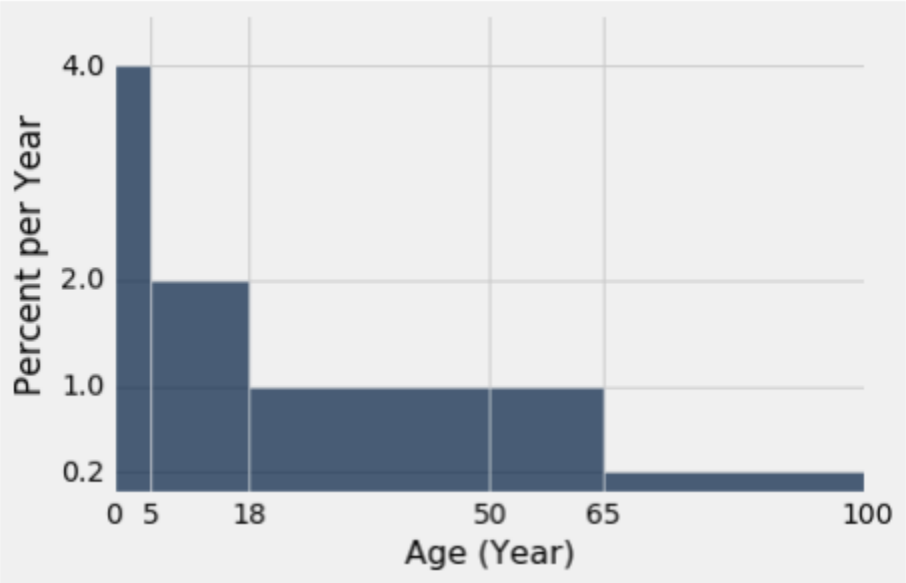
_____(nba._____('age')) max(nba.column('age'))

The three-letter prefix of the team which has the highest paid player with the position center (C) in the NBA. You may assume there is only one such player.

centers = nba._____('position', _____) centers = nba.where('position', are.equal_to('C'))

centers._____(_____).column(_____).item(0)

centers.sort('salary', descending=True).column('player').item(0)



The percent of people in the 0-5 bin is two times the percent of people in the 5-18 bin." This quoted statement is:

(i) True because ----------

(ii) False because ----------

(ii) False because the area of the 0-5 bar is less than the area of the 5-18 bar

Comparison

Comparison Operators

The result of a comparison expression is a **`bool`** value

```
x = 2                    y = 3
```

Assignment statements

```
x > 1                    x > y                    y >= 3

x == y                   x != 2                   2 < x < 5
```

Comparison expressions

(Demo)

Combining Comparisons

Boolean operators can be applied to **`bool`** values

```
a = True      b = False
```

Evaluate to True

```
not b          a or b          a and not b
```

```
a and b        not (a or b)      b and b
```

Evaluate to False

(Demo)

Predicates Make Comparisons

The result of calling a predicate function, such as `are.above(3)`, is a function that performs a comparison

(Demo)

Random Selection

Random Selection

`np.random.choice`

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

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

(Demo)

Discussion Question

```
d = np.arange(6) + 1
```

What results from evaluating the following 2 expressions? Are they the same? Do they describe the same process?

```
np.random.choice(d, 1000) + np.random.choice(d, 1000)
```

```
2 * np.random.choice(d, 1000)
```

Control Statements

Control Statements

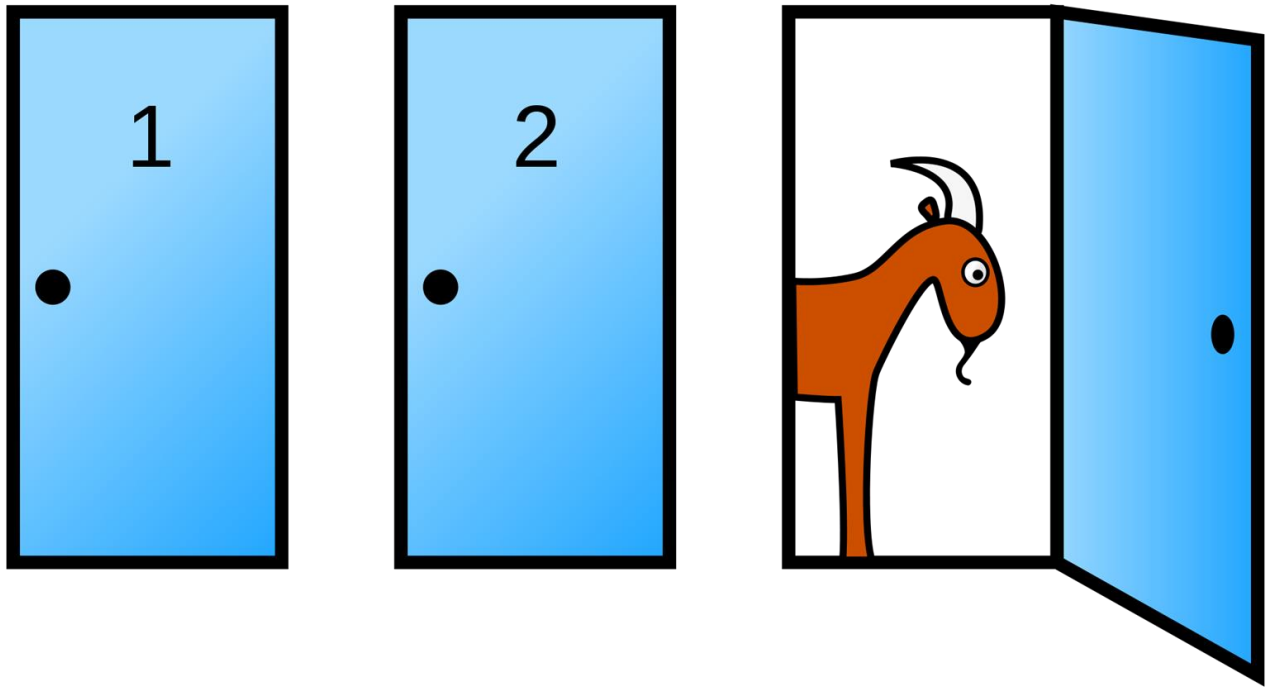
These statements *control* the sequence of computations that are performed in a program

- The keywords **`if`** and **`for`** begin control statements
- The purpose of **`if`** is to define functions that choose different behavior based on their arguments
- The purpose of **`for`** is to perform a computation for every element in a list or array

(Demo)

The Monty Hall Problem

Monty Hall Problem



Team up and DIY!

- Time: 20 min
- Additional 10% grade on your project grade!
- Expected output --------->

trial	guess	exposed	remaining
1	second goat	first goat	car
2	second goat	first goat	car
3	second goat	first goat	car
4	second goat	first goat	car
5	second goat	first goat	car
6	first goat	second goat	car
7	second goat	first goat	car
8	first goat	second goat	car
9	second goat	first goat	car
10	car	second goat	first goat

Probability

Probability

- 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

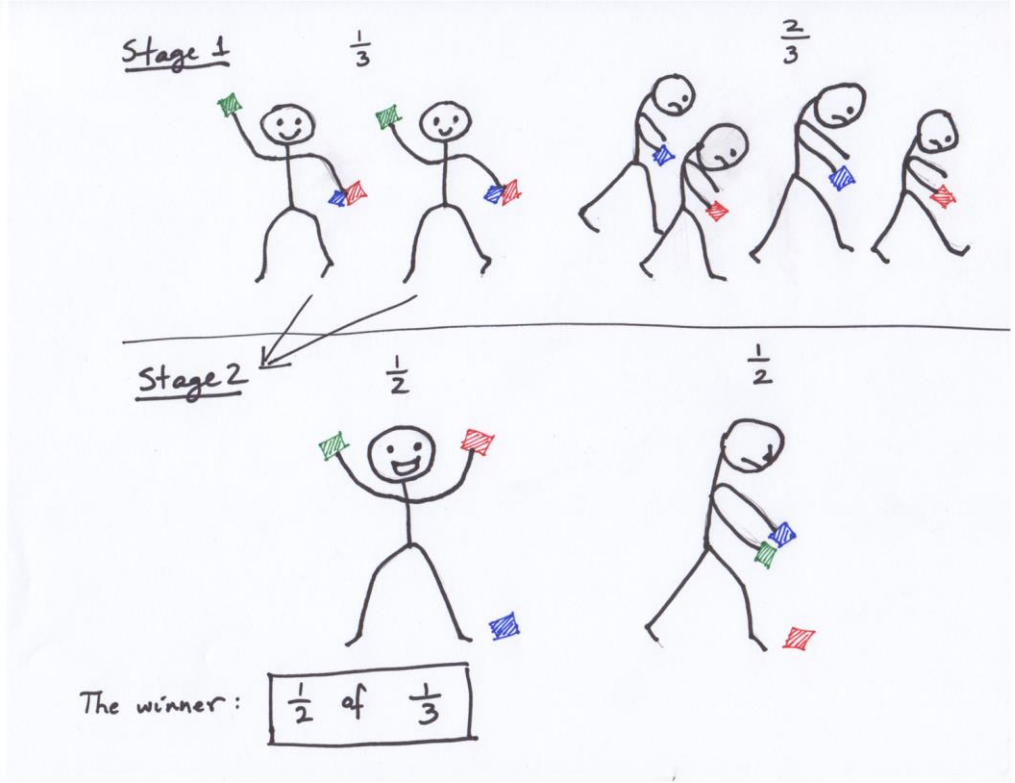
Equally Likely Outcomes

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

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

(Demo)

Fraction of a Fraction



(Demo)

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

(Demo)

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) = (½) x (½) x (½) = ⅛
  - P(at least one head) = 1 - P(TTT) = ⅞ = 87.5%

- In 10 tosses:
  - 1 - (½)**10
  - 99.9%

(Demo)

Sampling

Deterministic versus Probabilistic

- **Deterministic**: All data is known beforehand
  - Once you start the system, you know exactly what is going to happen.
  - *Example*: Predicting the amount of money in a bank account.
- **Probabilistic**: Element of chance is involved
  - You know the likelihood that something will happen, but you don't know when it will happen.
  - *Example:* Roll a die until it comes up '5'.

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

(Demo)

Sample of Convenience

- Type of non-probability sampling that involves the sample being drawn from that part of the population that is close to hand.
- 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
  - 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
- In some cases, the probability distribution can be worked out mathematically without ever generating (or simulating) the random quantity

(Demo)

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

(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

Large Random Samples

If the sample size is large,

then the empirical distribution of a uniform random sample

resembles the distribution of the population,

with high probability