Computer Science

CSC380: Principles of Data Science

Probability 2

Xinchen Yu

Announcements

HW1 has been out

- D2L -> Content
- Due next Friday, Jan 26 by 11:59 pm

Participation policy (5 points)

- Each office hour: + 1 point
- Answering question in the lecture: + 1point
- Answering question on Piazza: + 1 point
- Asking question (related to course materials) on Piazza: +0.5 point

Note: It is your responsibility to ensure the TA or instructor enter your participation points on gradescope during the office hour or after each lecture. Instructors will not award you these points at a later date, do not email instructors about getting points at a later date (for example, if you forget to ask the TA to enter your office hour points on gradescope).

Review

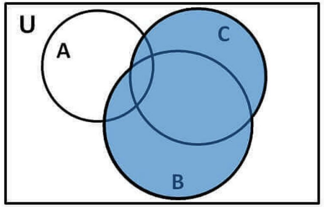
- What is probability?
- Axioms
- Event = set $\Rightarrow$ use set theory!
- Set theory + axiom 3 is quite useful
- Draw diagrams
- Lots of jargons

- Make your own cheatsheet.

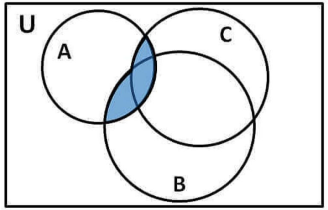
Review

- $A \cap (B \cup C) = (A \cap \mathrm{B}) \cup (A \cap \mathrm{C})$

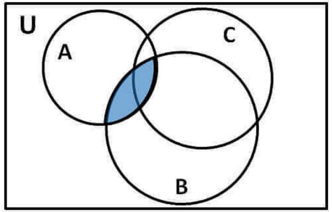
distributive law by Venn diagram



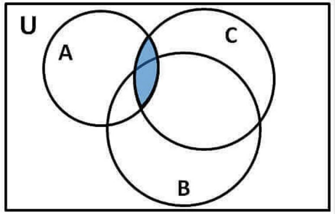
(B ∪ C)



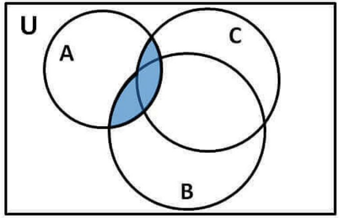
A ∩ (B ∪ C)



(A ∩ B)

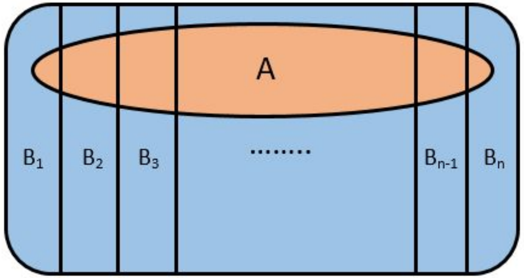


(A ∩ B)



(A ∩ B) ∪ (A ∩ C)

Review



- $A \cap (B \cup C) = (A \cap \mathrm{B}) \cup (A \cap \mathrm{C})$
- $A = A \cap \Omega = A \cap (\cup_i B_i) = \cup_i (A \cap B_i)$
  $= \mathrm{A} \cap (B_1 \cup B_2 \cup B_3 \ldots \cup B_n)$
  $= (\mathrm{A} \cap B_1) \cup (\mathrm{A} \cap B_2) \cup (\mathrm{A} \cap B_3) \ldots \cup (\mathrm{A} \cap B_n)$

<u>Law of total probability</u>*:* Let A be an event. For any events $\mathrm{B}_1, \mathrm{B}_2, \ldots$ that partitions Ω, we have

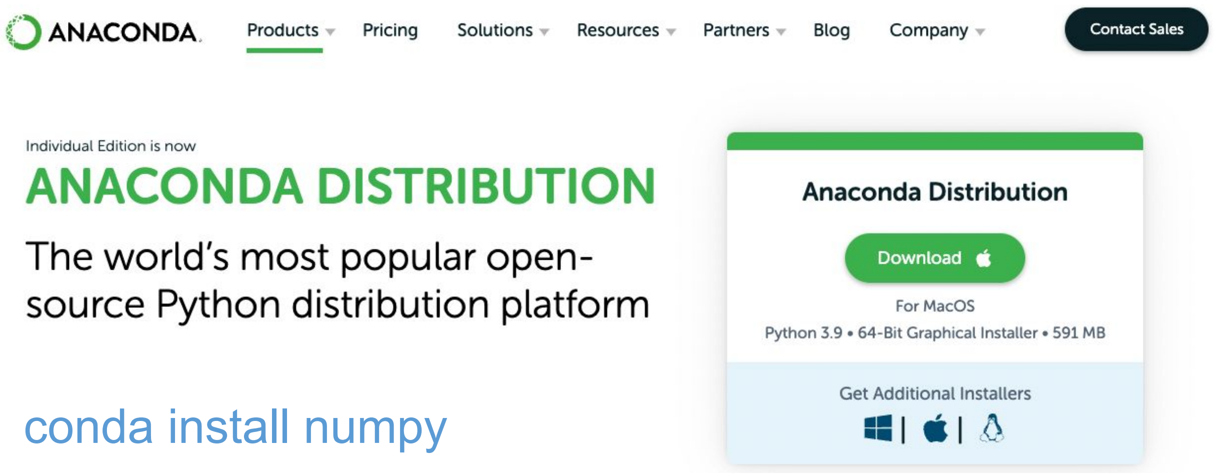
$$P(A) = \sum_i P(A \cap B_i)$$

Overview

- Conditional probability
- Independence

Numpy Library

Package containing many useful numerical functions…



conda install numpy

If you use pip:
pip install numpy

…we are interested in numpy.random at the moment

numpy.random

numpy.random.randint

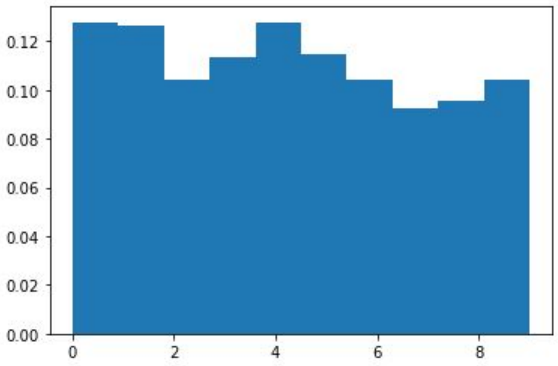
numpy.random.randint(*low*, *high=None*, *size=None*, *dtype='l'*)

Return random integers from *low* (inclusive) to *high* (exclusive).

Return random integers from the "discrete uniform" distribution of the specified dtype in the "half-open" interval [*low*, *high*). If *high* is None (the default), then results are from [0, *low*).

Sample a discrete uniform random variable,

```
import matplotlib.pyplot as plt
X = np.random.randint(0,10,1000)
count, bins, ignored = plt.hist(X, 10, density=True)
plt.show()
```



- **Caution** Interval is [low,high) and upper bound is **exclusive**
- `Size` argument accepts tuples for sampling ndarrays (multidimentional arrays)

numpy.random

Allows sampling from many common distributions

Set (global) random seed as,

```
import numpy as np

seed = 12345
np.random.seed(seed)
```

- ☺ easier to debug (otherwise, you may have ‘stochastic’ bug)
- ☹️ can be risky

  E.g., buy into the result based on a particular seed, publish a report.
  … turns out, you get a widely different result if you use a different seed!

  Recommendation: change the seed every now and then

Conditional Probability

Random Events and Probability

Consider: What is the probability of having two numbers sum to 6?

```
import numpy as np
for n in [10,100,1_000,10_000,100_000]:
    res_dice1 = np.random.randint(1,6+1,size=n)
    res_dice2 = np.random.randint(1,6+1,size=n)
    res = [(res_dice1[i], res_dice2[i]) for i in range(len(res_dice1))]

    cnt = len(list(filter(lambda x: x[0] + x[1] == 6, res)))
    print("n=%6d, result: %.4f " % (n, cnt/n))
```

```
n=      10, result: 0.1000
n=     100, result: 0.1200
n=    1000, result: 0.1350
n=   10000, result: 0.1365
n=  100000, result: 0.1388
n= 1000000, result: 0.1385
```

```
n=      10, result: 0.1000
n=     100, result: 0.1900
n=    1000, result: 0.1540
n=   10000, result: 0.1366
n=  100000, result: 0.1371
n= 1000000, result: 0.1394
```

every time you run, you get a different result

however, the number seems to **converge** to 0.138-0.139

There seems to be a precise value that it will converge to.. what is it?

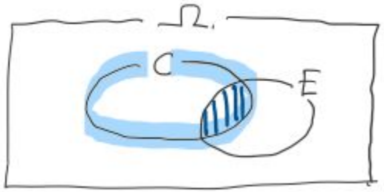
Conditional Probability

- Suppose I roll two dice secretly and tell you that <u>one of the dice is 2</u>. C
- **<u>In this situation</u>**, find the probability of two dice summing to 6. E

```
import numpy as np
for n in [10,100,1000,10_000,100_000, 1_000_000]:
    res_dice1 = np.random.randint(6,size=n) + 1
    res_dice2 = np.random.randint(6,size=n) + 1
    res = [(res_dice1[i], res_dice2[i]) for i in range(len(res_dice1))]

    conditioned = list(filter(lambda x: x[0] == 2 or x[1] == 2, res))
    n_eff = len(conditioned)

    cnt = len(list(filter(lambda x: x[0] + x[1] == 6, conditioned)))
    print("n=%9d, n_eff=%9d, result: %.4f " % (n, n_eff, cnt/n_eff))
```



Without conditioning, it was 0.138.

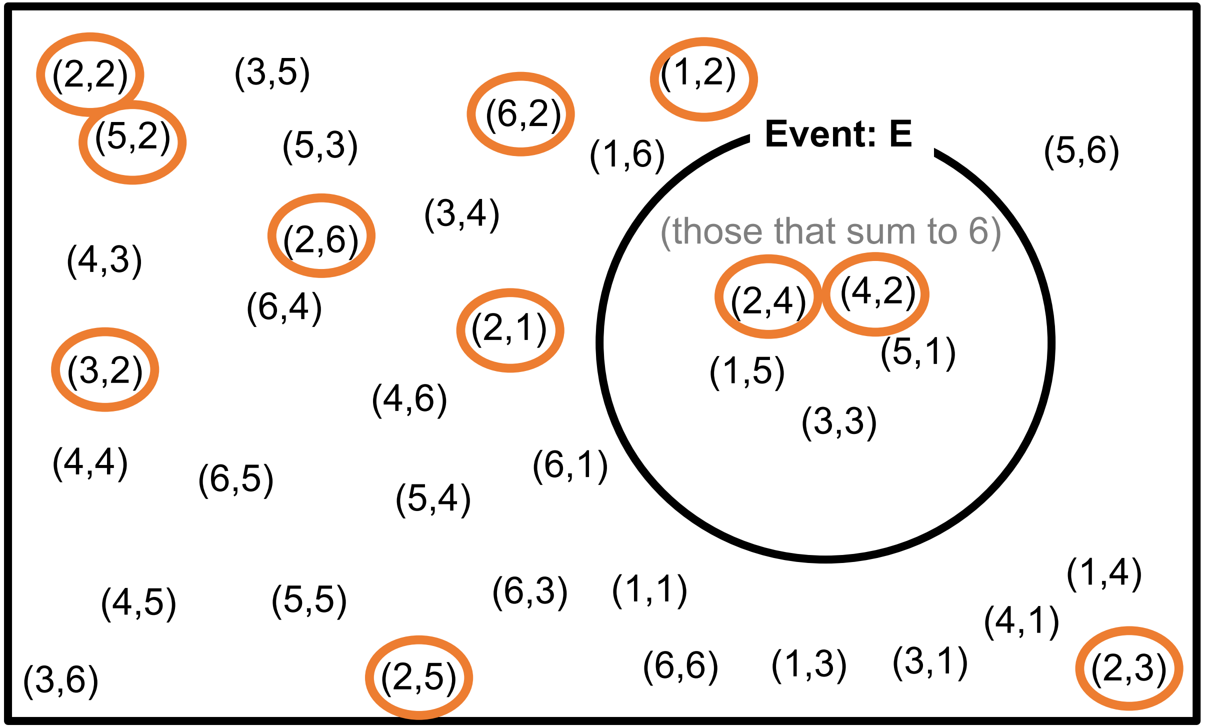
```
n=       10, n_eff=        4, result: 0.0000
n=      100, n_eff=       32, result: 0.2500
n=     1000, n_eff=      300, result: 0.1733
n=    10000, n_eff=     3002, result: 0.1742
n=   100000, n_eff=    30590, result: 0.1823
n=  1000000, n_eff=   305616, result: 0.1818
```

```
n=       10, n_eff=        3, result: 0.3333
n=      100, n_eff=       32, result: 0.0625
n=     1000, n_eff=      343, result: 0.2245
n=    10000, n_eff=     3062, result: 0.1897
n=   100000, n_eff=    30651, result: 0.1811
n=  1000000, n_eff=   305580, result: 0.1808
```

There seems to be a precise value that it will converge to.. what is it?

Random Events and Probability

What is the probability of having two numbers sum to 6 given one of dice is 2?



Each outcome is equally likely.
by the **independence**
(will learn this concept later)
=> 1/36

sum to 6:
=> 5

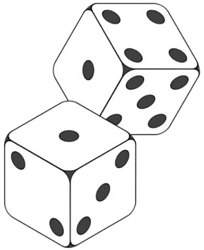
one of dice is 2:
=> 11

sum to 6 and one of dice is 2:
=> 2

answer:
2/11 = 0.181818….

Conditional Probability

Two fair dice example

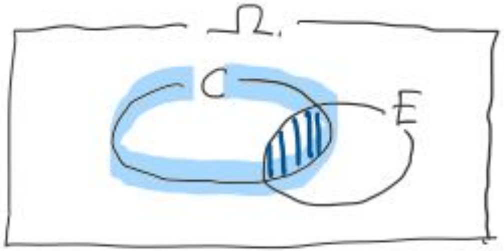


- Find the probability of one of the dice is 2 (event C) and two dice summing to 6 (E)

$$P(E \cap C)$$

- I secretly tell you one of the dice is 2, find the probability of two dice summing to 6.

$$\frac{P(E \cap C)}{P(C)}$$



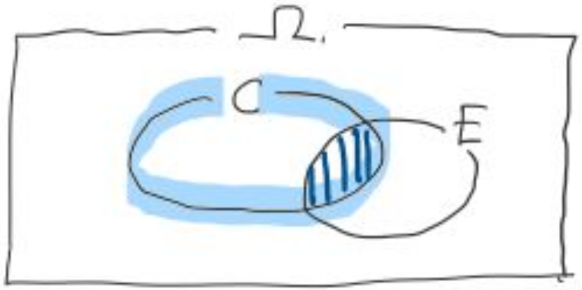
Conditional Probability

- Two fair dice example:
  - Suppose I roll two dice and secretly tell you that one of the dice is 2. C
  - **In this situation**, find the probability of two dice summing to 6. E

- Turns out, such a probability can be computed by $\frac{P(E \cap C)}{P(C)}$
- It's like "zooming in" to the condition.
- This happens a lot in practice, so let's give it a notation:

$$P(E|C) := \frac{P(E \cap C)}{P(C)}$$

Say: probability of "E given C", "E conditioned on C"



"it's the ratio"

Conditional Probability

Q: Conditional probability P(A|B) could be undefined. When?

- A: The denominator can be 0 already. In this case, numerator is also 0!

Note $P(A|B) \neq P(B|A)$ in general!

$$P(A|B) \coloneqq \frac{P(A \cap B)}{P(B)}$$

E.g., throw a fair die. X := outcome. A = {X=4}, B = {X is even}

Question: P(A | B) = P(B | A) ?

- $\mathrm{P(A)} = 1/6$
- $\mathrm{P(B)} = 1/2$
- $\mathrm{P\ (A \cap B)} = 1/6$
- Therefore, $\mathrm{P(A|B)} = 1/3$, $\mathrm{P(B|A)} = 1$

Conditional Probability

Chain rule

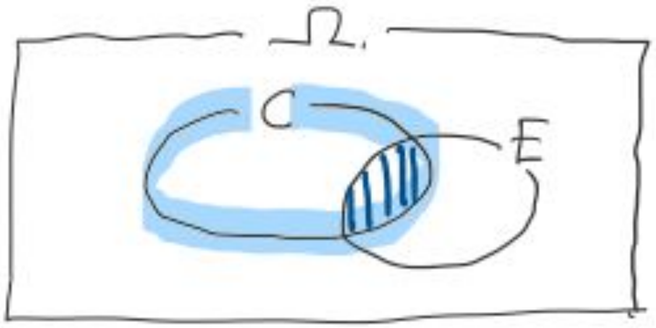
- $P(A \cap B) = P(A|B)P(B)$ ←just a rearrangement of definition: $P(A|B) \coloneqq \frac{P(A \cap B)}{P(B)}$

- $P(A \cap B \cap C) = P(A|B \cap C)P(B \cap C) = P(A|B \cap C)P(B|C)P(C)$

- $P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1) \prod_{i=2}^{n} P(E_i | \cap_{j=1}^{i-1} E_j)$ valid for any ordering!

Conditional Probability

- $P(E \cap C) = P(E|C)P(C) = P(C|E)P(E)$

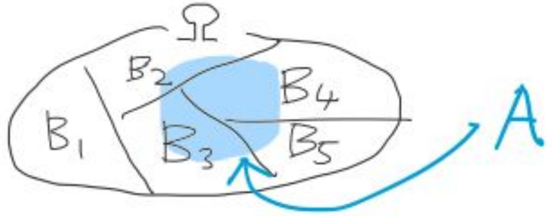


"it's the ratio"

Conditional Probability

Recall: let A be an event. For events $\mathrm{B}_1, \mathrm{B}_2, \ldots$ that partitions Ω, we have

$$P(A) = \sum_i P(A \cap B_i)$$



Law of total probability: If $A \in \mathcal{F}$ and $\{B_i \in \mathcal{F}\}_i$ partitions Ω, then

$$P(A) = \sum_i P(A, B_i) = \sum_i P(B_i) P(A|B_i)$$

$$= \sum_i P(A) P(B_i|A) \qquad \text{(by definition)}$$

Shortcut:
P(A,B) := P(A ∩ B)

Conditional Probability

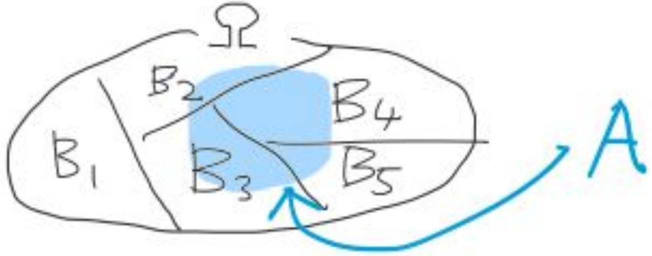
Law of total probability: If $A \in \mathcal{F}$ and $\{B_i \in \mathcal{F}\}_i$ partitions Ω, then

$$P(A) = \sum_i P(A, B_i) = \sum_i P(B_i) P(A|B_i)$$

$$= \sum_i P(A) P(B_i|A) \qquad \text{(by definition)}$$

If we divide both sides by P(A):

$$1 = \sum_i P(B_i|A)$$



Conditional Probability: an example

$$P(A) = \sum_i P(A \cap B_i)$$

A: customer (100)

B: fill gas

- B_1: unleaded (30)
- B_2: mid grade (30)
- B_3: premium (40)

30, 1/3   30, 1/2   40, 1/8

Q: what's the probability that the customer is a student?

$$\mathrm{P(A = student)}$$
$$= \mathrm{P(A = student, B = B_1) + P(A = student, B = B_2) + P(A = student, B = B_3)}$$
$$= \mathrm{P(A = student | B = B_1) P(B = B_1) + P(A = student | B = B_2) P(B = B_2) +}$$
$$\mathrm{P(A = student | B = B_3)} P\mathrm{(B = B_3)}$$

Conditional Probability: an example

- $P(A) = \sum_i P(A, B_i) = \sum_i P(B_i)P(A|B_i)$

$$\mathrm{P(A = student)} = \mathrm{P(A = student|B = B_1)P(B = B_1) + P(A = student|B = B_2)P(B = B_2) + P(A = student|B = B_3)}P\mathrm{(B = B_3)}$$

$$\mathrm{P(A = student)} = 1/3\times30/100 + 1/2\times30/100 + 1/8\times40/100$$

- $\sum_i P(B_i|A) = 1$

$$\mathrm{P(B_1|A = student) + P(B_2|A = student) + P(B_3|A = student)} = \frac{10}{10+15+5} + \frac{15}{10+15+5} + \frac{5}{10+15+5} = 1$$

30, ⅓: 10   30, ½: 15   40, ⅛: 5

Conditional Probability

The Public Health Department gives us the following information:

- A test for the disease yields a positive result (+) 90% of the time when the disease is present (Y)

  $P(+ \mid Y) = 0.9$

- A test for the disease yields a positive result 1% of the time when the disease is not present (N)

  $P(+ \mid N) = 0.01$

- One person in 1,000 has the disease.

  $P(Y) = 0.001$

Q: What is the probability that a person with positive test has the disease? $P(Y \mid +)$?

Pick a person uniformly at random from the population. Apply the test. When test=+, what is the probability of this person having the disease (Y) ?

Conditional Probability

What we know:

P(+ | Y) = 0.9
P(+ | N) = 0.01
P(Y) = 0.001

$\Rightarrow$

P(- | Y) = 0.1
P(- | N) = 0.99
P(N) = 0.999

Question: P(Y | +) $= \dfrac{P(Y,+)}{P(+)}$

$$P(+) = P(+,Y) + P(+,N)$$

$$P(+,Y) = P(+|\,Y)P(Y)$$

$$P(+,N) = P(+|\,N)P(N)$$

Law of total probability

$$P(A) = \sum_i P(A,B_i) = \sum_i P(B_i)P(A|B_i)$$

The answer is 0.0826…

Terminology

When we have two events A and B…

- Conditional probability: $P(A|B), P(A^c|B), P(B|A)$ etc.

- Joint probability: $P(A, B)$ or $P(A^c, B)$ or ...

- Marginal probability: $P(A)$ or $P(A^c)$

Conditional Probability

$P(+ \mid Y) = 0.9$
$P(+ \mid N) = 0.01$
$P(Y) = 0.001$

Tip: Make a table of **joint probabilities**

Each cell is P(column event ∩ row event) = P(T=t ∩ D=d) = P(T= t | D= d) P(D=d)

	Test = +	Test = -	
Disease=Y	$0.9 \cdot 0.001 = 0.0009$	$0.1 \cdot 0.001 = 0.0001$	0.001
Disease=N	$0.01 \cdot 0.999 = 0.00999$	$0.99 \cdot 0.999 = 0.98901$	0.999
	0.01089	0.98911	

$P(test = +)$

Workflow:

- make a table, then fill in the cells.
- write down the target P(A|B) all in terms of joint probabilities and marginal probabilities.

Conditional Probability

We can directly calculate:

$$P(Y|+) = \frac{P(Y,+)}{P(+)} = \frac{P(+|Y)P(Y)}{P(+)}$$

Bayes rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

proof: definition and definition!

$\Longrightarrow$ particularly useful in practice: infer P(A|B) given P(B|A)!

P(A): **prior** probability
P(A|B): **posterior** probability

e.g., A='dice sum to 6', B='one of the die is 2'
e.g., A='disease=Y', B='test=+'