Computer Science

CSC380: Principles of Data Science

Probability Primer

Chicheng Zhang

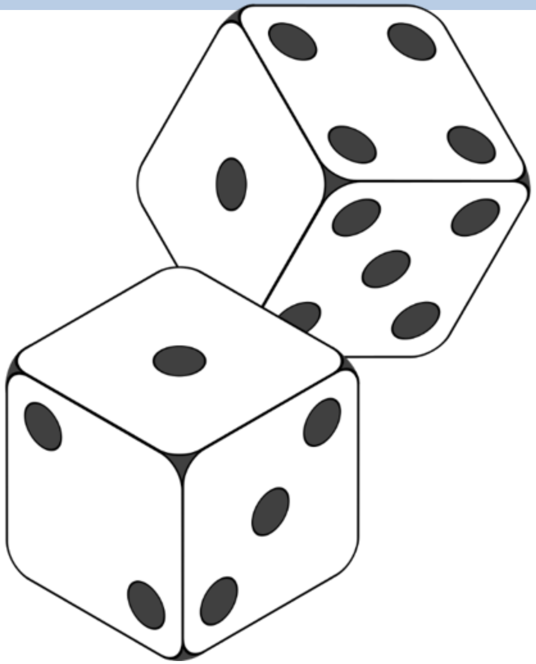
Acknowledgement: Built on Jason Pacheco, Kwang-Sung Jun's slides

Administrative Items

- **My office hours:** Tuesdays 3:30-4:30pm, Gould-Simpson 720 (in person, before Feb 28)

- Homework 1
  - Will be out Thursday Jan 19
  - Due Friday Jan 27, 11:59pm
  - Reminder: you are allowed to work individually or in pairs (see syllabus for detailed policy)

Random Events and Probability



Suppose we roll <u>two fair dice</u>…

- What are the possible outcomes?
- What is the *chance* of rolling two **even** numbers?
- What is the *chance* of having two numbers sum to 6?
- *Given the observation that one die rolls 1*, what is the chance of the second die also rolling 1?

*…this is a **random process**.*

How to mathematically formulate outcomes and characterize how likely they are?

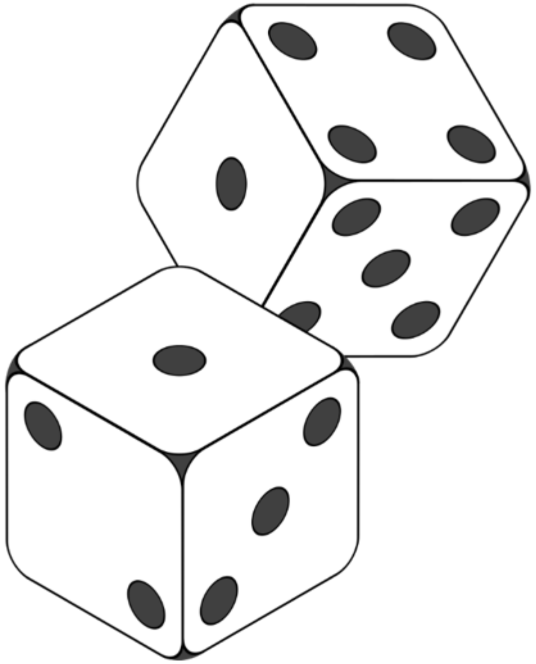
Probability: intuition

- Probability of a random event

  $\approx$

  Simulate the random process n times, the fraction of times this event happens

- How large should n be?
- Simulation results vary from trials?



Background: Numpy in Python

Numpy: numerical computing package

```
import numpy as np
np.random.randint(1,1+6,size=10)
=> array([5, 4, 1, 1, 1, 5, 5, 2, 4, 6])
```

randint(low,high,size)
: generate `size` random numbers in {low, low+1, …, high-1}

Numpy array

- Replaces python's **list** in numpy.
- More numerical functionality
- It's a 'vector' in mathematics.

```
a=np.array([1,2]); b=np.array([4,5])
a+b
⇒ np.array([5,7]) // elementwise addition
a @ b
⇒ 14          // inner product
```

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
```

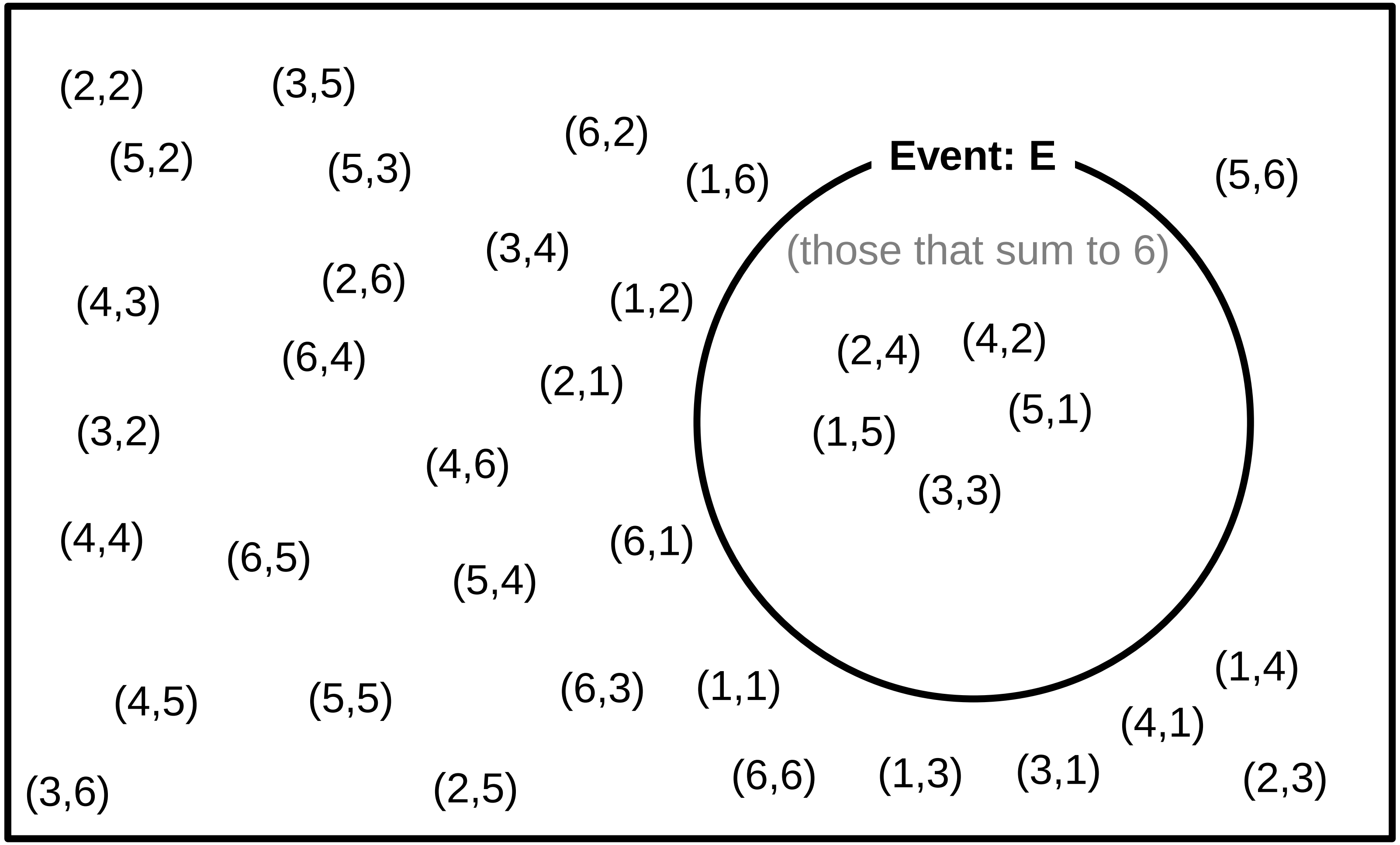
every time you run, you get a different result

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?

Random Events and Probability

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



Each outcome is equally likely by **independence**
(will learn this concept later)
And thus all have probability:
=> 1/36

# of outcomes that sum to 6:
=> 5

answer:
(1/36) * 5 = 0.13888..

Mathematics of Probability

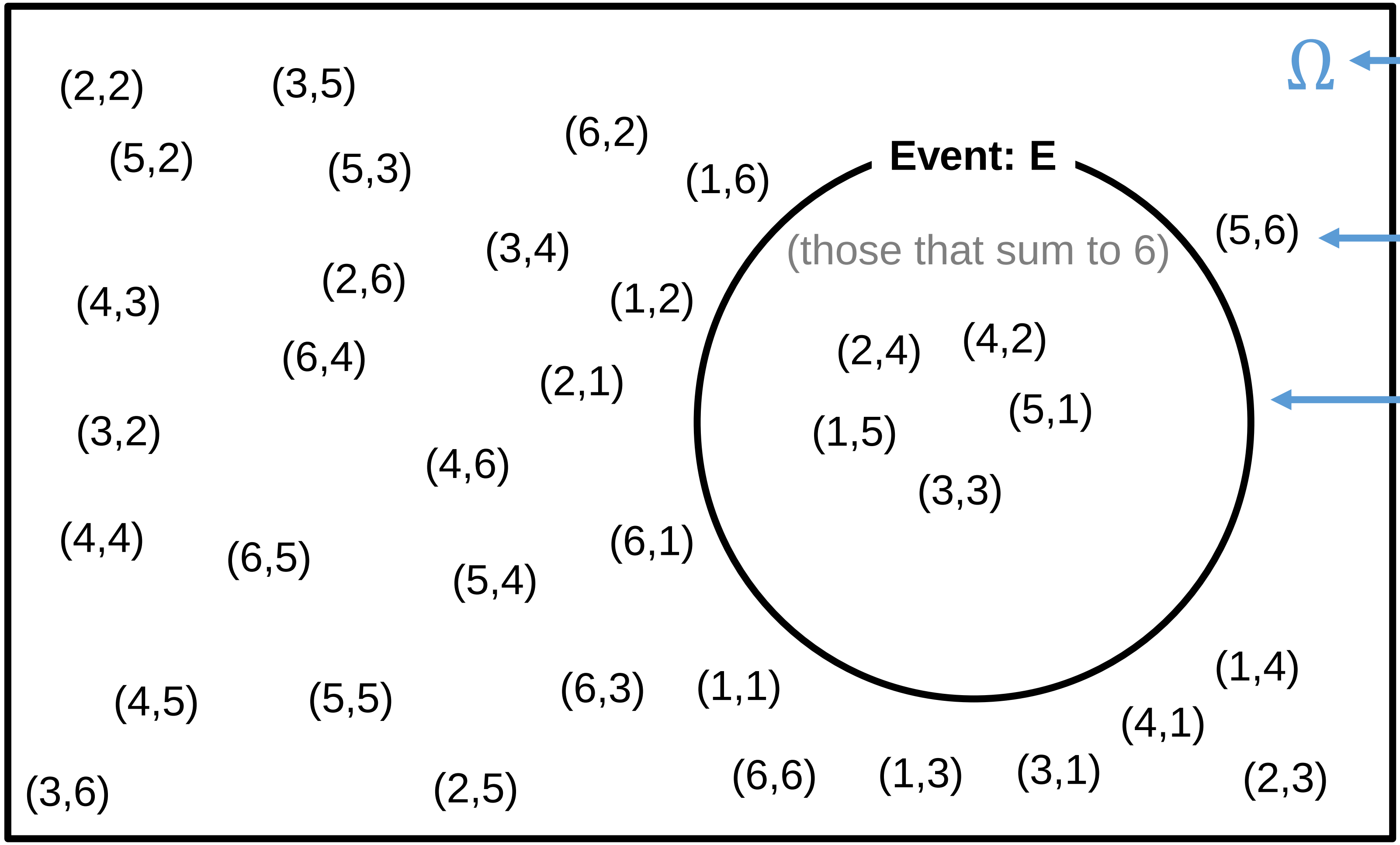
- **Probability** is a real-world phenomenon.

- But under what mathematical framework can we formulate **probability** so we can solve practical problems?
  - e.g., weather prediction, predicting the election outcome

- <u>**Disclaimer**</u>: not all mathematics correspond to real-world phenomenon (e.g., Banach–Tarski paradox). Fortunately, we will not talk about this in our lecture ☺

Random Events and Probability

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



Outcome space consists of all possible outcomes

Each **outcome** $\omega \in \Omega$.

Event is a collection of outcomes and a subset of the outcome space

$$E \subseteq \Omega$$

Note: outcome = sample

Random Events and Probability

Some examples of events…

Q: how many such pairs?

9

- Both even numbers

$$E^{\text{even}} = \{(2,2),(2,4),\ldots,(6,4),(6,6)\}$$

- The sum of both dice is even,

$$E^{\text{sum even}} = \{(1,1),(1,3),(1,5),\ldots,(2,2),(2,4),\ldots\}$$

- The sum is greater than 12,

$$E^{\text{sum}>12} = \emptyset$$

We can talk about impossible outcomes

Random Events and Probability

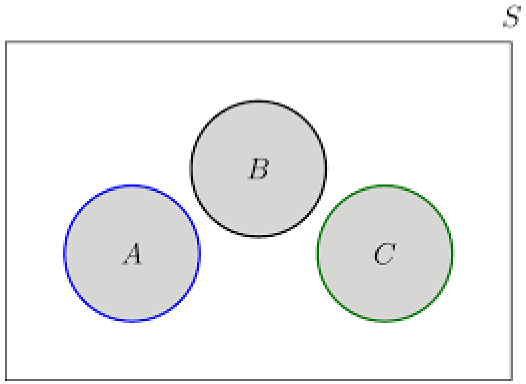
But, what is probability, really?

(e.g., can explain the probability of seeing an event when throwing two dice)

Mathematicians have found a set of conditions that 'makes sense'.

- Probability is a map P. ⇒ i.e., takes in an event, spits out a real value
- P must map events to a real value in interval [0,1].
- P is a (valid) **probability distribution** if it satisfies the following **axioms of probability**,
  1. For any event E, $P(\mathrm{E}) \geq 0$
  2. $P(\Omega) = 1$
  3. For any *finite* or *countably infinite* sequence of disjoint events $E_1, E_2, E_3, \ldots$

$$P\left(\bigcup_{i \geq 1} E_i\right) = \sum_{i \geq 1} P(E_i)$$



disjoint: intersection is empty

https://www.probabilitycourse.com/chapter1/1_2_2_set_operations.php

Background: Countable vs Uncountable

- two kinds of infinite sets
  - **countably infinite**: the kind of set that you can ”enumerate” the elements. For example, the set of integers = {0, -1, 1, -2, 2, …}
  - **uncountably infinite**: the kind of set that (provably) you cannot ”enumerate” the elements. E.g., the set of all real numbers

- “**enumerable**” is perhaps more intuitive than “countable”, but countable is more common.

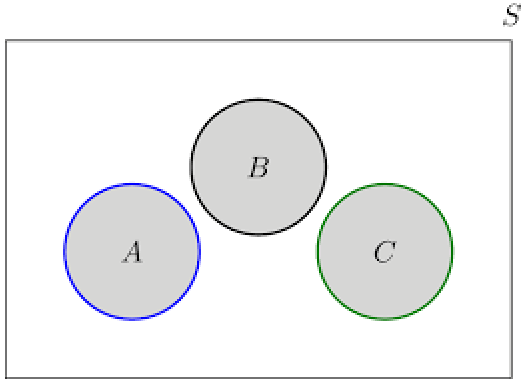
Random Events and Probability

- Many properties follows (i.e., can be proved mathematically)

$$
\begin{aligned}
\mathbb{P}(\emptyset) &= 0 \\
A \subset B &\implies \mathbb{P}(A) \leq \mathbb{P}(B) \\
0 \leq \mathbb{P}(A) &\leq 1 \\
\mathbb{P}(A^c) &= 1 - \mathbb{P}(A) \\
A \bigcap B = \emptyset &\implies \mathbb{P}\left(A \bigcup B\right) = \mathbb{P}(A) + \mathbb{P}(B).
\end{aligned}
$$

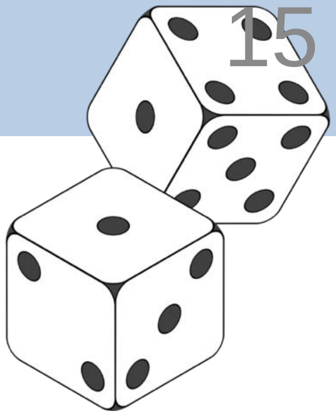
E.g., throw a die. A= getting 1, B=getting an odd number

E.g., A= getting 1, B=getting 3 or 5



(I recommend that you maintain your own version of cheat sheet!)

Random Events and Probability



Special case

If each outcome is equally likely, and sample space is finite, then the probability of event is:

$$P(E) = \frac{|E|}{|\Omega|}$$

Number of elements in event set

Number of possible outcomes (36)

This is called uniform probability distribution

Q: What axiom we are using?
=> Axiom 3

(Fair) Dice Example: Probability that we roll 2 even numbers,

$$P(\{(2,2),(2,4),\dots,(6,6)\}) = P(\{(2,2)\}) + P(\{2,4\}) + \cdots + P(\{6,6\})$$
$$= \frac{1}{36} + \cdots + \frac{1}{36} = \frac{9}{36}$$

9 Possible outcomes, each with equal probability of occurring

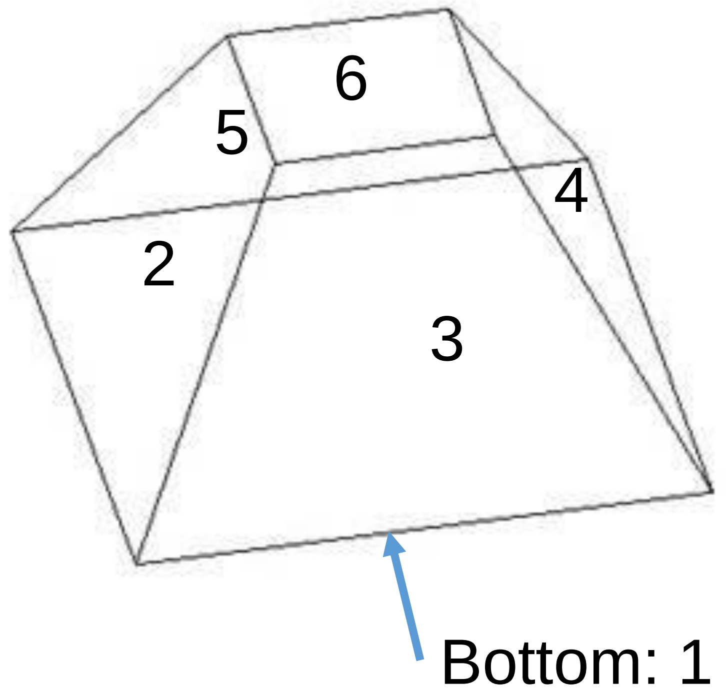
Unfair die example

- Let A be the outcome of a single throw.
- P(A=1) << P(A=2) = … = P(A=5) << P(A=6)

e.g., 0.1 0.15 0.15. 0.3

- Probabilities of throwing two of these dice are not easy to compute anymore!

- Will come back to this later.



Set Theory

Two dice example: *Suppose*

E_1 : *First* die equals 1 E_2 : *Second* die equals 1

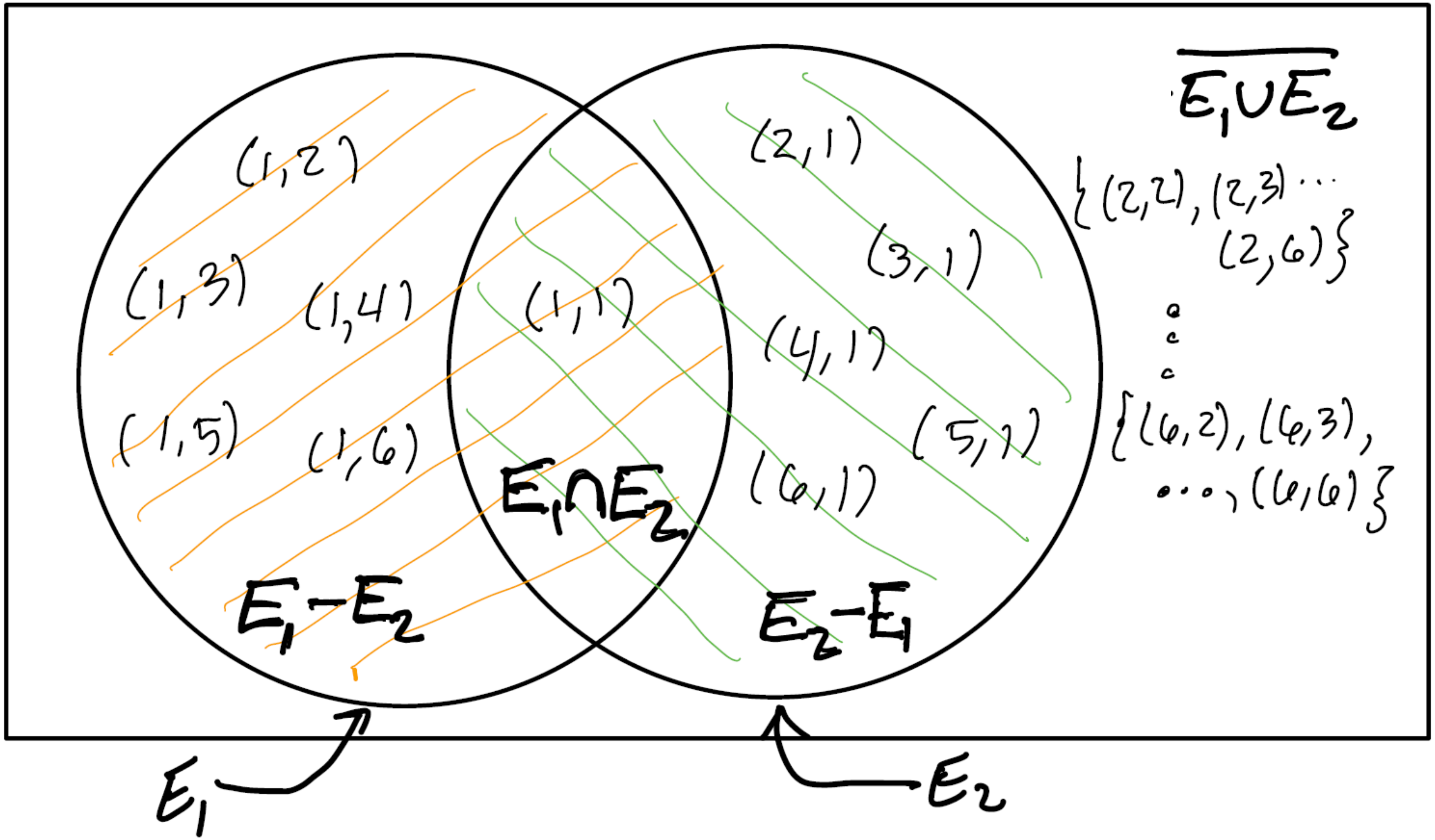
$$E_1 = \{(1,1),(1,2),\ldots,(1,6)\} \qquad E_2 = \{(1,1),(2,1),\ldots,(6,1)\}$$

Operators on events:

Operation	Value	Interpretation
$E_1 \cup E_2$	$\{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,1)\}$	Any die rolls 1
$E_1 \cap E_2$	$\{(1,1)\}$	Both dice roll 1
$E_1 \setminus E_2$ $\boxed{(= E_1 - E_2 \coloneqq E_1 \cap E_2^c)}$	$\{(1,2),(1,3),(1,4),(1,5),(1,6)\}$	Only the first die rolls 1
$\overline{E_1 \cup E_2}$ $\boxed{(= (E_1 \cup E_2)^c)}$	$\{(2,2),(2,3),\ldots,(2,6),(3,2),\ldots,(6,6)\}$	No die rolls 1

Set Theory

Can interpret these operations as a Venn diagram…



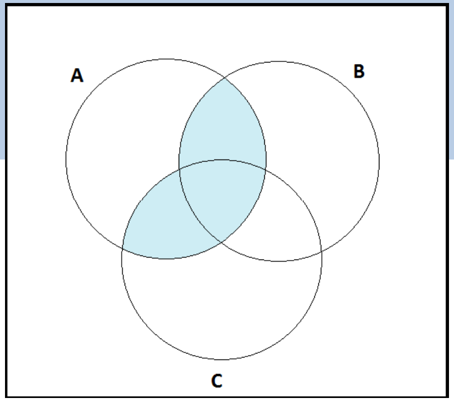
Set Theory

- Set theory vs probability theory (Watkins book Sec 5.6)

Event Language	Set Language	Set Notation
sample space	universal set	Ω
event	subset	$A, B, C, \cdots$
outcome	element	ω
impossible event	empty set	$\emptyset$

Event Language	Set Language	Set Notation
not A	A complement	A^c
A or B	A union B	$A \cup B$
A and B	A intersect B	$A \cap B$
A and B are mutually exclusive	A and B are disjoint	$A \cap B = \emptyset$
if A then B	A is a subset of B	$A \subset B$

Set Theory



More results

- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$. // distributive law

  $A \cap (\cup_i B_i) = \cup_i (A \cap B_i)$, $A \cup (\cap_i B_i) = \cap_i (A \cup B_i)$

- $\neg(\bigcup_n A_n) = \bigcap_n \neg A_n$, $\neg(\bigcap_n A_n) = \bigcup_n \neg A_n$ DEMORGAN

  Notation: $\neg A := A^c$

  Special case: $\neg(A \cup B) = \neg A \cap \neg B$

- $B = \Omega \cap B = (A \cup \neg A) \cap B = (A \cap B) \cup (\neg A \cap B)$ // by distributive law

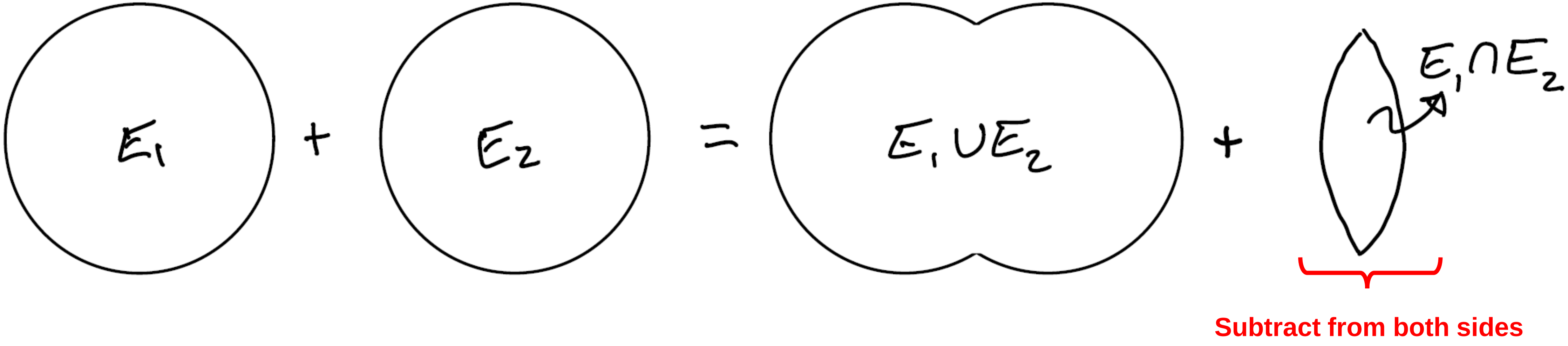
TIP: always draw pictures to visualize these identities!

For more, see https://math.libretexts.org/Courses/Mount_Royal_University/MATH_1150%3A_Mathematical_Reasoning/2%3A_Basic_Concepts_of_Sets/2.3%3A_Properties_of_Sets

Lemma: (inclusion-exclusion rule) *For any two events* E_1 *and* E_2,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Graphical Proof:

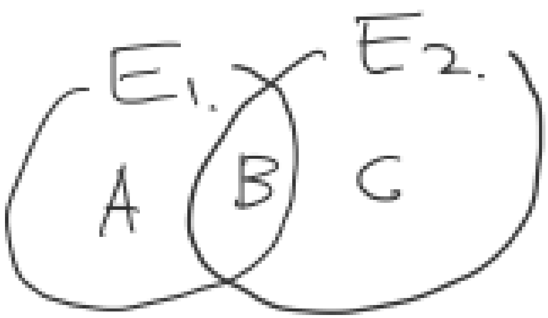


Alternative Proof

Lemma: (inclusion-exclusion rule) *For <u>any</u> two events* E_1 *and* E_2,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Formal proof:



$A = E_1 - (E_1 \cap E_2)$

$B = E_1 \cap E_2$

$C = E_2 - (E_1 \cap E_2)$.

$$\begin{aligned}
&P(E_1 \cup E_2) \\
&= P(A \cup B \cup C) \\
&= P(A) + P(B) + P(C) && \text{(by axiom 3)} \\
&= P(A) + P(B) + P(B) + P(C) - P(B) \\
&= P(A \cup B) + P(B \cup C) - P(B) && \text{(by axiom 3)}
\end{aligned}$$

Exercise

Quiz candidate

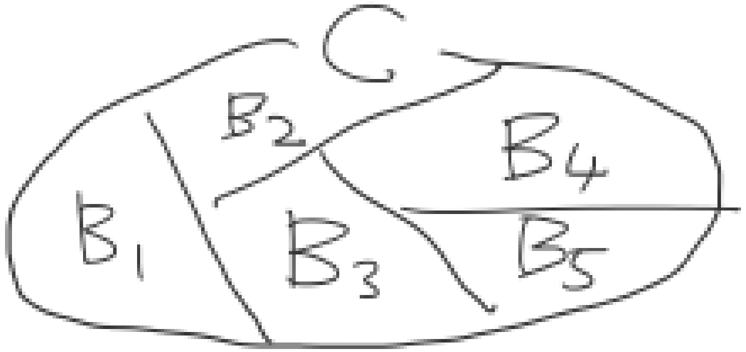
- Consider rolling two fair dice
- E_1: two dice sum to 6
- E_2: second die is even
- Compute the numerical value of $P(E_1 \cup E_2)$. Hint: Use inclusion-exclusion rule.

- $P(E_1) = 5/36$ $\quad (E_1 = \{(1,5), (2,4), \ldots, (5,1)\})$
- $P(E_2) = 3/6 = 1/2$
- $P(E_1 \cap E_2) = 2/36$ $\quad (E_1 \cap E_2 = \{(2,4), (4,2)\})$

answer: 21/36

Random Events and Probability

[Def] The set of events $\{B_i\}_{i=1}^n$ **partitions** $C \Leftrightarrow \cup_i B_i = C$ and $B_1, B_2, \ldots$ are disjoint. Here, n can be infinity.



<u>Law of total probability</u>*:* Let A be an event. For events $B_1, B_2, \ldots$ that partitions Ω, we have

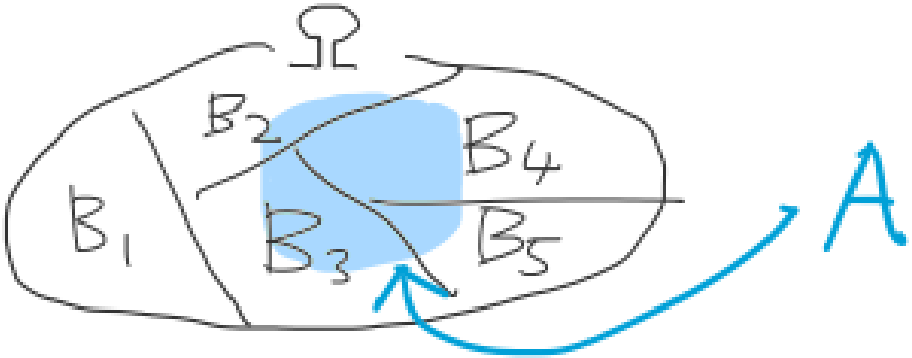
$$P(A) = \sum_i P(A \cap B_i)$$

Now, $\{A \cap B_i\}_{i=1}^n$ partitions A

Q: Why is this true?

A: Axiom 3!

$$A = A \cap \Omega = A \cap (\cup_i B_i) = \cup_i (A \cap B_i)$$



Random Events and Probability

<u>Law of total probability</u>*:* Let A be an event. For any events $B_1, B_2, \ldots$ that partitions Ω, we have

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

Example Roll two fair dice. Let X be the <u>outcome of the first die</u>. Let Y be the <u>sum of both dice</u>. What is the probability that both dice sum to 6 (i.e., Y=6)?

quiz candidate

$\mathrm{P}(A, B) := P(A \cap B)$

$$p(Y=6) = \sum_{x=1}^{6} p(Y=6, X=x)$$

$$= p(Y=6, X=1) + p(Y=6, X=2) + \ldots + p(Y=6, X=6)$$

$$= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + 0 = \frac{5}{36}$$

Summary So Far

- Most of the rules we learned is basically
  set theory + axiom 3

So, here is a generic workflow for computing P(A):

1. Use set theory and slice and dice A into a manageable partition of A where P(each piece of partition) is easy to compute.
2. Apply Axiom 3.

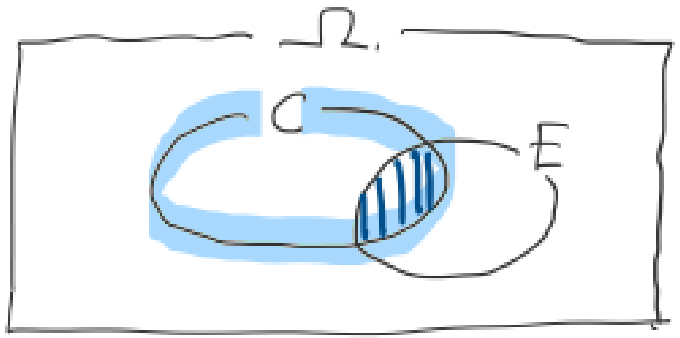
Next lecture: Conditional Probability

- Two fair dice example:
  - Suppose I roll two dice secretly and tell you that one of the dice is 2. C
  - **Given this situation**, 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))
```



compare:
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
```

More concise implementation


```
import numpy as np
for n in [10,100,1000,10_000,100_000, 1_000_000]:
    res = np.random.randint(1,1+6,size=(2,n))
    idx = (res[0,:] == 2) | (res[1,:] == 2)
    conditioned = res[:,idx]
    n_eff = conditioned.shape[1]

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

2 x n array

Length n, boolean array

2 x n_eff integer array

.shape returns `(#rows,#cols)`

Sum() sums up the boolean array

There is a quite a bit of tricks like this in numpy. You will get used to it over time!

CSC380: Principles of Data Science

Probability Primer 2

Administrative Item

- HW1 out; due on Jan 27
  - Recall combinations

5.4.3 Combinations

In the case that the order does not matter, a **combination** is a subset from a finite set. Write

$$\binom{n}{k}$$

(see the Watkins book)

HW01

Problem 2: Coinflips

Suppose we flip a fair coin 10 times. What is the probability that the following events occur: I recommend that you use the code like Problem 1 to debug your answers (but this debugging itself is not part of the evaluation).

a) The number of heads and the number of tails are equal

Recall:

Special case

Assume each outcome is equally likely, and sample space is finite, then the probability of event is:

$$P(E) = \frac{|E|}{|\Omega|}$$

Number of elements in event set

Number of possible outcomes

HW01

Problem 2: Coinflips

Suppose we flip a fair coin 10 times. What is the probability that the following events occur: I recommend that you use the code like Problem 1 to debug your answers (but this debugging itself is not part of the evaluation).

c) The number of heads and the number of tails are equal, but now with the assumption that the head probability is .2 (unfair coin).

Hint:

- Try what the result should look like, if you flip a fair coin twice.
- Code up the simulation to verify if your answer is right.

To simulate unfair coin:

numpy.random.rand() < 0.2 will be \`True\` with probability 0.2 and \`False’ w.p. 0.8

Note: Still, your answer will be correct only if you show your mathematical work.

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.

Numpy Library

Package containing many useful numerical functions…

NumPy

CONDA

If you use `conda`, you can install NumPy from the `defaults` or `conda-forge` channels:

```
# Best practice, use an environment rather than install in the base env
conda create -n my-env
conda activate my-env
# If you want to install from conda-forge
conda config --env --add channels conda-forge
# The actual install command
conda install numpy
```

PIP

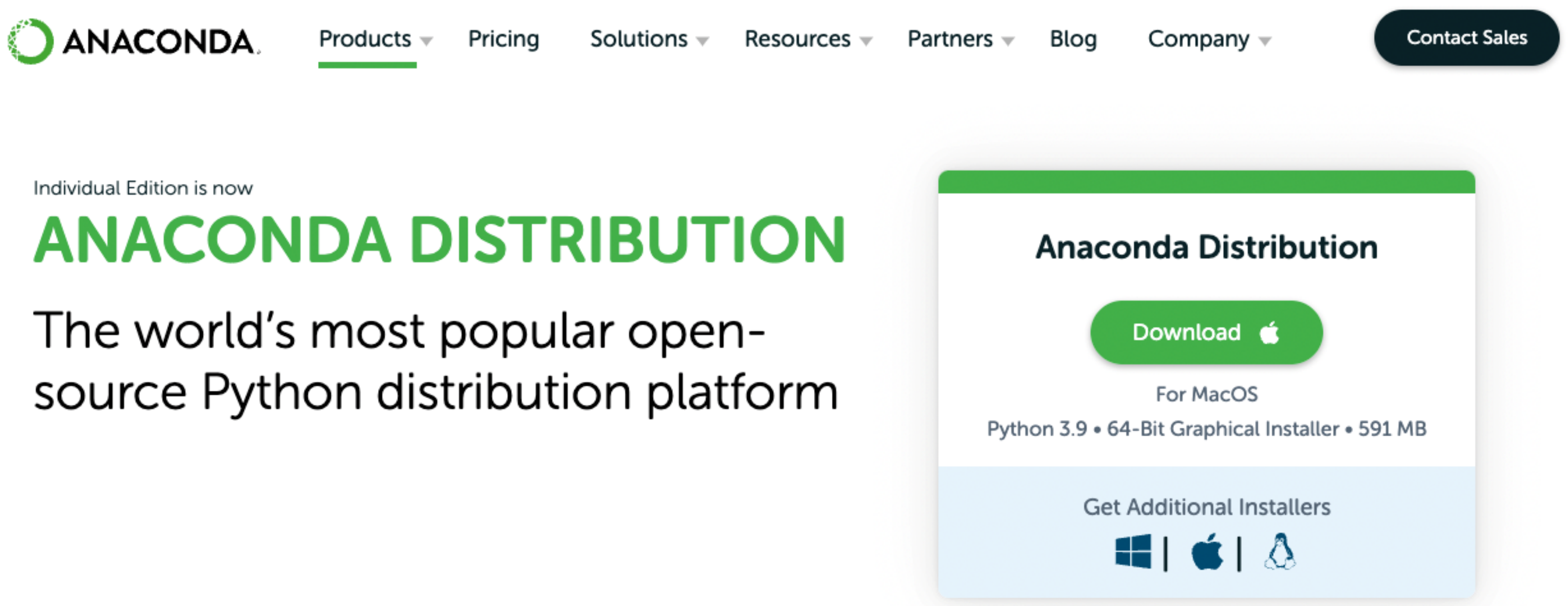
If you use `pip`, you can install NumPy with:

```
pip install numpy
```

…we are interested in numpy.random at the moment

Numpy Library

Package containing many useful numerical functions…



conda install numpy

If you use pip:
pip install numpy

numpy.random

- Lightweight library for sampling random variables
- Supports most standard discrete and continuous probability distributions
- Also handles random permutations of lists
- Imported along with Numpy as,

```
import numpy as np
```

(We can even do `import numpy.random as ra`)

- Functions accessible via np.random.(functionname)
- There are multiple random number generators… distinguishing them and seeding them can get a bit confusing…

Docs: https://numpy.org/doc/1.16/reference/routines.random.html

numpy.random

numpy.random.randint

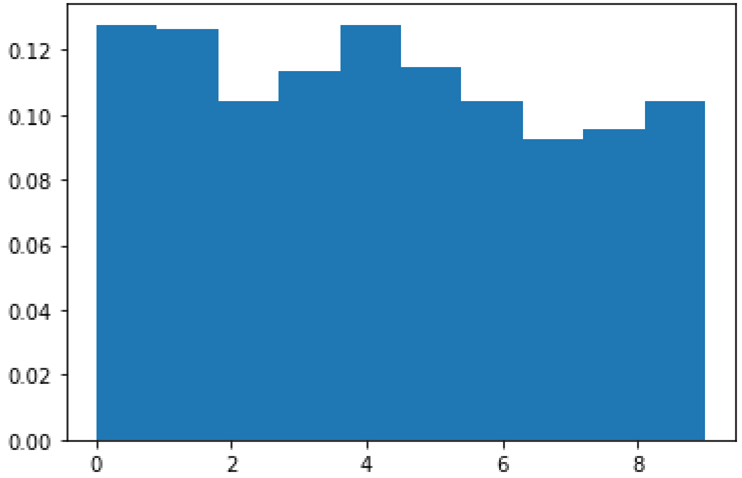
numpy.random.randint(*low*, *high=None*, *size=None*, *dtype='l'*)

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**
- Most calls (**but not all**) in numpy involving intervals follow this pattern
- `Size` argument accepts tuples for sampling ndarrays (multidimensional 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

numpy.random

Another good practice:

- Better to create new instance of the Random Number Generator (RNG)

```
mystream = numpy.random.RandomState(seed=3)
mystream.randint(1,1+6,size=10)
```

- Useful when you want reproducibility for <u>data shuffling</u> and <u>algorithms</u> separately.

Overview

- Conditional probability
- Independence
- Discrete distributions; probability mass function
- Continuous distributions; probability density function
- Some more on python

Conditional Probability

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

```
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))
```

compare:
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
```

More concise implementation


```
import numpy as np
for n in [10,100,1000,10_000,100_000, 1_000_000]:
    res = np.random.randint(1,1+6,size=(2,n))
    idx = (res[0,:] == 2) | (res[1,:] == 2)
    conditioned = res[:,idx]
    n_eff = conditioned.shape[1]

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

2 by n array

Length n, boolean array

2 by n_eff integer array

.shape returns `(#rows,#cols)`

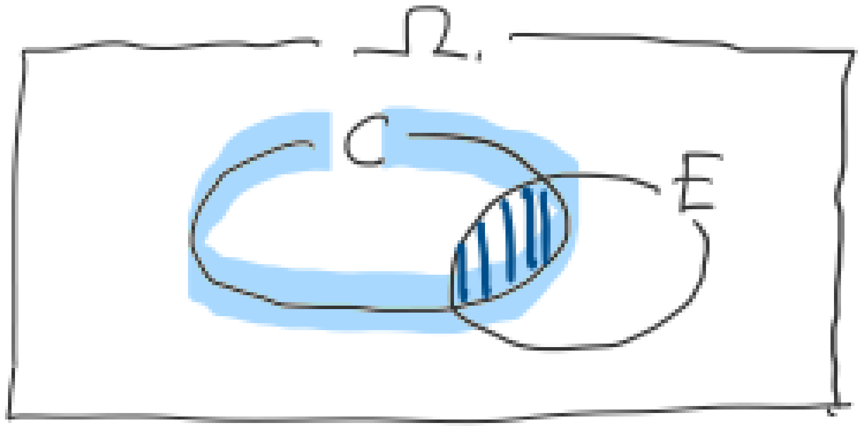
Sum() sums up the boolean array

There is a quite a bit of tricks like this in numpy. You will get used to it over time!

Conditional Probability

- Two fair dice example:
  - Suppose I roll two dice secretly and 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 $P(A|B) := \frac{P(A \cap B)}{P(B)}$ 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!

E.g., throw a fair die. X := outcome.
<u>Question</u>: P(X=4 | X is even) = P(X is even | X = 4) ?

No! 1/3 1

Conditional Probability

Chain rule

- $P(A \cap B) = P(A|B)P(B)$ ←just a rearrangement of definition
- $P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C)$
- $P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1) \prod_{i=2}^{n} P(E_i | \cap_{j=1}^{i-1} E_j)$ valid for any ordering!

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

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

$$= \sum_i P(A)P(B_i|A)$$ (by definition)

Shorthand:
$P(A, B) := P(A \cap B)$

Conditional Probability

[WJ:Ex.6.9] 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

$P(\text{test}=+ \mid \text{disease}=Y) = 0.9$

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

$P(\text{test}=+ \mid \text{disease}=N) = 0.01$

- One person in 1,000 has the disease.

$P(\text{disease}=Y) = 0.001$

Q: What is the probability that a person with positive test has the disease?

$P(\text{disease}=Y \mid \text{test}=+)$

Conditional Probability

What we know:

P(test=+ | D=Y) = 0.9

P(test=+ | D=N) = 0.01

P(D=Y) = 0.001

$\Rightarrow$

P(test=- | D=Y) = 0.1

P(test=- | D=N) = 0.99

P(D=N) = 0.999

Question: P(D=Y | test=+)

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

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

$$P(test = +, D = Y) = P(test = + | D = Y)P(D = Y)$$

$$P(test = +, D = N) = P(test = + | D = N)P(D = N)$$

The answer is 0.0826…

CAVEAT: P(test=+ | D=Y) = 0.9 ≠ P(D=Y | test=+)

Also: P(D=Y) = 0.001 vs P(D=Y | test=+)

Caveat

Q: What is the probability that a person with positive test has the disease?

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

P(disease=Y | test=+)

If a person took the test **because of having a symptom**, the actual probability (which is another conditional probability) will be different from our calculated answer.

For homework and exams, we will pretend that the person was chosen uniformly at random.

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

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(disease=Y | test=+)

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

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=+'

Independence

- Informally, given two events A and B, they are **independent** if the probability of A is not affected by whether B is true or false (and vice versa)
  - E.g., A = “die1=1” and B=“die2=1” are independent.

    $\Longrightarrow$ we know that the probability of die1 being 1 would not be changed just because die2=1.

- Mathematically, this can be written as $P(A|B) = P(A)$ or $P(B|A) = P(B)$.

- E.g., A = “die1=6” and B=“two dice sum to 6” are **not independent**.

  quiz candidate

  $\because$ intuitively, when B is true, A can never happen! So, P(A|B)=0 but P(A) = 1/6.

- E.g., A = “die1=1” and B=“two dice sum to 6” are **not independent**.

  quiz candidate

  $\because$ P(A) = 1/6 = 0.166... . However, P(A|B) = 1/5 = 0.2

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

More examples

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

- Q: A = "die1=1" and B="two dice sum to 5". Independent?

No

$\because$ P(A) = 1/6 , P(A|B) = 1/4 = .25

- Q: A = "die1=even" and B="two dice sum to 5". Independent?

Yes

$\because$ P(A) = 1/2 , P(A|B) = 2/4 = 1/2

Independence

[Def] Two events A and B are **independent** if

$$P(A,B) = P(A)P(B)$$

$A \perp B$ means A and B are independent

"joint probability is product of two marginal probabilities"

=> note: symmetric!

(skipping the following..)

Also, a set of events $\{A_i \in \mathcal{F}\}_{i=1}^{n}$ (n can be ∞) are **mutually independent** if

for every J $\subseteq \{1, \dots, n\}$, we have $P\left(\cap_{i\in J} A_i\right) = \prod_{i\in J} P(A_i)$

(∃ a notion of 'pairwise' independence, but not much useful, so we omit it here)

Independence

- Ex) recall two fair dice
  - We took it for granted that P((1,1)) is 1/36.
  - But why is it true, really?
  - To be rigorous,

$$P(die1 = 1, die2 = 1) = P(die1 = 1)P(die2 = 1) = \frac{1}{6} \cdot \frac{1}{6}$$

  due to independence.

or, …= P(die1=1 | die2=1) * P(die2=1) = P(die1=1) * P(die2=1)

- E.g., two biased coin **<u>C1</u>** and **<u>C2</u>**. Suppose P(C1=H) = 0.3 and P(C2=H) = 0.4. Compute the probability of P(C1=H,C2=T).

$0.3 \cdot 0.6 = 0.18$

quiz candidate

Computer Science

CSC380: Principles of Data Science

Probability Primer 3

Review

Axiom 3:

For any *finite* or *countably infinite* sequence of disjoint events $E_1, E_2, E_3, \ldots$, $P\left(\bigcup_{i\geq 1} E_i\right) = \sum_{i\geq 1} P(E_i)$

Inclusion-exclusion rule: $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

Law of total probability: For events $\mathrm{B}_1, \mathrm{B}_2, \ldots$ that partitions Ω, $P(A) = \sum_i P(A \cap B_i)$

Conditional probability: $P(E|C) \coloneqq \dfrac{P(E \cap C)}{P(C)}$ ($P(A|B) \neq P(B|A)$ in general)

Probability chain rule: $P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C)$

Bayes' rule: $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$

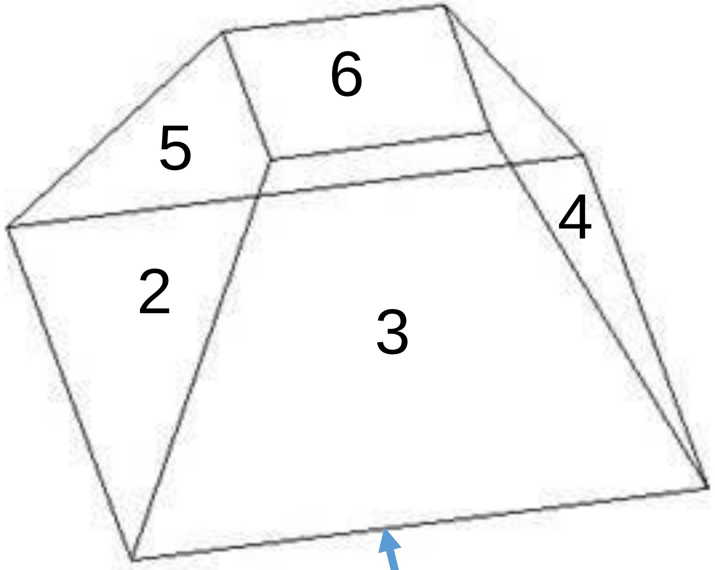
Independence: (definition) A and B are independent if $P(A, B) = P(A)P(B)$

(property) A and B are independent if and only if $P(A|B) = P(A)$ (or $P(B|A) = P(B)$)

Independence

- Ex) Unfair die
  - Let A be the outcome of a single throw.
  - P(A=1) << P(A=2) = … = P(A=5) << P(A=6)

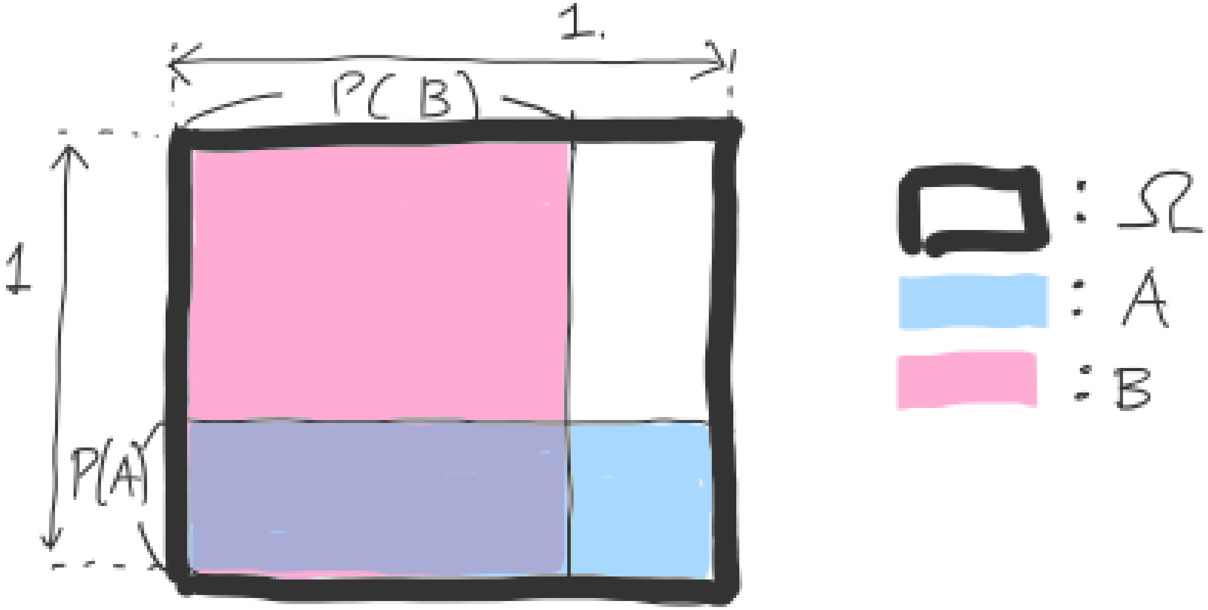
say, 0.1 0.15 0.15. 0.3



- Throw this die twice. What's the probability of observing (1,3)?
  - P((1,3)) = .1*0.15 = 0.015
- Similarly, P((6,3)) = 0.3*0.15 = 0.045

Independence

- Suppose: area = probability



Verify:
P(A,B) = P(A)P(B)

(or, show that
P(A|B) = P(A) or
P(B|A) = P(B))

- **Independence** is different from being **disjoint**.

Q: def'n of disjoint?

- **Exercise**: if A and B are disjoint (and P(A), P(B) > 0), then A and B are not independent. (hint: use the definition of independence)

quiz candidate

- Note: if A and B are not disjoint, then A and B may or may not be independent

Example: Dependent Coin Flips

- First coin (X1): fair coin
- Second coin (X2):
  - if X1=H, throw a **fair** coin.
  - If X1=T, throw an **unfair** coin P(H) = 0.2, P(T) = 0.8

- Q: Are X1=H and X2=H independent or not?

P(X1=H) = ____ 0.5

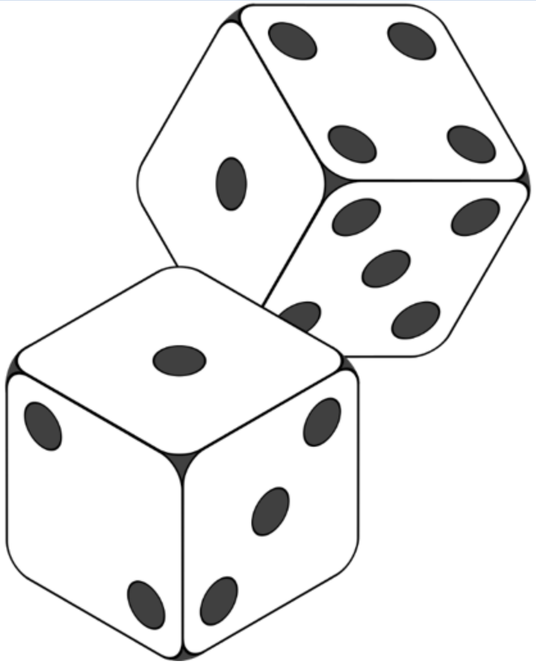
P(X2=H) = ____ = P(X2=H,X1=H) + P(X2=H,X1=T) = 0.25 + 0.1 = 0.35

P(X1=H, X2=H) = ____ 0.25

P(X1=H)*P(X2=H) = 0.175

Quiz candidate

Random Variables and Probability



Suppose we are interested in probabilities about the <u>sum of dice</u>…

<u>Option 1</u> Let E_i be event that the sum equals *i*

Two dice example:

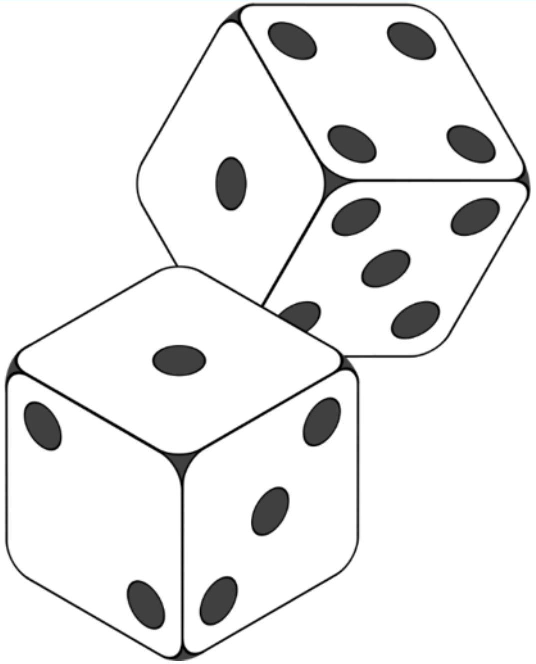
$$E_2 = \{(1,1)\} \quad E_3 = \{(1,2),(2,1)\} \quad E_4 = \{(1,3),(2,2),(3,1)\}$$

$$E_5 = \{(1,4),(2,3),(3,2),(4,1)\} \quad E_6 = \{(1,5),(2,4),(3,3),(4,2),(5,1)\}$$

Enumerate all possible outcomes obtaining the desired sum.
Gets cumbersome for N>2 dice…

Random Variables and Probability

Suppose we are interested in probabilities about the sum of dice…



Option 2 Give it a name

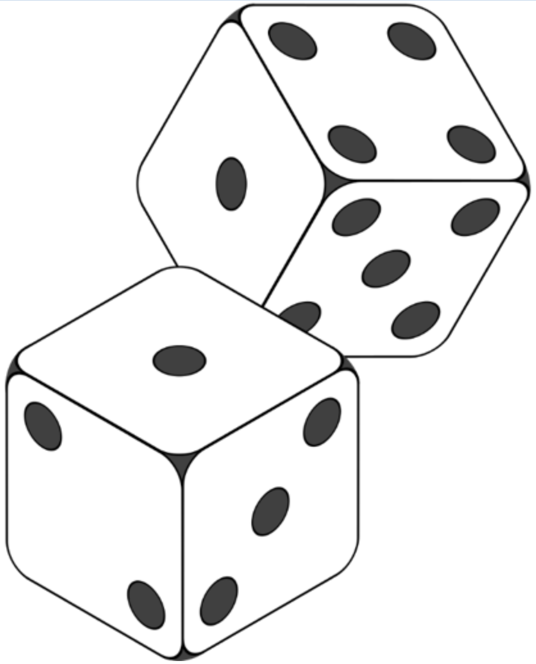
Let X be the sum of two dice.

We can say the event "X = i" to mean E_i.

X is called random variable.

(formal definition, next slide)

Random Variables and Probability



Suppose we are interested in a distribution over the sum of dice…

Option 2 Use a function of sample space…

Definition *A* **random variable** $X(\omega)$ *for* $\omega \in \Omega$ *is a real-valued function* $X : \Omega \to \mathbb{R}$.

(skipping; too technical)

Random Variables and Probability

- Obviously, all the laws/rules about events applies to RVs.

The ***law of total probability*** for random variable is,

$$P(Y = y) = \sum_{x} P(Y = y, X = x)$$

for all x: P(X=x) >0

… you will also see people write down $p(Y) = \sum_{x} p(Y, X = x)$

This means $p(Y = y) = \sum_x p(Y = y, X = x)$ for all y

but don't do: $p(Y) = \sum_x p(Y, X)$

*Lower case p often has a very slight difference from P, but we don't care in our class.

Conditional Probability

$$P(Y=y)=\sum_{x} P(Y=y, X=x)$$

Also works for conditional probabilities,

$$p(Y \mid Z)=\sum_{x} p(Y, X=x \mid Z)$$

Rule: Any rules about the probability still works for the conditional probabilities!!

(just make sure you add <u>the conditioning part</u> for every p()!)

Conditional Probability

Conditional probability $p(X \mid Y) = \frac{p(X,Y)}{p(Y)}$

Chain rule: $p(X,Y) = p(X|Y)p(Y)$

Bayes rule: $p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)}$

Conditional probability version

$$p(X|Y,Z) = \frac{p(X,Y|Z)}{p(Y|Z)}$$

↑ there is no 'double' conditioning

$$p(X,Y|Z) = p(X|Y,Z)p(Y|Z)$$

Try it now

$$p(X|Y,Z) = \frac{p(Y|X,Z)p(X|Z)}{p(Y|Z)}$$

I recommend that you verify the correctness by yourself!

Tabular Calculations for Random Variables

Tabular representation of two binary RVs (join probability)

Use K-by-K probability table for K-valued discrete RVs

	Y: y_1	y_2	$P(X)$
X: x_1	0.04	0.36	0.4 ($P(x_1)$)
x_2	0.30	0.30	0.6 ($P(x_2)$)
$P(Y)$	0.34 ($P(y_1)$)	0.66 ($P(y_2)$)	

$P(y_1)=P(x_1,y_1)+P(x_2,y_1)$
$P(y_2)=P(x_1,y_2)+P(x_2,y_2)$
[i.e., sum down columns]

$P(x_1)=P(x_1,y_1)+P(x_1,y_2)$
$P(x_2)=P(x_2,y_1)+P(x_2,y_2)$
[i.e., sum across rows]

Tabular Calculations for Random Variables

We don't care about event $Y=y_2$

		Y	
		y_1	y_2
X	x_1	0.04	Censored!
	x_2	0.30	
		0.34	

$P(y_1)$

Being lazy.. Same as $P(X|Y=y_1)$

$P(X \mid y_1)=?$

Tabular Calculations for Random Variables

		$Y=y_1$
X	x_1	0.04
	x_2	0.30

0.34

$P(y_1)$

⇨

0.04 / 0.34

0.30 / 0.34

$P(X|y_1)$

These sum to one:
A conditional probability is still a 'probability'.

Independence

Definition *Two random variables X and Y are <u>independent</u> given if and only if*

$$p(X = x, Y = y) = p(X = x)p(Y = y)$$

for all values x and y, and we say $X \perp Y$.

- From now on, we will just write it down as $p(X, Y) = p(X)p(Y)$
- Property: X and Y are independent if and only if $p(X) = p(X|Y)$ (or $p(Y) = p(Y|X)$)

➢ N RVs are independent if

$$p(X_1, \dots, X_N) = \prod_{i=1}^{N} p(X_i)$$

(Again, for all the possible values $x_1, \dots, x_N$)

Conditional Independence

Definition *Two random variables X and Y are* *<u>conditionally independent</u>* *given Z if and only if,*

$$p(X = x, Y = y \mid Z = z) = p(X = x \mid Z = z) p(Y = y \mid Z = z)$$

for all values x, y, and z, and we say that $X \perp Y \mid Z$.

Then, property: $p(X = x \mid Y = y, Z = z) = p(X = x \mid Z = z)$

- N RVs conditionally independent, given Z, if and only if:

$$p(X_1, \ldots, X_N \mid Z) = \prod_{i=1}^{N} p(X_i \mid Z)$$

- Also symmetric: $X \perp Y \mid Z \Leftrightarrow Y \perp X \mid Z$

Caveat: $X \perp Y \neq X \perp Y | Z$

Distribution

- If X is a random variable, then we can talk about its ‘distribution’
- **Distribution**: the set of values X can take and the probability assigned to each value.
- Examples:

X_1: unfair coin

value	prob.
1	0.2
2	0.8

X_2: unfair die

value	prob.
1	0.1
2	0.15
3	0.15
4	0.15
5	0.15
6	0.3

- Such a table can be viewed as a function f(x). This is called **probability mass function (PMF).**

Distribution

[Definition] *A* ***discrete*** *random variable takes on only a finite or countably infinite number of values.*

The case of **continuous** random variable will be discussed later!

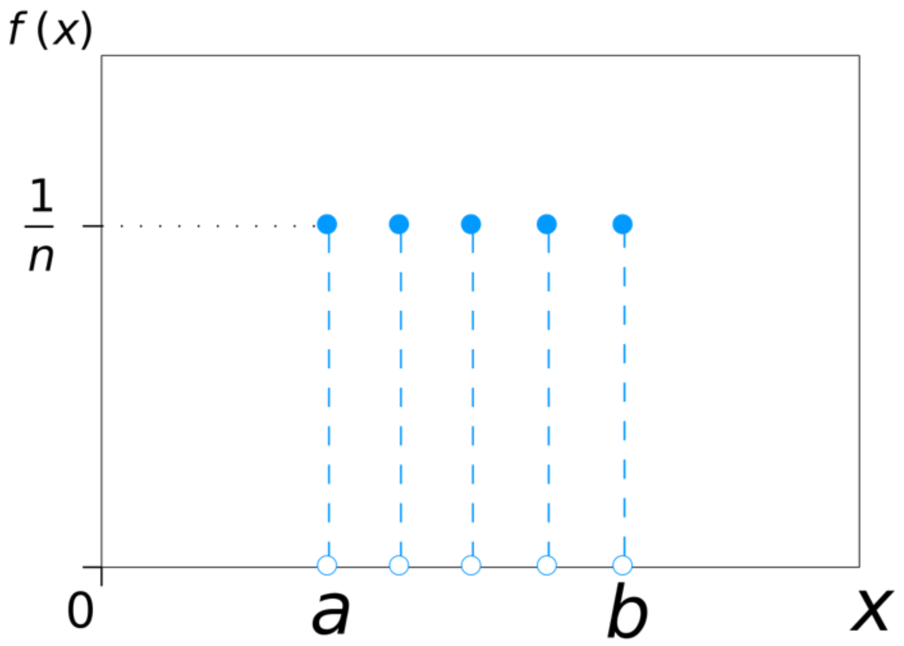
Useful Discrete Distributions

Generalization of fair die with N-faced die. Its PMF is:

$$p(X = k) = \frac{1}{N}$$

More generally, we define a set of numbers $\{v_1, v_2, \dots, v_N\}$

Discrete Uniform PMF



$$\text{Uniform}(X{=}k;\ \{v_1, v_2, \dots, v_N\}) = \begin{cases} \frac{1}{N} & \text{if } k \in \{v_1, v_2, \dots, v_N\} \\ 0 & \text{o.w.} \end{cases}$$

↑ it's like P(X=k) but being explicit about 'what' distribution X follows.

numpy.random

To generate a sample from a uniform discrete distribution,

```
numpy.random.choice([2,5,6])
out: 2
```

Bernoulli distribution

Bernoulli *a.k.a. the* ***coinflip*** *distribution on binary RVs* $X \in \{0,1\}$

PMF: $p(X=x) = \pi^x (1-\pi)^{1-x}$

Where π *is the probability of* ***success*** *(e.g., heads)*

Suppose we flip N independent coins $X_1, X_2, \ldots, X_N$, what is the distribution over their sum $Y = \sum_{i=1}^N X_i$

Num. "successes" out of N trials

Num. ways to obtain k successes out of N

Binomial Dist. $p(Y=k) = \binom{N}{k} \pi^k (1-\pi)^{N-k}$

Useful Discrete Distributions

Binomial Dist. $$p(Y=k) = \binom{N}{k} \pi^k (1-\pi)^{N-k}$$

Why is this true?
Say N=5. Compute p(Y=3)

p(HTTHH) = $\pi(1-\pi)(1-\pi)\pi\pi$

p(TTHHH) =$(1-\pi)(1-\pi)\pi\pi\pi$

…

The values are the same: $\pi^3(1-\pi)^2$!

By axiom 3, just add up $\pi^3(1-\pi)^2$ over all possible outcomes with the # of H is 3.
⇒ count: N choose k!

You'll use the same argument for HW1

numpy.random

numpy.random.binomial

numpy.random.binomial(*n*, *p*, *size=None*)

Binomial PMF

$$p(Y=k) = \binom{N}{k}\pi^k(1-\pi)^{N-k}$$

Example A company drills 9 wild-cat oil exploration wells, each with an estimated probability of success of 0.1. What is the probability of all nine wells fail?

Answer this by simulating 20,000 trials…

```
N = 20000
p = 0.1
wells = 9
X = np.random.binomial(wells, p, N)
odds = sum( X == 0 )/N
odds
```

```
0.38685
```

array([2,0,3,1,…,0])

(each∈ {0,…,9})

numpy.random

- If you want to compute it exactly,

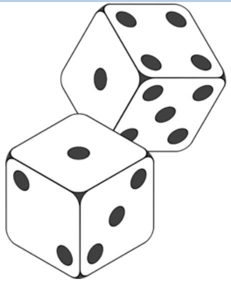
```
import scipy as sp
sp.special.binom(N,0) * (0.1**0) * (0.9**9)
=> out: 0.3874…
```

binom(n,k) returns n choose k

Scipy: sort of an extension of numpy for more specialized numerical computation.

Useful Discrete Distributions

Categorical Distribution on integer-valued RV $\mathrm{X} \in \{1, \dots, \mathrm{K}\}$ that takes $X = k$ with probability π_k.



K-sided biased die

$$p(X) = \prod_{k=1}^{K} \pi_k^{\mathbf{I}(X=k)} \quad \text{or} \quad p(X) = \sum_{k=1}^{K} \mathbf{I}(X = k) \cdot \pi_k$$

equivalent to:

$$p(X = x) = \prod_{k=1}^{K} \pi_k^{\mathrm{I}(x=k)}$$

Indicator function

$$\mathbf{I}(A) := \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{othewise} \end{cases}$$

X is a vector, then X_k is its k-th component

Can also use **one-hot** vector representation,

$$X \in \{0,1\}^K \quad \text{where} \quad \sum_{k=1}^{K} X_k = 1 \quad \text{then} \quad p(X) = \prod_{k=1}^{K} \pi_k^{X_k}$$

numpy.random

```
numpy.random.choice([2,5,6],p=[.5,.3,.2])
out: 2
```

Useful Discrete Distributions

What if we count outcomes of *n* i.i.d. **categorical RVs** $Y_1, \ldots, Y_n \in \{1, \ldots, K\}$?

Multinomial *Distribution of* $(X_1, \ldots, X_K)$ *where* $X_k = \sum_{i=1}^{n} \mathbf{I}(Y_i = k)$ *is the count of item k. Note:* $\sum_k X_k = n$.

$$p(x_1, \ldots, x_K) = \binom{n}{x_1 x_2 \ldots x_K} \prod_{k=1}^{K} \pi_k^{x_k}$$

Number of ways to partition N objects into K groups of size $x_1, \ldots, x_K$:

$$\binom{n}{x_1 x_2 \ldots x_K} = \frac{n!}{x_1! x_2! \ldots x_K!}$$

Q: what are the parameters of the multinomial distribution?

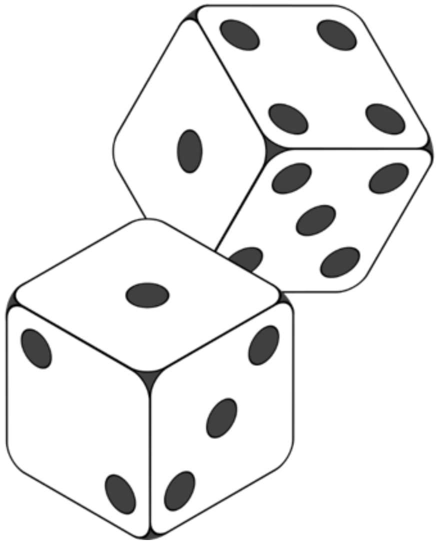
numpy.random

numpy.random.multinomial

numpy.random.multinomial(*n*, *pvals*, *size=None*)

Draw samples from a multinomial distribution.

pvals: list of probability parameters that sums to 1.

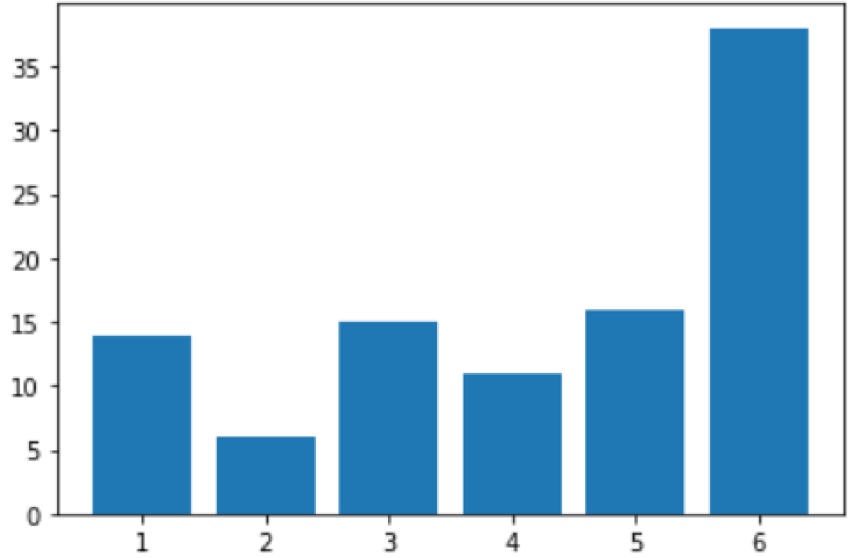


Example Simulate 100 throws of a *"loaded"* die that has 3X the chance of rolling 6, and equal chance for remaining numbers.

(import matplotlib.pyplot as plt)

```
N = 100
p_unnorm = np.array([1,1,1,1,1,3])
p = p_unnorm / sum(p_unnorm) # normalize
X = np.random.multinomial(N, p)
plt.bar(np.arange(6) + 1, X)
plt.show()
```

(X is a vector of counts; length 6)



Note: Probability vector <u>must</u> be a valid PMF (nonnegative, normalized a.k.a sum to 1)

numpy.random

How to simulate Bernoulli? Categorical?

Bernoulli is equivalent to a single draw from a binomial,

```
X = np.random.binomial(n=1, p=0.5) # fair coin flip
print(X)
```

```
0
```

Categorical is equivalent to a single draw from a multinomial,

```
X = np.random.multinomial(1, [0.5, 0.5]) # also a fair coin flip
print(X)
```

```
[0 1]
```

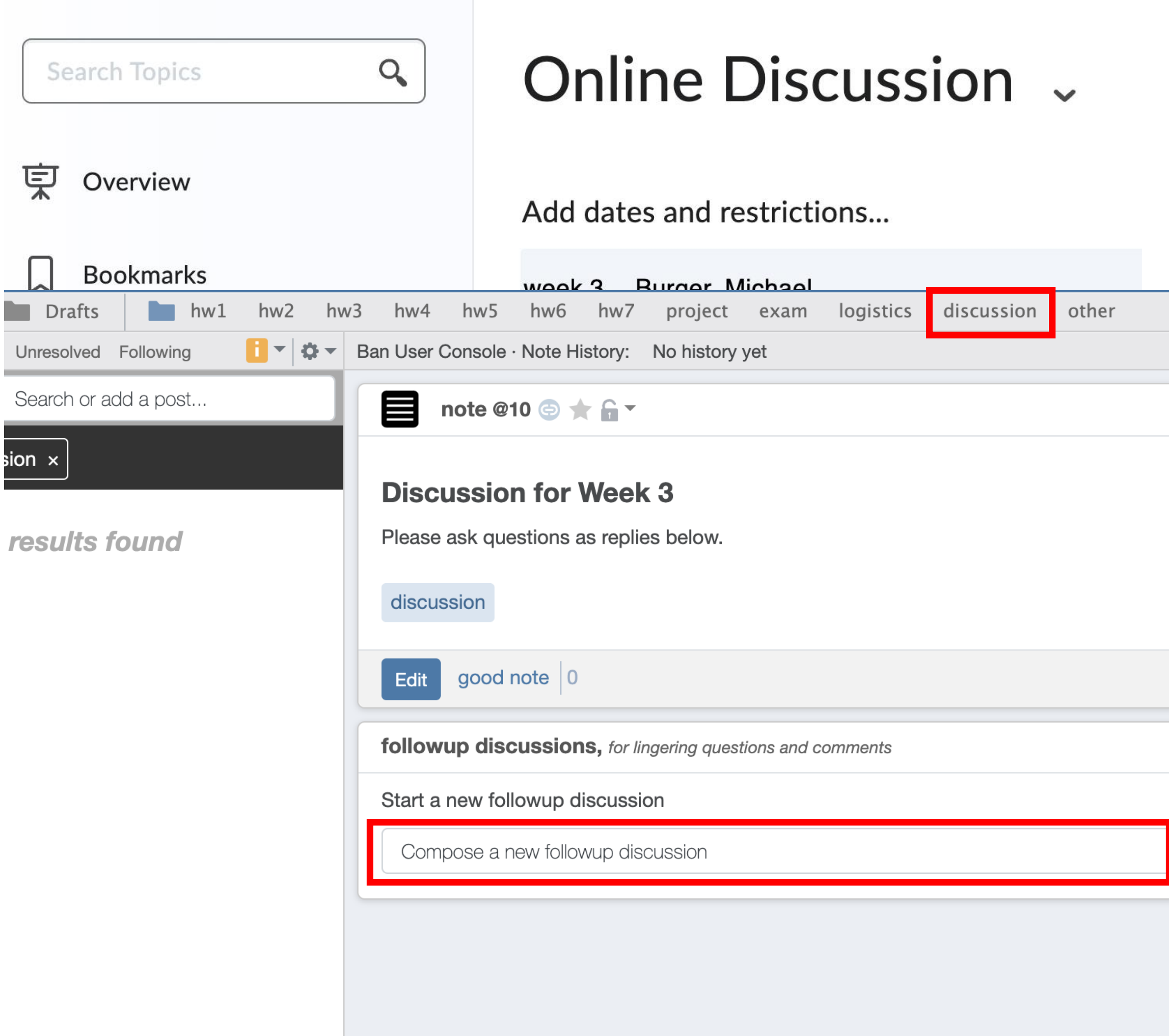
Computer Science

CSC380: Principles of Data Science

Probability Primer 4

Online Discussion

- Each student asks one question **by Thursday** of the week in piazza.
- Other students answers the questions.
- Questions remained unanswered will be addressed by me.



Random Variable Examples

- X: an outcome of a die.
  - $R_1 = I\{X\ is\ even\}$
  - $R_2 = I\{X = 1\}$

Random variable induces a partition of the outcome space!

$\{R_1 = 1\} \Leftrightarrow \{2,4,6\}$

$\{R_1 = 0\} \Leftrightarrow \{1,3,5\}$

$\{R_2 = 1\} \Leftrightarrow \{1\}$

$\{R_2 = 0\} \Leftrightarrow \{2,3,4,5,6\}$

- X_1, X_2: outcomes of two dice
  - $R_3 = X_1 + X_2$
  - $R_4 = \frac{(X_1+X_2)}{2}$
  - $R_5 = I\{X_1 = 1\}$

$\{R_5 = 1\} \Leftrightarrow \{(1,1),(1,2),\ldots,(1,6)\}$

$$\begin{aligned}\{R_5 = 0\} \Leftrightarrow \{&(2,1),(2,2),\ldots,(2,6),\\ &(3,1),(3,2),\ldots,(3,6),\\ &\ldots\\ &(6,1),(6,2),\ldots,(6,6)\}\end{aligned}$$

Q: what distribution does R_5 follow with what parameter?

Continuous Probability

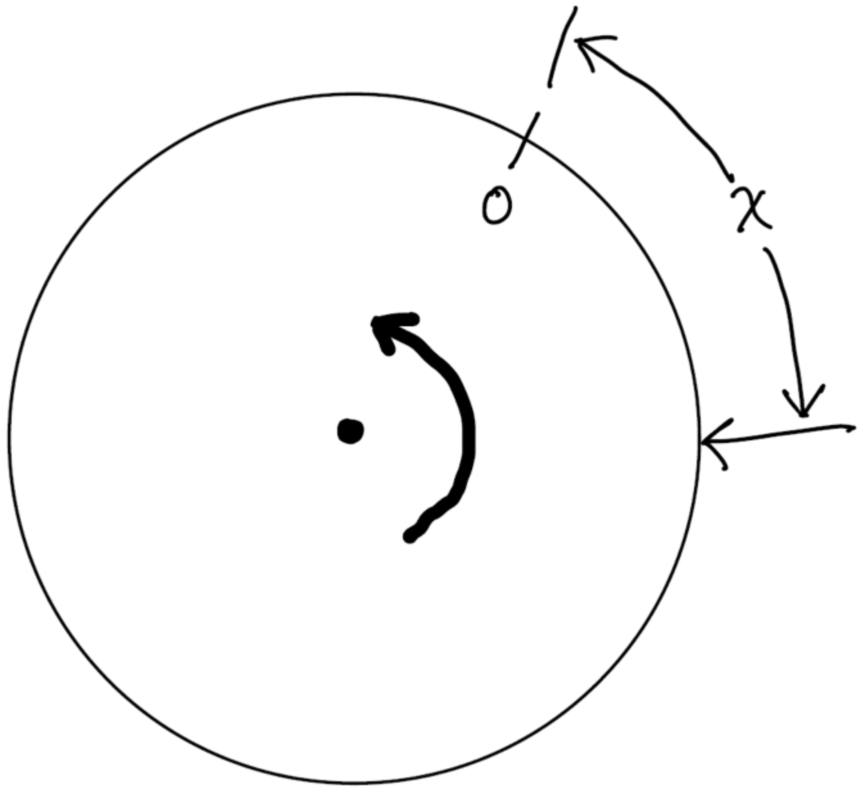
(TV show spin the wheel)

Continuous Probability

Experiment Spin continuous wheel and measure X displacement from 0

Say the circumference is 1.

Outcome space Ω is all points (real numbers) in $(0,1]$



Question Assuming uniform distribution, what is $P(X = x)$?

Proof

<u>Goal</u>: Show P(X=x) = 0

- Say the displacement X is in $(0,1]$
- Let ϵ be a very small number.
- Let $I_k = ((k-1)\epsilon, k\epsilon]$

Q: how many such intervals fit into (0,1]?

- Let $j(x)$ be k such that $x \in I_k$
- $P(X = x) \leq P\left(X \in I_{j(x)}\right) = \epsilon$

(say $1/\epsilon$ is an integer)

Q: why is it ϵ?

- **We can make ϵ as small as we want! ⇒ P(X=x) must be 0.**

Continuous Probability

Alternative proof

<u>Claim</u>: Uniform distribution on (0,1] satisfies $P(X = x) = 0$.

- Suppose $p(X = x) = \pi > 0$ for every $x \in [0,1)$
- Let $S(k)$ be set of any k *distinct* points in [0,1). Then,

$$P(x \in S(k)) = k\pi$$

- By setting $k = \left\lceil \frac{1}{\pi} \right\rceil + 1$, $P\big(x \in S(k)\big) > 1$ => CONTRADICTION!!

In (uncountably) infinite sample space, an event may be **possible** but may have zero “probability”

*Assign probability to **<u>intervals</u>**, not individual outcomes.*

Continuous Probability

Maybe, it’s not so weird.

- Q1: Probability that Usain Bolt will run 100m with time exactly 9.58?

- Q2: Probability that Usain Bolt will run 100m with time exactly 9.58912312850982349871239428710298395723409809182309812098?

in reality, we never work with a precise real number. we work with intervals!!

Continuous Probability

we could try to convince ourselves that it is sensible.

… or we could just accept this oddity…

Or, Use a Heuristic Argument

- Computers are inherently dealing with discrete numbers anyways.

- Imagine you consider extremely fine grained intervals like
  … (-ϵ,0], (0, ϵ], (ϵ, 2ϵ], …
  where $\epsilon = 10^{-300}$

- The outcome space is real, but we only talk about the events that are union of those intervals.

Continuous Probability

Definition *The* *cumulative distribution function (CDF) of a real-valued continuous RV X is the function given by,*

$$F(x) = P(X \leq x)$$

Let x be multiple of ϵ

$I_k \coloneqq \big((k-1)\epsilon, k\epsilon\big] \Rightarrow$ " base intervals"

$j(x) \coloneqq \frac{x}{\epsilon} \Rightarrow$ "index of the interval containing x"

$F(x) = \sum_{k=-\infty}^{j(x)} P(X \in I_k)$

➢ Can easily measure probability of closed intervals,

$$P(a < X \leq b) = F(b) - F(a)$$

$\sum_{k=j(a)+1}^{j(b)} P(X \in I_k)$

When $\epsilon \to 0$ this does not necessarily become 0

➢ If $F(X)$ is differentiable then,

$$\frac{F(x) - F(x-\epsilon)}{x - (x-\epsilon)} = \frac{P(X \in I_{j(x)})}{\epsilon}$$

$$\sum_{k=-\infty}^{j(t)} f(x) \cdot (x - (x-\epsilon))$$

$$f(x) = \frac{dF(x)}{dx} \quad \text{and} \quad F(t) = \int_{-\infty}^{t} f(x)\, dx$$

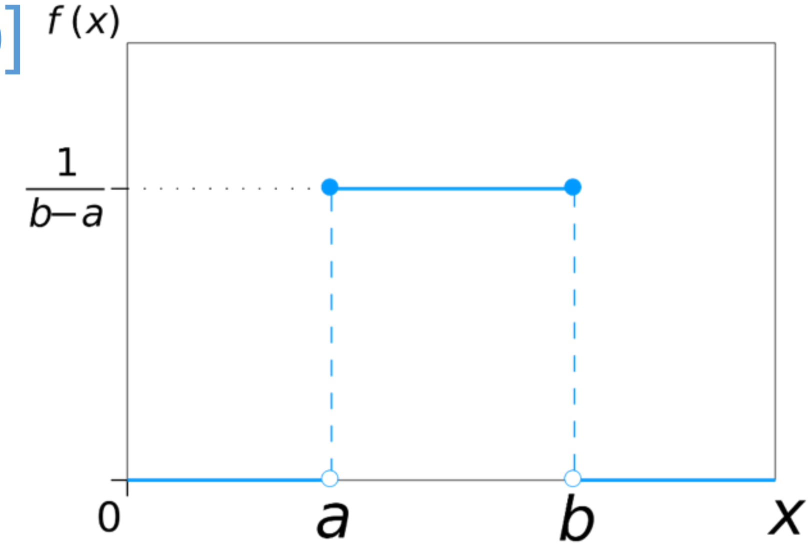
Where $f(x)$ is called the ***probability density function (PDF)***

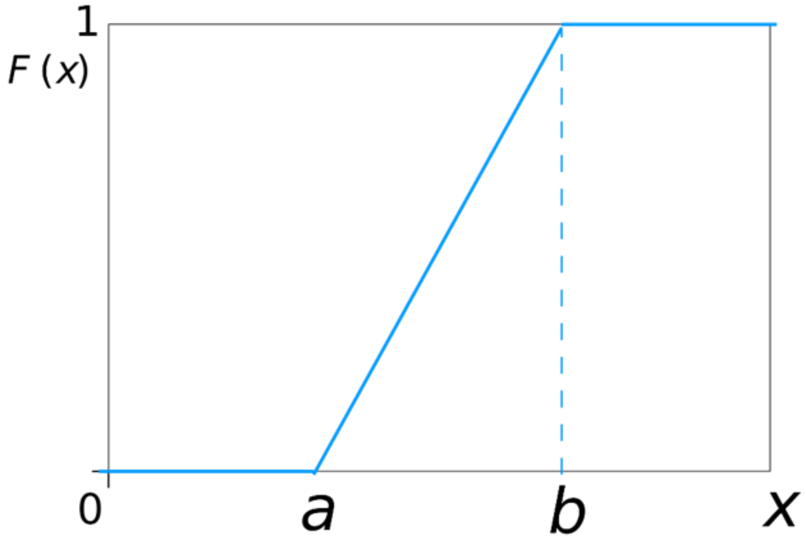
Fundamental Theorem of Calculus

Useful Continuous Distributions

Uniform distribution on interval $[a, b]$: Uniform[a,b]

$$p(x) = \begin{cases} 0 & \text{if } x \leq a, \\ \frac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{if } b \leq x \end{cases} \qquad P(X \leq x) = \begin{cases} 0 & \text{if } x \leq a, \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b, \\ 1 & \text{if } b \leq x \end{cases}$$





Notation:
p(x) for the PDF function
P(A) for the probability

Again, PDF function $\neq$ probability

Continuous Probability

Most definitions for discrete RVs hold, replacing sum with integral…

Law of Total Probability for continuous distributions,

$$p(x) = \int_y p(x,y)\,dy \qquad \frac{P\left(X \in I_{j(x)}\right)}{\epsilon} = \sum_{j=-\infty}^{\infty} \frac{P\left(X \in I_{j(x)}, Y \in I_{j(y)}\right)}{\epsilon^2} \cdot \epsilon$$

All the rules apply when replacing PMF with PDF…

Conditional PDF:

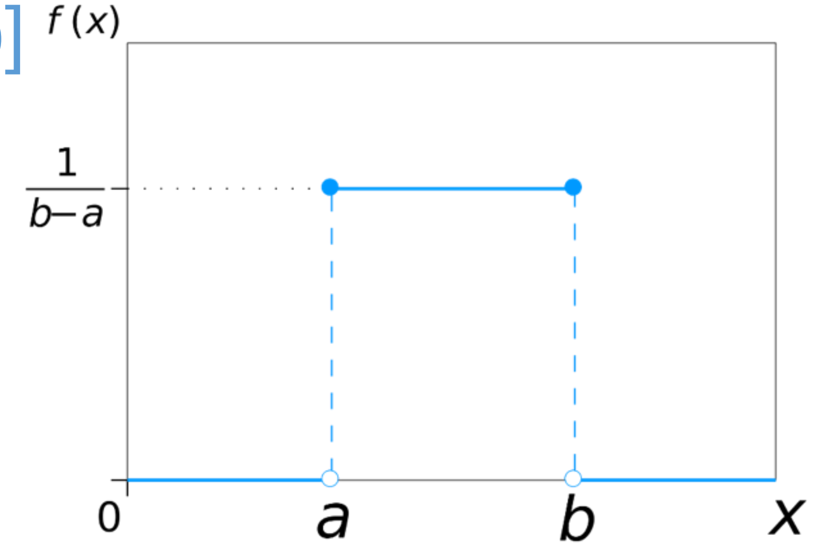
$$p(X \mid Y) = \frac{p(X,Y)}{p(Y)} = \frac{p(X,Y)}{\int p(x,Y)\,dx}$$

Probability Chain Rule:

$$p(X,Y) = p(Y)p(X \mid Y)$$

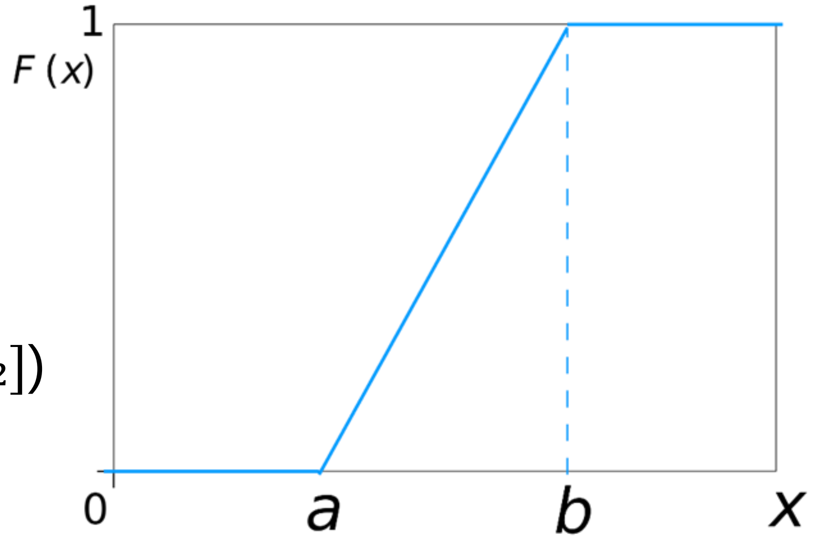
Useful Continuous Distributions

Uniform distribution on interval $[a,b]$: Uniform[a,b]



$$p(x) = \begin{cases} 0 & \text{if } x \le a, \\ \frac{1}{b-a} & \text{if } a \le x \le b, \\ 0 & \text{if } b \le x \end{cases} \qquad P(X \le x) = \begin{cases} 0 & \text{if } x \le a, \\ \frac{x-a}{b-a} & \text{if } a \le x \le b, \\ 1 & \text{if } b \le x \end{cases}$$

Suppose $X \sim U(0,1)$ and we are told $X \le \frac{1}{2}$ what is the conditional distribution?



$$P(X \le x \mid X \le \tfrac{1}{2}) = F(x) \text{ of Uniform}[0, ½]$$

(i.e., $P(Y \le x)$ where $Y\sim\text{Uniform}[0, ½]$)

*Holds generally: Uniform distr. is **closed** under conditioning*

Useful Continuous Distributions

numpy.random.uniform

numpy.random.uniform(*low=0.0, high=1.0, size=None*)

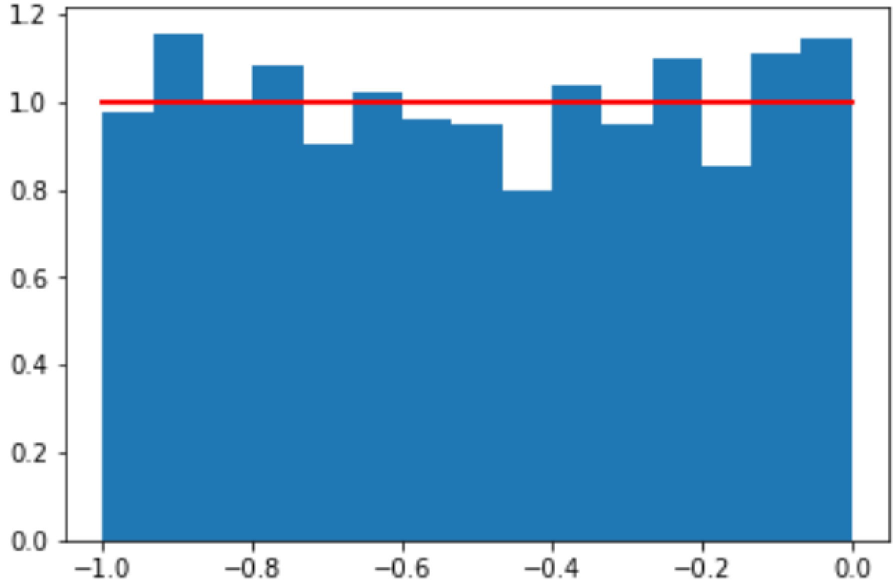
Draw samples from a uniform distribution.

Samples are uniformly distributed over the half-open interval `[low, high)` (includes low, but excludes high). In other words, any value within the given interval is equally likely to be drawn by **uniform**.

Example Draw 1,000 samples from a uniform on [-1,0),

redline: PDF of uniform distr.

```
a = -1
b = 0
N = 1000
X = np.random.uniform(a,b,N)
count, bins, ignored = plt.hist(X, 15, density=True)
plt.plot(bins, np.ones_like(bins), linewidth=2, color='r')
plt.show()
```



Notation

- $X \sim D$ X follows distribution D
- E.g., $X \sim \text{Uniform}[0,1]$

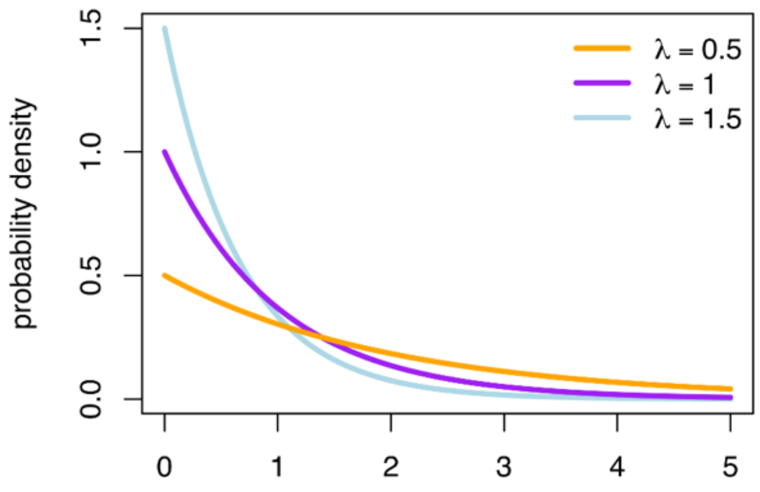
Useful Continuous Distributions

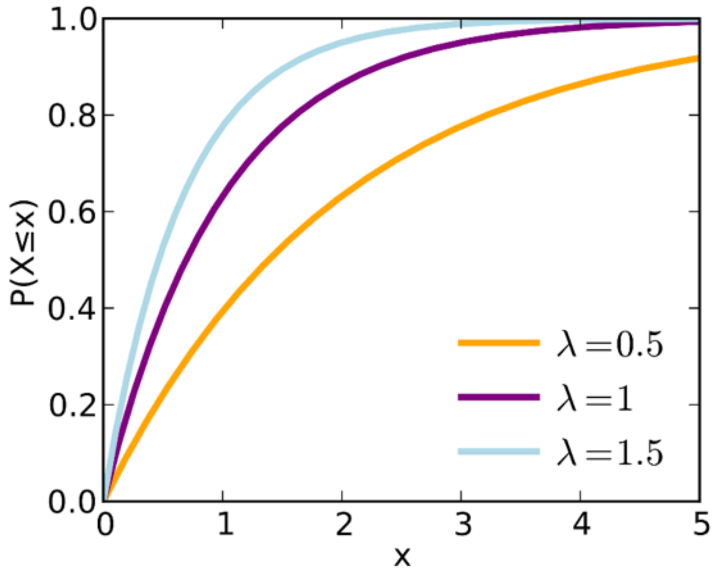
Exponential distribution with scale λ,

$$p(x) = \lambda e^{-\lambda x} \qquad\qquad P(x) = 1 - e^{-\lambda x}$$

for X>0.

'waiting time' often follows an exponential distribution.





numpy.random

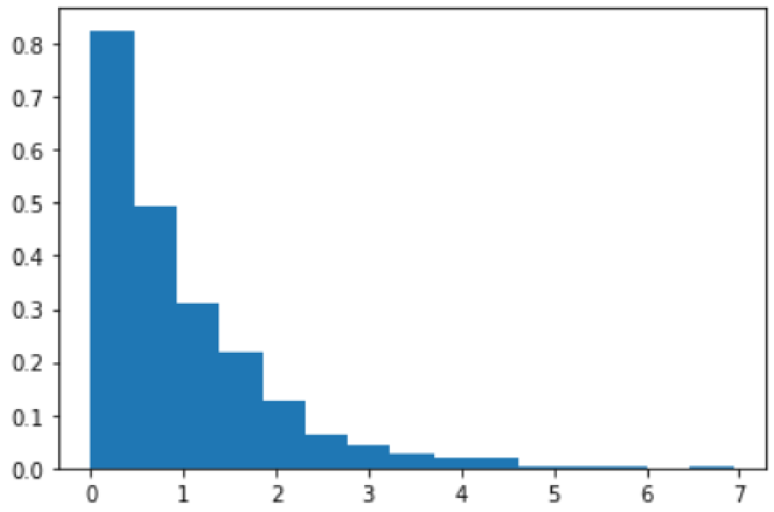
numpy.random.exponential

numpy.random.exponential(*scale=1.0, size=None*)

scale = λ

Example Draw 1,000 samples from exponential with $\lambda = 1.0$

```
lam = 1.0
N = 1000
X = np.random.exponential(lam, N)
count, bins, ignored = plt.hist(X, 15, density=True)
plt.show()
```

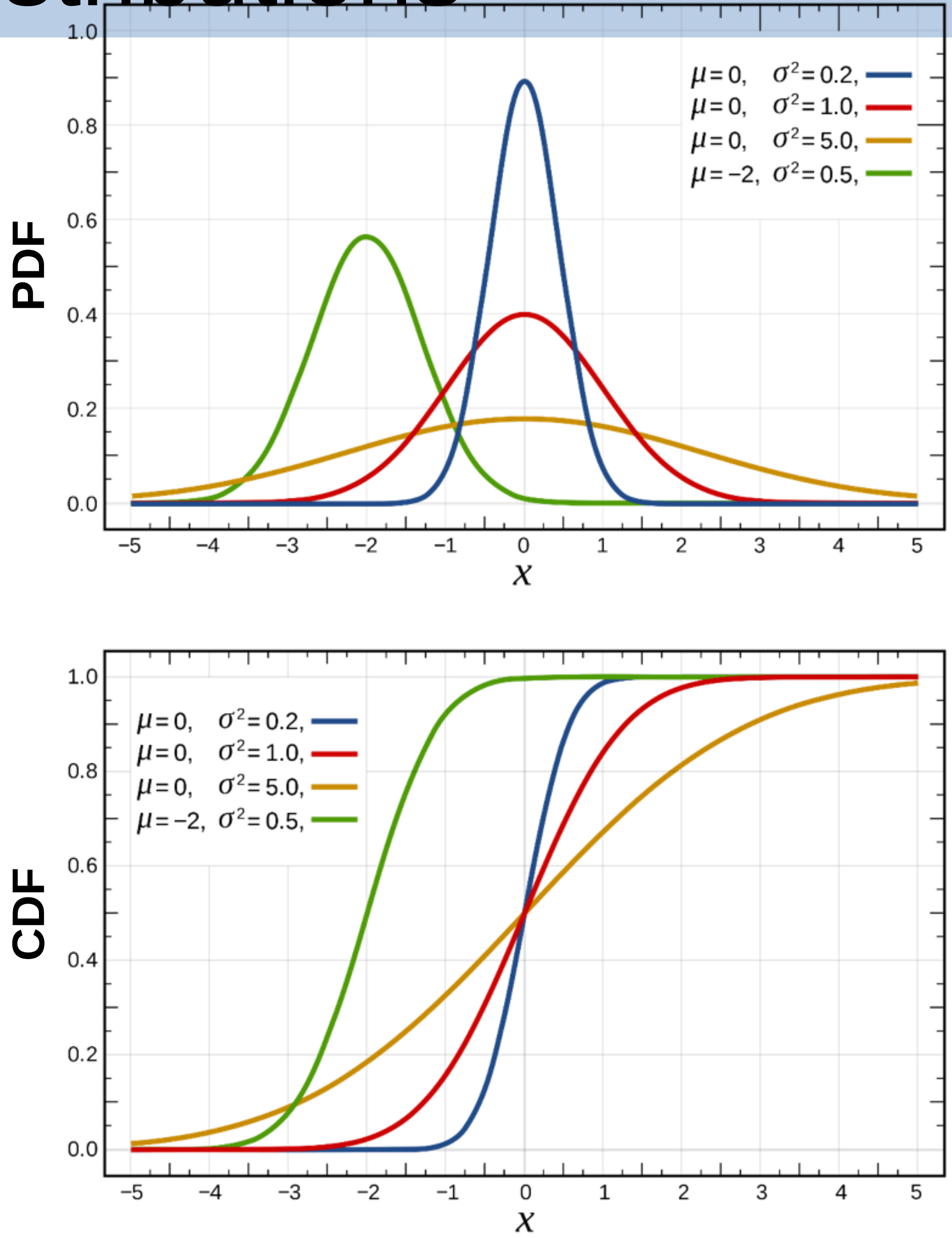


Useful Continuous Distributions

Gaussian (a.k.a. Normal) distribution with mean mean (location) μ and variance (scale) σ^2 parameters,

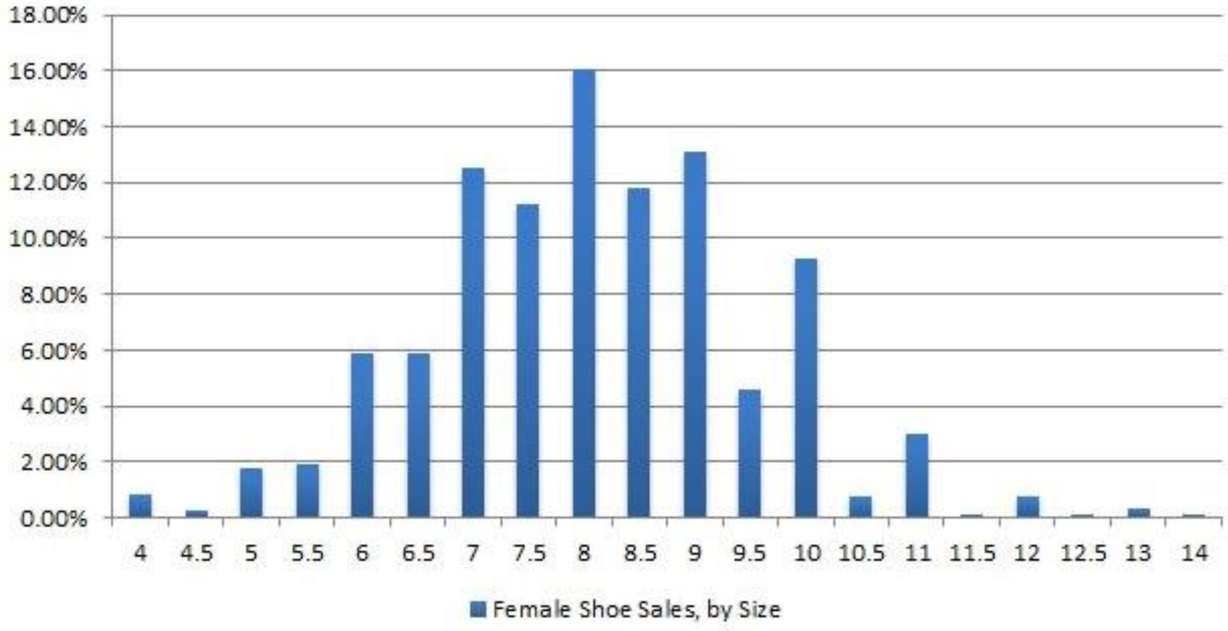
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

Compactly, $X \sim \mathcal{N}(\mu, \sigma^2)$

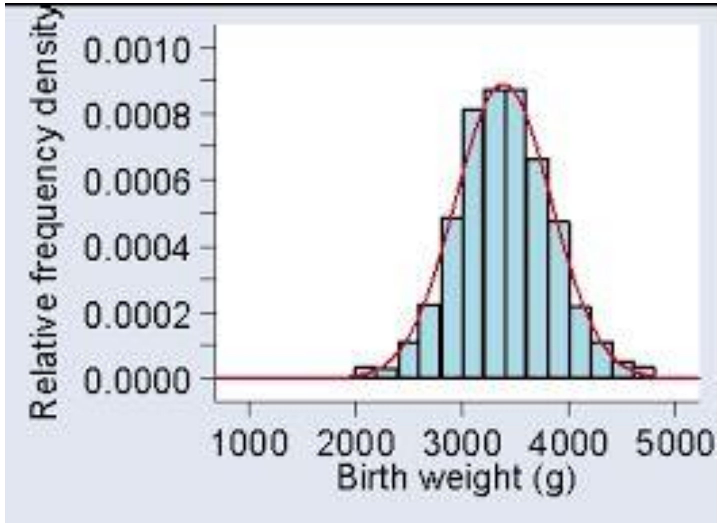


Things that follow Gaussian

<u>Female shoe size</u>



<u>Birth Weight</u>



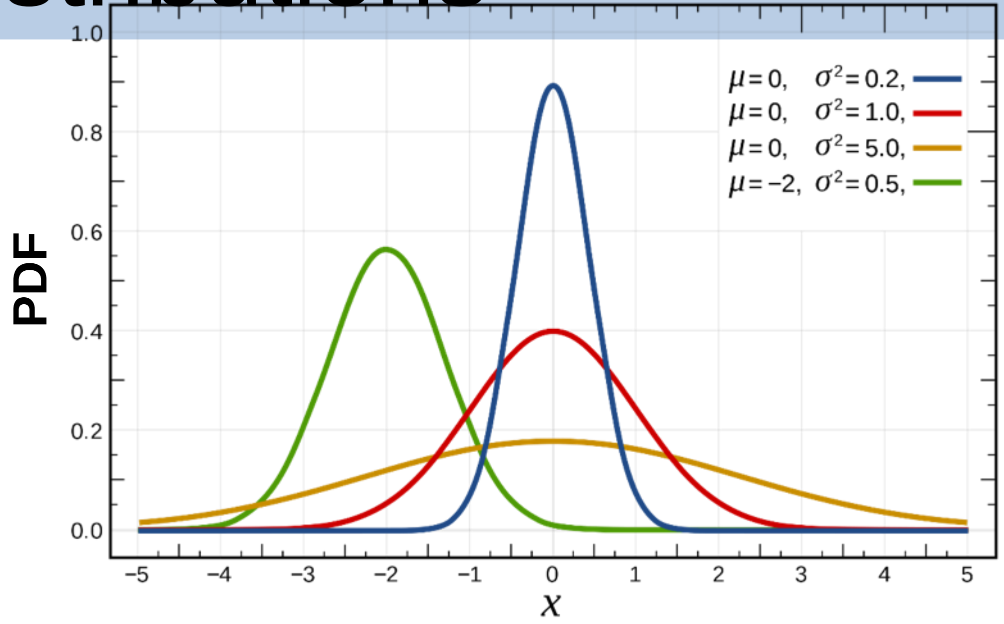
(From https://studiousguy.com/real-life-examples-normal-distribution/)

Useful Continuous Distributions

Gaussian (a.k.a. Normal) distribution with mean mean (location) μ and variance σ^2 parameters,

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

Compactly, $X \sim \mathcal{N}(\mu, \sigma^2)$



Useful Properties

Quiz candidate

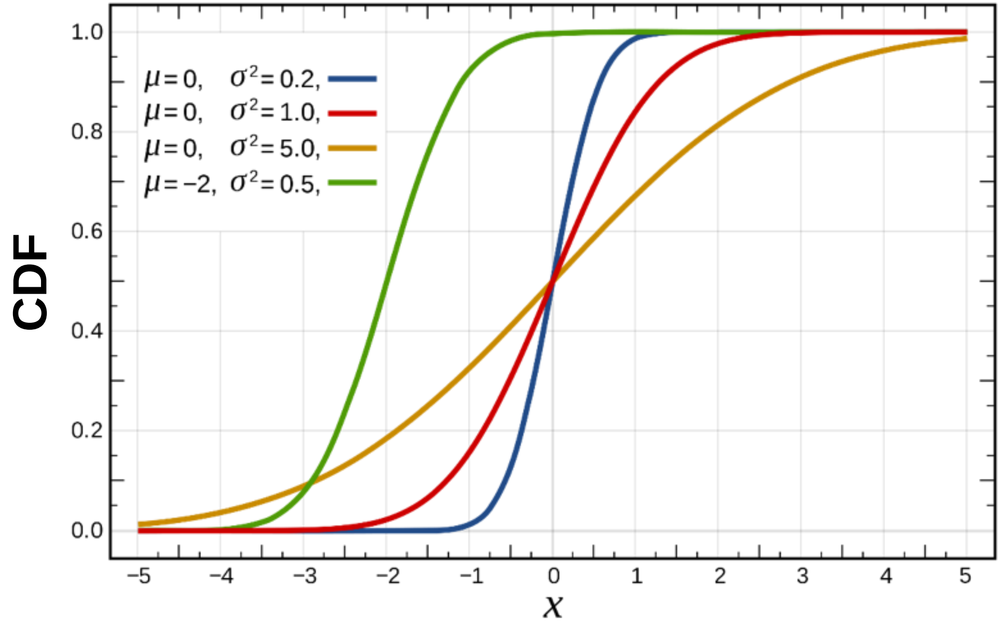
- Closed under additivity:

$$X \sim \mathcal{N}(\mu_x, \sigma_x^2) \qquad Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$$

$$X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

- Closed under affine transformation (a and b constant):

$$aX + b \sim \mathcal{N}(a\mu_x + b, a^2\sigma_x^2)$$



numpy.random

numpy.random.normal

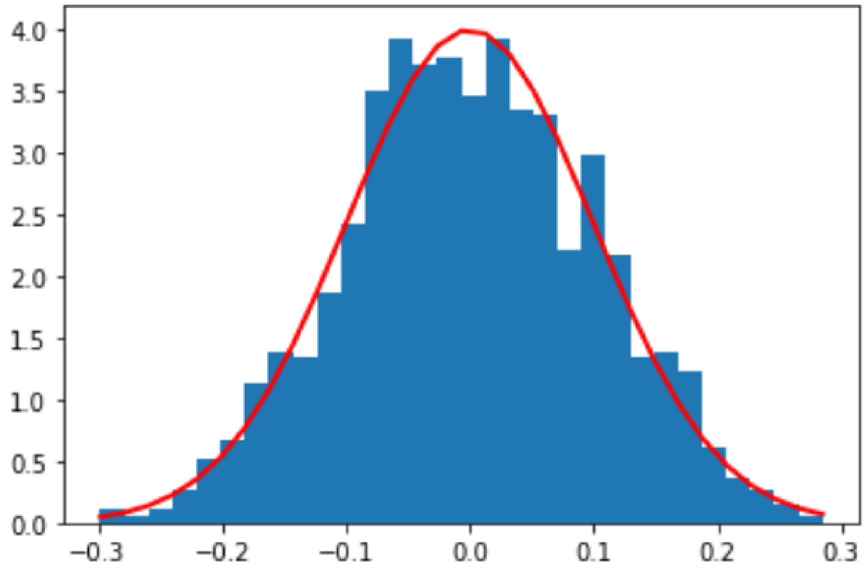
scale = $\sqrt{\sigma^2}$

numpy.random.normal(*loc=0.0, scale=1.0, size=None*)

Draw random samples from a normal (Gaussian) distribution.

Example Sample zero-mean gaussian with scale 0.1,

```
mu, sigma = 0, 0.1 # mean and standard deviation
X = np.random.normal(mu, sigma, 1000)
count, bins, ignored = plt.hist(X, 30, density=True)
plt.plot(bins, 1/(sigma * np.sqrt(2 * np.pi)) *
         np.exp( - (bins - mu)**2 / (2 * sigma**2) ),
         linewidth=2, color='r')
plt.show()
```



bins: length 31, consisting of boundary points

numpy.random

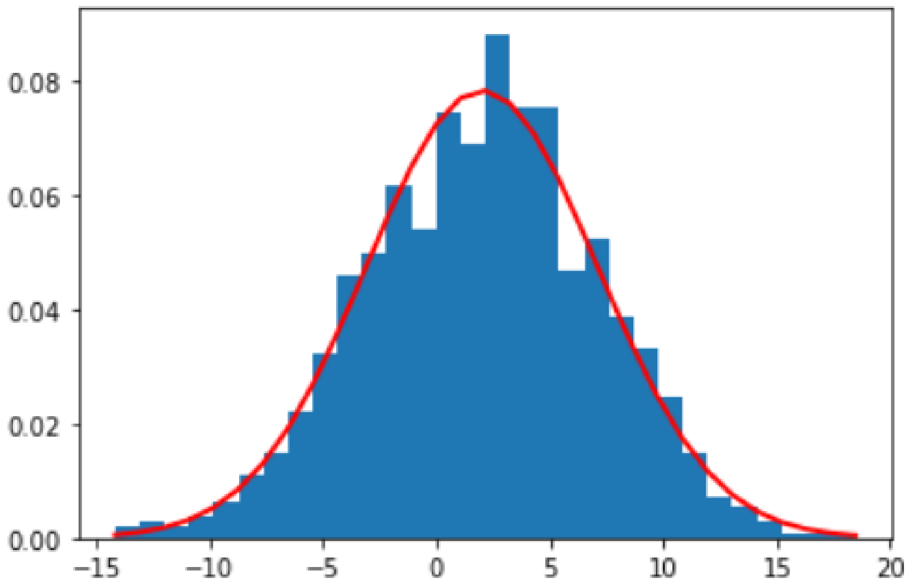
Gaussians are closed under additivity

Example *Add two Gaussian RVs,*

$$X \sim \mathcal{N}(\mu_x, \sigma_x^2) \qquad Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$$

$$X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

```
mu_x, sigma_x = 0, 1
mu_y, sigma_y = 2, 5
X = np.random.normal(mu_x, sigma_x, 1000)
Y = np.random.normal(mu_y, sigma_y, 1000)
Z = X+Y
count, bins, ignored = plt.hist(Z, 30, density=True)
mu_z = mu_x + mu_y
sig_z_sq = sigma_x**2 + sigma_y**2
plt.plot(bins, 1/(np.sqrt(sig_z_sq * 2 * np.pi)) *
         np.exp( - (bins - mu_z)**2 / (2 * sig_z_sq) ),
         linewidth=2, color='r')
plt.show()
```



Property extends to a sequence of Gaussian RVs,

$$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \qquad \sum_i X_i \sim \mathcal{N}(\cdot)$$

Recap

Useful discrete distributions

- Bernoulli → "Coinflip Distribution"
- Binomial → Multiple Bernoulli draws
- Categorical / Multinomial → One / Many die rolls

Continuous probability

- P(X=x) = 0 <u>does not mean you won't see x</u>
- Probabilities assigned to *intervals* via CDF P(X > x)
- PDF measures probability *density* of single points p(X=x) >= 0

Useful continuous distributions

- Exponential → waiting time.
- Univariate / Multivariate Gaussian → Probably most ubiquitous distribution
- There are a lot more we will touch on later in the course…

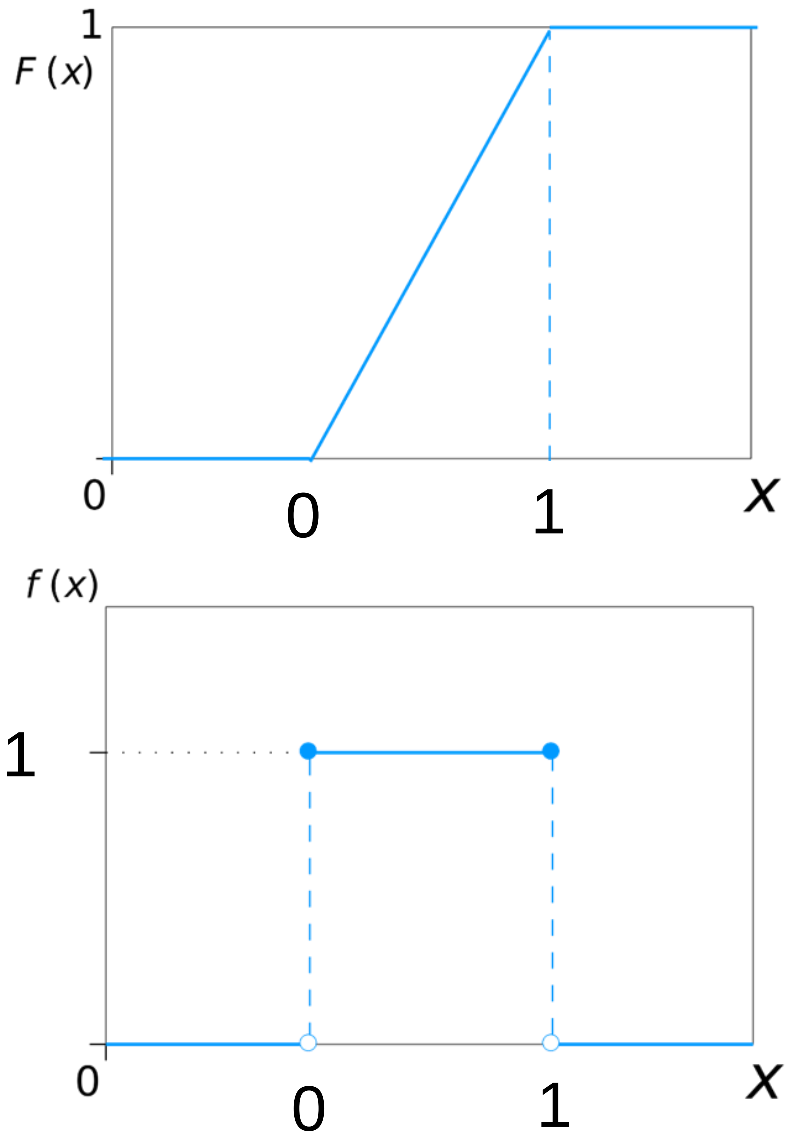
CSC380: Principles of Data Science

Probability Primer 5

Review: Continuous Random Variable

- Probability can be assigned to intervals
- Define CDF: $F(x) := P(X \leq x)$
- Then, PDF: $f(x) := p(X = x) := F'(x)$ // the slope at $F(x)$
- $P(X \in [a, b])$ = F(b) – F(a) // area under the PDF curve

Uniform[0,1]



<u>Another viewpoint</u>

- A continuous distribution is defined by PDF $f(x)$ whose area under the curve is 1
- Then, we can compute $P(X \in [a, b])$ by computing the area under the curve on [a,b].

Note:

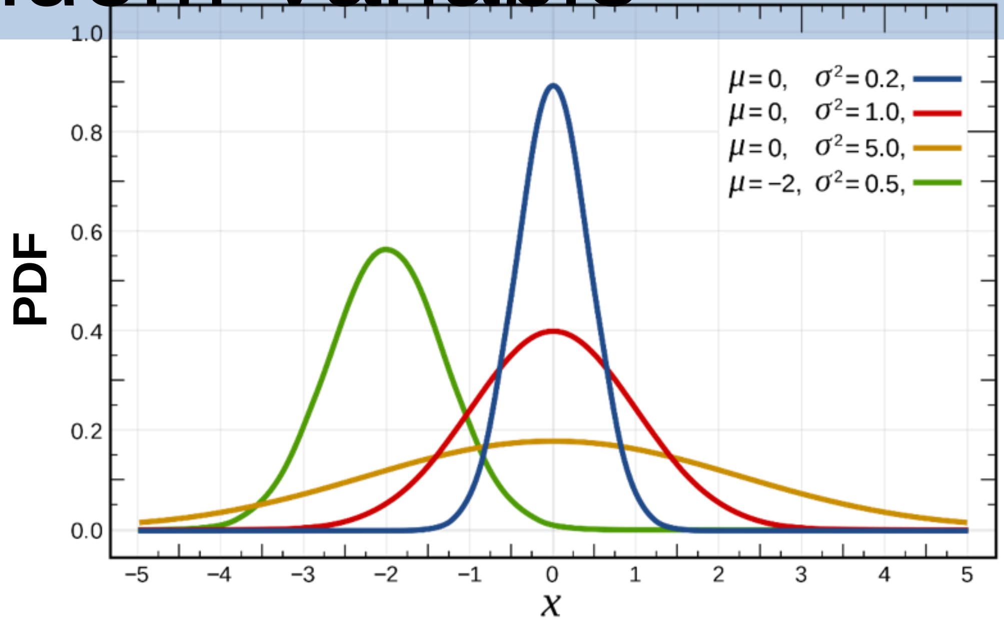
$$P(X \in [a, b]) = P(X \in (a, b]) = P\big(X \in [a, b)\big) = P\big(X \in (a, b)\big)$$

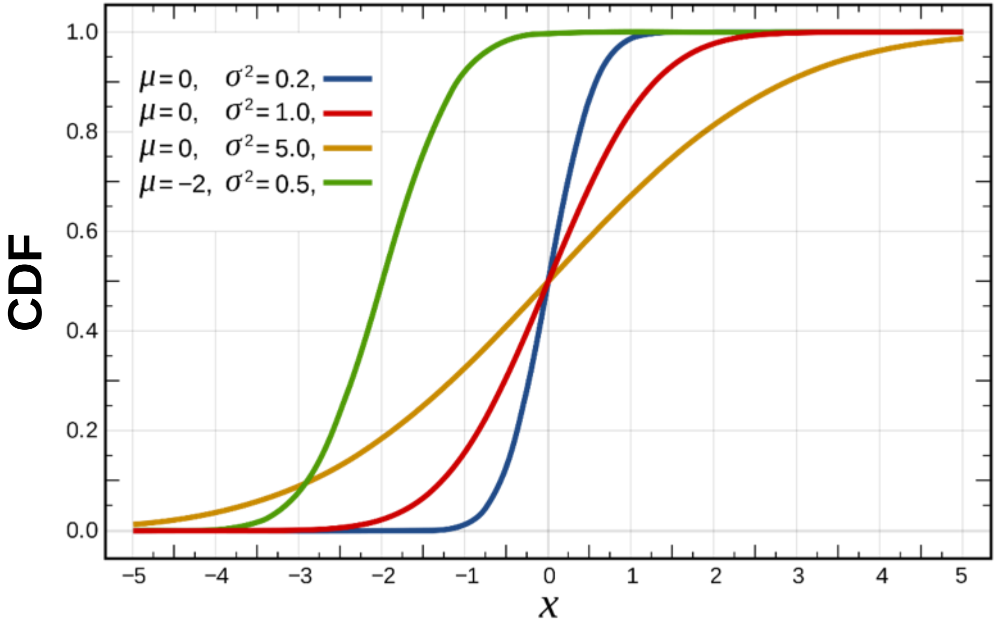
Review: Continuous Random Variable

Gaussian (a.k.a. Normal) distribution with mean mean (location) μ and variance (scale) σ^2 parameters,

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

Compactly, $X \sim \mathcal{N}(\mu, \sigma^2)$





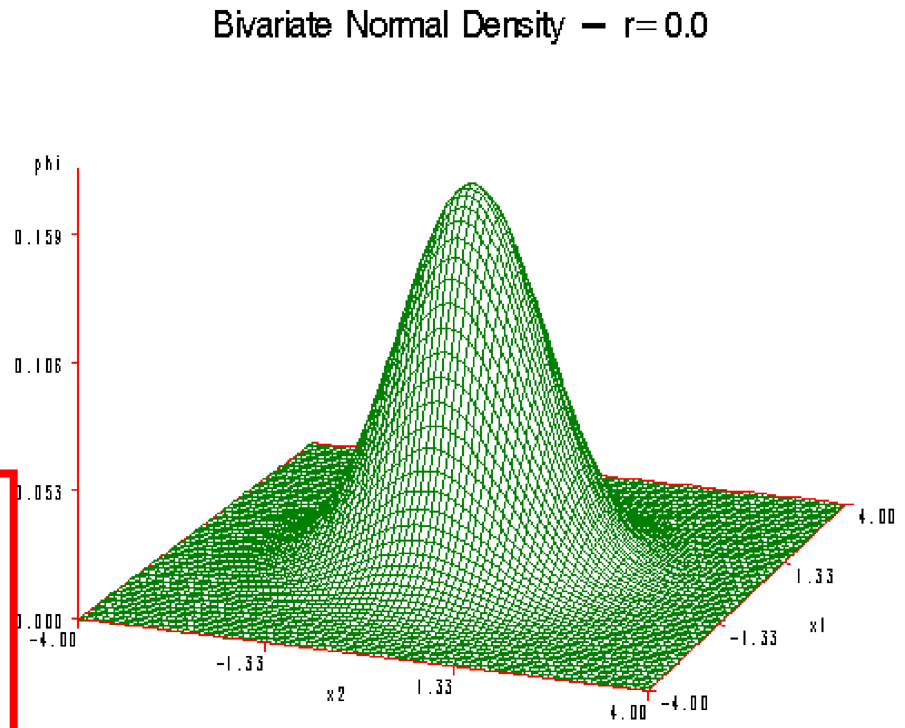
Useful Continuous Distributions

Multivariate Gaussian On RV $X \in \mathcal{R}^d$ with mean $\mu \in \mathcal{R}^d$ and positive semidefinite covariance matrix $\Sigma \in \mathcal{R}^{d \times d}$,

$\Leftrightarrow x^T \Sigma x \geq 0, \forall x$

$$p(x) = \underline{|2\pi\Sigma|^{-1/2}} \exp -\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)$$

$|A|$: matrix determinant of A



Useful Properties

- Closed under additivity (same as univariate case)
- Closed under affine transformations,

$$AX + b \sim \mathcal{N}(A\mu_x + b, A\Sigma A^T)$$

Where $A \in \mathcal{R}^{m \times d}$ and $b \in \mathcal{R}^m$ (output dimensions may change)

- Closed under conditioning and marginalization

let's cover this when needed..

The volume under the surface on a set A
= The probability of observing one of those outcomes in A!!
(i.e., $P(X \in A)$)

Moments of Random Variables

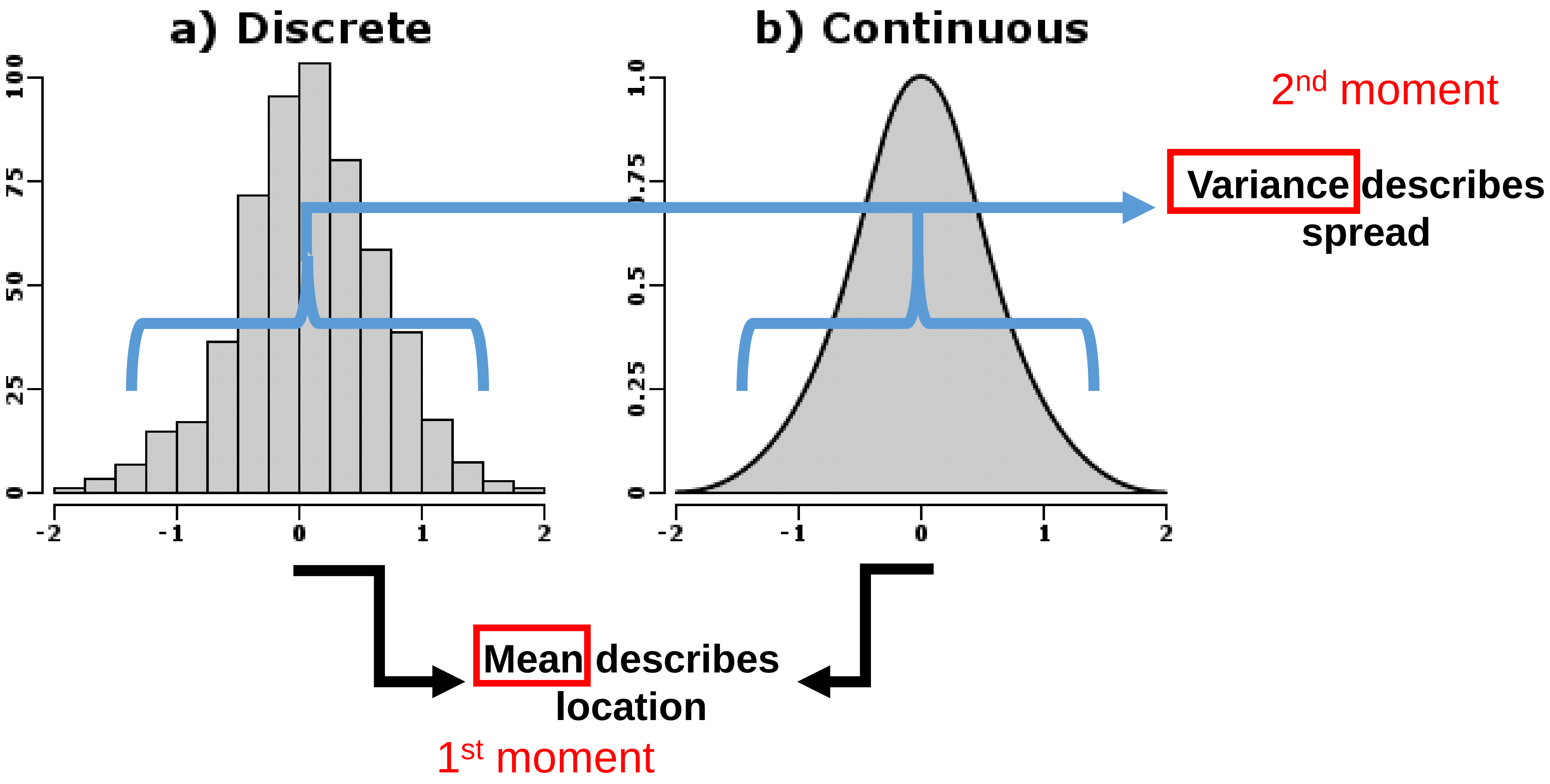
(informal introduction)

Properties of a RV are characterized by its distribution / PMF / PDF
But there are "summary" numbers capturing important characteristics
This is called "**moments**".

Moment ordinal	Moment			Cumulant	
	Raw	Central	Standardized	Raw	Normalized
1	Mean	0	0	Mean	N/A
2	–	Variance	1	Variance	1
~~3~~	~~–~~	~~–~~	~~Skewness~~	~~–~~	~~Skewness~~
~~4~~	~~–~~	~~–~~	~~(Non-excess or historical) kurtosis~~		~~Excess kurtosis~~

(Wikipedia)

Moments of Random Variables



Moments characterize properties of the distribution “shape”

Mean = Expectation = Expected Value

Definition *The* *expectation* *of a discrete RV* X*, denoted by* $\mathbf{E}[X]$*, is:*

(with PMF) $$\mathbf{E}[X] = \sum_x x \cdot p(X = x)$$

Summation over all values in domain of X

- **Effectively, a weighted average**: *each outcome weighted by probability of occurring*

Some people call it **average** rather than **mean**, but I wouldn't.

$\Rightarrow$ average is a particular 'operator': $\frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} x$

$\Rightarrow$ in data science, average is something about the data, not the distribution behind the data

Expected Value

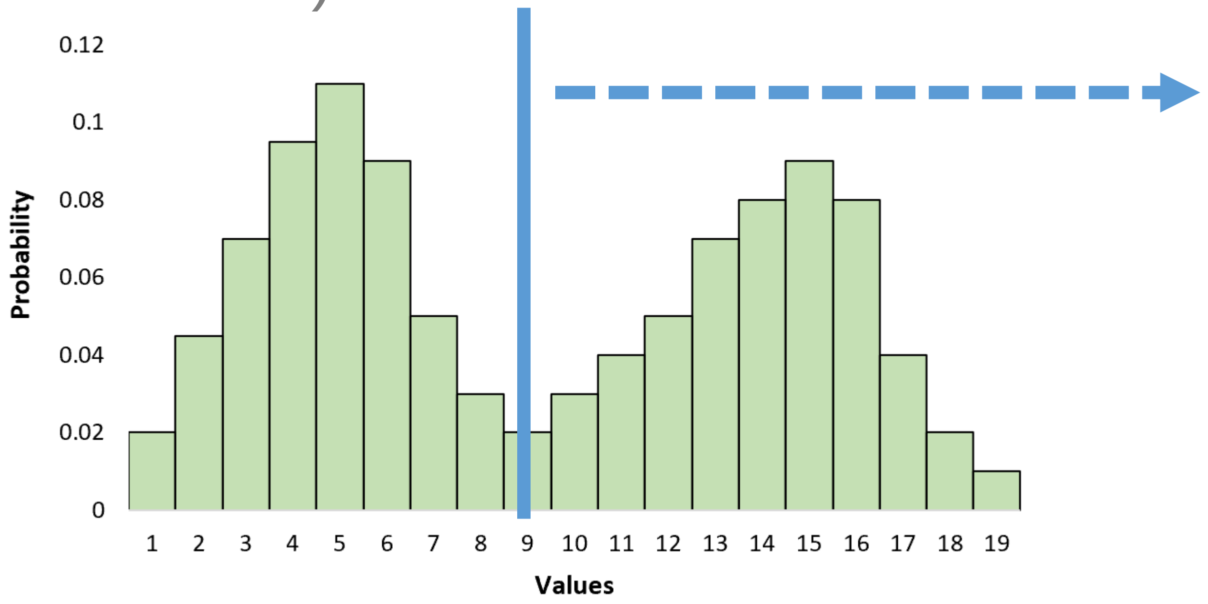
Example Let X be the sum of two fair dice, compute $\mathbf{E}[X]$:

	count	prob.
2: (1,1)	1	1/36
3: (1,2), (2,1)	2	2/36
…	..	…
6: (1,5), (2,4), (3,3), (4,2), (5,1)	5	5/36
7: (1,6), (2,5), (3,4), (4,3), (5,2), (6,1)	6	6/36
8: (2,6), (3,5), (4,4), (5,3), (6,2)	5	5/36
…	…	…
12: (6,6)	1	1/36

Expectation: $2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} +$ ... $+ 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} +$ ... $+ 12 \cdot \frac{1}{36}$ =7

Expected Value

(discrete RV)



Expected value is not always a high probability event…

…in fact, it may not even be a feasible value…

Example Let X be the result of a fair die, then:

$$\mathbf{E}[X] = \frac{1}{6} \cdot (1+2+3+4+5+6) = 3.5$$

Can't actually roll 3.5

Expected Value

Theorem (Linearity of Expectations) *For any finite collection of discrete RVs* $X_1, X_2, \ldots, X_N$ *with finite expectations,*

$$\mathbf{E}\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} \mathbf{E}[X_i]$$

E.g. for two RVs X and Y
$\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$

you do not need an independence!

Example Throw two fair dice. What is the expected sum? Let X and Y be the outcome of the first and second die, respectively. Then,

$$\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y] = 3.5 + 3.5 = 7$$

Expected Value

Before proving the theorem, a useful property:

$$\mathbf{E}[f(X,Y)] = \sum_x \sum_y f(x,y)\, P(X = x, Y = y)$$

Or simply, $\mathbf{E}[f(Z)] = \sum_z f(z) P(Z = z)$

Let $Z = f(X,Y)$

$$\sum_z z\, P(Z = z)$$

$$= \sum_z z \sum_x \sum_y P(Z = z, X = x, Y = y)$$

$$= \sum_z \sum_x \sum_y z \cdot P(Z = z, X = x, Y = y) \mathbf{I}\{f(x,y) = z\}$$

$$= \sum_x \sum_y f(x,y) \sum_z P(Z = z, X = x, Y = y)$$

$$= \sum_x \sum_y f(x,y)\, P(X = x, Y = y)$$

law of total probability (applied in the opposite way)

Expected Value

Proof: $E[X+Y] = E[X] + E[Y]$

$$\mathbf{E}[X+Y] = \sum_i \sum_j (i+j) p(X=i, Y=j)$$

Sum is linear operator

$$= \sum_i \sum_j i \cdot p(X=i, Y=j) + \sum_i \sum_j j \cdot p(X=i, Y=j)$$

Sum is linear operator

$$= \sum_i i \sum_j p(X=i, Y=j) + \sum_j j \sum_i p(X=i, Y=j)$$

Law of Total Probability

$$= \sum_i i \cdot p(X=i) + \sum_j j \cdot p(Y=j)$$

By definition of Expectation

$$= \mathbf{E}[X] + \mathbf{E}[Y]$$

Expected Value

Theorem For any random variable X and constant c,

$$\mathbf{E}[cX] = c\mathbf{E}[X]$$

Example Let X and Y be the outcome of two fair dice, then:

$$\begin{aligned} \mathbf{E}[2(X+Y)] &= \mathbf{E}[2X] + \mathbf{E}[2Y] \\ &= 2\mathbf{E}[X] + 2\mathbf{E}[Y] \\ &= 2 \cdot 3.5 + 2 \cdot 3.5 = 14 \end{aligned}$$

Caveat: c has to be a constant, not a random variable!
E.g., X: outcome of a fair die, c: outcome of another fair die

Linearity

In mathematics, a linear map or linear function $f(x)$ is a function that satisfies the two properties:[1]

- Additivity: $f(x + y) = f(x) + f(y)$.
- Homogeneity of degree 1: $f(\alpha x) = \alpha f(x)$ for all α.

So, expectation is a linear function/operator!

We will just say ”linearity of expectation”

Expected Value

Definition *The <u>conditional expectation</u> of a discrete RV* X, *given* Y *is:*

cf. $\mathbf{E}[X] = \sum_x x \cdot p(X = x)$

$$\mathbf{E}[X \mid Y = y] = \sum_x x\, p(X = x \mid Y = y)$$

Example Roll two fair dice. X_1: first die outcome, Y: sum of two dice

quiz candidate

$$\mathbf{E}[X_1 \mid Y = 5] = \sum_{x=1}^{4} x\, p(X_1 = x \mid Y = 5)$$

$$= \sum_{x=1}^{4} x \frac{p(X_1 = x, Y = 5)}{p(Y = 5)} = \sum_{x=1}^{4} x \frac{1/36}{4/36} = \frac{5}{2}$$

Conditional expectation follows properties of expectation (linearity, etc.)

Expected Value

Example: Two fair dice.

Y=outcome of die 1
X=sum of two dice

X|Y=1 ~ U{2,3,4,5,6,7} E[X|Y=1] = 4.5 $P(Y=1)=\frac{1}{6}$

X|Y=2 ~ U{3,4,5,6,7,8} E[X|Y=2] = 5.5 $P(Y=2)=\frac{1}{6}$

E[X|Y=3] = 6.5

… E[X|Y=4] = 7.5 …

E[X|Y=5] = 8.5

X|Y=6 ~ U{7,8,9,10,11,12} E[X|Y=6] = 9.5 $P(Y=6)=\frac{1}{6}$

$\mathbf{E}_X[X|Y]$ is a random variable:
$\mathbf{E}_X[X|Y] \sim U\{4.5, 5.5, 6.5, 7.5, 8.5, 9.5\}$

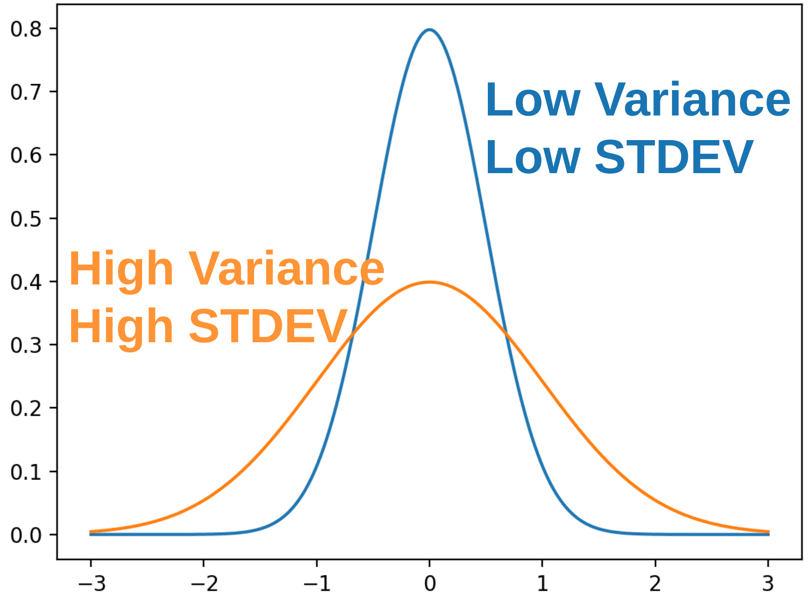
Expectation is 7
=> $\mathrm{E}_Y\left[\mathbf{E}_X[X|Y]\right] = 7$
=> coincides with $\mathbf{E}[X]$ we computed before!

Variance

Definition The variance of a RV X is defined as,

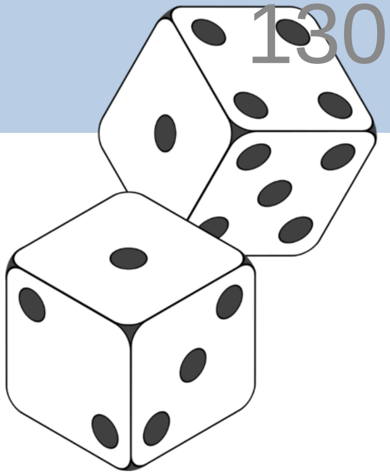
$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2]$$

The standard deviation (STDEV) is $\sigma[X] = \sqrt{\mathbf{Var}[X]}$.



- *Describes the "spread" of a distribution*
- *Describes uncertainty of outcome*
- *STDEV is in original units (more intuitive), variance is in units^2*
- *Variance is more mathematically useful than STDEV*

Variance



Example Let X be the result of a fair six-sided die.

The variance is then,

$$\mathrm{Var}(X) = \sum_{i=1}^{6} \frac{1}{6}\left(i - \frac{7}{2}\right)^2$$

$$= \frac{1}{6}\left((-5/2)^2 + (-3/2)^2 + (-1/2)^2 + (1/2)^2 + (3/2)^2 + (5/2)^2\right)$$

$$= \frac{35}{12} \approx 2.92.$$

The STDEV is $\sqrt{\mathrm{Var}(X)} \approx 1.71$, which suggests we should expect outcomes to vary around the mean of 3.5 by $\pm$ 1.71

Variance

Lemma An equivalent form of variance is:

$$\mathbf{Var}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$$

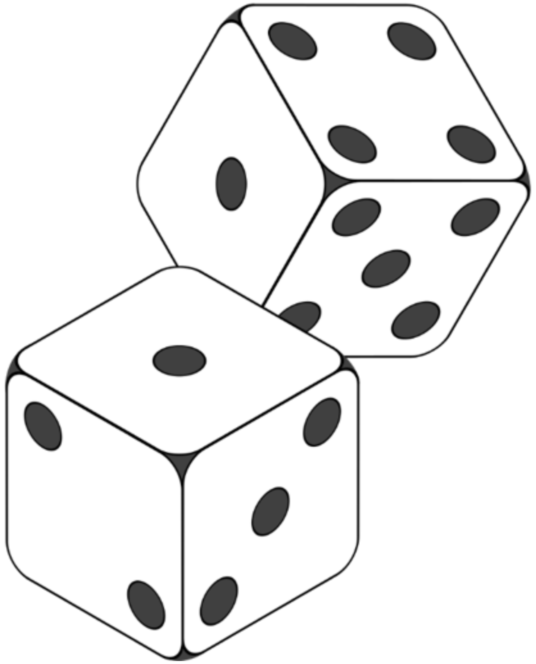
Proof

$$\begin{aligned}
\mathbf{E}[(X - \mathbf{E}[X])^2] &= \mathbf{E}[X^2 - 2X\mathbf{E}[X] + \mathbf{E}[X]^2] && \textbf{(Expand it)} \\
&= \mathbf{E}[X^2] - 2\mathbf{E}[X]\mathbf{E}[X] + \mathbf{E}[X]^2 && \textbf{(Linearity of expectations)} \\
&= \mathbf{E}[X^2] - 2\mathbf{E}[X]^2 + \mathbf{E}[X]^2 && \textbf{(Algebra)} \\
&= \mathbf{E}[X^2] - \mathbf{E}[X]^2 && \textbf{(Algebra)}
\end{aligned}$$

Variance

Example General form of variance for a fair **n-sided** fair die,

$$
\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}\left(X^2\right) - (\mathrm{E}(X))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} i\right)^2 \\
&= \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 \\
&= \frac{n^2-1}{12}.
\end{aligned}
$$



Variance

- If c is a constant, $Var[cX] = c^2Var[X]$

- Important that c has to be a constant here!

Moments of Useful Discrete Distributions

Bernoulli *A.k.a. the* ***coinflip*** *distribution on* <u>*binary*</u> *RVs* $X \in \{0, 1\}$

$$p(X) = \pi^X (1-\pi)^{(1-X)}$$

Where π *is the probability of* ***success*** *(i.e., heads), and also the mean*

$$\mathbf{E}[X] = \pi \cdot 1 + (1-\pi) \cdot 0 = \pi \qquad \mathbf{Var}[X] = \pi(1-\pi)$$

Binomial Sum of N independent coinflips,

$$p(Y = k) = \binom{N}{k} \pi^k (1-\pi)^{N-k}$$

With moments,

$$\mathbf{E}[Y] = N \cdot \pi \qquad \mathbf{Var}[Y] = N\pi(1-\pi)$$

^: (by linearity of expectation)

Moments of Useful Discrete Distributions

Multinomial distribution: Let $X_1, \ldots, X_K$ be the count of *N* independent categorical RVs

$$p(x_1, \ldots, x_K) = \frac{N!}{x_1! x_2! \ldots x_K!} \prod_{k=1}^{K} \pi_k^{x_k}$$

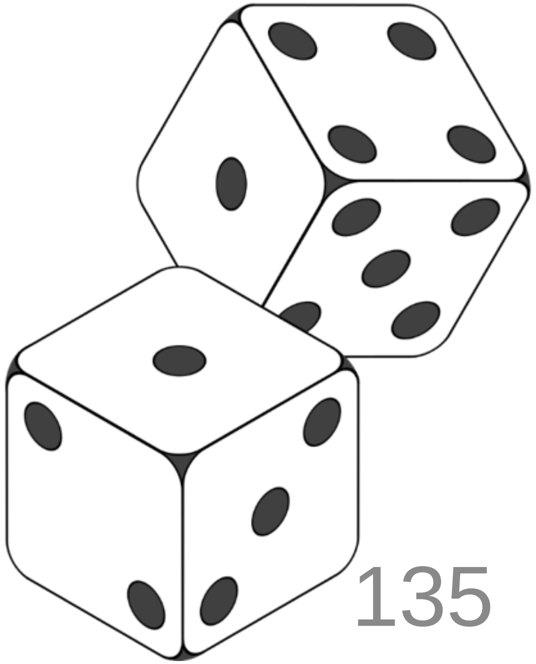
Where parameter $\pi \in [0,1]^K$ is a probability vector,

$$\sum_{k=1}^{K} \pi_k = 1$$

Marginal moments are given by,

$$\mathbf{E}[X_k] = N\pi_k \qquad \mathbf{Var}[X_k] = N\pi_k(1-\pi_k)$$

Moments are similar to Binomial, but over K outcomes



Covariance

Definition *The covariance of two RVs* X *and* Y *is defined as,*

$$\mathbf{Cov}(X,Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$$

Question *What is Cov(X,X)?*

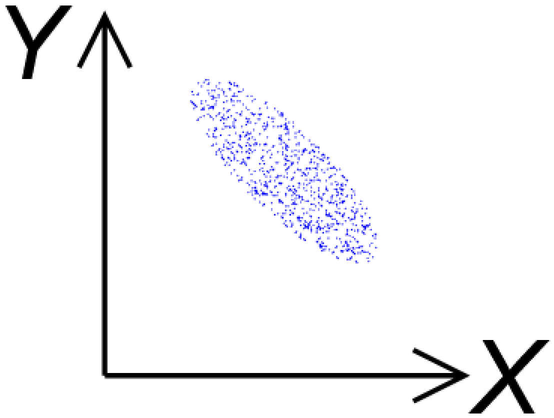
Answer *Cov(X,X) = Var(X)*

Covariance

Definition *The* *<u>covariance</u>* *of two RVs* X *and* Y *is defined as,*

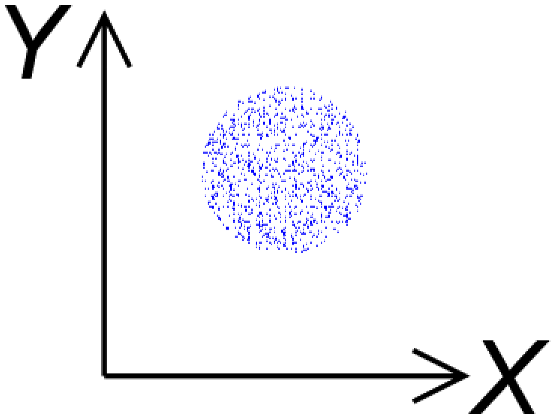
$$\mathbf{Cov}(X,Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$$

Measures the <u>linear relationship</u> between X and Y

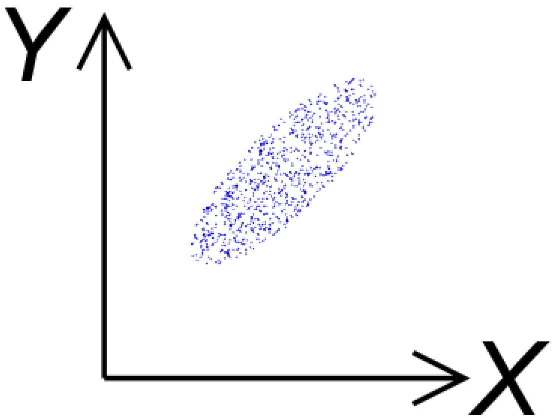


$\operatorname{cov}(X,Y) < 0$

inversely proportional



$\operatorname{cov}(X,Y) \approx 0$



$\operatorname{cov}(X,Y) > 0$

proportional

Example: height vs weight

Covariance

- A shortcut to compute covariance.
- $Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$

$$= E\left[XY - X \cdot E[Y] - Y \cdot E[X] + E[X]E[Y]\right]$$

$$= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y]$$

$$= E[XY] - E[X]E[Y]$$

- Safety check: $Cov(X,X) = E[XX] - E[X]E[X] = Var(X)$

Covariance

Lemma *For any two RVs* X *and* Y *,*

$$\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}(X,Y)$$

=> *variance is* *<u>not a linear operator</u>*.

Proof $\mathbf{Var}[X+Y] = \mathbf{E}[(X+Y-\mathbf{E}[X+Y])^2]$

(Linearity of expt.) $= \mathbf{E}[(X+Y-\mathbf{E}[X]-\mathbf{E}[Y])^2]$

(Distributive property) $= \mathbf{E}[(X-\mathbf{E}[X])^2 + (Y-\mathbf{E}[Y])^2 + 2(X-\mathbf{E}[X])(Y-\mathbf{E}[Y])]$

(Linearity of expt.) $= \mathbf{E}[(X-\mathbf{E}[X])^2] + \mathbf{E}[(Y-\mathbf{E}[Y])^2] + 2\mathbf{E}[(X-\mathbf{E}[X])(Y-\mathbf{E}[Y])]$

(Definition of Var / Cov) $= \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}(X,Y)$

Correlation

Definition *The correlation of two RVs X and Y is given by,*

$$\mathbf{Corr}(X,Y) = \frac{\mathbf{Cov}(X,Y)}{\sigma_X \sigma_Y} \quad \textit{where} \quad \sigma_X = \sqrt{\mathbf{Var}(X)}$$

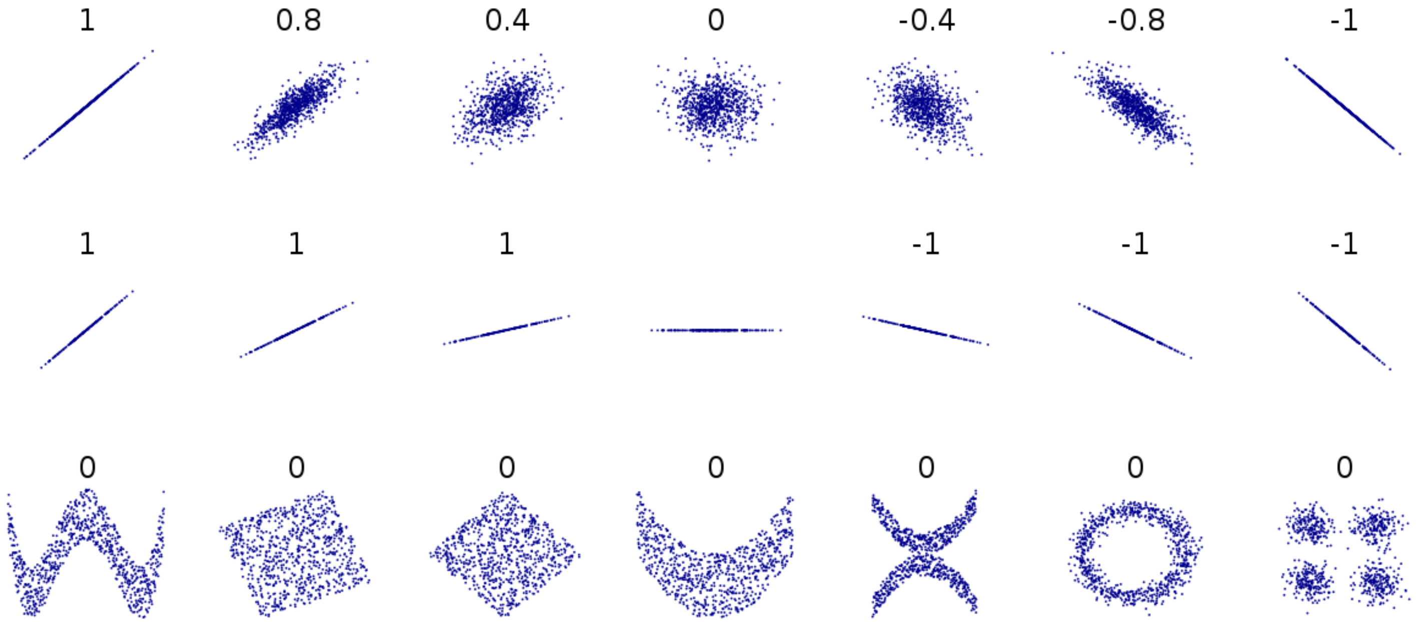
Normalized version of covariance!
⇒ Always between -1 and 1

Useful when you are interested in how X and Y are related, independent of the individual variability.
⇒ $Cov(cX, dY) \neq Cov(X, Y)$ **but** $Corr(cX, dY) = Corr(X, Y)$

Correlation

Definition *The correlation of two RVs X and Y is given by,*

$$\mathbf{Corr}(X,Y) = \frac{\mathbf{Cov}(X,Y)}{\sigma_X \sigma_Y} \quad \textit{where} \quad \sigma_X = \sqrt{\mathbf{Var}(X)}$$



Like covariance, only expresses linear relationships!

Computer Science

CSC380: Principles of Data Science

Probability Primer 6

Kwang-Sung Jun

TA: Yang Hong, Tuan Nguyen

based on Jason Pacheco’s slides

Independence and Moments

Theorem: *If* $X \perp Y$ *then* $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$.

Comparison: $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$ regardless of independence!

Independence and Moments

Theorem: *If* $X \perp Y$ *then* $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$.

Proof:

$$\mathbf{E}[XY] = \sum_x \sum_y (x \cdot y) p(X = x, Y = y)$$

$$= \sum_x \sum_y (x \cdot y) p(X = x) p(Y = y) \quad \textbf{(Independence)}$$

$$= \left(\sum_x x \cdot p(X = x) \right) \left(\sum_y y \cdot p(Y = y) \right) = \mathbf{E}[X]\mathbf{E}[Y] \quad \textbf{(Linearity of Sum)}$$

Example *Let* $X_1, X_2 \in \{1, \ldots, 6\}$ *be RVs representing the result of rolling two fair standard dice.* *What is the mean of their product?*

$$\mathbf{E}[X_1 X_2] = \mathbf{E}[X_1]\mathbf{E}[X_2] = 3.5^2 = 12.25$$

Independence and Moments

Question: *What is the variance of their sum (recall independence)?*

$$\mathbf{Var}[X_1 + X_2] = \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2\mathbf{Cov}(X_1, X_2)$$

$$= \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2\mathbf{E}[(X_1 - \mathbf{E}[X_1])(X_2 - \mathbf{E}[X_2])]$$

$$= \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2\mathbf{E}[(X_1 - \mathbf{E}[X_1])]\mathbf{E}[(X_2 - \mathbf{E}[X_2])]$$ $Y_1 \perp Y_2 \Rightarrow f(Y_1) \perp f(Y_2)$

$$= \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2\left(\mathbf{E}[X_1] - \mathbf{E}[X_1]\right)\left(\mathbf{E}[X_2] - \mathbf{E}[X_2]\right)$$

$$= \mathbf{Var}[X_1] + \mathbf{Var}[X_2]$$

Independence and Moments

Recall that for any two RVs X and Y variance is not a linear function,

$$\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}(X,Y)$$

<u>If X and Y are independent then they have zero covariance</u>,

$$\mathbf{Cov}(X,Y) = 0$$

Thus, $$\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y]$$

And, for a collection of independent RVs $X_1, X_2, \ldots, X_N$ we have,

$$\mathbf{Var}(\sum_{i=1}^N X_i) = \sum_{i=1}^N \mathbf{Var}(X_i)$$

Q: Is variance is a linear operator under independence?

A: No! $Var(cX) \neq c\,Var(X)$ for a constant c. Rather, $Var(cX) = c^2 Var(X)$.

Example: Independent Gaussian RVs

Let X and Y be **independent** Gaussian random variables with,

$$X \sim \mathcal{N}(\mu_x, \sigma_x^2) \qquad\qquad Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$$

(Property of Gaussian: $\mathrm{E}[X] = \mu_x, Var[X] = \sigma_x^2$)

What is the variance of their sum?

$$\mathbf{Var}(X+Y) = \mathbf{Var}(X) + \mathbf{Var}(Y) = \sigma_x^2 + \sigma_y^2$$

What is the mean of their product?

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y] = \mu_x \mu_y$$

Suppose X and Y are **dependent**, what is the mean of their sum?

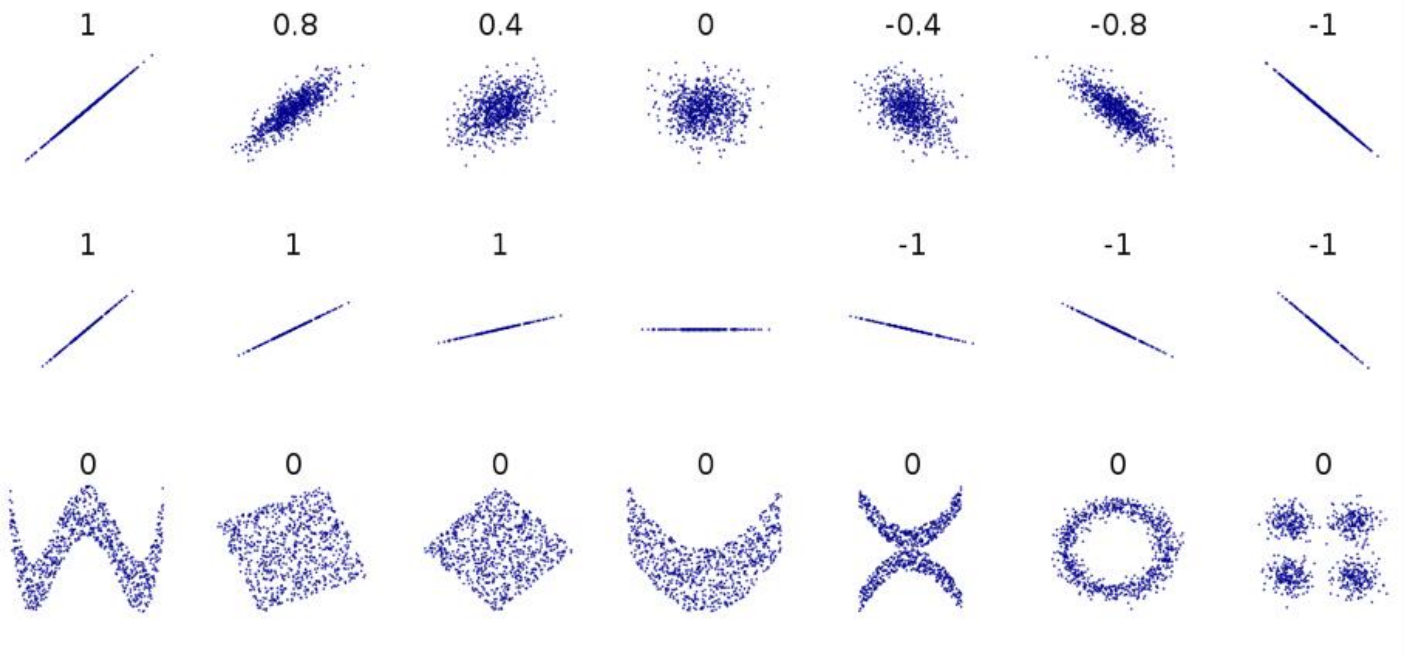
$$\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y] = \mu_x + \mu_y$$

Independence and Moments

From previous slide If X and Y are independent random variables, then:

$$\mathbf{Cov}(X,Y) = 0$$

The reverse is not true! $(\mathbf{Cov}(X,Y) = 0) \not\Rightarrow X \perp Y$



Counter Example

- Let X, Z be independent random variable that is -1 or $+1$ with probability 0.5.
- Let $Y = Z \cdot I\{X = 1\}$
- Claim: Cov(X,Y)=0 but X and Y are dependent.

Recall: $Cov(X,Y) = E[XY] - E[X]E[Y]$

E[X] = 0

E[Y] = 0

$$
\begin{aligned}
E[XY] &= (-1) \cdot 0 \cdot P(X = -1) \\
&+ 1 \cdot 1 \cdot P(X = 1, Y = 1) \\
&+ 1 \cdot (-1) \cdot P(X = 1, Y = -1) \\
&= 0.
\end{aligned}
$$

Q: how to check independence between X and Y?

P(Y=1 | X=-1) = 0, P(Y=1) = .25 ⇒ not independent

Moments of Continuous RVs

Replace all sums with integrals,

$$\mathbf{E}[X] = \int xp(x)\, dx \qquad\qquad \mathbf{Var}[X] = \int (x - \mathbf{E}[X])^2 p(x)\, dx$$

- All properties push through, as you would expect (e.g. law of total expectation, conditional expectation, etc.)

(and use PDF p(x) instead of PMF P(X=x))

Review

We have covered a lot of ground on probability in short time…

Discrete Random Processes

- Definition of sample space / random events
- Axioms of probability
- Uniform probability of random event
- Random Variables
- Fundamental rules of probability (chain rule, conditional, law of total probability)

Probability Distributions

- Useful discrete probability mass functions'
- Introduction to continuous probability
- Useful probability density functions

Moments / Independence

- Expected Value
- Linearity / Law of total expectation
- Variance, Covariance, Corr.
- Dependent / Independent RVs

Exercise

<u>*Question:*</u> *Roll two dice and let their outcomes be* $X_1, X_2 \in \{1, \ldots, 6\}$ *for die 1 and die 2, respectively. Recall the definition of conditional probability,*

$$p(X_1 \mid X_2) = \frac{p(X_1, X_2)}{p(X_2)}$$

Which of the following are true?

a) $p(X_1 = 1 | X_2 = 1) > p(X_1 = 1)$

b) $\boxed{p(X_1 = 1 | X_2 = 1) = p(X_1 = 1)}$ Outcome of die 2 doesn't *affect* die 1

c) $p(X_1 = 1 | X_2 = 1) < p(X_1 = 1)$

Exercise

<u>Question:</u> *Let* $X_1 \in \{1, \ldots, 6\}$ *be outcome of die 1, as before. Now let* $X_3 \in \{2, 3, \ldots, 12\}$ *be the* sum of both dice. Which of the following are true?

$\boxed{\text{a) } p(X_1 = 1 | X_3 = 3) > p(X_1 = 1)}$

b) $p(X_1 = 1 | X_3 = 3) = p(X_1 = 1)$

c) $p(X_1 = 1 | X_3 = 3) < p(X_1 = 1)$

Only 2 ways to get $X_3 = 3$ *, each with equal probability:*

$$(X_1 = 1, X_2 = 2) \quad \textit{or} \quad (X_1 = 2, X_2 = 1)$$

so

$$p(X_1 = 1 \mid X_3 = 3) = \frac{1}{2} > \frac{1}{6} = p(X_1 = 1)$$