



CSC380: Principles of Data Science

Probability Primer

Chicheng Zhang

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)

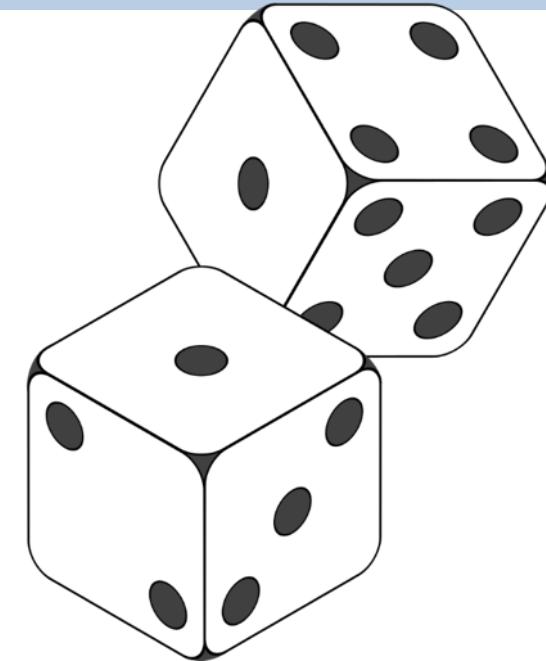
Random Events and Probability

Suppose we roll two fair dice...

- 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?



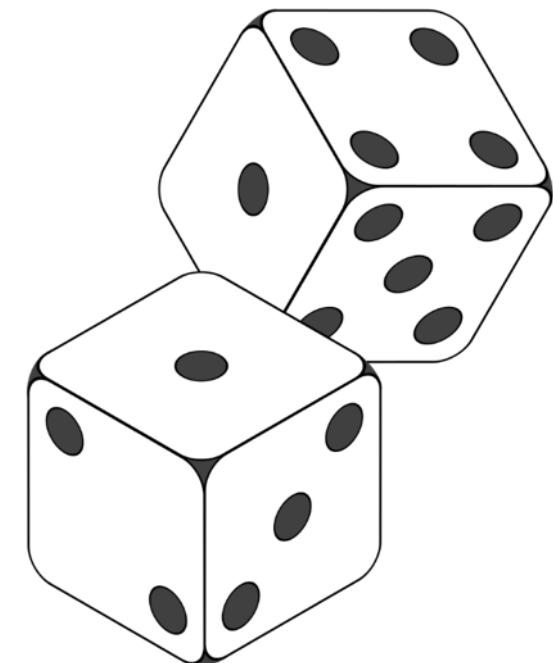
Probability: intuition

- 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
```

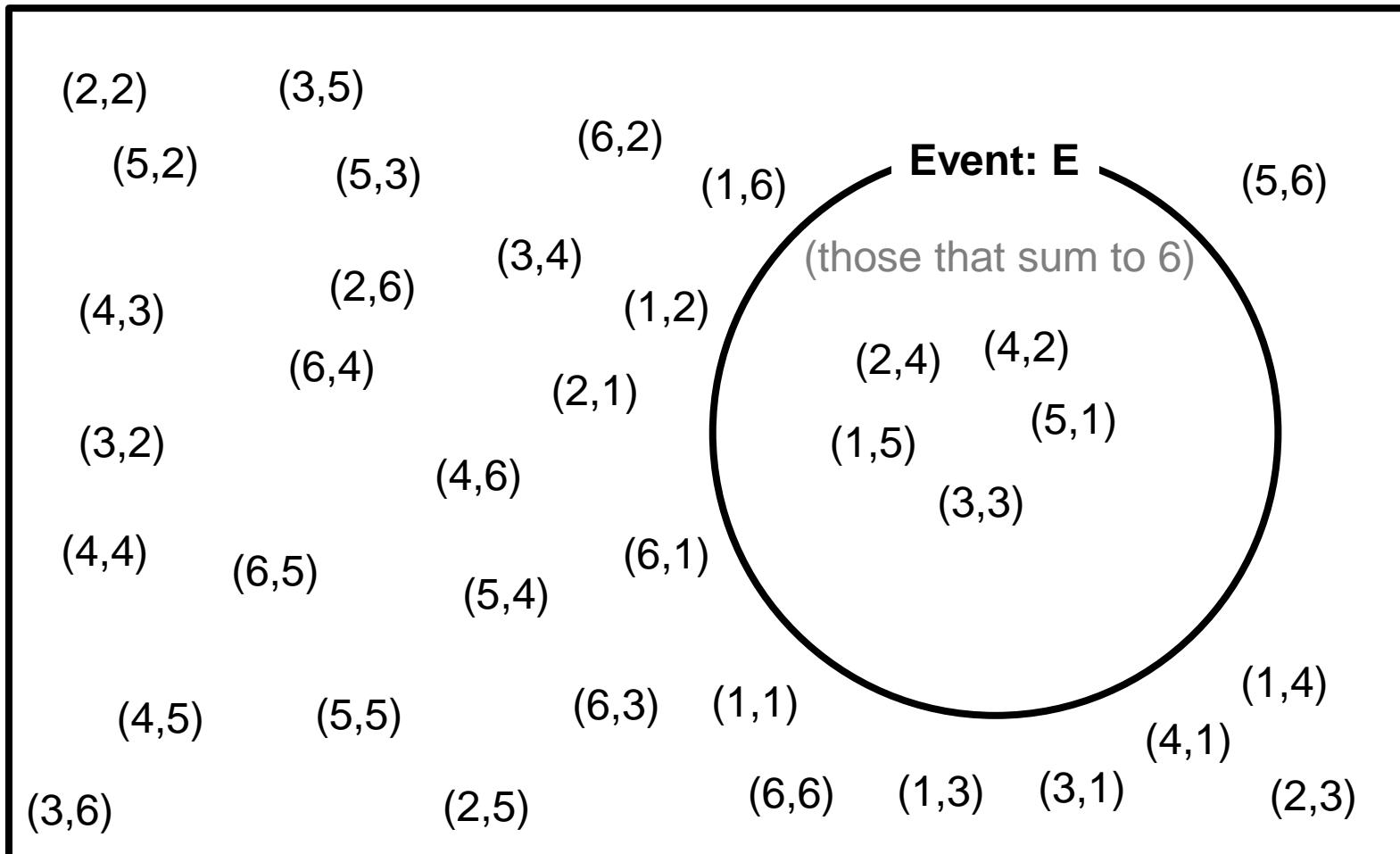
every time you run, you
get a different result

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:
 $\Rightarrow 1/36$

of outcomes that sum to 6:
 $\Rightarrow 5$

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

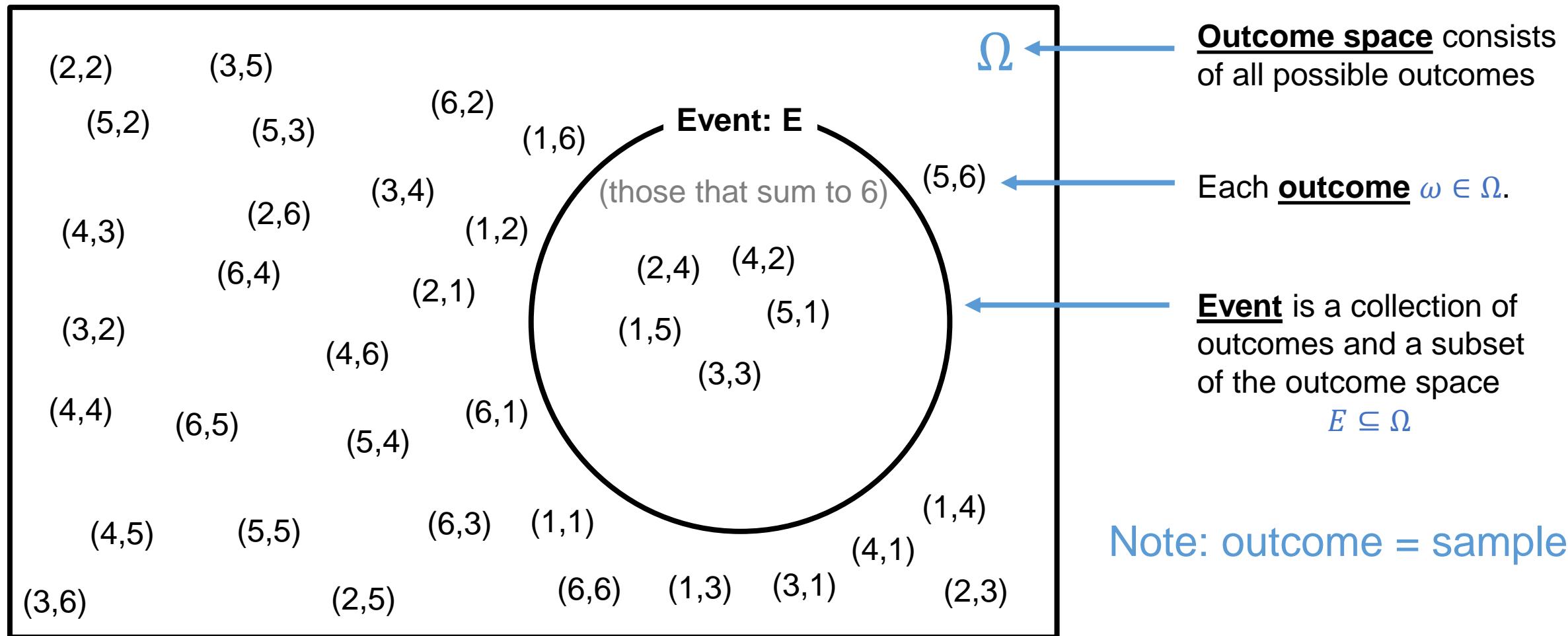
Mathematics of Probability

8

- **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
- **Disclaimer:** 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?



Random Events and Probability

Some examples of events...

- Both even numbers

Q: how many such pairs?

9

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

- The sum of both dice is even,

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

- 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?

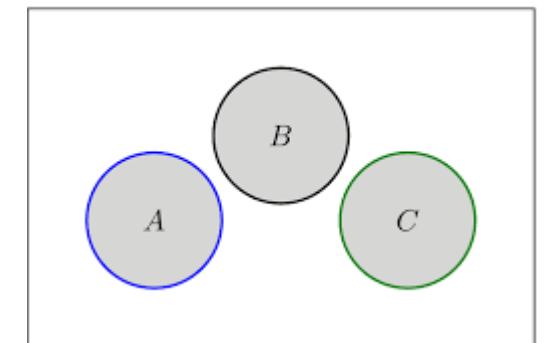
(e.g., can explain the probability of seeing an event when throwing two dice)

Mathematicians have found a set of conditions that ‘makes sense’.

- Probability is a map P . \Rightarrow 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 , $P(E) \geq 0$
2. $P(\Omega) = 1$
3. For any *finite or countably infinite* sequence of disjoint events E_1, E_2, E_3, \dots

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



disjoint: intersection is empty

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)

$$\mathbb{P}(\emptyset) = 0$$

$$A \subset B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$$

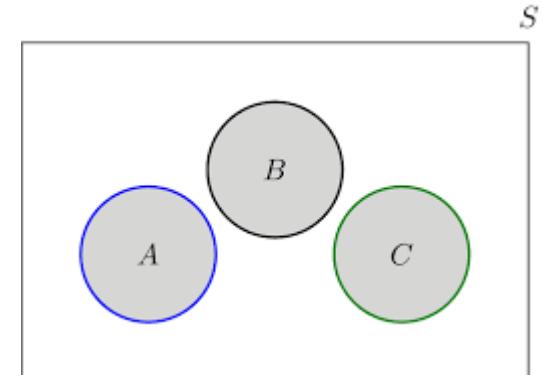
E.g., throw a die. A= getting 1, B=getting an odd number

$$0 \leq \mathbb{P}(A) \leq 1$$

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$

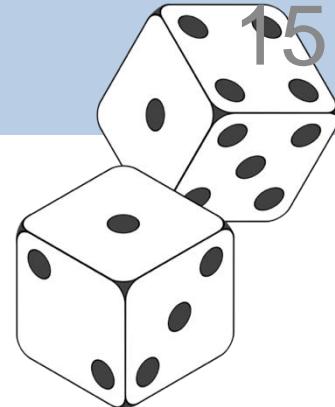
$$A \cap B = \emptyset \implies \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

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,

$$\begin{aligned} P(\{(2,2), (2,4), \dots, (6,6)\}) &= P(\{(2,2)\}) + P(\{2,4\}) + \dots + P(\{6,6\}) \\ &= \frac{1}{36} + \dots + \frac{1}{36} = \frac{9}{36} \end{aligned}$$

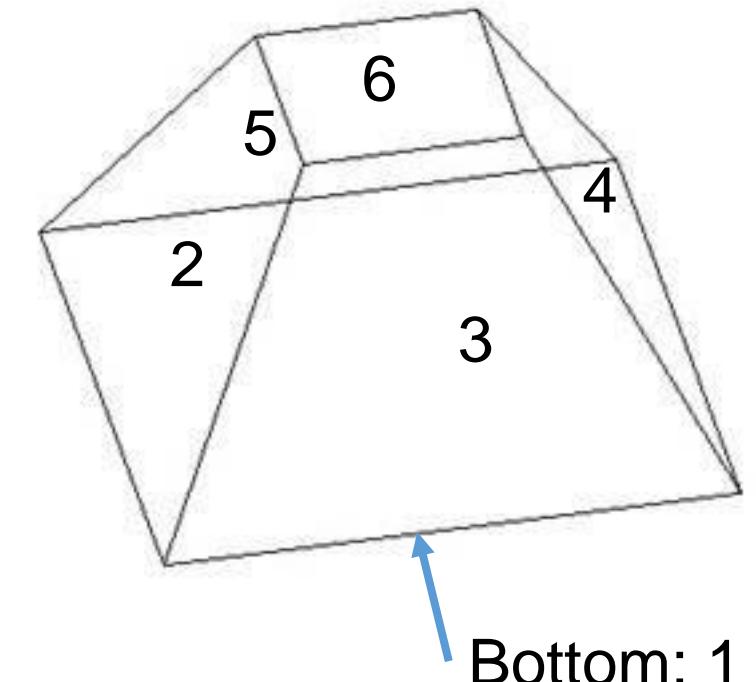
9 Possible outcomes, each with
equal probability of occurring

Unfair die example

- Let A be the outcome of a single throw.
- $P(A=1) \ll P(A=2) = \dots = P(A=5) \ll P(A=6)$

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

$$E_1 = \{(1, 1), (1, 2), \dots, (1, 6)\}$$

E_2 : Second die equals 1

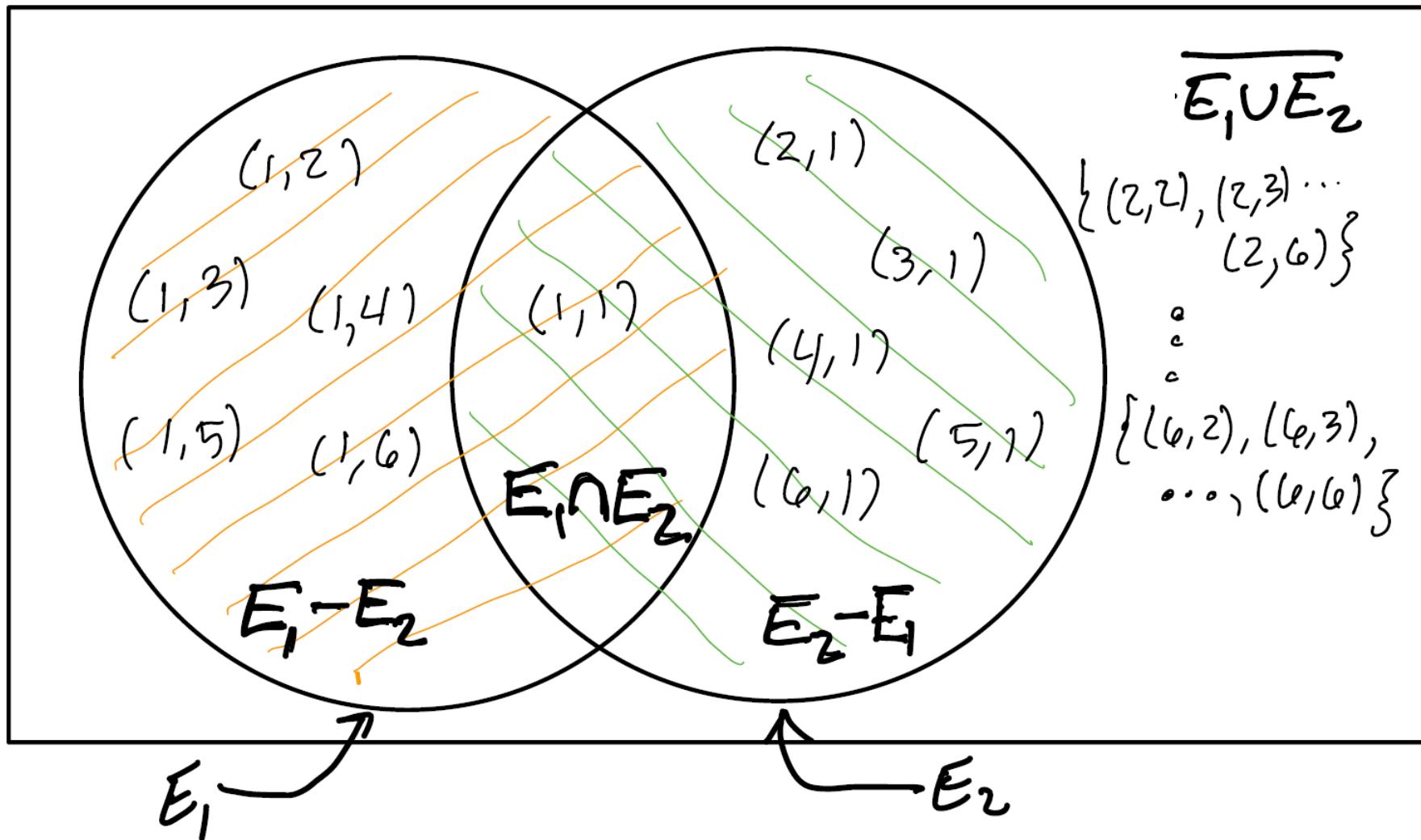
$$E_2 = \{(1, 1), (2, 1), \dots, (6, 1)\}$$

Operators on events:

Operation	Value	Interpretation
$E_1 \cup E_2$	$\{(1, 1), (1, 2), \dots, (1, 6), (2, 1), \dots, (6, 1)\}$	Any die rolls 1
$E_1 \cap E_2$	$\{(1, 1)\}$	Both dice roll 1
$E_1 \setminus E_2$ <small>(= $E_1 - E_2$:= $E_1 \cap E_2^c$)</small>	$\{(1, 2), (1, 3), (1, 4), (1, 5), (1, 6)\}$	Only the first die rolls 1
$\overline{E_1 \cup E_2}$ <small>(= $(E_1 \cup E_2)^c$)</small>	$\{(2, 2), (2, 3), \dots, (2, 6), (3, 2), \dots, (6, 6)\}$	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)

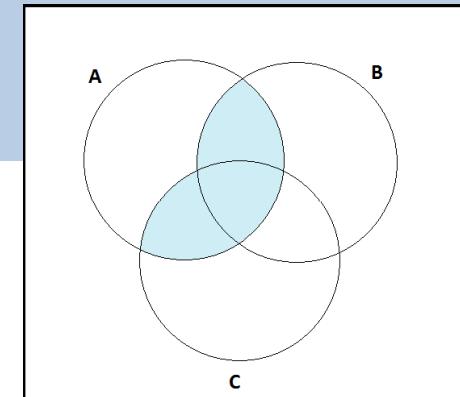
Event Language	Set Language	Set Notation
sample space	universal set	Ω
event	subset	A, B, C, \dots
outcome	element	ω
impossible event	empty set	\emptyset

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

20

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

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

- $\neg(\cup_n A_n) = \cap_n \neg A_n$, $\neg(\cap_n A_n) = \cup_n \neg A_n$ DEMORGAN

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

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

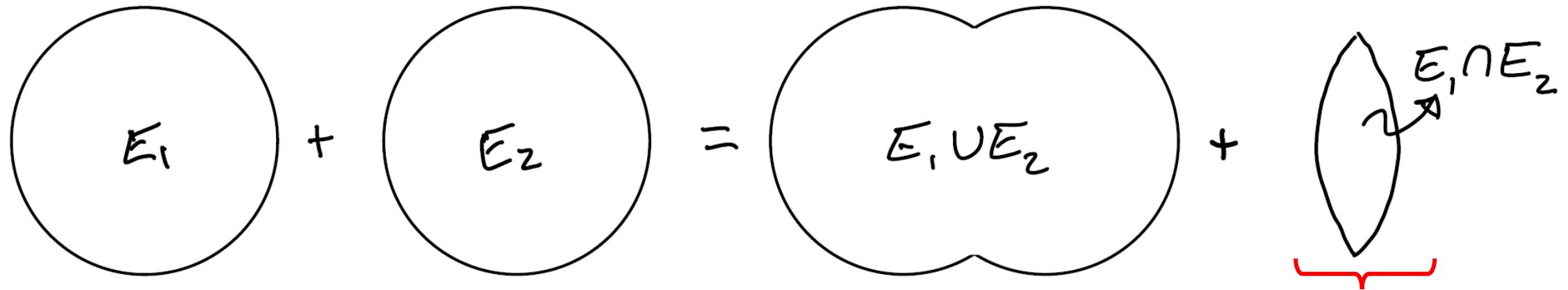
- $B = \Omega \cap B = (A \cup \neg A) \cap B = (A \cap B) \cup (\neg A \cap B)$ // by distributive law

TIP: always draw pictures to visualize these identities!

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:



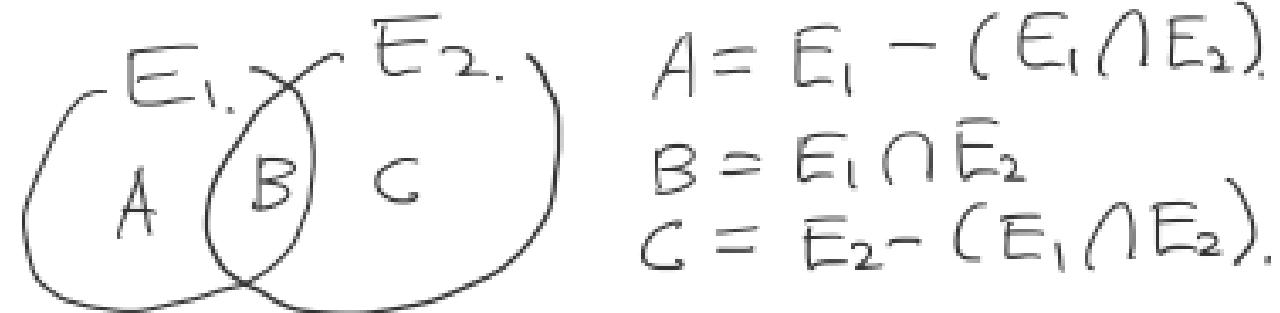
Subtract from both sides

Alternative Proof

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

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

Formal proof:



$$\begin{aligned} A &= E_1 - (E_1 \cap E_2) \\ B &= E_1 \cap \bar{E}_2 \\ C &= \bar{E}_2 - (E_1 \cap E_2). \end{aligned}$$

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

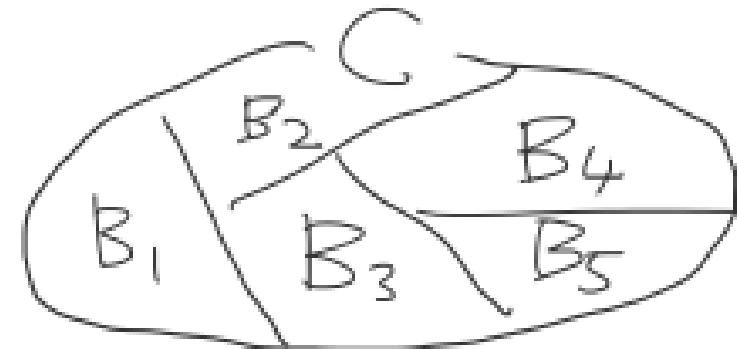
Exercise

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

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, \dots are disjoint.
Here, n can be infinity.



Law of total probability: Let A be an event. For events B_1, B_2, \dots that partitions Ω , we have

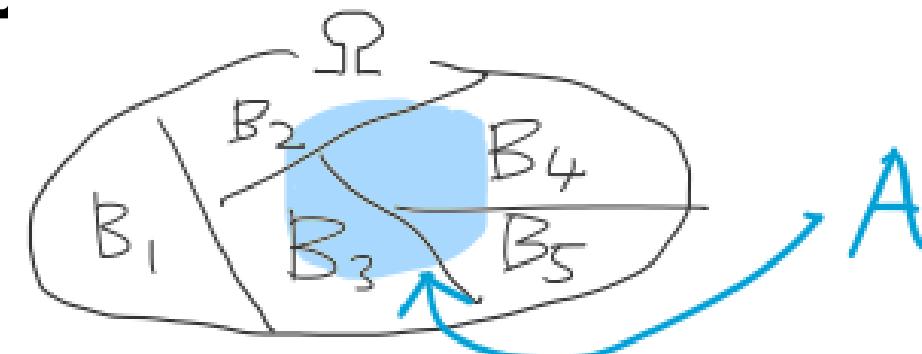
$$P(A) = \sum_i P(A \cap B_i)$$

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

Law of total probability: Let A be an event. For any events B_1, B_2, \dots that partitions Ω , we have

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

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

quiz candidate

$$\begin{aligned} p(Y = 6) &= \sum_{x=1}^6 p(Y = 6, X = x) \\ &= p(Y = 6, X = 1) + p(Y = 6, X = 2) + \dots + p(Y = 6, X = 6) \\ &= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + 0 = \frac{5}{36} \end{aligned}$$

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

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 $P(\text{each piece of partition})$ 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.
- Given this situation,** find the probability of two dice summing to 6.

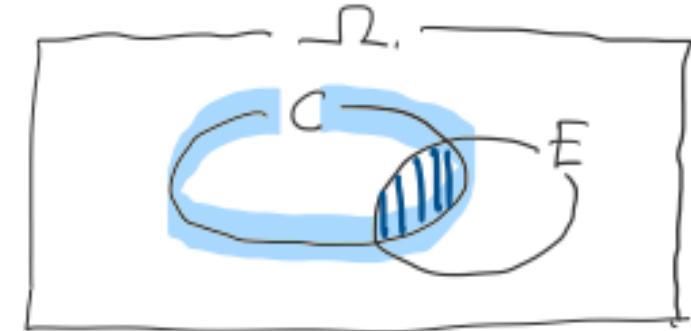
```
import numpy as np
for n in [10,100,1000,10_000,100_000, 1_000_000]:
    res_dice1 = np.random.randint(6,size=n) + 1
    res_dice2 = np.random.randint(6,size=n) + 1
    res = [(res_dice1[i], res_dice2[i]) for i in range(len(res_dice1))]
```

```
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))
```

```
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
```



compare:
without conditioning,
it was 0.138..

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 Items

- Participation credits (10pts)
 - Q&A on Piazza
- Reading quiz (~3 times in total) – I will announce the reading materials in advance
- Posting weekly review questions on Piazza
 - Questions for materials last week
 - I will ask a subset of you to do so every week & rotate

Administrative Items

- 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 .3 (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() < 0.3` will be ‘True’ with probability 0.3 and ‘False’ w.p. 0.7

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

- Problem 3 c) uses the concept of random variables
- We will get to that in the beginning of next lecture, at the latest

- 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...



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

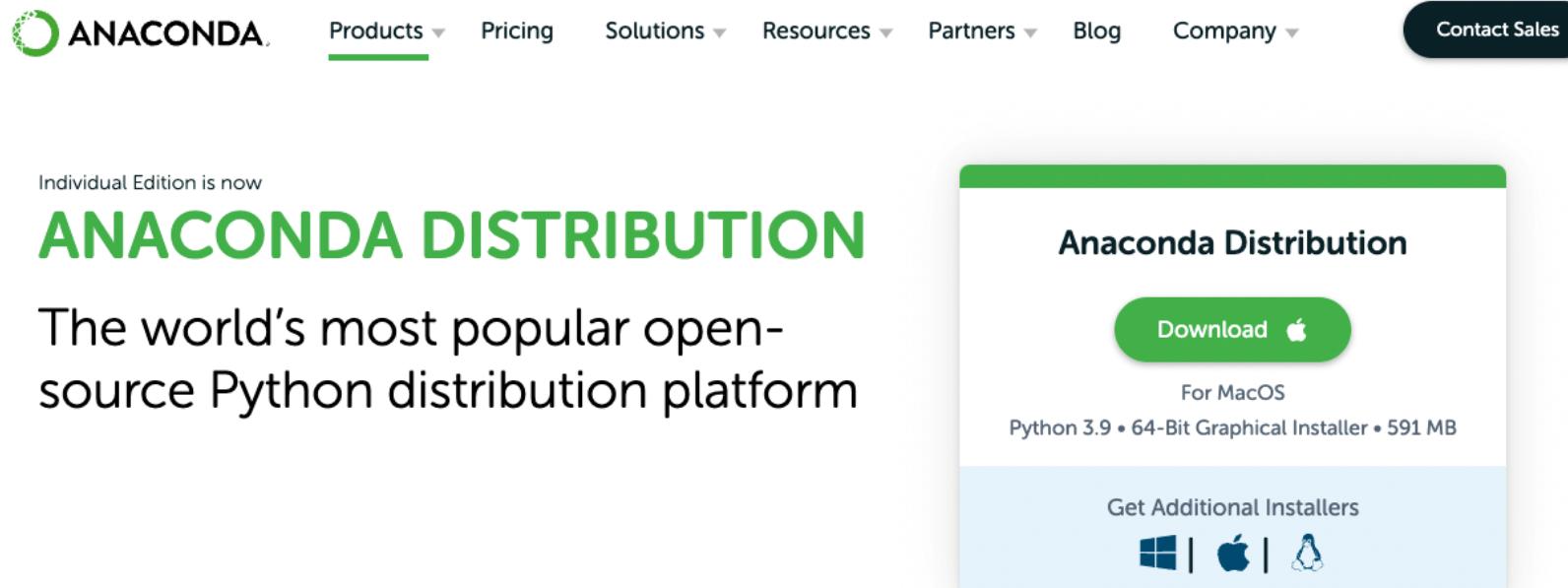
If you use `pip`, you can install NumPy with:

```
pip install numpy
```

...we are interested in `numpy.random` at the moment

Numpy Library

Package containing many useful numerical functions...



The screenshot shows the Anaconda website's main navigation bar with links for Products, Pricing, Solutions, Resources, Partners, Blog, Company, and Contact Sales. Below the navigation, a message states "Individual Edition is now ANACONDA DISTRIBUTION". The main heading "ANACONDA DISTRIBUTION" is displayed in large green letters. Below it, the text "The world's most popular open-source Python distribution platform" is written. To the right, there is a callout box for the "Anaconda Distribution" featuring a "Download" button with an Apple icon, a link for "For MacOS", and a note about "Python 3.9 • 64-Bit Graphical Installer • 591 MB". At the bottom of the callout, there is a link "Get Additional Installers" followed by icons for Windows, Mac, and Linux.

`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

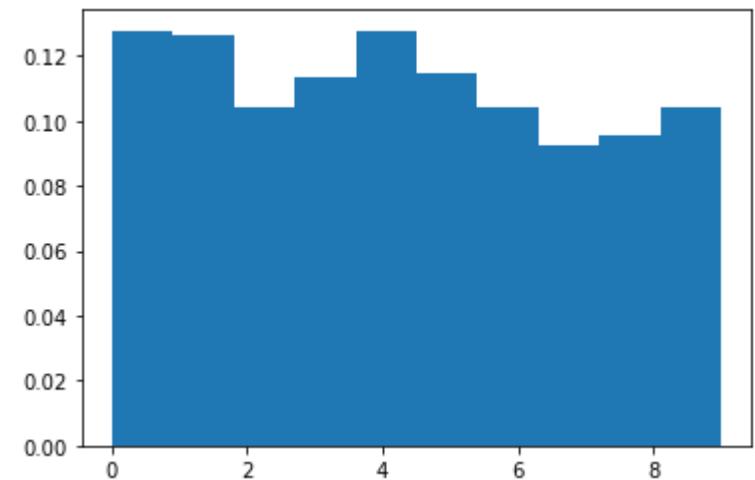
```
numpy.random.randint(low, high=None, size=None, dtype='l')
```

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 data shuffling and algorithms 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. C
- Given this situation**, find the probability of two dice summing to 6. E

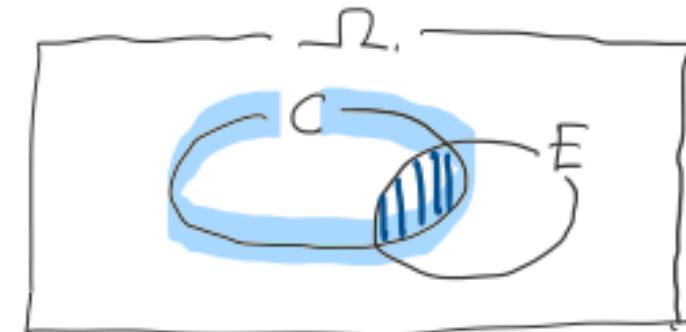
```
import numpy as np
for n in [10,100,1000,10_000,100_000, 1_000_000]:
    res_dice1 = np.random.randint(6,size=n) + 1
    res_dice2 = np.random.randint(6,size=n) + 1
    res = [(res_dice1[i], res_dice2[i]) for i in range(len(res_dice1))]
```

```
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))
```

```
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
```



compare:
without conditioning,
it was 0.138..

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))
```

Ex:

res[0,:]	2	1	2
res[1,:]	3	4	4

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!

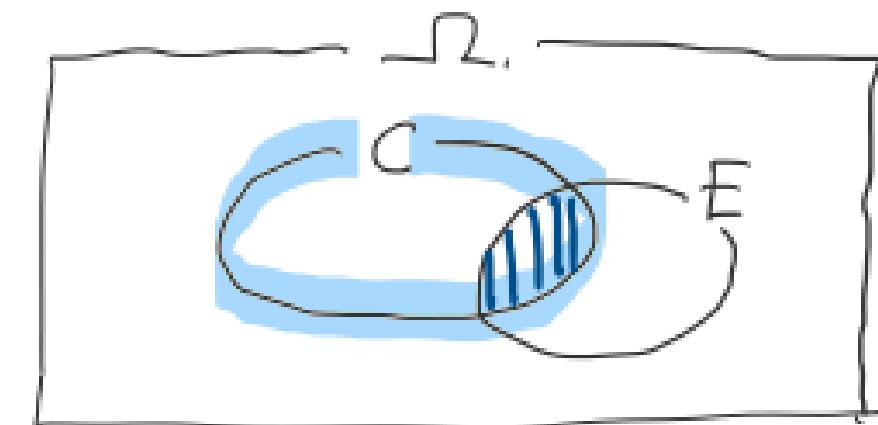
Conditional Probability

- 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, $A = \{X=4\}$, $B = \{X \text{ is even}\}$

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

- $P(A) = 1/6$, $P(B) = 1/2$
- $A \cap B = \{X=4\} \Rightarrow P(A \cap B) = 1/6$
- Therefore, $P(A|B) = 1/3$, $P(B|A) = 1/2$

Conditional Probability

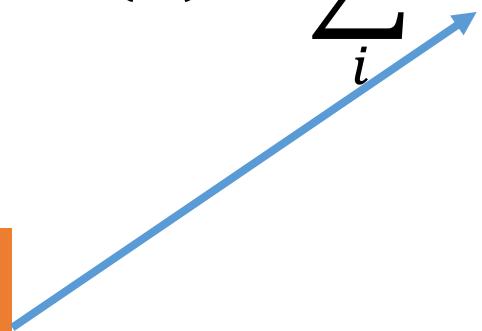
Chain rule

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

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

$$\begin{aligned}
 P(A) &= \sum_i P(A, B_i) = \sum_i P(B_i)P(A|B_i) \\
 &= \sum_i P(A)P(B_i|A) \quad \text{(by definition)}
 \end{aligned}$$

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
 $P(\text{test}=+ | \text{disease}=Y) = 0.9$
- A test for the disease yields a positive result 1% of the time when the disease is not present
 $P(\text{test}=+ | \text{disease}=N) = 0.01$
- 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?

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

Conditional Probability

What we know:

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

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

$$P(D=Y) = 0.001$$

$$P(\text{test} = - \mid D=Y) = 0.1$$

$$\Rightarrow P(\text{test} = - \mid D=N) = 0.99$$

$$P(D=N) = 0.999$$

Question: $P(D=Y \mid \text{test} = +)$

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

(on avg) out of 100,000 people,
90 have disease & tested +
999 not have disease & tested +

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

$$P(\text{test} = +, D = Y) = P(\text{test} = + \mid D = Y)P(D = Y) = \frac{90}{100000}$$

$$P(\text{test} = +, D = N) = P(\text{test} = + \mid D = N)P(D = N) = \frac{999}{100000}$$

The answer is 0.0826...

CAVEAT: $P(\text{test} = + \mid D=Y) = 0.9 \neq P(D=Y \mid \text{test} = +)$

Also: $P(D=Y) = 0.001$ vs $P(D=Y \mid \text{test} = +)$

Interpretation

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?)

$P(\text{disease}=Y \mid \text{test}=+)$ = 0.0826 \Rightarrow unsuitable for general, large-scale screening

Caveat: 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 $P(\text{column event} \cap \text{row event}) = P(T=t \cap D=d) = P(T=t | D=d) P(D=d)$

	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	

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.

$P(\text{test} = +)$

Conditional Probability

We can directly calculate:

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

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

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

Bayes rule

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

proof: definition and definition!

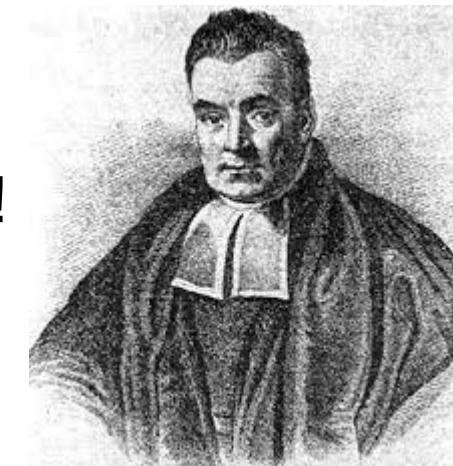
⇒ particularly useful in practice: infer $P(A|B)$ given $P(B|A)$!

$P(A)$: **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., rolling two fair dice, $A = \text{"die1=1"}$ and $B = \text{"die2=1"}$ are independent.
 \Rightarrow 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., $A = \text{"die1=6"}$ and $B = \text{"two dice sum to 6"}$ are not independent. quiz candidate
 $\because P(A) = 1/6$; and intuitively, when B is true, A can never happen
 formally: $P(A,B)=0$, $P(B)=5/36 \Rightarrow P(A|B)=0$ quiz candidate
- E.g., $A = \text{"die1=1"}$ and $B = \text{"two dice sum to 6"}$ are not independent.
 $\because P(A) = 1/6$. However, $P(A|B) = 1/5$

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

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!

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

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

(\exists a notion of ‘pairwise’ independence, but not much useful, so we omit it here)

More examples

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

56

- Rolling two fair dice
- Q: A = “die1=1” and B=“two dice sum to 5”. Independent?

No

$$\because P(A) = 1/6, P(B) = 4/36 = 1/9,$$

$$P(A,B) = 1/36 \neq P(A) P(B)$$

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

Yes

$$\because P(A) = 1/2, P(B) = 4/36 = 1/9$$

$$P(A,B) = 2/36 = 1/18 = P(A) P(B)$$

- 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,

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

due to independence.

or, ... = $P(\text{die1}=1 | \text{die2}=1) * P(\text{die2}=1) = P(\text{die1}=1) * P(\text{die2}=1)$

- E.g., two biased coins C1 and C2. Suppose $P(C1=H) = 0.3$ and $P(C2=H) = 0.4$. Compute the probability of $P(C1=H, C2=T)$.

(this may be useful in HW1, P2)

$0.3 \cdot 0.6 = 0.18$

quiz candidate



CSC380: Principles of Data Science

Probability Primer 3

Administrative Items

Office hour times

- Chicheng Zhang: Tuesdays 3:30-4:30pm, Gould-Simpson 720 (before Feb 28)
- Kyoungseok Jang: Tuesdays 3:30-4:30pm, Gould-Simpson 732 (after Feb 28)
- Saiful Islam Salim: Wednesdays 10-11am, Gould-Simpson 856
- Yinan Li: Mondays 12:45- 1:45pm, Gould-Simpson 856
- Sayyed Faraz Mohseni: Fridays 12-1pm, Gould-Simpson 837

Review

Axiom 3 (additivity):

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

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 B_1, B_2, \dots that partitions Ω ,

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

Conditional probability:

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

$(P(A|B) \neq P(B|A) \text{ in general})$

Probability chain rule:

$$P(A, B, C) = P(A|B, C)P(B|C)P(C) \quad (\Rightarrow P(A, B|C) = P(A|B, C)P(B|C))$$

Bayes' rule:

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

Independence:

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

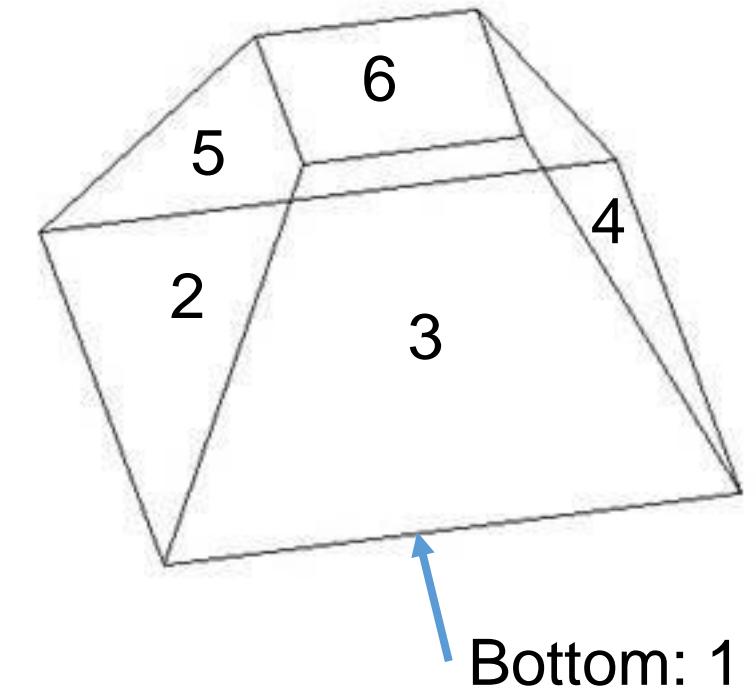
(property) A and B are independent if and only if $P(A|B) = P(A)$ (or $P(B|A) = P(B)$)

Independence

- Ex) Unfair die
 - Let A be the outcome of a single throw.
 - $P(A=1) \ll P(A=2) = \dots = P(A=5) \ll P(A=6)$

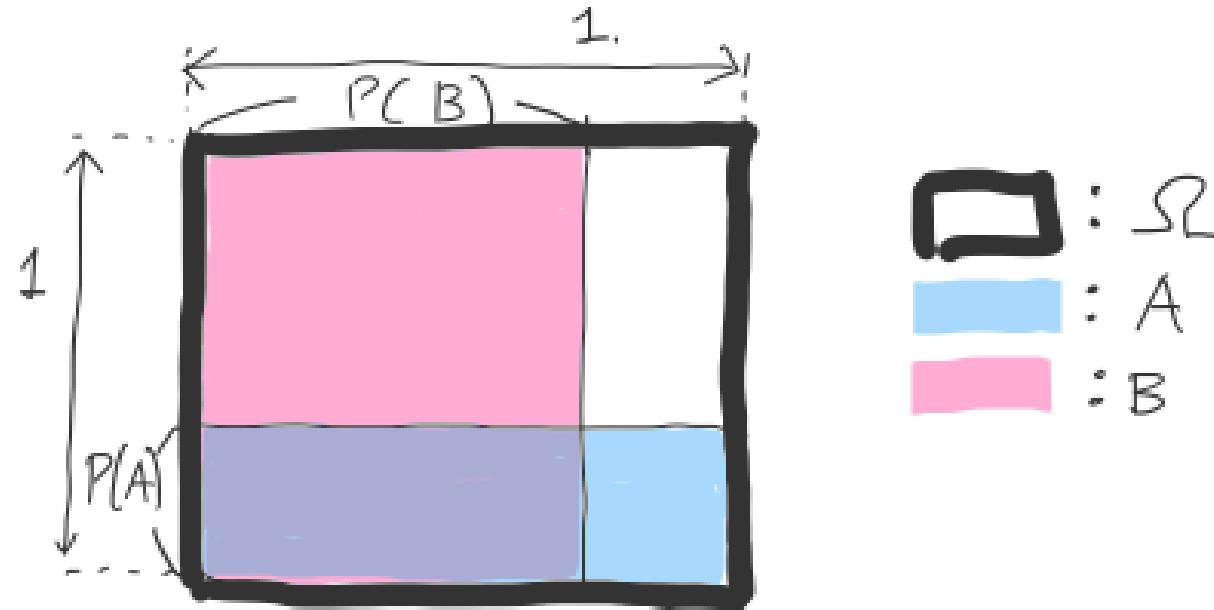
say, 0.1 0.15 0.15. 0.3

- 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

$$P(A|B) = P(A) \text{ or } P(B|A) = P(B)$$

- 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 (X_1): fair coin
- Second coin (X_2):
 - if $X_1=H$, throw a fair coin.
 - If $X_1=T$, throw an unfair coin $P(H) = 0.2$, $P(T) = 0.8$
- Q: Are $X_1=H$ and $X_2=H$ independent or not?

$$P(X_1=H) = \underline{\hspace{2cm}} \quad 0.5$$

$$P(X_2=H) = \underline{\hspace{2cm}} = P(X_2=H, X_1=H) + P(X_2=H, X_1=T) = 0.25 + 0.1 = 0.35$$

$$P(X_1=H, X_2=H) = \underline{\hspace{2cm}} \quad 0.25$$

$$P(X_1=H) * P(X_2=H) = 0.175$$

Quiz candidate

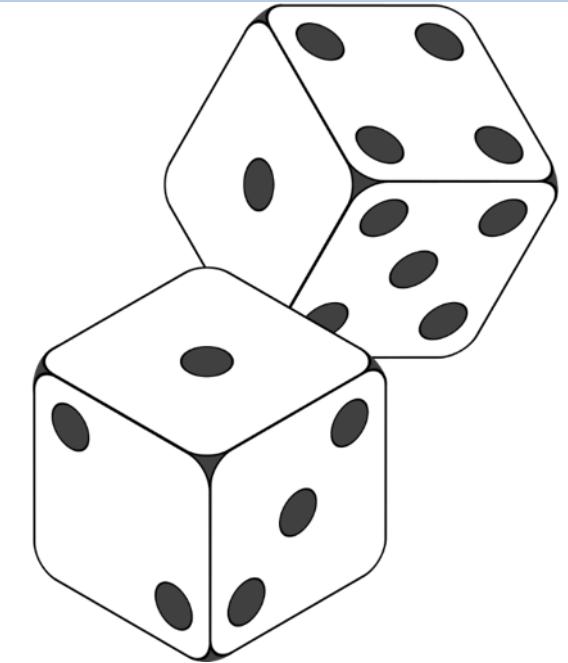
Random Variables and Probability

64

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

Option 1 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...

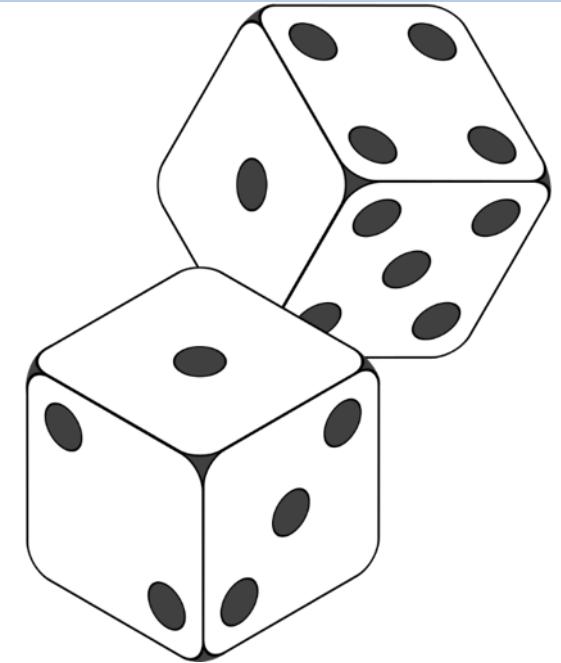
Random Variables and Probability

65

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 (abbrev. RV).

(formal definition, next slide)

Allows for modular constructions (e.g. $X = X_1 + X_2$), where X_1, X_2 are two dice outcomes

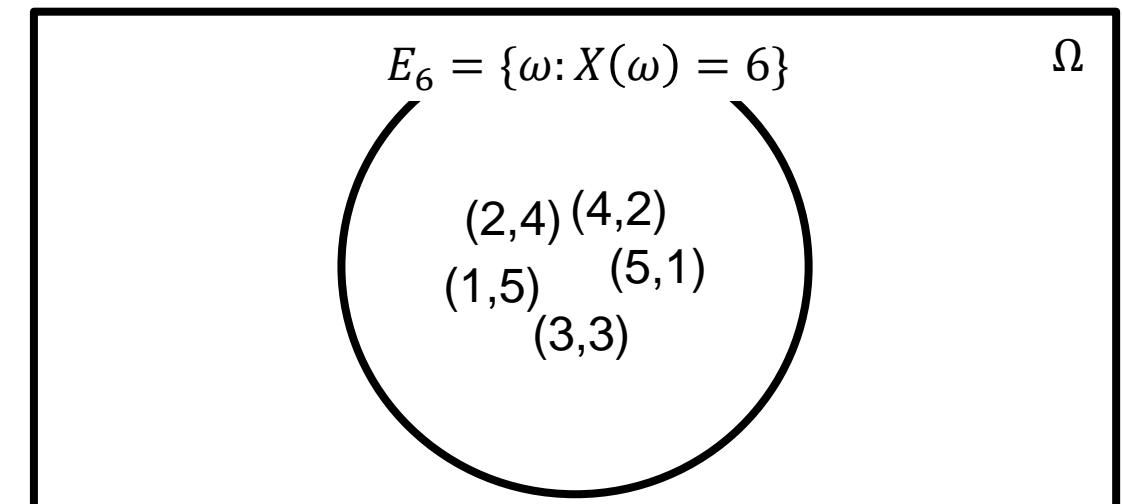
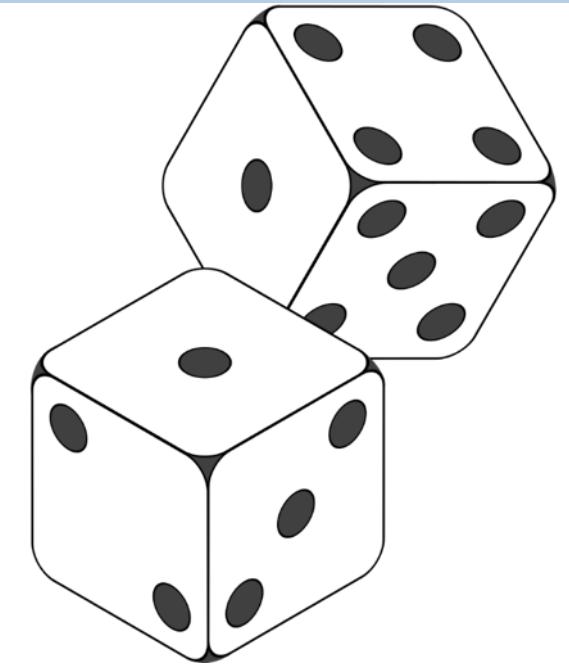
Random Variables and Probability

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 $X : \Omega \rightarrow \mathbb{R}$.

(skipping; too technical)



Random Variables and Probability

- Obviously, all the laws/rules about events apply 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 $p(Y) = \sum_x p(Y, X = x)$

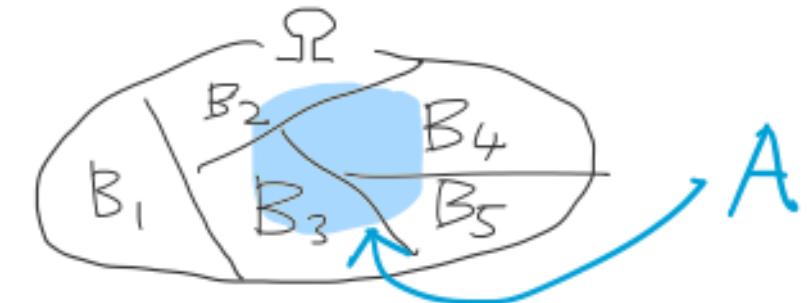
This means $P(Y = y) = \sum_x P(Y = y, X = x)$ for all y , abbrev. $P(Y = \cdot) = \sum_x P(Y = \cdot, X = x)$

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

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

Conditional Probability

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



Law of Total Probability also works for conditional probabilities,

$$p(Y | Z) = \sum_x p(Y, X = x | Z)$$

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

(just make sure you add the conditioning part for every $p()$!)

Conditional Probability

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

(This means $P(X = x | Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$ for all x, y ;
The rest of the rules also use this convention)

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 (think of Z as some *background evidence*)

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

↑ 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!

Random variables in probability: notation

- HW1, Problem 3(c), hint 4

$$\begin{aligned} & P(W, R_{i+j+1} \mid S = (i, j)) \\ &= P(W \mid R_{i+j+1}, S = (i, j)) \cdot P(R_{i+j+1} \mid S = (i, j)) \end{aligned}$$

What this means: for any r ,

$$\begin{aligned} & P(W, R_{i+j+1} = r \mid S = (i, j)) \\ &= P(W \mid R_{i+j+1} = r, S = (i, j)) \cdot P(R_{i+j+1} = r \mid S = (i, j)) \end{aligned}$$

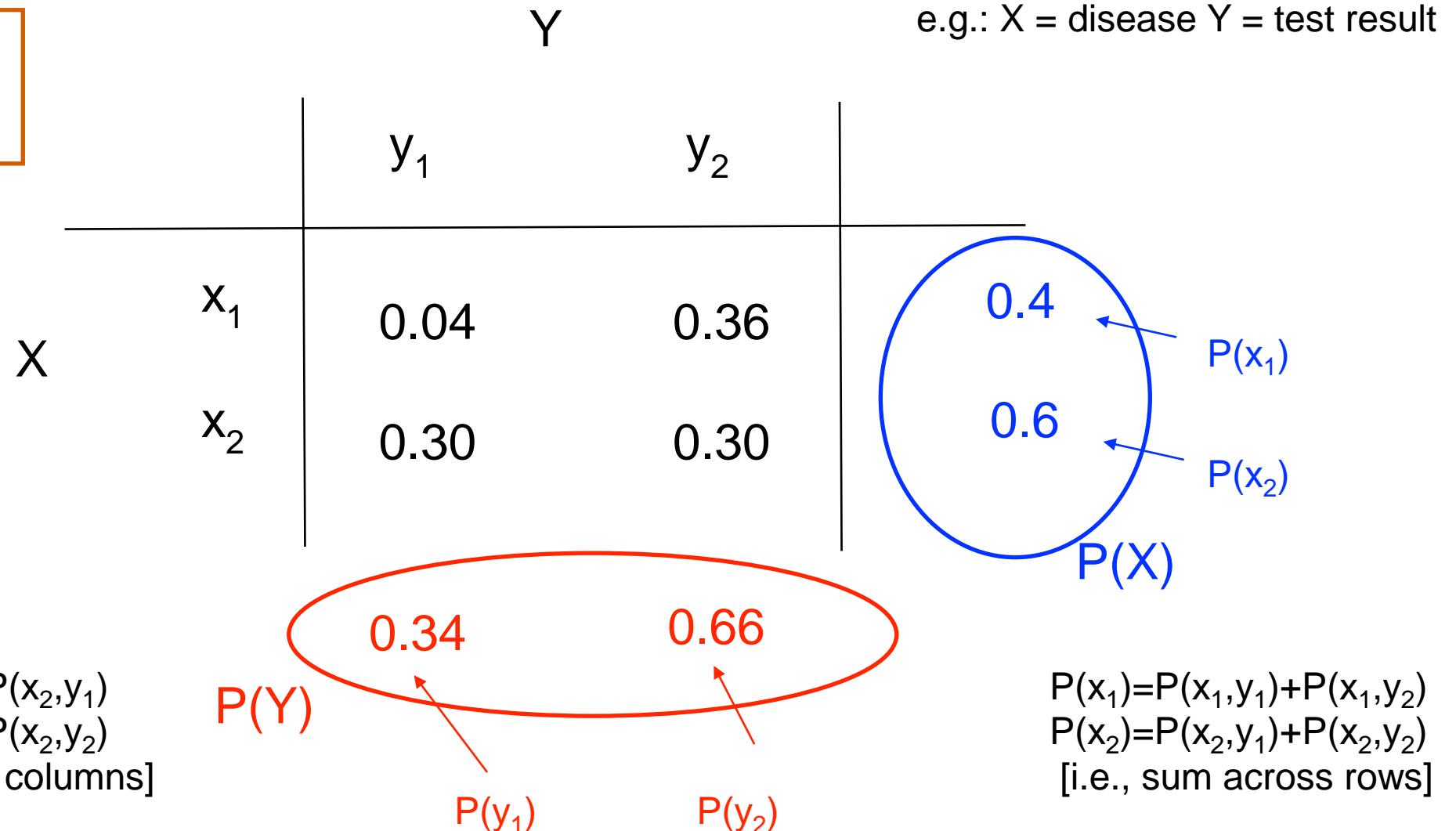
This is because: $P(A, B \mid C) = P(A \mid B, C) \cdot P(B \mid C)$ for any events A, B, C

Tabular Calculations for Random Variables

71

Tabular representation of two binary RVs (join probability)

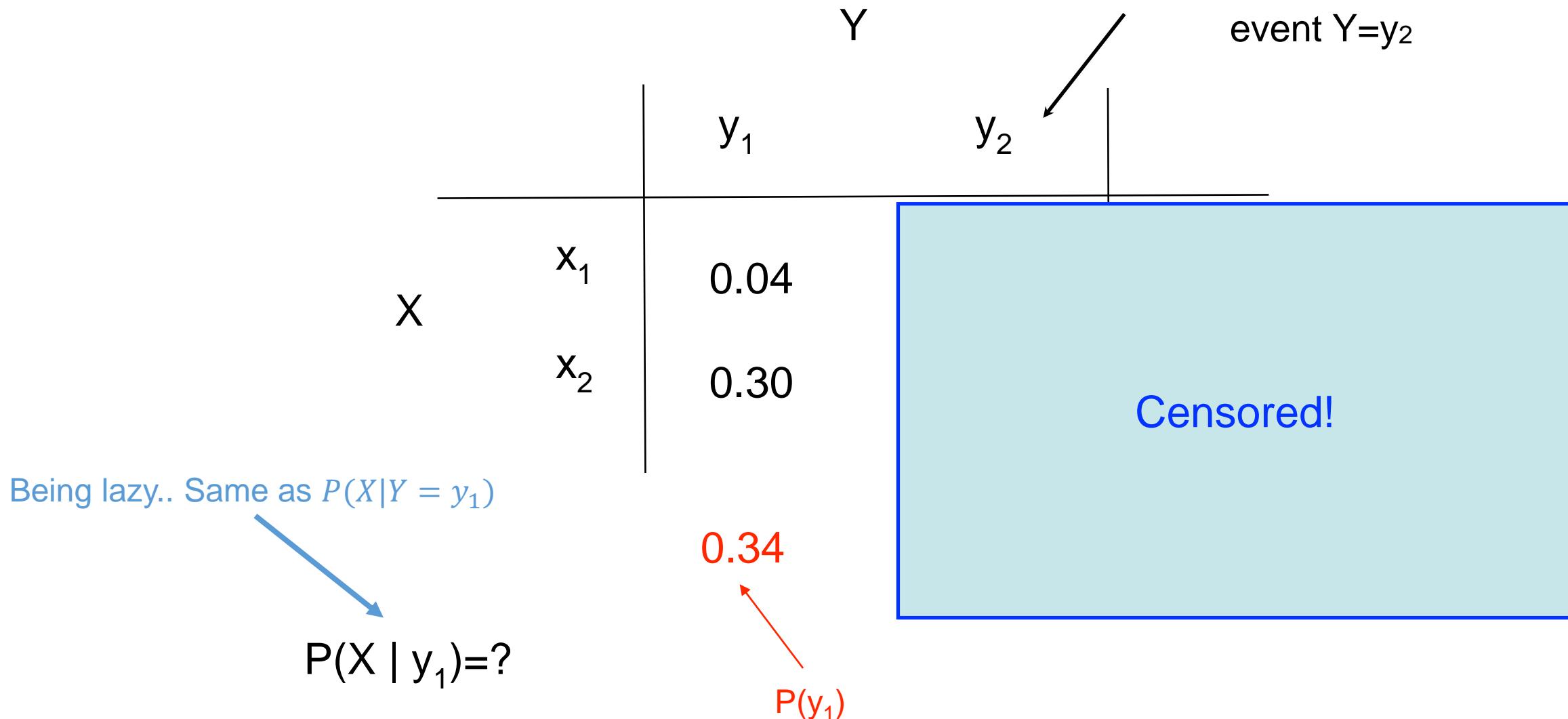
Use K-by-K probability table for K-valued discrete RVs



Tabular Calculations for Random Variables

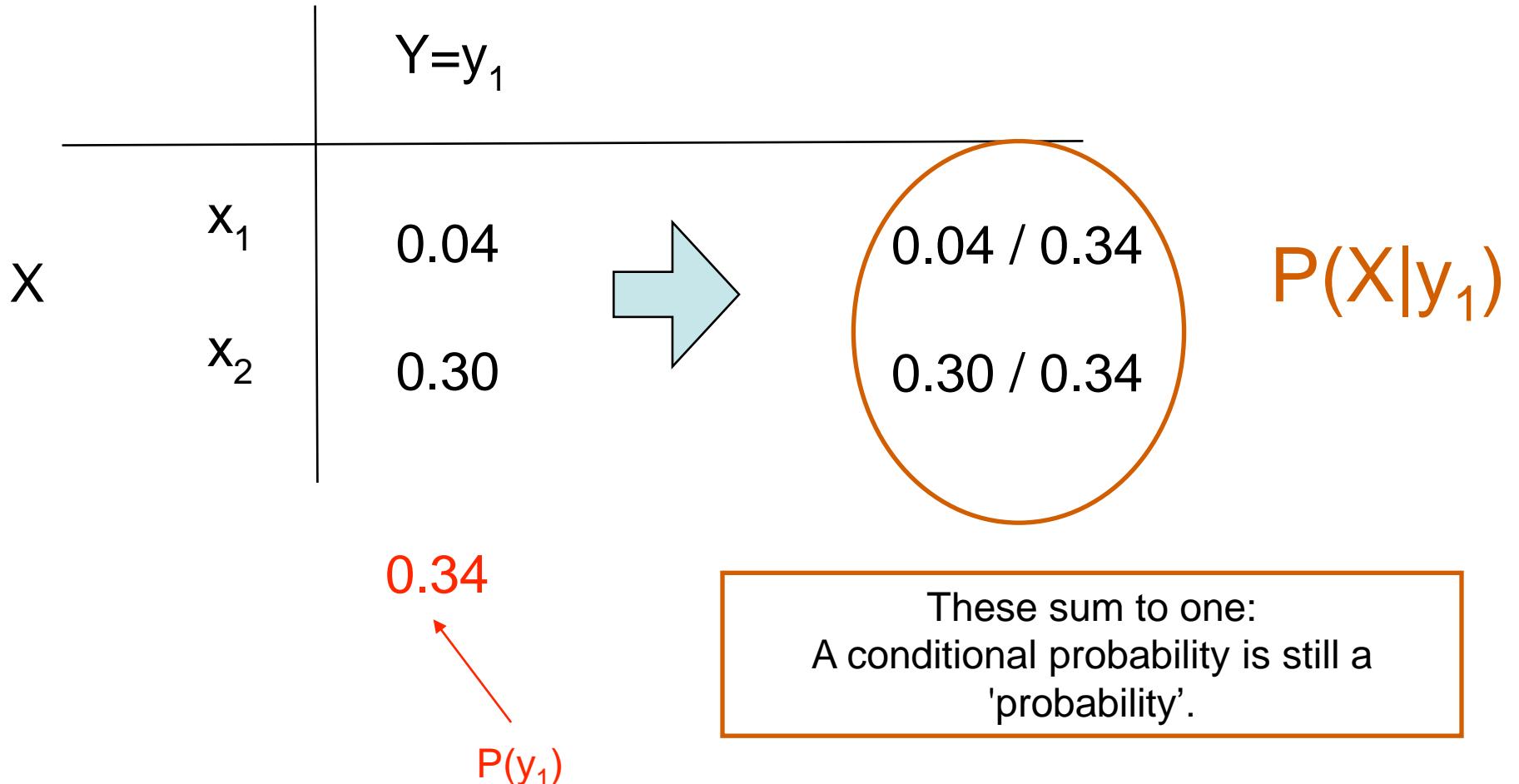
72

e.g.: $X = \text{disease}$ $Y = \text{test result}$



Tabular Calculations for Random Variables

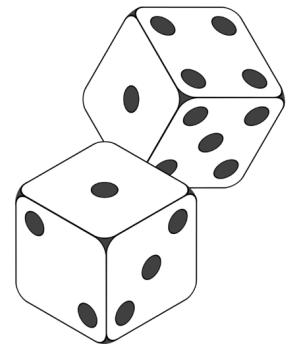
73



Definition Two random variables X and Y are independent 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)

Definition Two random variables X and Y are conditionally independent given Z if and only if,

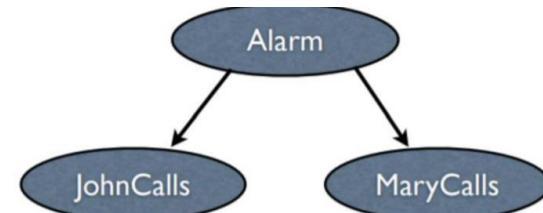
$$p(X = x, Y = y \mid Z = z) = p(X = x \mid Z = z)p(Y = y \mid Z = z)$$

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:

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



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

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

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 X_2 : unfair die

value	prob.
1	0.2
2	0.8

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

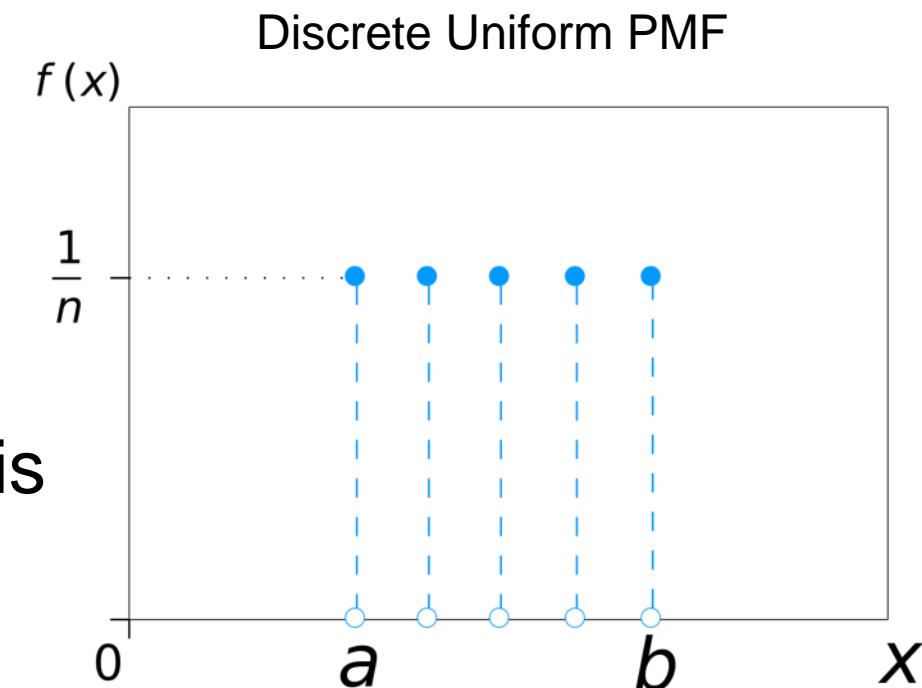
Generalization of fair die with N-faced die. Its PMF is:

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

More generally, consider $S = \{v_1, v_2, \dots, v_N\}$; X is drawn from the uniform distribution of S

\Leftrightarrow

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



numpy.random

To generate a sample from a uniform discrete distribution,

```
random.choice(a, size=None, replace=True, p=None)
```

Generates a random sample from a given 1-D array

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

```
out: 2
```

Bernoulli distribution

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

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

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

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

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

Num. “successes” out of N trials Num. ways to obtain k successes out of N



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)

$$p(\text{HTTHH}) = \pi(1 - \pi)(1 - \pi)\pi\pi$$

$$p(\text{TTHHH}) = (1 - \pi)(1 - \pi)\pi\pi\pi$$

...

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 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 $\in \{0,\dots,9\}$)



numpy.random

- If you want to compute it exactly,

```
from scipy.stats import binom  
binom.pmf(0, 9, 0.1)
```

=> 0.3874204890000001

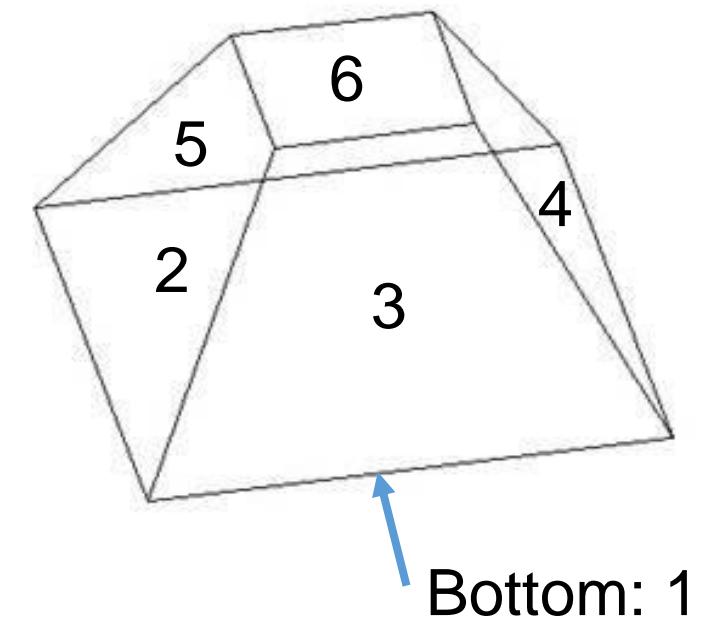
Scipy: sort of an extension of numpy for more specialized numerical computation.

Useful Discrete Distributions

85

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

In other words,

$$\begin{cases} P(X = 1) = \pi_1 \\ \dots \\ P(X = k) = \pi_k \end{cases}$$


Useful Discrete Distributions

Categorical distribution: equivalent forms of PMF

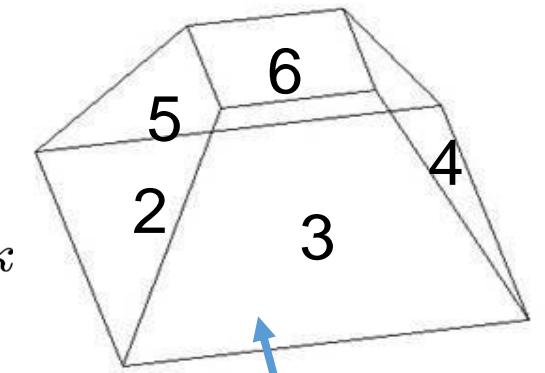
$$p(X) = \prod_{k=1}^K \pi_k^{\mathbf{I}(X=k)} \quad \text{or} \quad p(X) = \sum_{k=1}^K \mathbf{I}(X = k) \cdot \pi_k$$

equivalent to:

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

Indicator function

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



Bottom: 1

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

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

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

$$\text{e.g. } P(X = (1, 0, 0, 0, 0, 0)) = 0.1$$

numpy.random

- np.random.choice(range(1,7),p=[.1, .15, .15, .15, .15, .3])
out: 6



CSC380: Principles of Data Science

Probability Primer 4

Announcements

- HW1 due Friday 11:59pm
 - Again, no late days
 - HW1 solution will be out on Saturday.
 - HW1 will be graded no later than Feb 3 (Friday)
 - **Note:** we will only take regrade requests that are within 7 *academic days* of grading
 - Please follow the submission instructions & video – otherwise points may be deducted
 - Two submission links (one for solutions, another for code)

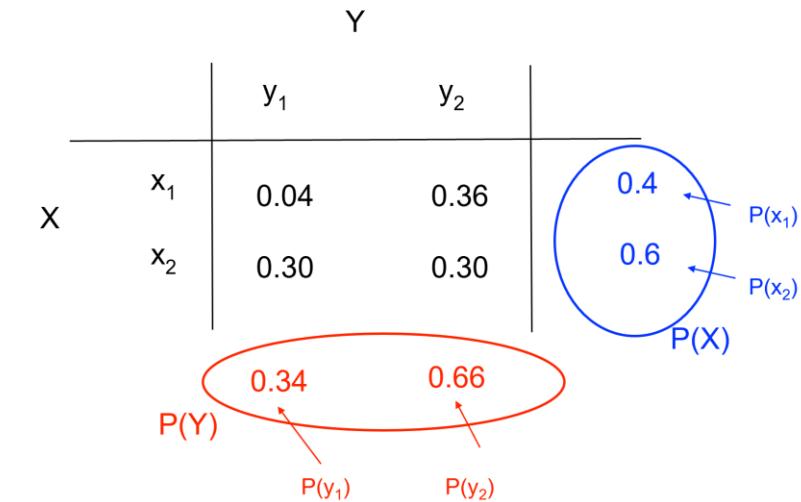
Weekly Review Question on Piazza

- For each week, I will ask a subset of you (about 5-10) to ask one question **by Monday** on Piazza, based on lectures in the past week
- Other students answer the questions (and can earn regular Piazza participation credit)
- Questions remained unanswered will be addressed by me **on Thursday.**

The screenshot shows the Piazza interface. At the top, there is a navigation bar with links for Drafts, hw1 through hw7, project, exam, logistics, discussion (which is highlighted with a red box), and other. Below the navigation bar is a search bar with the placeholder "Search or add a post...". The main content area has a header "Discussion for Week 3" and a sub-header "Please ask questions as replies below.". There is a "discussion" button with a blue background. Below that is a "Edit" button and a "good note | 0" link. A section titled "followup discussions, for lingering questions and comments" contains a "Start a new followup discussion" button and a red-bordered input field labeled "Compose a new followup discussion".

Terminology

- For pair of random variables (X, Y)
 - Joint probability distribution $P(X, Y)$
 - Marginal probability distribution $P(X)$, $P(Y)$
 - Obtaining $P(X)$ from $P(X, Y)$:
 - Using the law of total probability
 - “Marginalization” / “Marginalizing out Y ”
- Also generalizes to n random variables (X_1, \dots, X_n)



Last lecture: discrete random variables

92

- E.g. $X = \text{unfair die}$

- Probability mass function X : $p(x) = P(X = x)$

- Useful examples:

- Uniform: $U(\{\nu_1, \dots, \nu_N\})$
- Bernoulli: $\text{Ber}(\pi)$, $\pi \in [0,1]$
- Binomial: $\text{Bin}(n, \pi)$, $n \in \mathbb{N}$, $\pi \in [0,1]$
- Categorical: $\text{Cat}(\pi)$, $\pi = (\pi_1, \dots, \pi_K)$ (constraint: $\sum_k \pi_k = 1, \pi_k \geq 0$)
- Notation convention: $X \sim \text{Cat}(\pi)$: X follows / conforms to / is drawn from the $\text{Cat}(\pi)$ distribution

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

Multinomial distribution

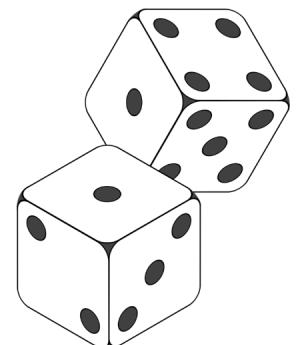
What if we count outcomes of n independent draws of categorical RVs $Y_1, \dots, Y_n \sim \text{Cat}(\pi)$?

e.g.

- Flip a coin $n=10$ times, how many times do we see H/T ($K=2$)? Binomial distribution
- Throw a die $n=10$ times, how many times do we see 1/2/../6? ($K=6$)

Multinomial distribution

- **Multinomial distribution:** Distribution of (X_1, \dots, X_K) where $X_k = \sum_{i=1}^n \mathbf{I}(Y_i = k)$ is the count of item k .
- Q: what are the parameters of the multinomial distribution?
 - n : number of trials, π : distribution of individual observation
 - Abbrev. $\text{Mult}(n, \pi)$



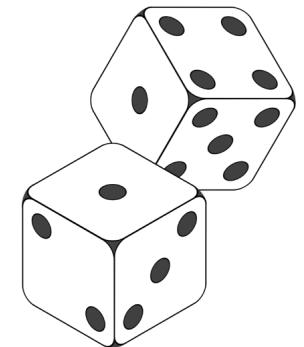
Multinomial distribution

- What is $\text{Mult}(n, \pi)$'s PMF, i.e.

$$P((X_1, \dots, X_K) = (x_1, \dots, x_K))$$

for any (x_1, \dots, x_k) ?

- e.g. $X \sim \text{Mult}(3, (0.5, 0.5))$, $P((X_1, X_2) = (x_1, x_2))$



- First observation: PMF is nonzero only if (*) $\sum_k x_k = n$ and x_1, \dots, x_K are nonnegative integers

- e.g. $P((X_1, X_2) = (1,3)) = 0$

- When (*) is satisfied,

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

- e.g. $P((X_1, X_2) = (1,2)) = \binom{3}{1 2} 0.5^1 0.5^2 = \binom{3}{2} \cdot 0.5^1 0.5^2$

Multinomial coefficient: number of strings of length n with x_1 1's, \dots , x_K K's

numpy.random

numpy.random.multinomial

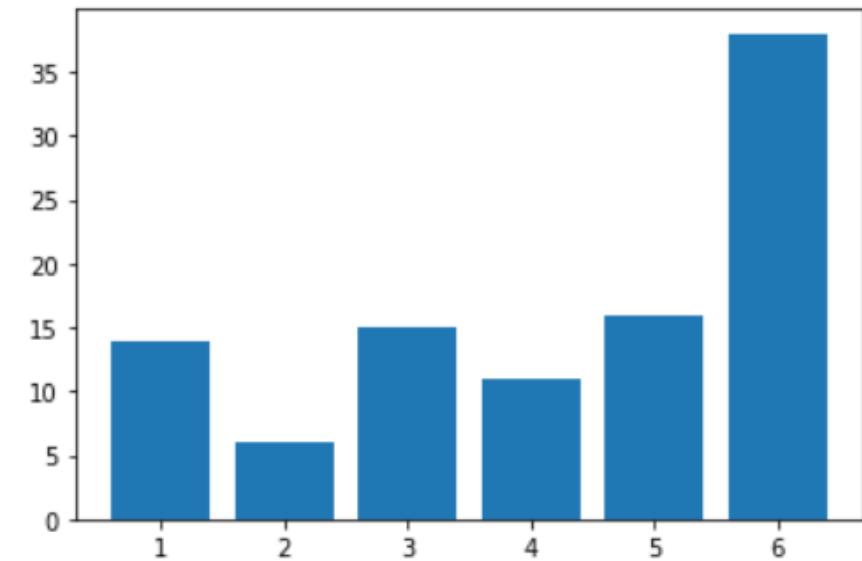
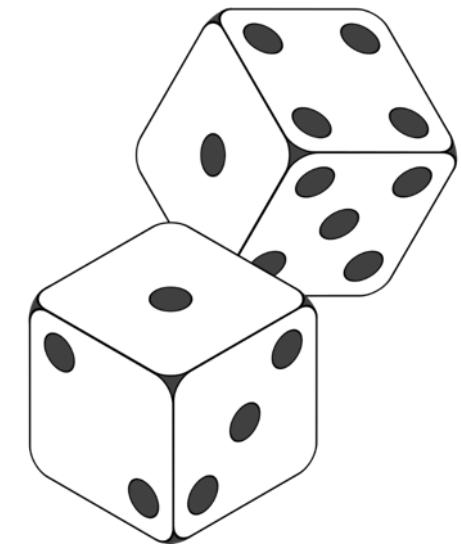
`numpy.random.multinomial(n, pvals, size=None)`

Draw samples from a multinomial distribution.

pvals: list of probability parameters that sums to 1 (π parameter).

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)  (X is a vector of counts;
plt.bar(np.arange(6) + 1, x)    length 6)
plt.show()
```



Note: pvals (π) must be a valid probability distribution ($\sum_k \pi_k = 1, \pi_k \geq 0$)

numpy.random

Alternative ways 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]
```

Random Variable Examples

- X : an outcome of a die.

- $R_1 = I\{X \text{ is even}\}$
- $R_2 = I\{X = 1\}$

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

- $R_3 = X_1 + X_2$
- $R_4 = \frac{(X_1 + X_2)}{2}$
- $R_5 = I\{X_1 = 1\}$

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

$$\{R_5 = 0\} \Leftrightarrow \{(2,1), (2,2), \dots, (2,6), (3,1), (3,2), \dots, (3,6),$$

...

$$(6,1), (6,2), \dots, (6,6)\}$$

Q: what distribution does R_5 follow with what parameter?

Continuous Probability



(TV show spin the wheel)

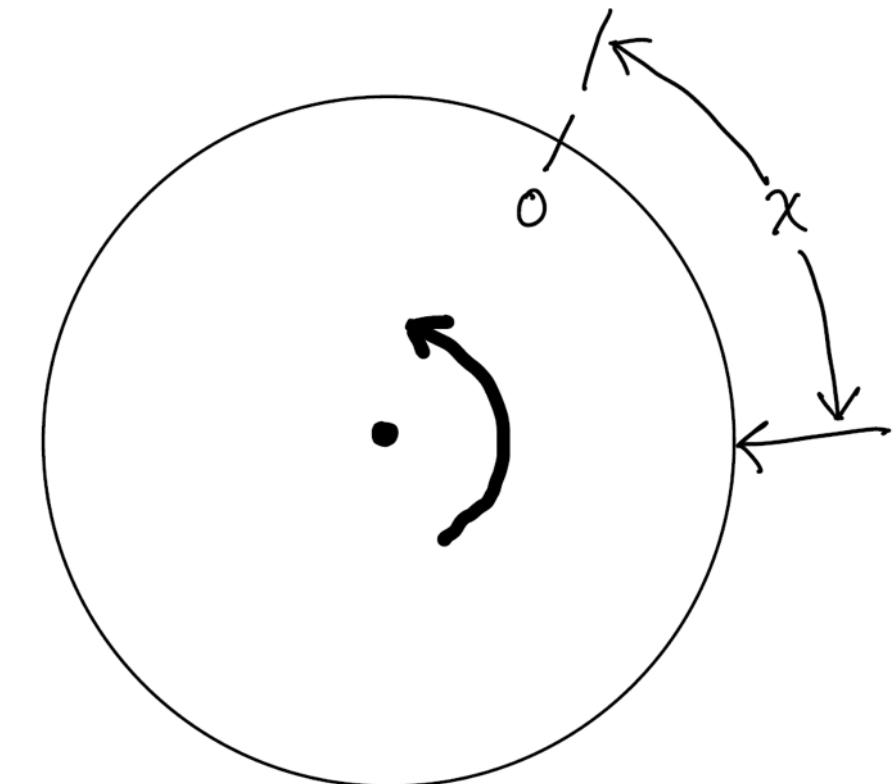
Continuous Probability

Experiment Spin continuous wheel and measure X displacement from O

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

Goal: Show $P(X=x) = 0$

- Say the displacement X is in $(0, 1]$
- Let N be a very large integer

- Let $I_k = \left(\frac{k-1}{N}, \frac{k}{N}\right]$

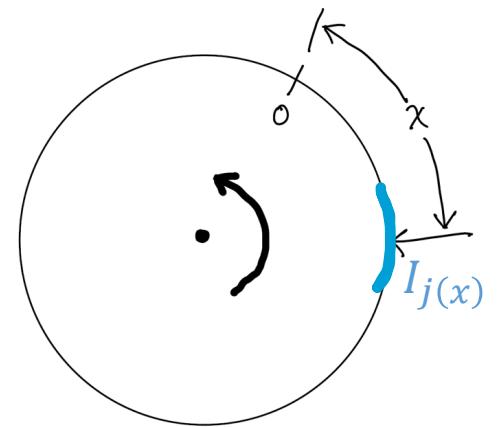
- Let $j(x)$ be k such that $x \in I_k$

- $P(X = x) \leq P(X \in I_{j(x)}) = \frac{1}{N}$

- We can make N as large as we want! $\Rightarrow P(X=x)$ must be 0.

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

Q: why is it $\frac{1}{N}$?



Continuous Probability

Maybe, it's not so weird.

- Q1: Probability that your house water usage tomorrow is 20.58 gallon?
- Q2: Probability that your house water usage tomorrow is 20.5891231285 gallon?



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

Or, Use a Heuristic Argument

- Computers are inherently dealing with discrete numbers anyways.
- Imagine you consider extremely fine grained intervals like
... $(-\epsilon, 0]$, $(0, \epsilon]$, $(\epsilon, 2\epsilon]$, ...
where $\epsilon = 10^{-300}$
- The outcome space is real, but we only talk about the events that are union of those intervals.

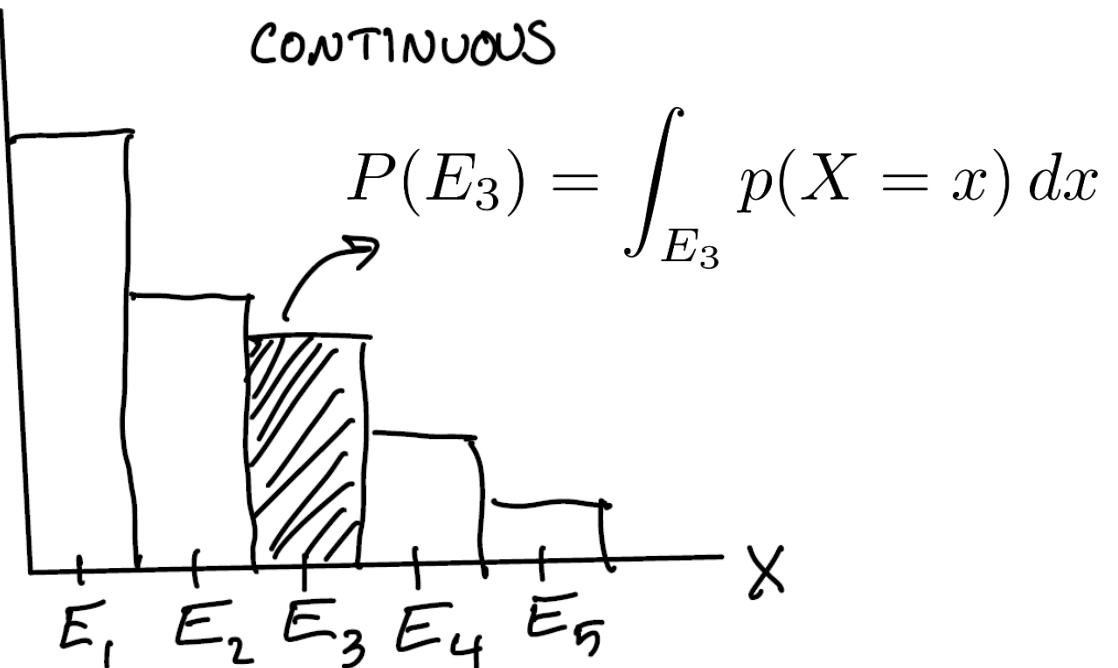
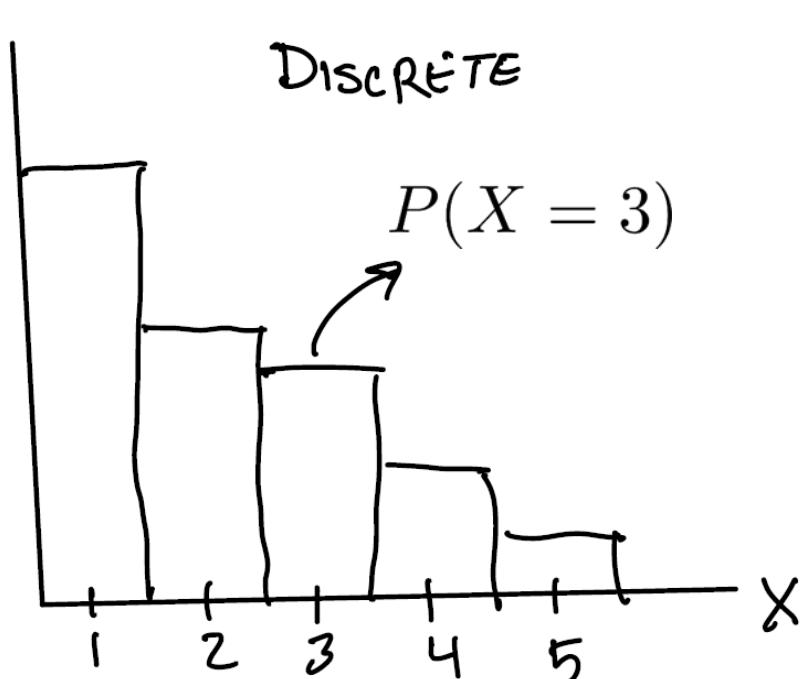
Continuous Probability

we could try to convince ourselves that it is sensible.

... or we could just accept this oddity...



Continuous Probability Distributions



- Events represented as intervals $a \leq X < b$ with probability,

$$P(a \leq X < b) = \int_a^b p(X = x) dx$$

- Specific outcomes have zero probability $P(X = x) = 0$
- But may have nonzero *probability density* $p(X = x) > 0$

Continuous Probability Distributions

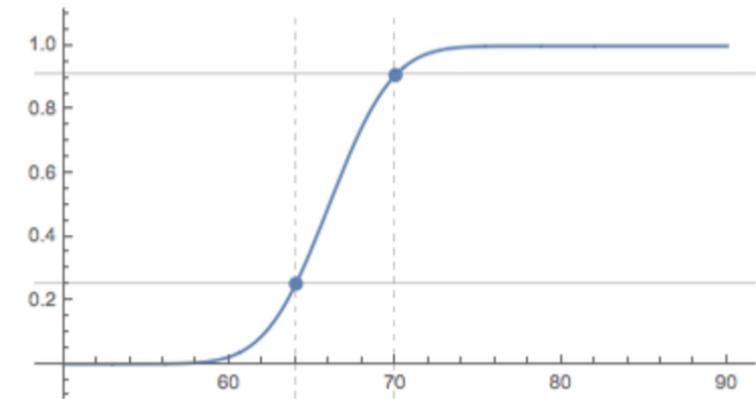
Definition *The cumulative distribution function (CDF) of an RV X is the function given by,*

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

Key properties:

F is monotonically increasing

$F(x)$ goes to 0/1 if x goes to $-\infty$ / $+\infty$



- Can easily measure probability of closed intervals,

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

e.g. $a = 64, b = 70$

Continuous Probability Distributions

- If $F(x)$ is differentiable then X is a continuous RV,

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

$p(x)$ is called X 's **probability density function (PDF)**

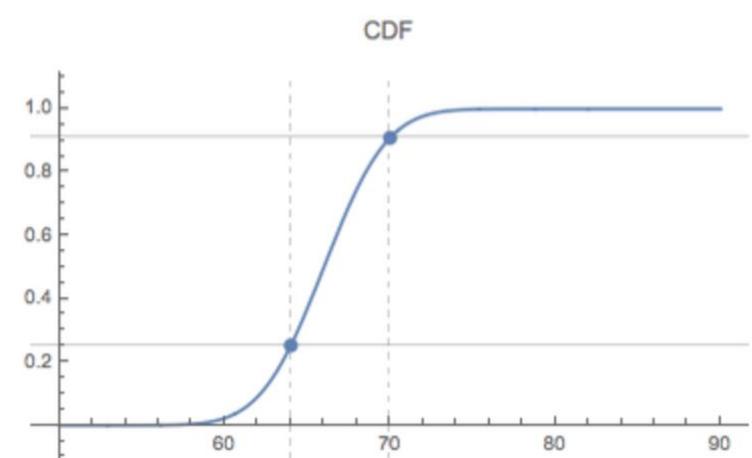
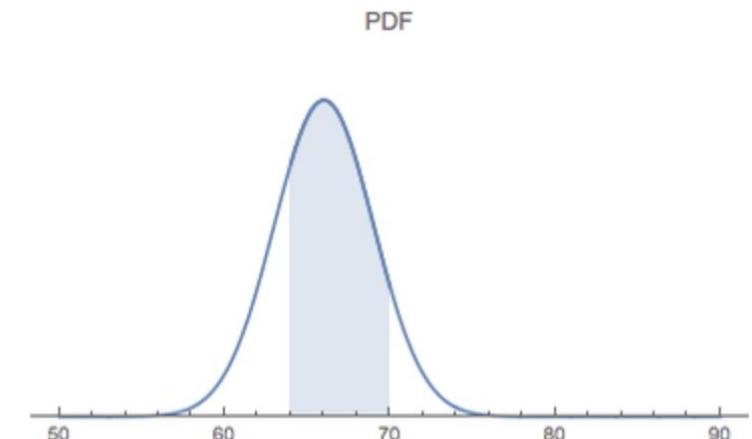
$$\approx \frac{F(x) - F(x-\epsilon)}{\epsilon} = \frac{P(X \in (x-\epsilon, x])}{\epsilon} \text{ when } \epsilon \rightarrow 0$$

Intuition: $p(x)$ characterizes how likely X takes values in *the neighborhood of x*

Observations:

- $p(x) \geq 0$ for all x
- $P(a < X \leq b) = F(b) - F(a) = \int_a^b p(x) dx$
- $\int_{-\infty}^{+\infty} p(x) dx = 1$

Fundamental Theorem
of Calculus



Notation

- For continuous RV X , we will use $p(X = x)$, $p(x)$, $p_X(x)$ to denote its PDF
 - Recall: $P(X = x)$ is not its PDF value (in fact, always 0)
- For discrete RV X , we will use $p(X = x)$, $p(x)$, $p_X(x)$ to denote its PMF
 - In this case, $p(X = x) = P(X = x)$
- General suggestion for HW / exams: to be extra safe, you can explicitly declare “we use $p(X = x)$ to denote the PDF of continuous RV X ”

Continuous Probability Distributions: Multivariate

108

- **Definition** A pair of random variables (X, Y) is said to be continuous if there exists a probability density function $p(x, y) \geq 0$, such that for any set A ,

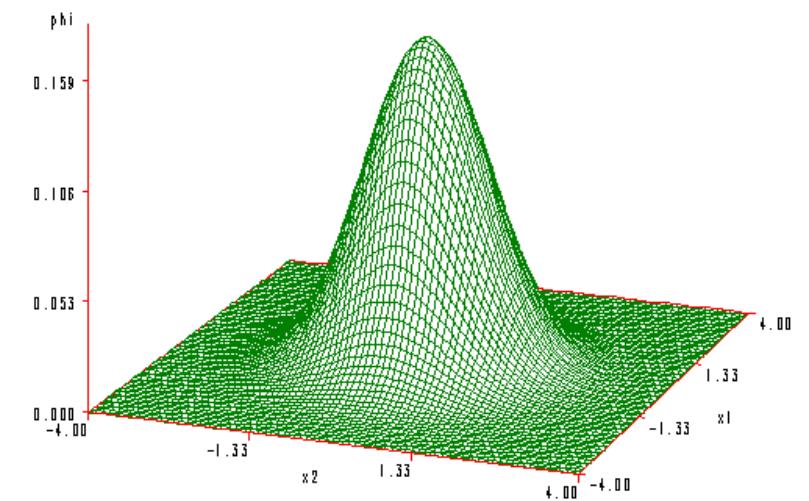
$$P((X, Y) \in A) = \int_A p(x, y) dx dy$$

Bivariate Normal Density - r=0.0

- Recall: for discrete (X, Y) ,

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

- Can generalize to a tuple of n random variables (random vector) (X_1, \dots, X_n)



Continuous Probability Distributions

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

Law of Total Probability for continuous distributions,

$$p(x) = \int_{\mathcal{Y}} p(x, y) dy$$

Recall: for discrete X

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

All the rules apply when replacing PMF with PDF...

Conditional PDF:

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

Probability Chain Rule:

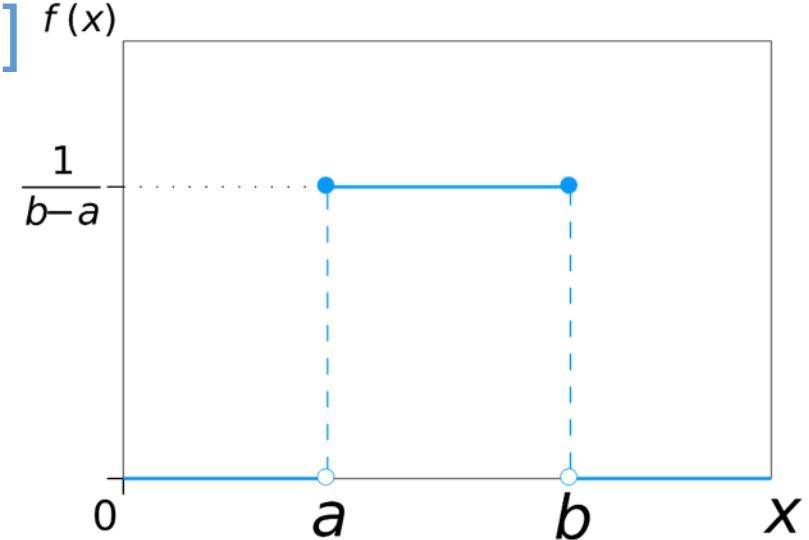
$$p(X, Y) = p(Y)p(X | Y)$$

Useful Continuous Distributions

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

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

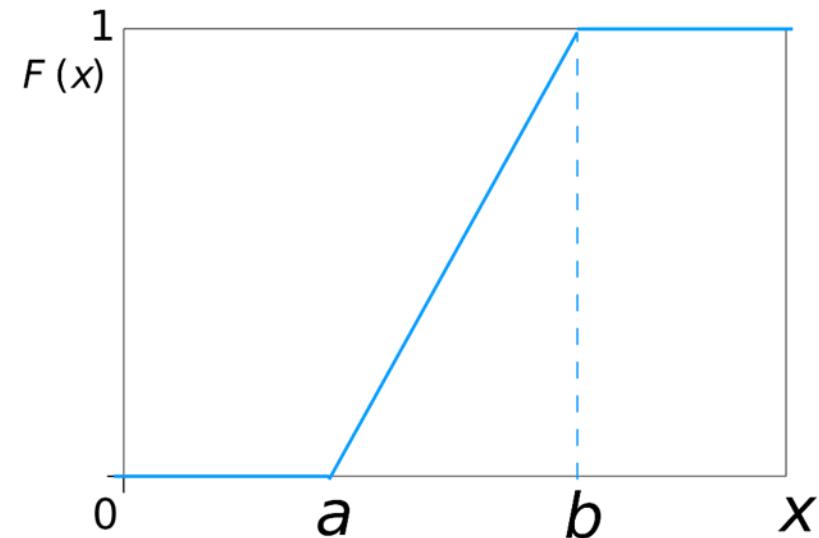
$$P(X \leq x) = \int_{-\infty}^x p(t) dt$$



Notation:

$p(x)$ for the PDF function at location x
 $P(A)$ for the probability of event A

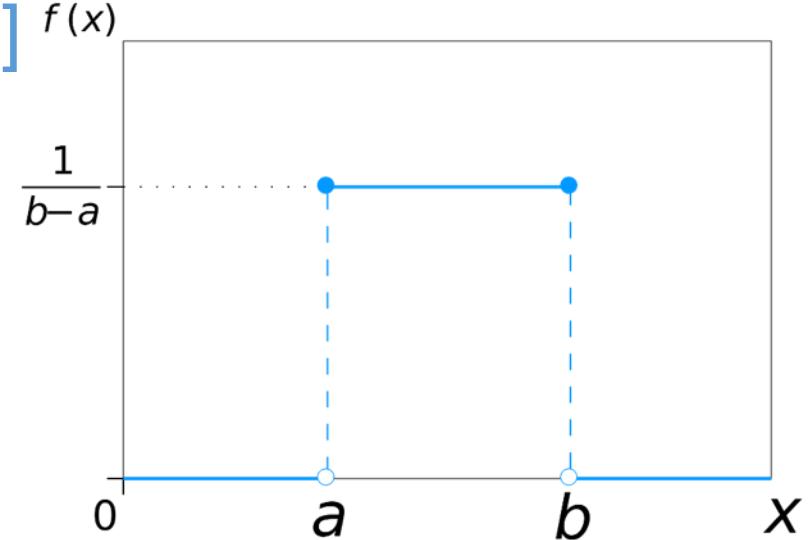
Again, PDF function \neq probability



Useful Continuous Distributions

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

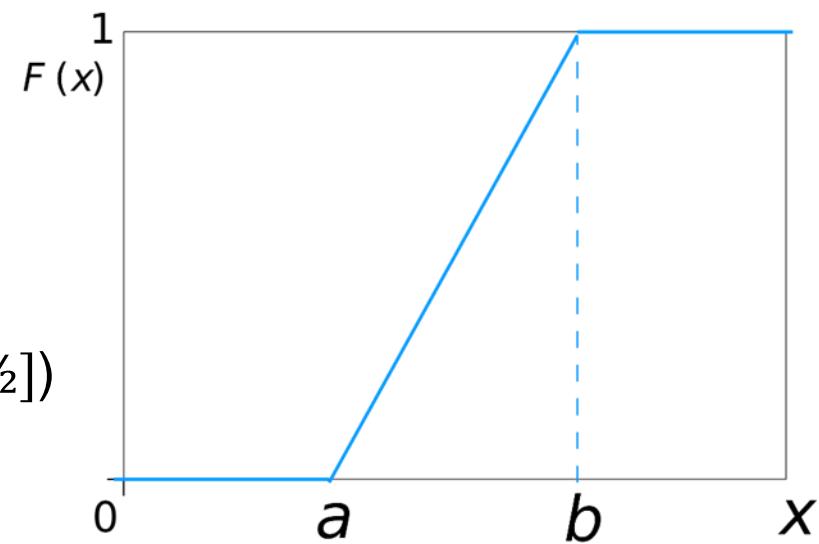
$$p(x) = \begin{cases} 0 & \text{if } x \leq a, \\ \frac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{if } b \leq x \end{cases} \quad P(X \leq x) = \begin{cases} 0 & \text{if } x \leq a, \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b, \\ 1 & \text{if } b \leq x \end{cases}$$



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

$$P(X \leq x \mid X \leq \frac{1}{2}) = F(x) \text{ of Uniform}[0, \frac{1}{2}] \quad (\text{i.e., } P(Y \leq x) \text{ where } Y \sim \text{Uniform}[0, \frac{1}{2}])$$

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)
```

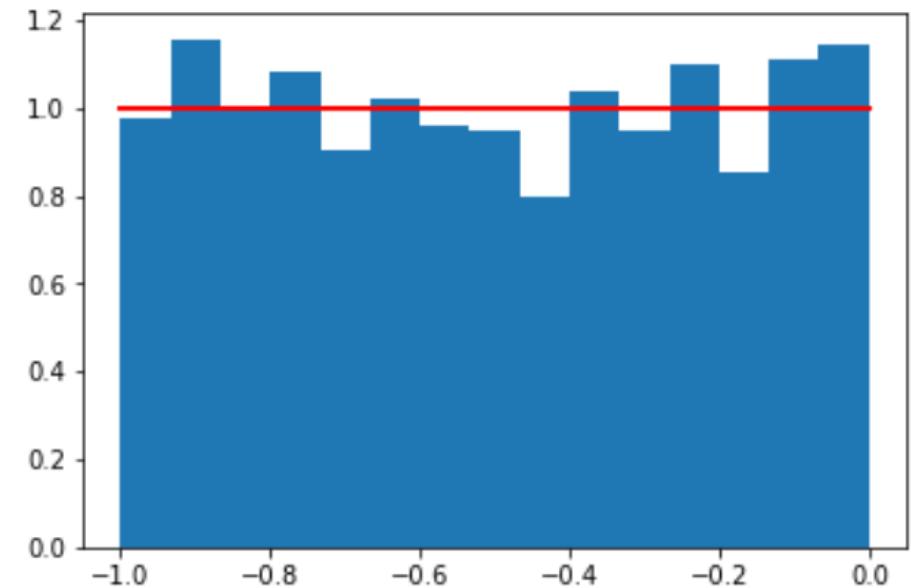
Draw samples from a uniform distribution.

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 distribution on [-1,0),

```
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()
```

redline: PDF of uniform distr.



Useful Continuous Distributions

Exponential distribution with scale λ ,

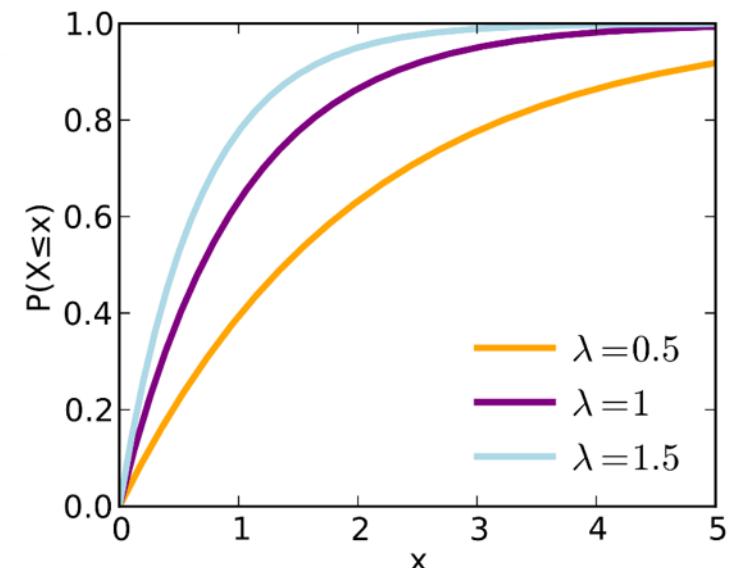
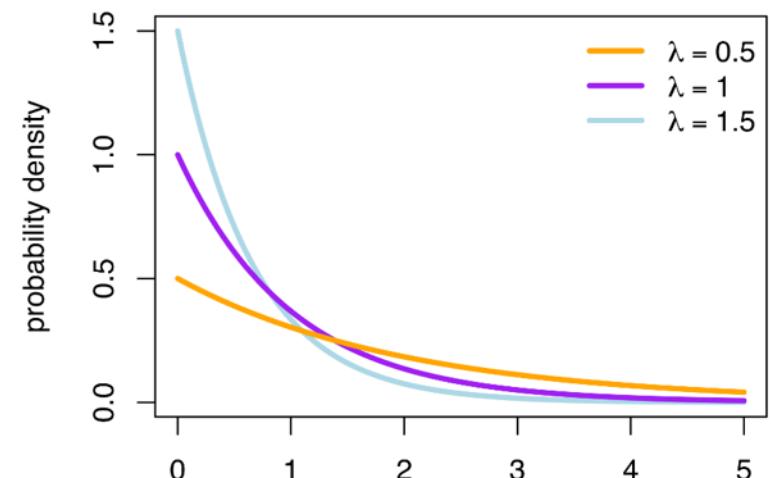
$$p(x) = \lambda e^{-\lambda x}$$

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

for $x > 0$.

Examples:

- Time between geyser eruptions
- Time between customers
- Lifetime of lightbulbs
- ..



numpy.random

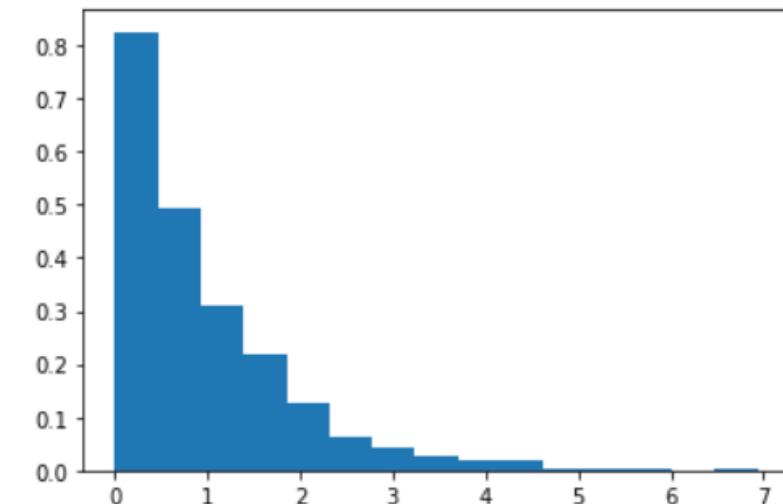
numpy.random.exponential

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

$$\text{scale} = \lambda$$

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

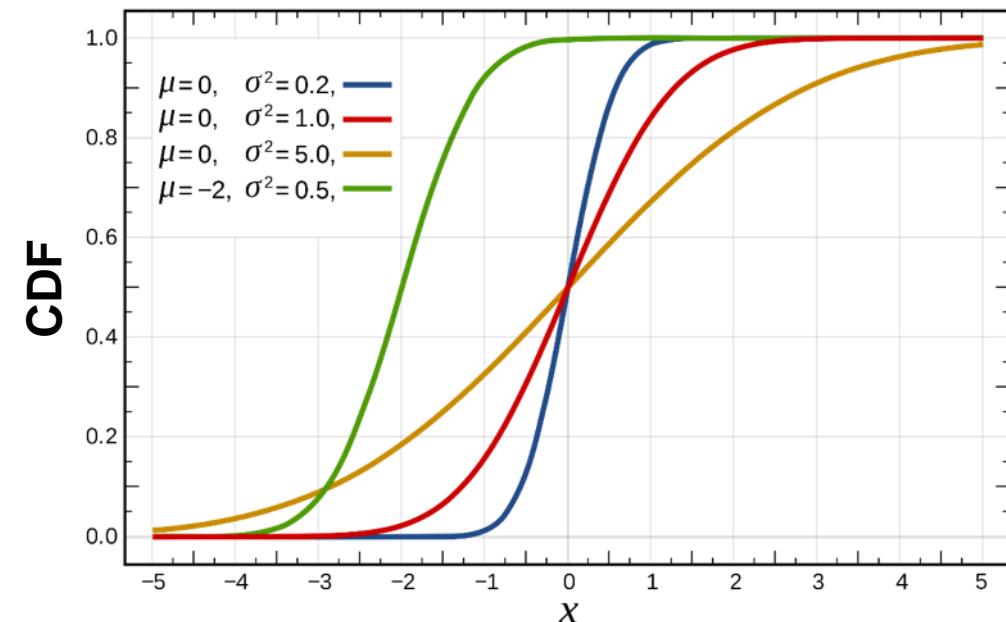
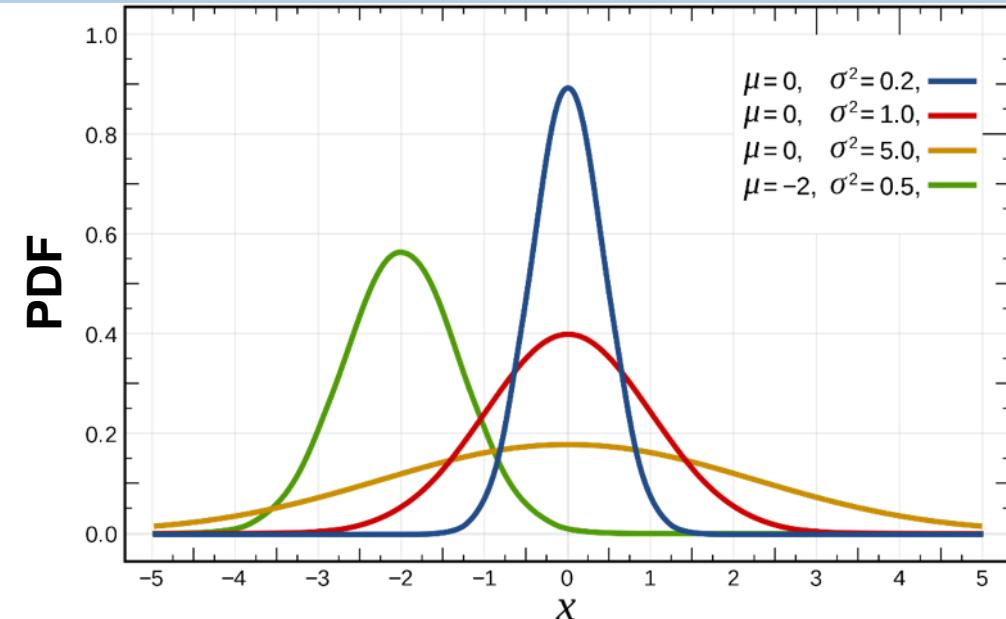
Gaussian (a.k.a. Normal) distribution with mean (location) μ and variance (scale) σ^2 parameters,

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

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

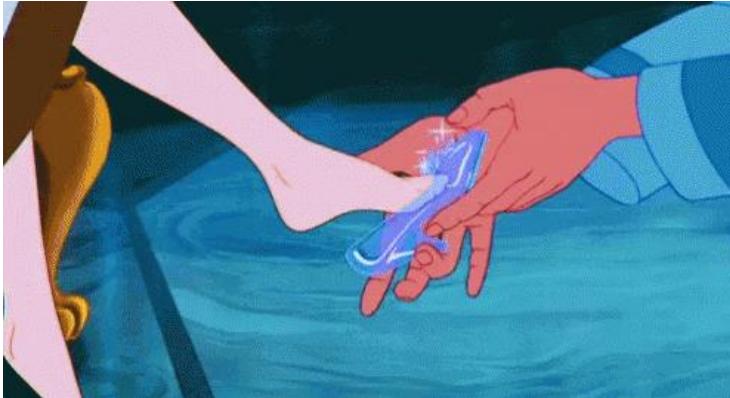
Observations:

- Larger $\sigma^2 \Rightarrow p(x)$ more “spread out”
- Larger $\mu \Rightarrow p(x)$ ’s center shifts to the right more

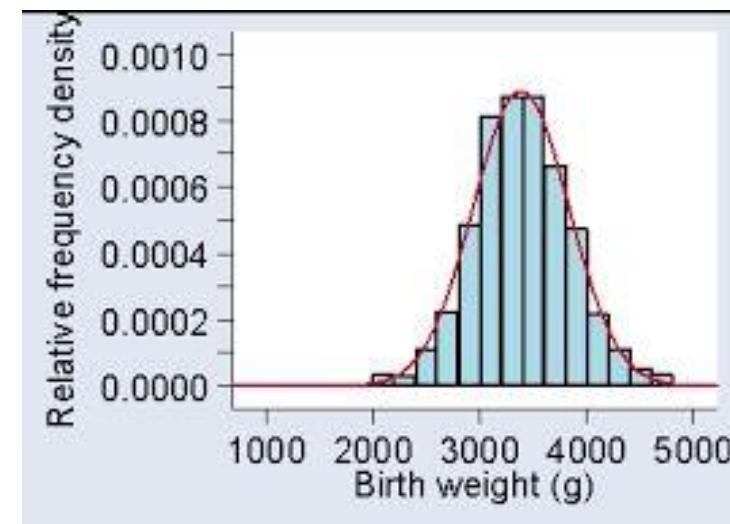
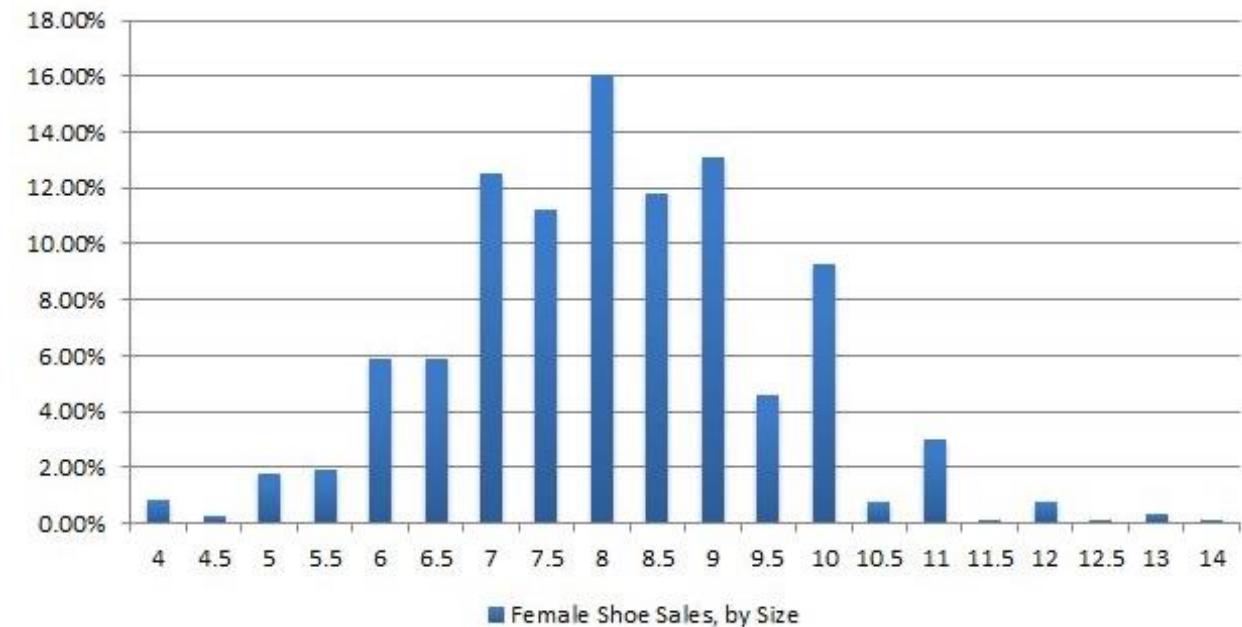


Things that follow Gaussian

Female shoe size



Birth Weight



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

Useful Continuous Distributions

117

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

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

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

Useful Properties Quiz candidate

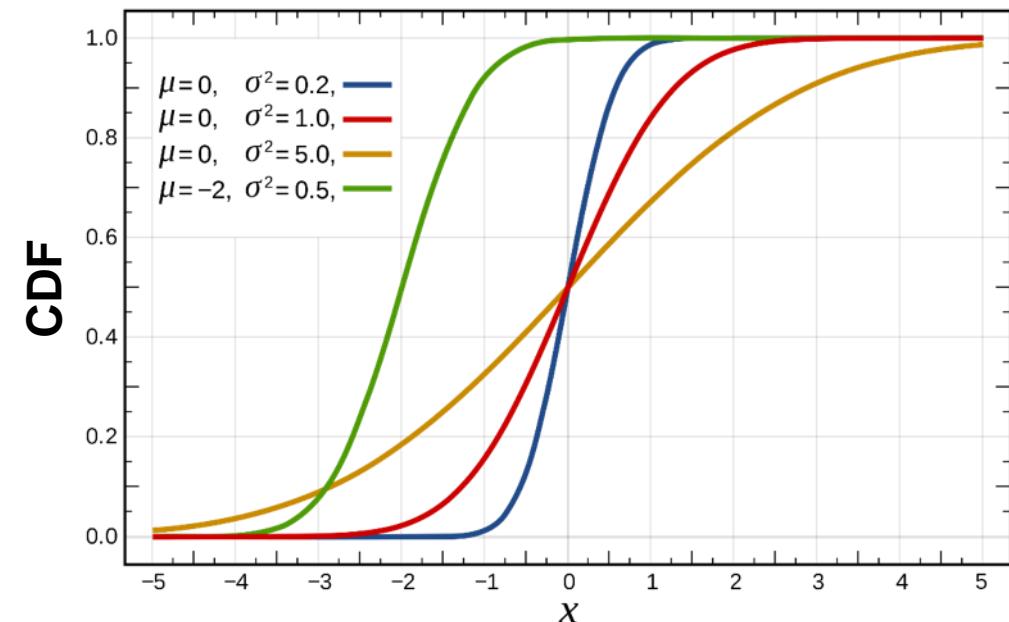
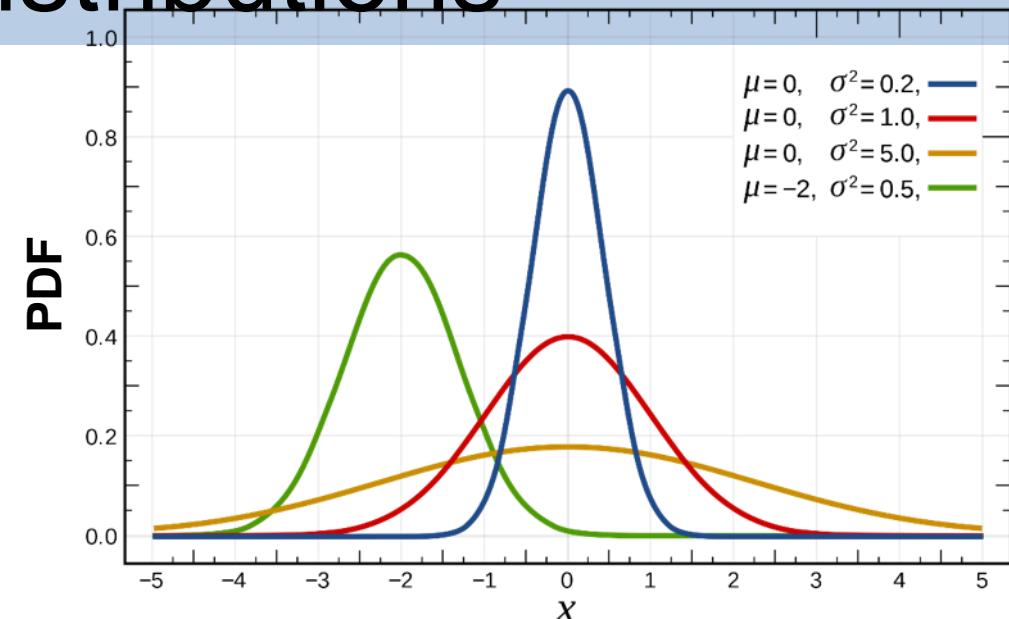
- Closed under addition w/ independence assumption:

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

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

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

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





CSC380: Principles of Data Science

Probability Primer 5

Announcement

- HW1 solution will be out Wednesday
- HW2 will be out today
- Discussion questions should be asked by tonight (piazza).

Recap

Useful discrete distributions

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

Continuous probability distributions

- $P(X = x) = 0$ for any x , but we are mostly interested in X taking values in intervals
- Probabilities assigned to *intervals* via CDF $P(X \leq x)$
- PDF measures probability *density* of single points $p(X = x)$

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...

Review: Continuous Random Variable

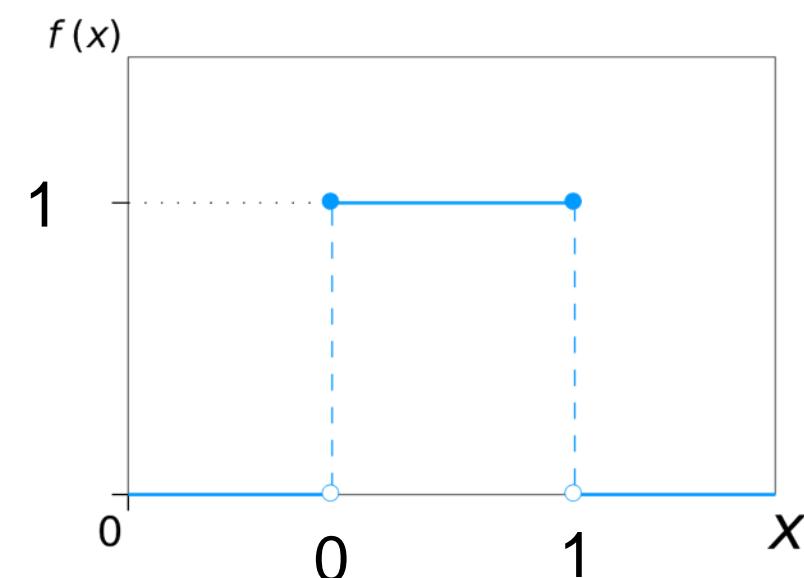
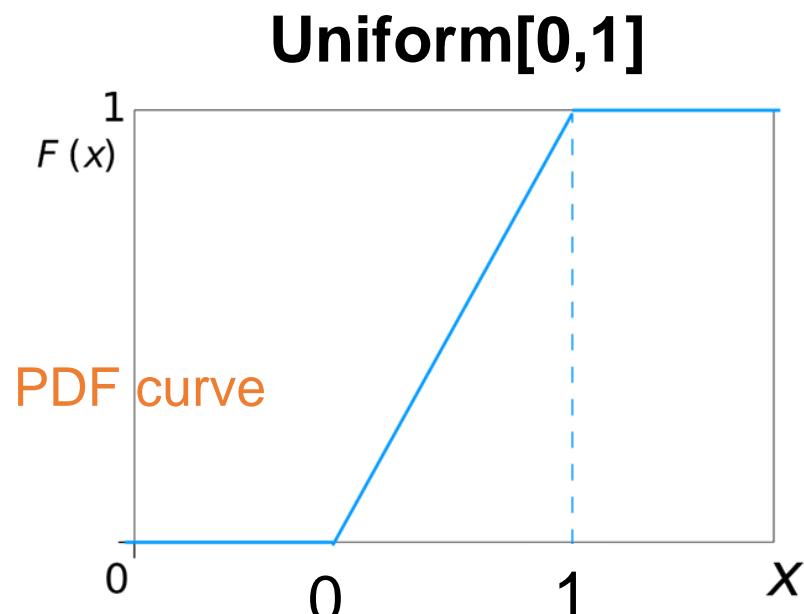
- Probability can be assigned to intervals
- Define CDF: $F(x) := P(X \leq x)$
- Then, PDF: $f(x) := p(X = x) := F'(x)$ // F 's slope at x
- $P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)dx$ // area under the PDF curve

Another viewpoint

- 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:

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



Review: Continuous Random Variable

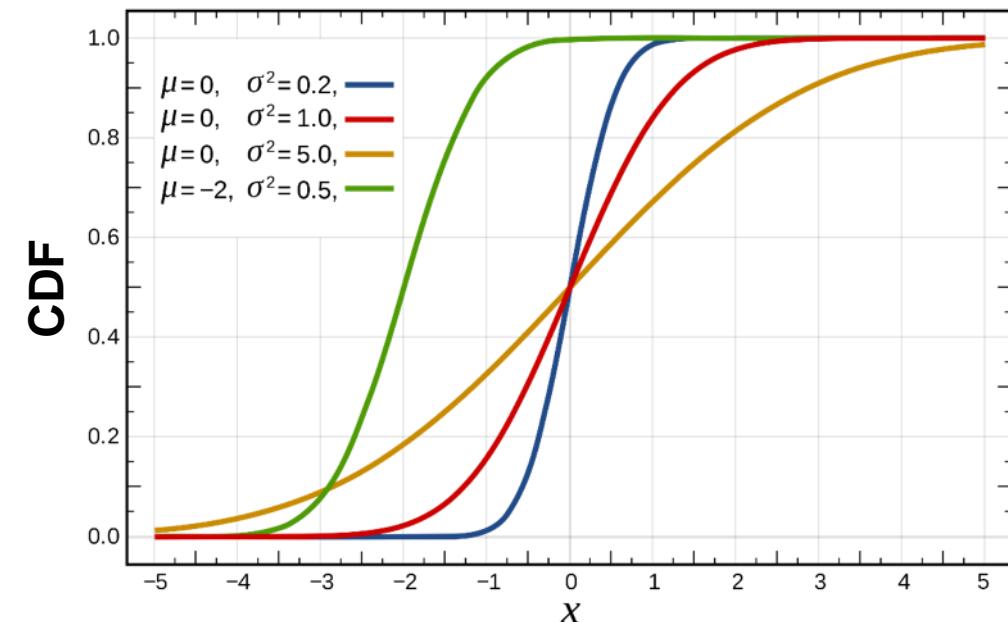
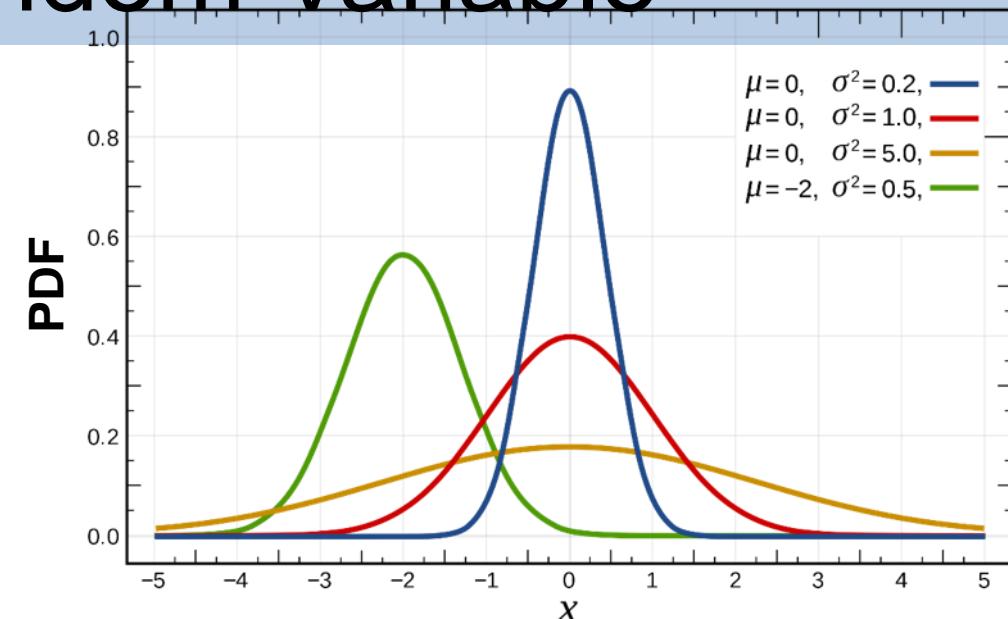
122

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

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

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

We will come back to the naming ‘mean’ / ‘variance’ in later parts of this lecture

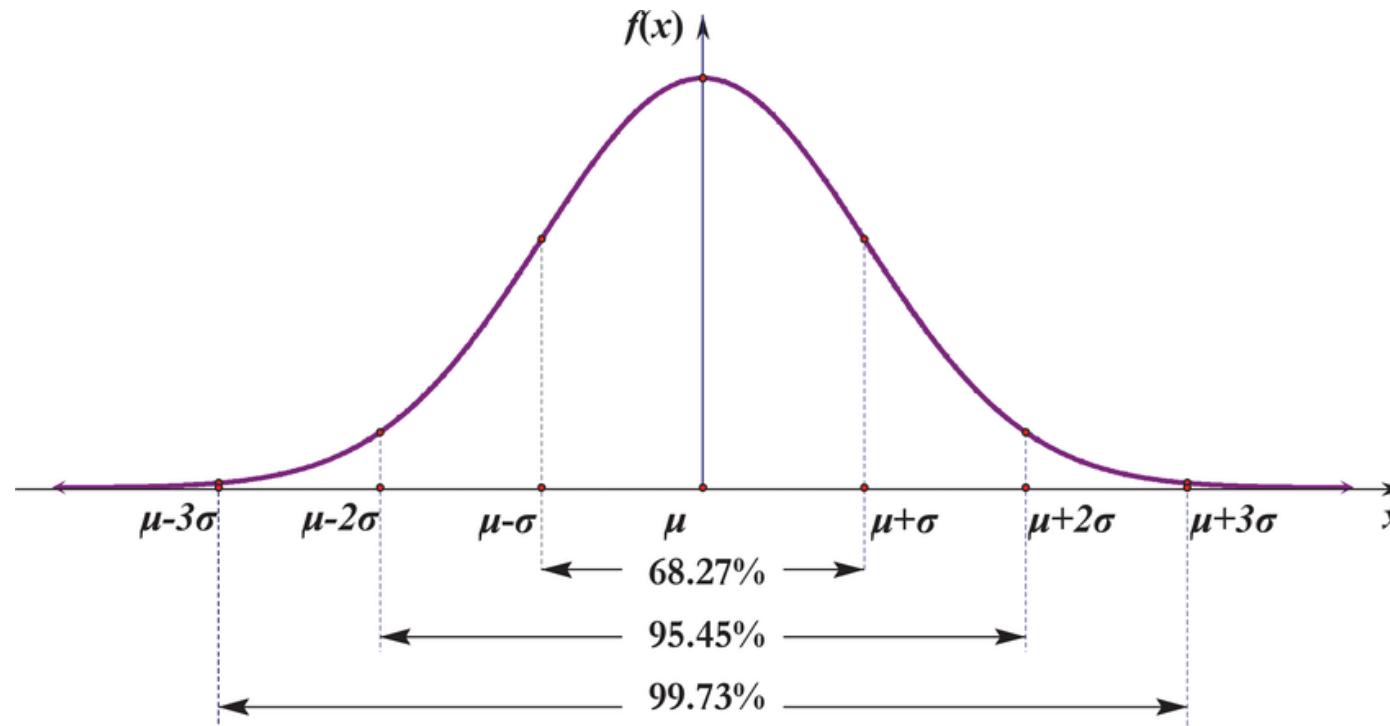


Gaussians are light-tailed

- Suppose $X \sim N(\mu, \sigma)$

Then, $P(X \in [\mu - 3\sigma, \mu + 3\sigma]) \geq 99.73\%$

Equivalently, $P(|X - \mu| > 3\sigma) \leq 0.27\%$



numpy.random

numpy.random.normal

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

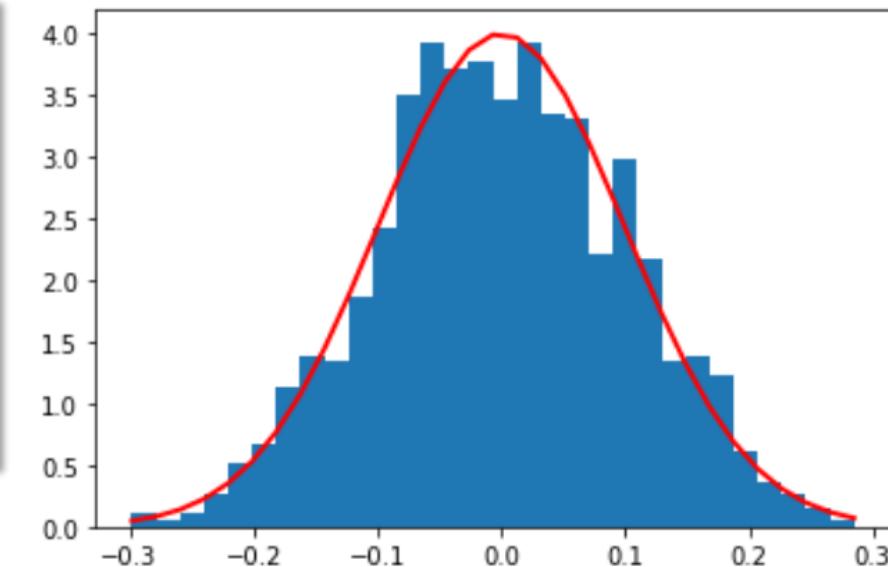
$$\text{scale} = \sqrt{\sigma^2}$$

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 independent Gaussian RVs,

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

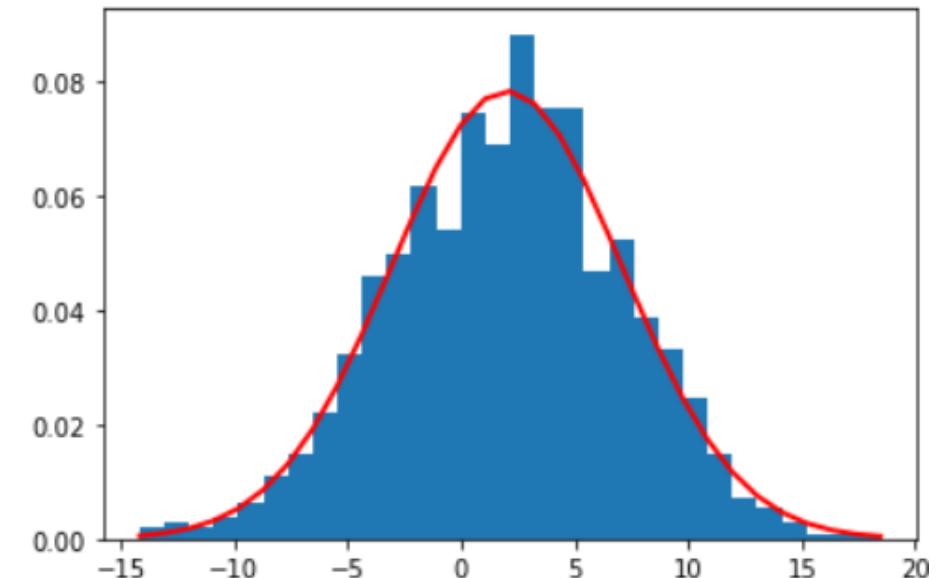
$$X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

```

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 independent Gaussian RVs,

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

Useful Continuous Distributions

Multivariate Gaussian On RV $X \in \mathcal{R}^d$

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

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

$|A|$: matrix determinant of A

Useful Properties

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

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

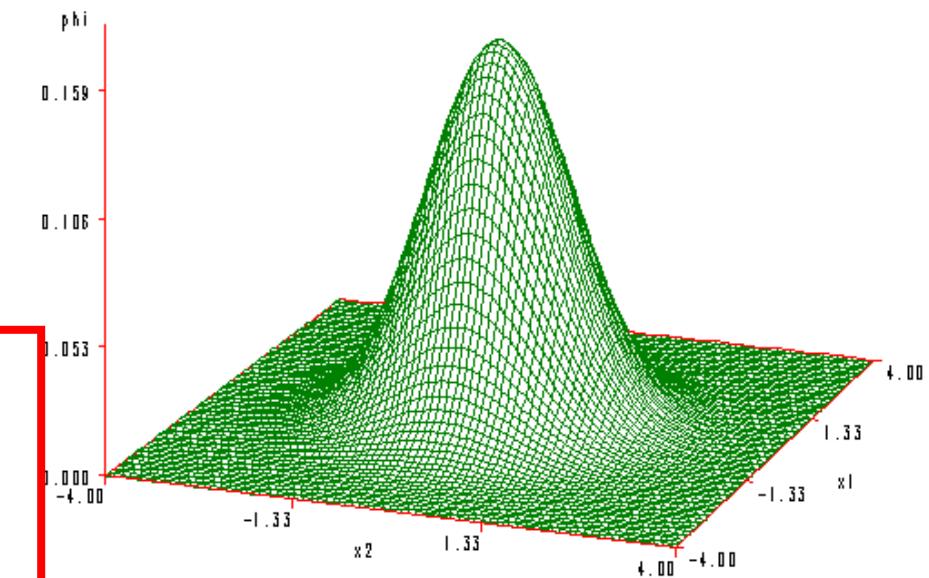
Where $A \in \mathcal{R}^{m \times d}$ and $b \in \mathcal{R}^m$ (output dimensions may change)

- Closed under conditioning and marginalization

let's cover this when needed..

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

Bivariate Normal Density – $r=0.0$



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)

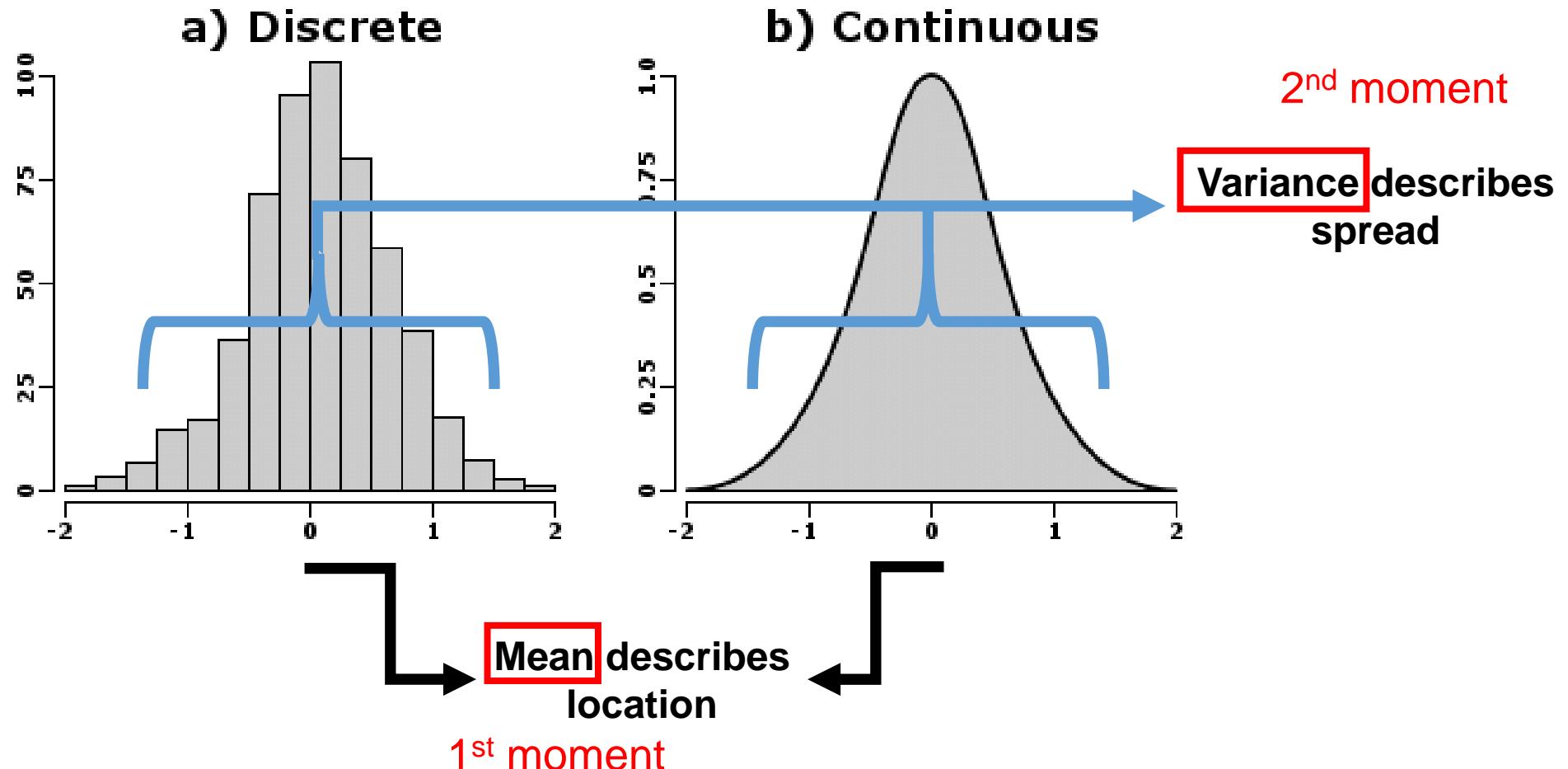
Properties of a RV are characterized by its distribution / PMF / PDF
 But there are “summary” numbers capturing important characteristics
 This is called “**moments**”.

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”

Expectation: a game-theoretic viewpoint

129

- Consider the following game:

- Flip an unfair coin X with PMF
 - You receive $\$X$

value	prob.
1	0.7
-1	0.3

- How much are you willing to pay to play the game?
 - As long as you pay $\leq \$0.4$ per game, your wealth will not decrease in the long run
 - ‘value of the game’ = $\$0.4$

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.

⇒ average is a particular ‘operator’: $\frac{1}{|x|} \sum_{x \in X} x$

⇒ in data science, average is something about the data, not the distribution behind the data

Expected Value

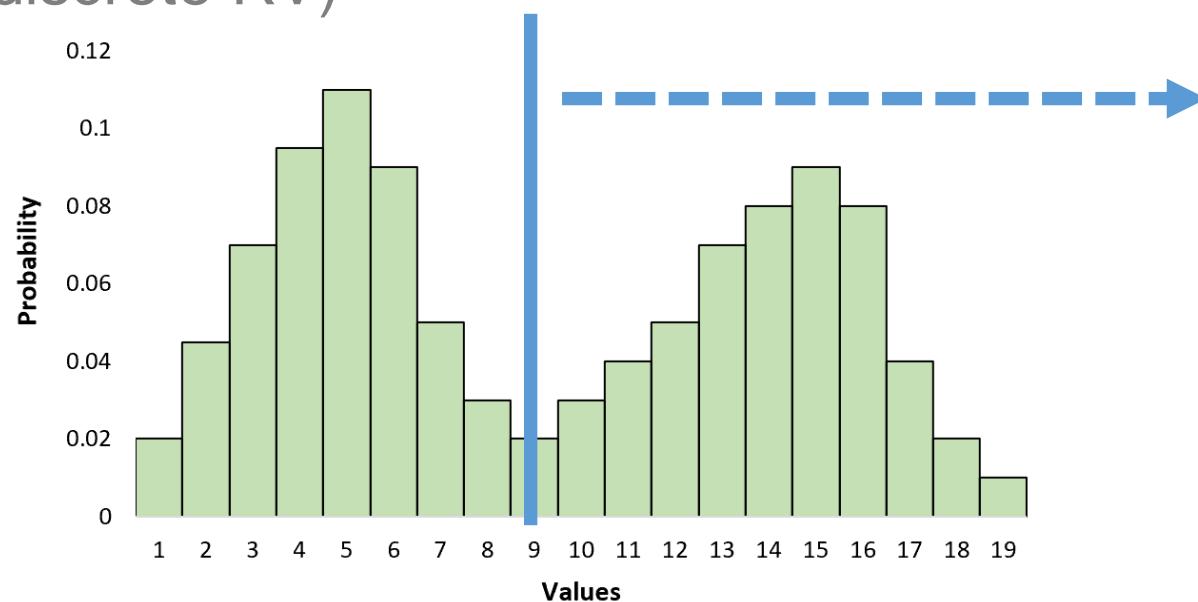
Example Let X be the sum of two fair dice, compute $E[X]$:

	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

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

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, \dots, 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 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:

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

Let $Z = f(X, Y)$

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

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

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

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

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

$$P(Z = z, X = x, Y = y) = \begin{cases} 0, & z \neq f(x, y) \\ P(X = x, Y = y), & z = f(x, y) \end{cases}$$

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: $f(ax) = a f(x)$ for all a .

So, expectation is a linear function/operator!

We will just say "linearity of expectation"

Expected Value

Definition *The conditional expectation of a discrete RV X , given Y is:*

$$\mathbf{E}[X \mid Y = y] = \sum_x x p(X = x \mid Y = y) \quad \text{cf. } \mathbf{E}[X] = \sum_x x \cdot p(X = x)$$

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

$$\begin{aligned} \mathbf{E}[X_1 \mid Y = 5] &= \sum_{x=1}^4 x p(X_1 = x \mid Y = 5) \\ &= \sum_{x=1}^4 x \frac{p(X_1 = x, Y = 5)}{p(Y = 5)} = \sum_{x=1}^4 x \frac{1/36}{4/36} = \frac{5}{2} \end{aligned}$$

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

Expected Value

Example: Two fair dice.

$Y = \text{outcome of first die}$
 $X = \text{sum of two dice}$

$$X|Y=1 \sim U\{2,3,4,5,6,7\}$$

$$E[X|Y=1] = 4.5 \quad P(Y=1) = \frac{1}{6}$$

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

$$X|Y=2 \sim U\{3,4,5,6,7,8\}$$

$$E[X|Y=2] = 5.5 \quad P(Y=2) = \frac{1}{6}$$

$$E[X|Y=3] = 6.5$$

...

$$E[X|Y=4] = 7.5 \quad \dots$$

$$E[X|Y=5] = 8.5$$

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

Expectation is 7
 $\Rightarrow E_Y[E_X[X|Y]] = 7$
 \Rightarrow coincides with $E[X]$ we
 computed before!
 Actually not a coincidence;
Law of total expectation

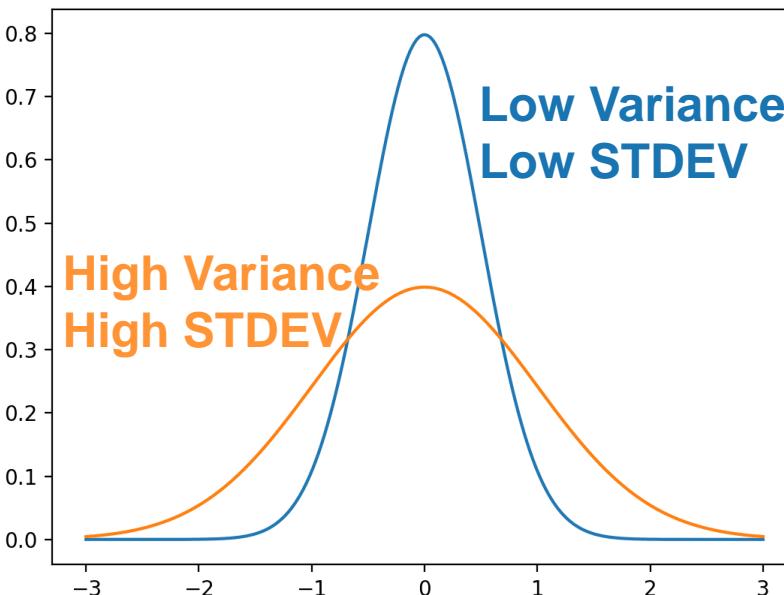


Variance

Definition The variance of a RV X is defined as,

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

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



- Describes the “spread” of a distribution
- Describes uncertainty of outcome
- STDEV is in original units (more intuitive), variance is in units²
- Variance is more mathematically useful than STDEV

Variance

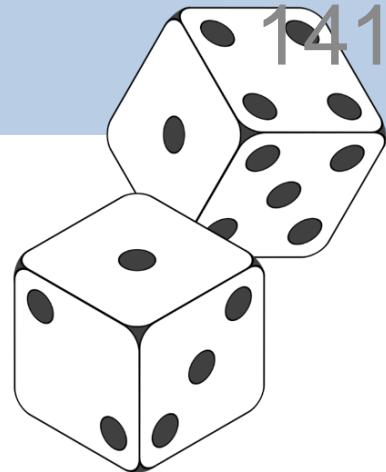
141

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

The variance is then,

$$\begin{aligned}\text{Var}(X) &= \sum_{i=1}^6 \frac{1}{6} \left(i - \frac{7}{2} \right)^2 \\ &= \frac{1}{6} \left((-5/2)^2 + (-3/2)^2 + (-1/2)^2 + (1/2)^2 + (3/2)^2 + (5/2)^2 \right) \\ &= \frac{35}{12} \approx 2.92.\end{aligned}$$

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



Variance

Lemma An equivalent form of variance is:

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

Proof

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



CSC380: Principles of Data Science

Probability Primer 6

Announcement

- Discussion this week
- Upload the questions by next Monday night.
- Please participate in the discussion.

Cumulative Distribution Functions

Definition *The cumulative distribution function (CDF) of an RV X is the function given by,*

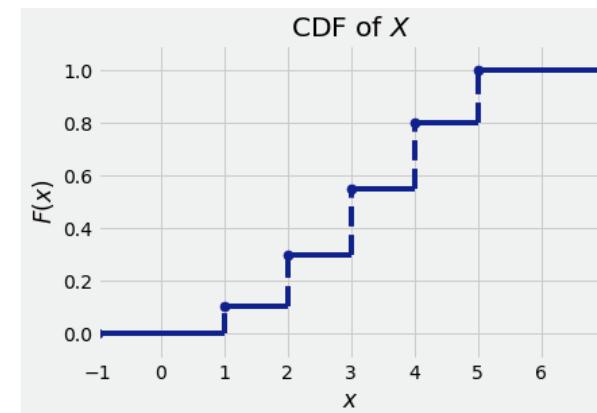
$$F(x) = P(X \leq x)$$

What does the CDF and PDF of a discrete RV look like?

i	1	2	3	4	5
$P(X = i)$	0.1	0.2	0.25	0.25	0.2

Example: X has PMF

X 's CDF:



X 's PDF: undefined in “points with positive probability mass”, 0 everywhere else

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

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

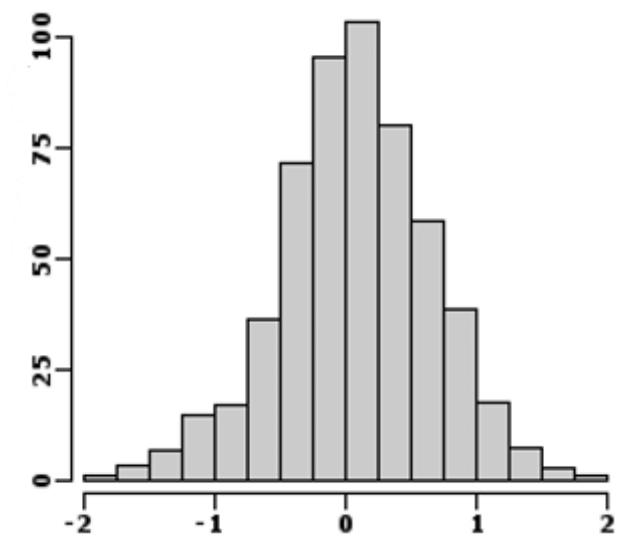
Summation over all
values in domain of X

1st order moment; location

Definition The variance of a RV X is defined as,

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

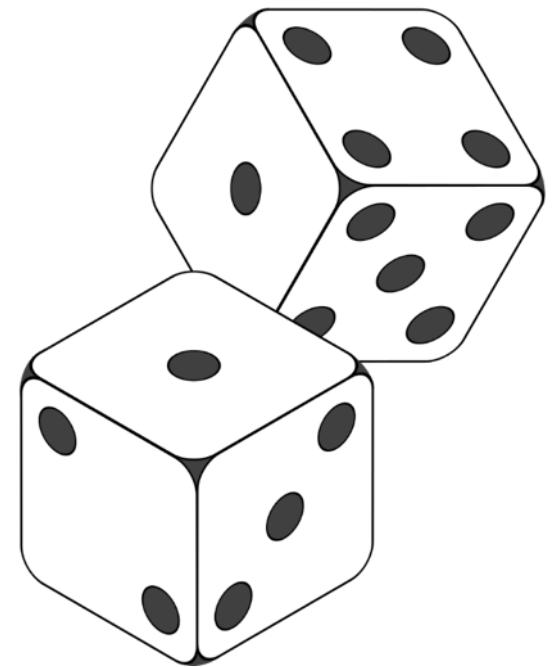
2nd order moment; spread



Variance

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

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



Variance

- If c is a constant, $\text{Var}[cX] = c^2 \text{Var}[X]$
 - Exercise: try to convince yourself why this is true
 - Hint: use $\text{E}[bX] = b \cdot \text{E}[X]$
- Important that c has to be a constant here!

Moments of Continuous RVs

Replace all sums with integrals,

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

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

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

Moments of Useful Discrete Distributions

150

Bernoulli A.k.a. the *coinflip* distribution on binary 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

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

Binomial Sum of N independent coinflips,

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

With moments,

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

^: (by linearity of expectation)



Moments of Useful Discrete Distributions

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

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

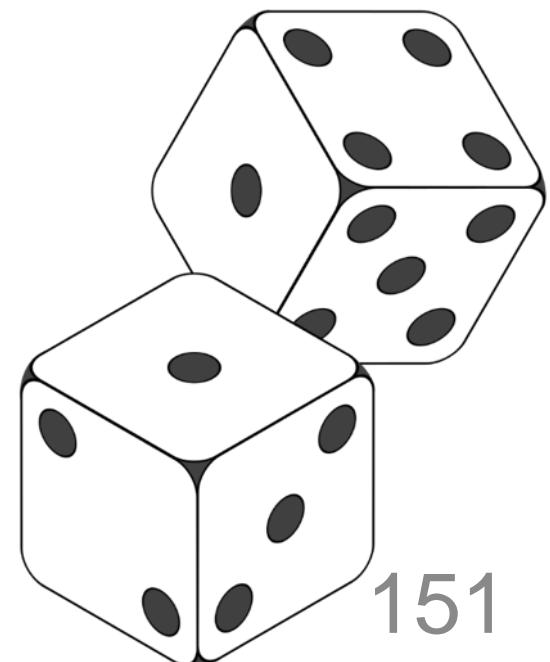
Where parameter $\pi \in [0, 1]^K$ is a probability vector,

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

Marginal moments are given by,

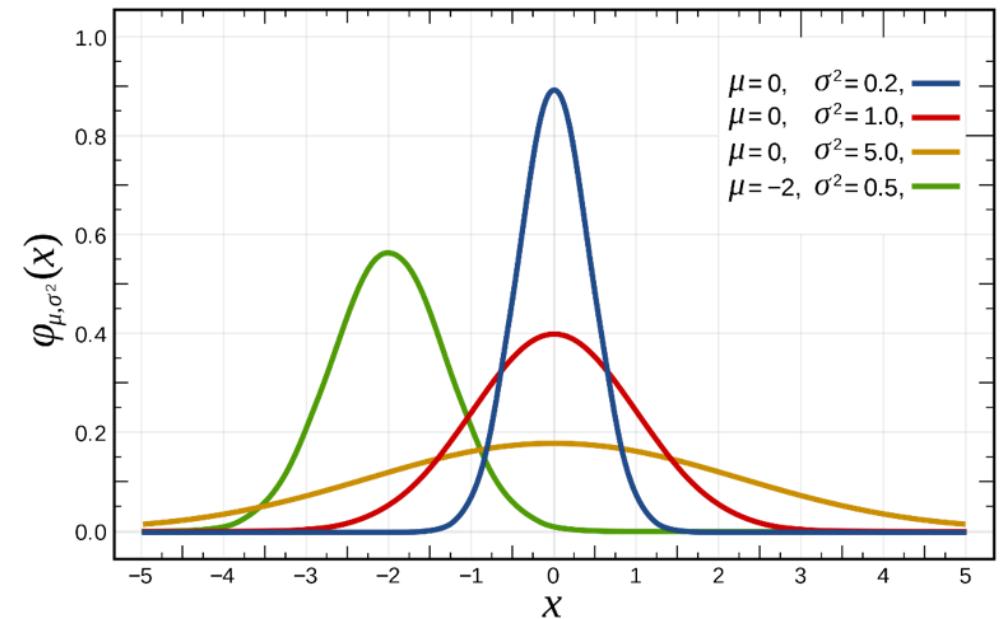
$$\mathbf{E}[X_k] = N\pi_k \quad \mathbf{Var}[X_k] = N\pi_k(1 - \pi_k)$$

*Moments are similar to Binomial, but over K outcomes
(In fact, $X_k \sim \text{Binomial}(N, \pi_k)$ – can you see why?)*



Moments of Gaussians

- $X \sim N(\mu, \sigma^2)$
- **Fact:** $E[X] = \mu$, $\text{var}(X) = \sigma^2$
- **Intuition:**
 - X 's probability density is distributed symmetrically around μ
 - Smaller $\sigma \Rightarrow$ sharper peak in PDF \Rightarrow smaller $\text{var}(X)$



$$E[X] = \int xp(x) dx$$

$$\text{Var}[X] = \int (x - E[X])^2 p(x) dx$$

Covariance

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

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

Question *What is $\text{Cov}(X, X)$?*

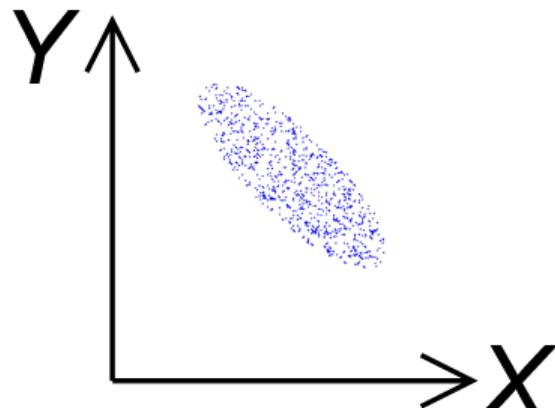
Answer $\text{Cov}(X, X) = \text{Var}(X)$

Covariance

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

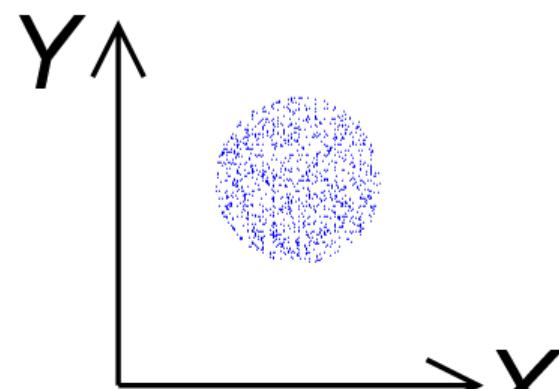
$$\text{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$$

Measures the linear relationship between X and Y

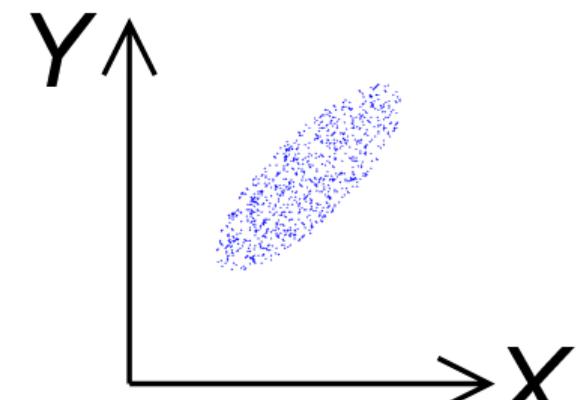


$$\text{cov}(X, Y) < 0$$

Negative relationship



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



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

Positive relationship

Example: height vs weight

Covariance

- A shortcut to compute covariance.

$$\begin{aligned}\bullet \quad Cov(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - X \cdot E[Y] - Y \cdot E[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \\ \bullet \quad \text{Sanity check: } Cov(X, X) &= E[XX] - E[X]E[X] = Var(X)\end{aligned}$$

Covariance

Lemma For any two RVs X and Y ,

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

\Rightarrow variance is not a linear operator.

Intuition of the $2 \text{Cov}(X, Y)$ “correction” term: $Y = X$; $Y = -X$

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

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

(Whole square formula) $= \mathbf{E}[(X - \mathbf{E}[X])^2 + (Y - \mathbf{E}[Y])^2 + 2(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$

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

Correlation

Definition *The Pearson correlation (abbrev. correlation) of X and Y is given by,*

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

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

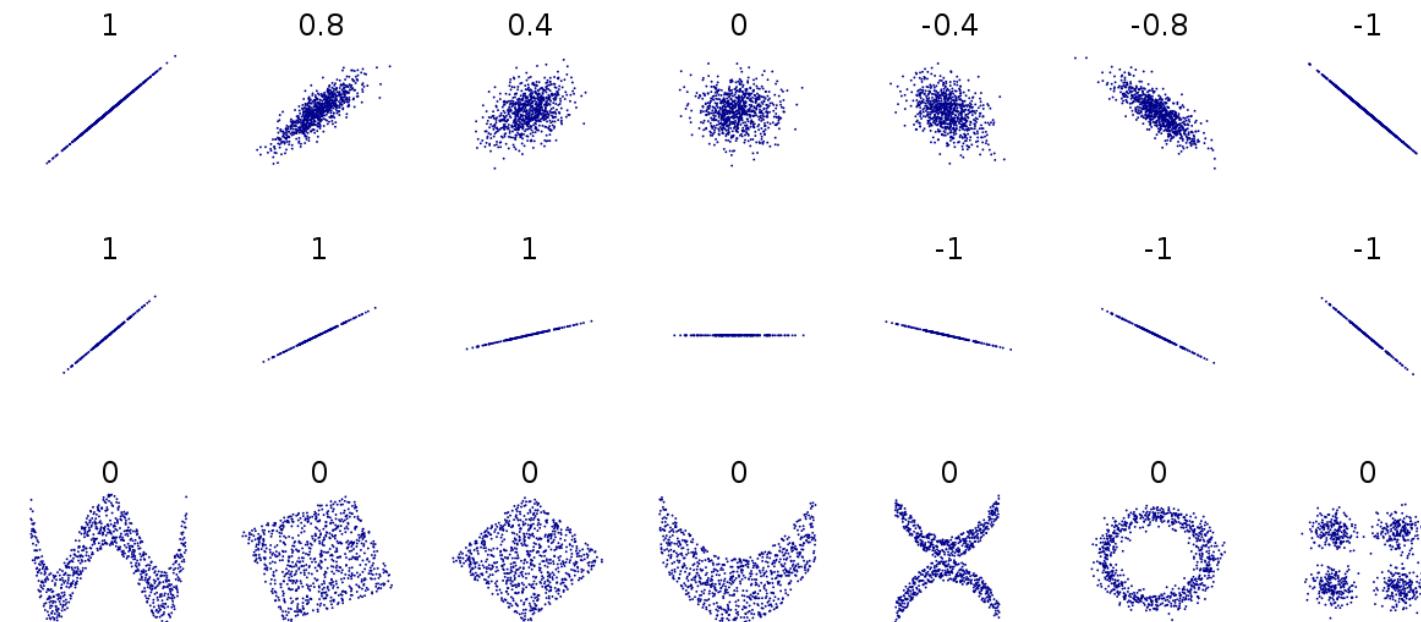


Useful when you are interested in how X and Y are related, independent of the individual variability.
⇒ $\text{Cov}(cX, dY) \neq \text{Cov}(X, Y)$ **but** $\text{Corr}(cX, dY) = \text{Corr}(X, Y)$

Correlation

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

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



Like covariance, only expresses linear relationships!

Independence and Moments

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

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

Independence and Moments

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

Proof:

$$\begin{aligned}
 \text{E}[XY] &= \sum_x \sum_y (x \cdot y) p(X = x, Y = y) \\
 &= \sum_x \sum_y (x \cdot y) p(X = x) p(Y = y) && (\text{Independence}) \\
 &= \left(\sum_x x \cdot p(X = x) \right) \left(\sum_y y \cdot p(Y = y) \right) = \text{E}[X]\text{E}[Y] && (\text{Linearity of Sum})
 \end{aligned}$$

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

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

Independence and Moments

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

- Further applications on this theorem
 - **Fact:** If $X \perp Y$, then for any functions f, g , $f(X) \perp g(Y)$
 - **Important Corollary:** If $X \perp Y$, then $\mathbf{E}[f(X)g(Y)] = \mathbf{E}[f(X)] \cdot \mathbf{E}[g(Y)]$
 - E.g. this implies that, if $X \perp Y$, $\mathbf{E}[X^2Y^3] = \mathbf{E}[X^2] \cdot \mathbf{E}[Y^3]$

Independence and Moments

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

$$\begin{aligned}\mathbf{Var}[X_1 + X_2] &= \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2\mathbf{Cov}(X_1, X_2) \\&= \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2\mathbf{E}[(X_1 - \mathbf{E}[X_1])(X_2 - \mathbf{E}[X_2])] \\&= \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2\mathbf{E}[(X_1 - \mathbf{E}[X_1])]\mathbf{E}[(X_2 - \mathbf{E}[X_2])] \quad Y_1 \perp Y_2 \Rightarrow f(Y_1) \perp g(Y_2) \\&= \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2(\mathbf{E}[X_1] - \mathbf{E}[X_1])(\mathbf{E}[X_2] - \mathbf{E}[X_2]) \\&= \mathbf{Var}[X_1] + \mathbf{Var}[X_2]\end{aligned}$$

Independence and Moments

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

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

If X and Y are independent then they have zero covariance,

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

Thus,

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$$

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

$$\text{Var}\left(\sum_{i=1}^N X_i\right) = \sum_{i=1}^N \text{Var}(X_i)$$

Q: Is variance is a linear operator under independence?

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

Example: Independent Gaussian RVs

Let X and Y be independent Gaussian random variables with,

$$X \sim \mathcal{N}(\mu_x, \sigma_x^2)$$

$$Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$$

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

What is the variance of their sum?

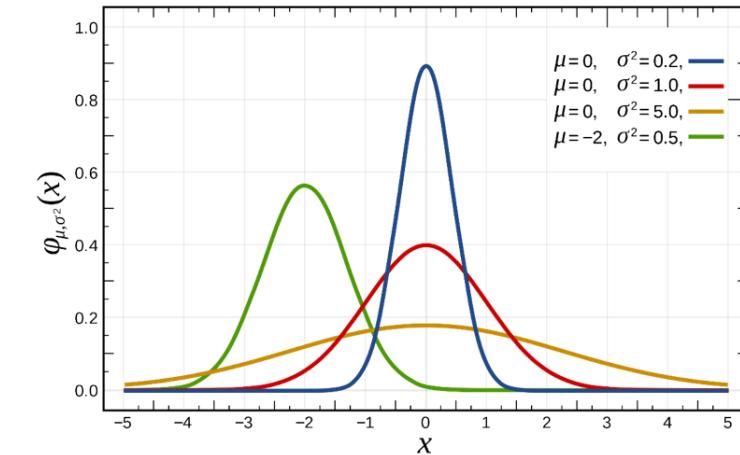
$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) = \sigma_x^2 + \sigma_y^2$$

What is the mean of their product?

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

What is the mean of their sum?

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

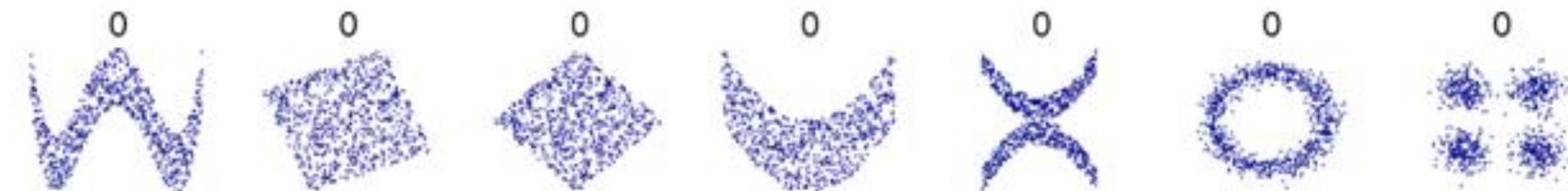


Independence and Moments

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

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

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



A Numerical Counterexample

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

Recall: $\text{Cov}(X, Y) = E[XY] - E[X] \cdot E[Y]$

$$E[X] = 0$$

$$E[Y] = 0$$

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

Q: how to check independence between X and Y ?

$$P(Y=1 \mid X=-1) = 0, P(Y=1) = .25 \Rightarrow \text{not independent}$$

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