Computer
Science

CSC535: Probabilistic Graphical Models

Probability Primer

Prof. Jason Pacheco

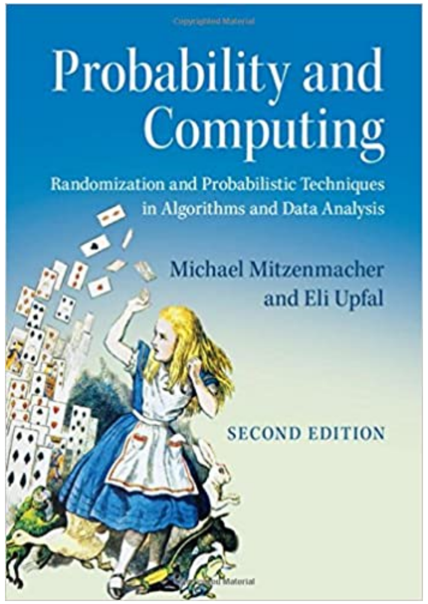
Administrivia

- Homework 1 will be out Mon 8/31, next week
- Reading: Murphy, Secs. 2.1 and 2.2
- Questions: Raise hand in Zoom or get my attention
- Lots of source material from this book…

Mitzenmacher, M. and Upfal, E.
"Probability and Computing"

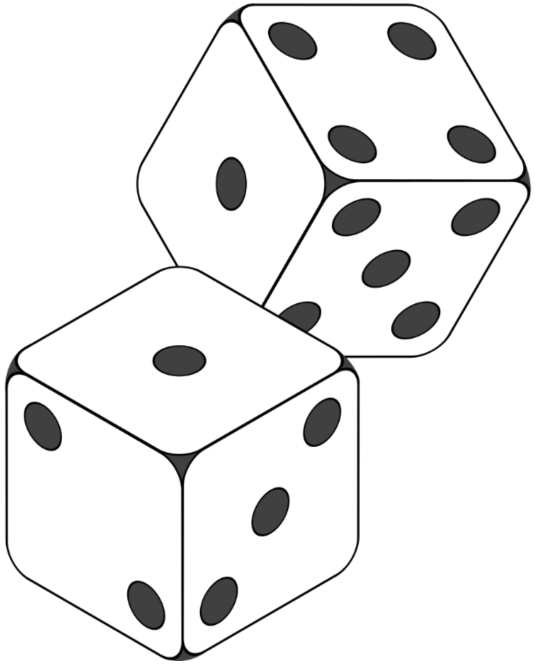
← First Edition Second Edition →

[(Lecture Slides from Eli Upfal's course)]



Random Events and Probability

Suppose we roll <u>two fair dice</u>…

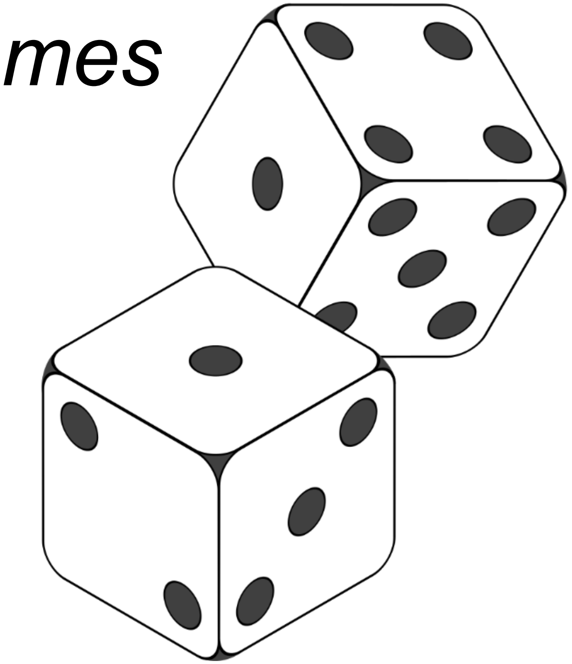


- What are the possible outcomes?
- What is the *probability* of rolling **even** numbers?
- What is the *probability* of rolling **odd** numbers?
- If one die rolls 1, then what is the probability of the second die also rolling 1?
- How to mathematically formulate outcomes and their probabilities?

*…this is an **experiment** or **random process**.*

Formulate as <u>probability space</u> having <u>3 components</u>

Random Events and Probability



1. A **sample space** Ω: *set of all possible outcomes* of the experiment.

Dice Example: All combinations of dice rolls,

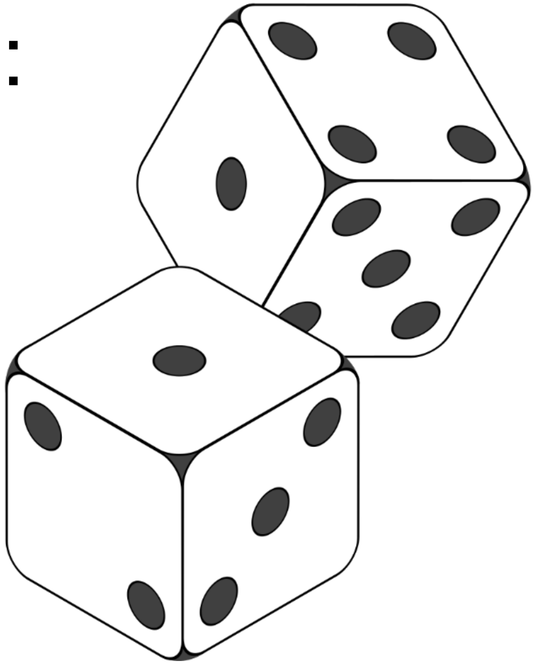
$$\Omega = \{(1,1),(1,2),\ldots,(6,5),(6,6)\}$$

2. An **event space** $\mathcal{F}$: Family of sets representing <u>allowable events</u>, where each set in $\mathcal{F}$ is a subset of the sample space Ω.

Dice Example: Event that we roll <u>even numbers</u>,

$$E = \{(2,2),(2,4),\ldots,(6,4),(6,6)\} \in \mathcal{F}$$

Random Events and Probability



3 A **probability function** $P : \mathcal{F} \to \mathbf{R}$ satisfying:

Axioms of Probability

1. For any event E, $0 \leq P(E) \leq 1$
2. $P(\Omega) = 1 \text{ and } P(\emptyset) = 0$
3. For any *finite* or *countably infinite* sequence of pairwise mutually disjoint events $E_1, E_2, E_3, \ldots$

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

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

$$P((2,2) \cup (2,4) \cup \ldots \cup (6,6)) = P((2,2)) + P((2,4)) + \ldots + P((6,6))$$

9 Possible outcomes, each with equal probability of occurring

$$= \frac{1}{36} + \frac{1}{36} + \ldots + \frac{1}{36} = \frac{9}{36}$$

Random Events and Probability

Some rules regarding set of event space $\mathcal{F}$*…*

- $\mathcal{F}$ must include $\emptyset$ and Ω
- $\mathcal{F}$ is **closed** under set operations, if $E_1, E_2 \in \mathcal{F}$ then:
  - $E_1 \cup E_2 \in \mathcal{F}$
  - $E_1 \cap E_2 \in \mathcal{F}$
  - $\overline{E_1} = \Omega - E_1 \in \mathcal{F}$

Random Events and Probability

Two dice example: *If* $E_1, E_2 \in \mathcal{F}$ *where,*

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

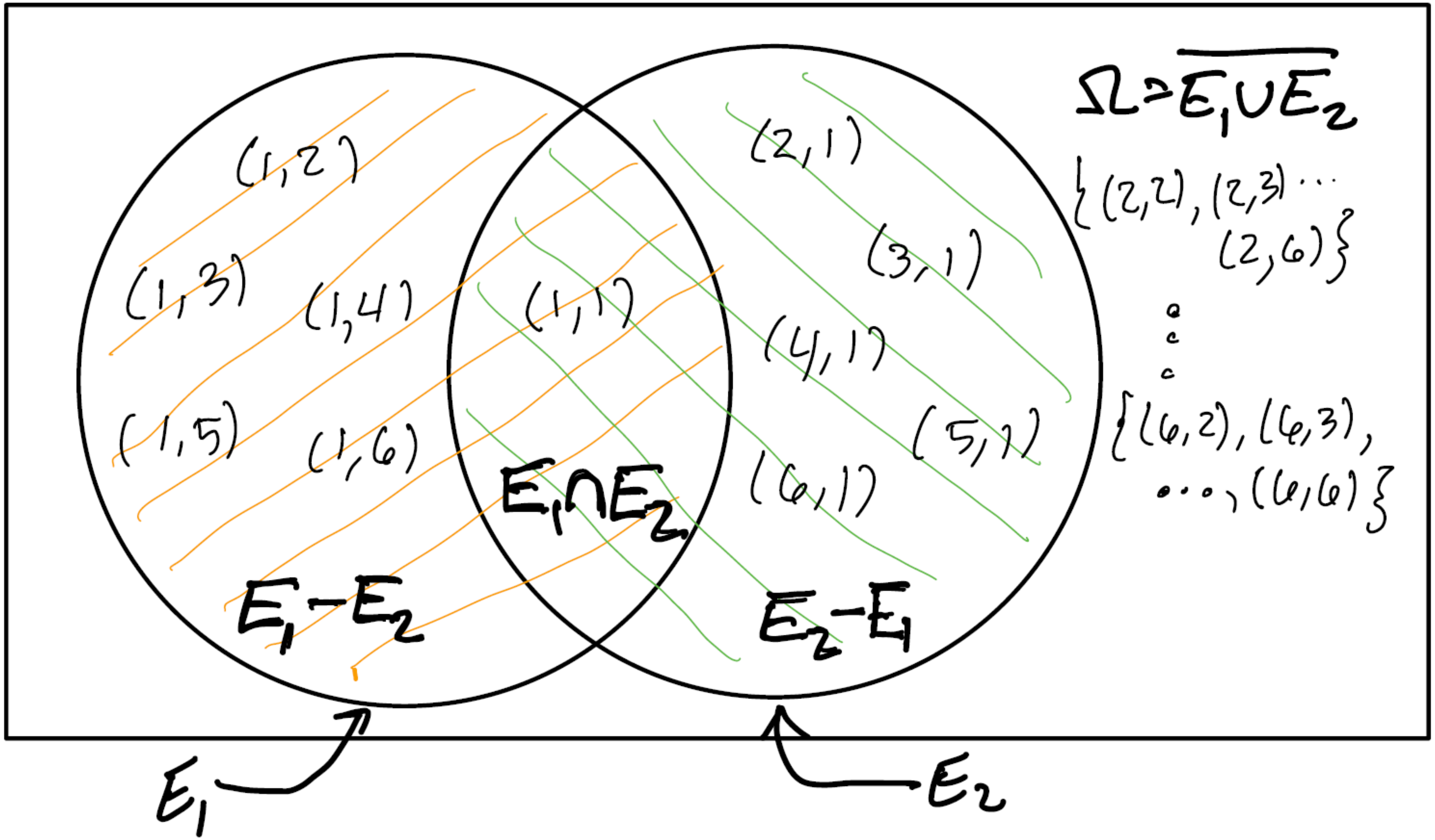
$$E_1 = \{(1,1),(1,2),\ldots,(1,6)\} \qquad E_2 = \{(1,1),(2,1),\ldots,(6,1)\}$$

Then we must include (at least) the following events…

Operation	Value	Interpretation
$E_1 \cup E_2$	$\{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,1)\}$	Any die rolls 1
$E_1 \cap E_2$	$\{(1,1)\}$	Both dice roll 1
$E_1 - E_2$	$\{(1,2),(1,3),(1,4),(1,5),(1,6)\}$	First die rolls 1 only
$\overline{E_1 \cup E_2}$	$\{(2,2),(2,3),\ldots,(2,6),(3,2),\ldots,(6,6)\}$	No die rolls 1

Random Events and Probability

Can interpret these operations as a Venn diagram…

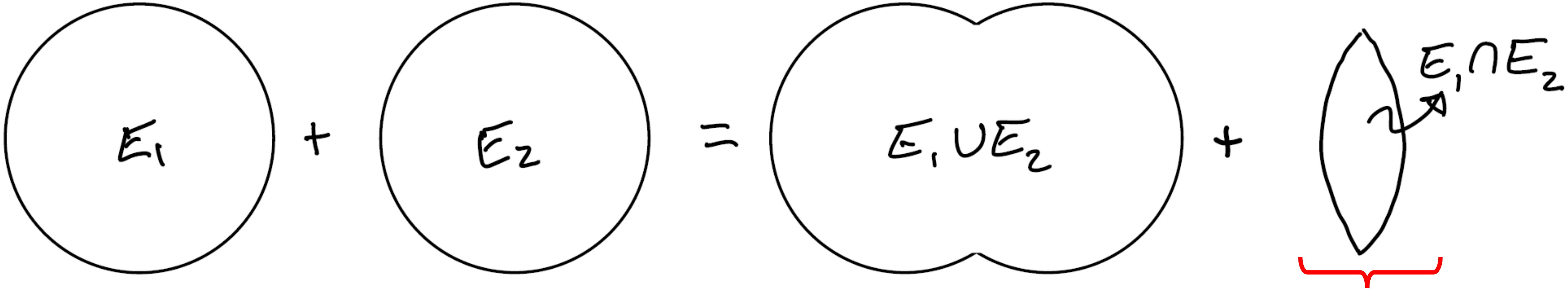


Random Events and Probability

Lemma: *For any two events* E_1 *and* E_2 *,*

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

Graphical Proof:



Subtract from both sides

Random Events and Probability

Lemma: *For <u>any</u> two events* E_1 *and* E_2 *,*

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

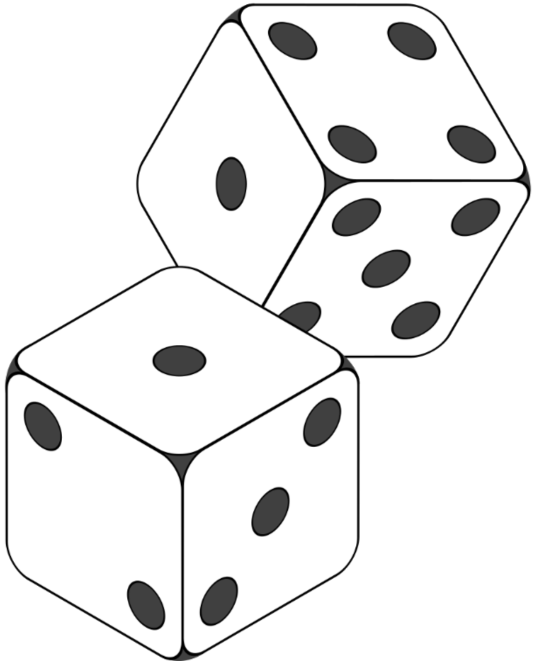
Proof:

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

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

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

Random Variables



Suppose we are interested in a distribution over the <u>sum of dice</u>…

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

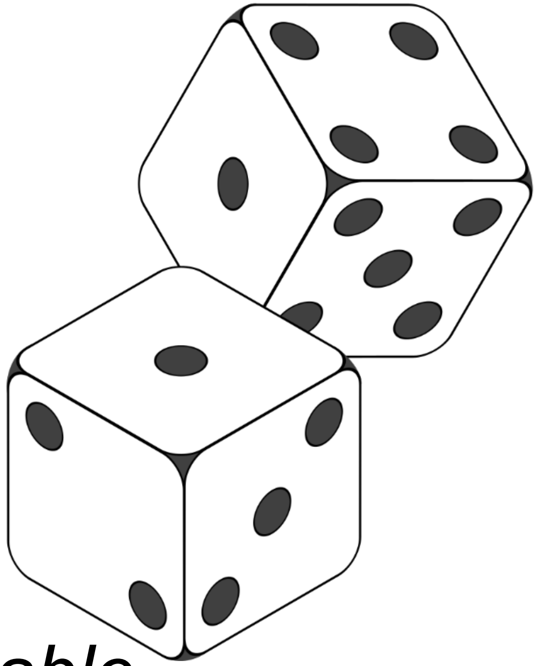
Two dice example:

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

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

Enumerate all possible means of obtaining desired sum. Gets cumbersome for N>2 dice…

Random Variables



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* $X : \Omega \rightarrow \mathbb{R}$*. A discrete random variable takes on only a finite or countably infinite number of values.*

For *discrete* RVs $X = x$ is an **event** with **probability mass function**:

$$p(X = x) = \sum_{\omega \in \Omega \,:\, X(\omega) = x} P(\omega)$$

Random Variables

Some notes on random variables (RVs)…

- We denote the RV by capital X and its realization by lowercase x
- Generally use shorthand X instead of $X(\omega)$
- Other common shorthand: $p(x) = p(X = x)$
- Any function $f(X)$ of an RV is also an RV, e.g. $Y(\omega) = f(X(\omega))$
- More shorthand: the joint distribution of RVs $p(X, Y) = p(X \cap Y)$
- We will use *“distribution”* loosely to refer to distributions, PMFs, probability density and cumulative distribution functions (defined later)

Fundamental Rules of Probability

Given two RVs X and Y the **conditional distribution** is:

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

Multiply both sides by $p(Y)$ to obtain the **probability chain rule**:

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

For N RVs $X_1, X_2, \ldots, X_N$:

$$p(X_1, X_2, \ldots, X_N) = p(X_1)p(X_2 \mid X_1) \ldots p(X_N \mid X_{N-1}, \ldots, X_1)$$

Chain rule valid for any ordering

$$= p(X_1) \prod_{i=2}^{N} p(X_i \mid X_{i-1}, \ldots, X_1)$$

Fundamental Rules of Probability

Law of total probability

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

Proof $$\sum_x p(Y, X = x) = \sum_x p(Y) p(X = x \mid Y) \quad \text{(chain rule)}$$

$$= p(Y) \sum_x p(X = x \mid Y) \quad \text{(distributive property)}$$

$$= p(Y) \quad \text{(axiom of probability)}$$

Generalization for conditionals:

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

Independence of RVs

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

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

Which of the following are true?

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

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

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

Independence of RVs

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

[✓ a) $p(X_1 = 1 | X_3 = 3) > p(X_1 = 1)$]

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

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

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

$(X_1 = 1, X_2 = 2)$ *or* $(X_1 = 2, X_2 = 1)$

so

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

Independence of RVs

Intuition…

Consider $P(B|A)$ where you want to bet on B

Should you pay to know A?

In general you would pay something for A if it changed your belief about B. In other words if,

$$P(B|A) \neq P(B)$$

Independence of RVs

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

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

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

Definition RVs $X_1, X_2, \ldots, X_N$ are <u>mutually independent</u> if and only if,

$$p(X_1 = x_1, \ldots, X_N = x_N) = \prod_{i=1}^{N} p(X_i = x_i)$$

- ➢ Independence is *symmetric:* $X \perp Y \Leftrightarrow Y \perp X$
- ➢ Equivalent definition of independence: $p(X \mid Y) = p(X)$

Independence of RVs

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

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

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

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

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

Shorthand notation
Implies for all *x, y, z*

- Equivalent def'n of conditional independence: $p(X \mid Y, Z) = p(X \mid Z)$
- Symmetric: $X \perp Y \mid Z \Leftrightarrow Y \perp X \mid Z$

Administrivia

- Homework 1
  - Out now, see course webpage or D2L
  - Due: Wed, 9/9
  - (Easy) Worth 4 points vs. standard 7
- Office hours:
  - Tue, 3-4:30pm (local time)
  - Will add Zoom meeting in D2L
  - Optional hours: Thurs, 9-10:30am (message me on Piazza before)

Recap

- A random process is modeled by a probability space $(\Omega, \mathcal{F}, P)$ where:
  - **Sample space** Ω is the set of all possible outcomes
  - **Event space** $\mathcal{F}$ is the set of **events**, each being a subset of Ω
  - **Probability function** P assigns a probability in $[0, 1]$ to each event

- Axioms of probability

1. For any event E, $0 \leq P(E) \leq 1$
2. $P(\Omega) = 1 \text{ and } P(\emptyset) = 0$
3. For any *finite* or *countably infinite* sequence of pairwise mutually disjoint events $E_1, E_2, E_3, \ldots$

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

- An event space must contain $\{\Omega, \emptyset\}$
- Must be closed under:
  - Complements
  - Countable unions
  - Countable intersections

Recap

- A **random variable** is a function of samples to real values: $X : \Omega \to \mathbb{R}$
- $X = x$ Is an event with probability: $p(X = x) = \sum_{\omega \in \Omega \,:\, X(\omega) = x} P(\omega)$
- Some fundamental rules of probability:
  - Conditional: $p(X \mid Y) = \frac{p(X,Y)}{p(Y)} = \frac{p(X,Y)}{\sum_x p(X=x,Y)}$
  - Law of total probability: $p(Y) = \sum_x p(Y, X = x)$
  - Probability chain rule: $p(X,Y) = p(Y)p(X \mid Y)$
- Independence of RVs:
  - Two RVs X & Y are independent iff: $p(X \mid Y) = p(X)$
  - Equivalently: $p(X,Y) = p(X)p(Y)$
  - X & Y are conditionally independent given Z iff: $p(X \mid Y, Z) = p(X \mid Z)$
  - Equivalently: $p(X, Y \mid Z) = p(X \mid Z)p(Y \mid Z)$

Outline

- Moments of (discrete) random variables
- Some useful discrete distributions
- Continuous probability

Outline

- Moments of (discrete) random variables
- Some useful discrete distributions
- Continuous probability

Moments of RVs

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

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

Summation over all values in domain of X

Example Let X be the sum of two fair dice, then:

$$\mathbf{E}[X] = \frac{1}{36} \cdot 2 + \frac{1}{36} \cdot 3 + \ldots + \frac{1}{36} \cdot 12 = 7$$

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

Corollary *For any constant c*
$\mathbf{E}[cX] = c\mathbf{E}[X]$

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

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

Moments of RVs

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

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

Example Roll two standard six-sided dice and let X be the result of the first die and let Y be the sum of both dice, then:

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

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

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

Moments of RVs

Definition *The <u>variance</u> of a RV X is defined as,*

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

(X-units)^2

The <u>standard deviation</u> is $\sigma[X] = \sqrt{\mathbf{Var}[X]}$.

(X-units)

Lemma An equivalent form of variance is:

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

Proof Keep in mind that $E[X]$ is a constant,

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

Moments of RVs

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

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

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

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

e.g. variance is not a linear operator.

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

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

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

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

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

Moments of RVs

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

Proof:

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

$$= \sum_x \sum_y (x \cdot y) p(X = x) p(Y = y)$$

(Independence)

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

(Linearity of Expectation)

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

$$\mathbf{E}[X_1 X_2] = \mathbf{E}[X_1]\mathbf{E}[X_2] = 3.5^2 = 12 \cdot \frac{1}{4}$$

Moments of RVs

Question: *What is the variance of their sum?*

$$\begin{aligned}\mathbf{Var}[X_1 + X_2] &= \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2\mathbf{Cov}(X_1, X_2) \\ &= \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2\mathbf{E}[(X_1 - \mathbf{E}[X_1])(X_2 - \mathbf{E}[X_2])] \\ &= \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2\mathbf{E}[(X_1 - \mathbf{E}[X_1])]\mathbf{E}[(X_2 - \mathbf{E}[X_2])] \\ &= \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2\left(\mathbf{E}[X_1] - \mathbf{E}[X_1]\right)\left(\mathbf{E}[X_2] - \mathbf{E}[X_2]\right) \\ &= \mathbf{Var}[X_1] + \mathbf{Var}[X_2]\end{aligned}$$

Theorem: *If* $X \perp Y$ *then* $\mathbf{Var}[X + Y] = \mathbf{Var}[X] + \mathbf{Var}[Y]$

Corollary: *If* $X \perp Y$ *then* $\mathbf{Cov}(X, Y) = 0$

Corollary: *For collection of RVs* $X_1, X_2, \ldots, X_N$ *:* $\mathbf{Var}(\sum_{i=1}^{N} X_i) = \sum_{i=1}^{N} \mathbf{Var}(X_i)$

Moments of RVs

Law of Total Expectation *Let X and Y be discrete RVs with finite expectations, then:*

$$\mathbf{E}[X] = \mathbf{E}_Y[\mathbf{E}_X[X \mid Y]]$$

Proof

$$\begin{aligned}
\mathbf{E}_Y[\mathbf{E}_X[X \mid Y]] &= \mathbf{E}_Y\left[\sum_x x \cdot p(x \mid Y)\right] \\
&= \sum_y \left[\sum_x x \cdot p(x \mid y)\right] \cdot p(y) && \textbf{(Definition of expectation)} \\
&= \sum_y \sum_x x \cdot p(x, y) && \textbf{(Probability chain rule)} \\
&= \sum_x x \sum_y \cdot p(x, y) && \textbf{(Linearity of expectations)} \\
&= \sum_x x \cdot p(x) = \mathbf{E}[X] && \textbf{(Law of total probability)}
\end{aligned}$$

Outline

- Moments of (discrete) random variables
- Some useful discrete distributions
- Continuous probability

Useful Discrete Distributions

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*** *(e.g. heads), and also the mean*

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

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

Binomial Dist.

$$p(\underbrace{Y = k}_{\text{Num. “successes” out of N trials}}) = \underbrace{\binom{N}{k}}_{\text{Num. ways to obtain k successes out of N}} \pi^k (1 - \pi)^{N-k}$$

Binomial Mean: $\mathbf{E}[Y] = N \cdot \pi$ **Sum of means for N indep. Bernoulli RVs**

Useful Discrete Distributions

Represent joint Bernoulli distribution as probability table…

For Binomial use K-by-K probability table.

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

$P(x_1)$

$P(x_2)$

$P(y_1)$

$P(y_2)$

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

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

Useful Discrete Distributions

We don’t care about event Y=y_2

Y

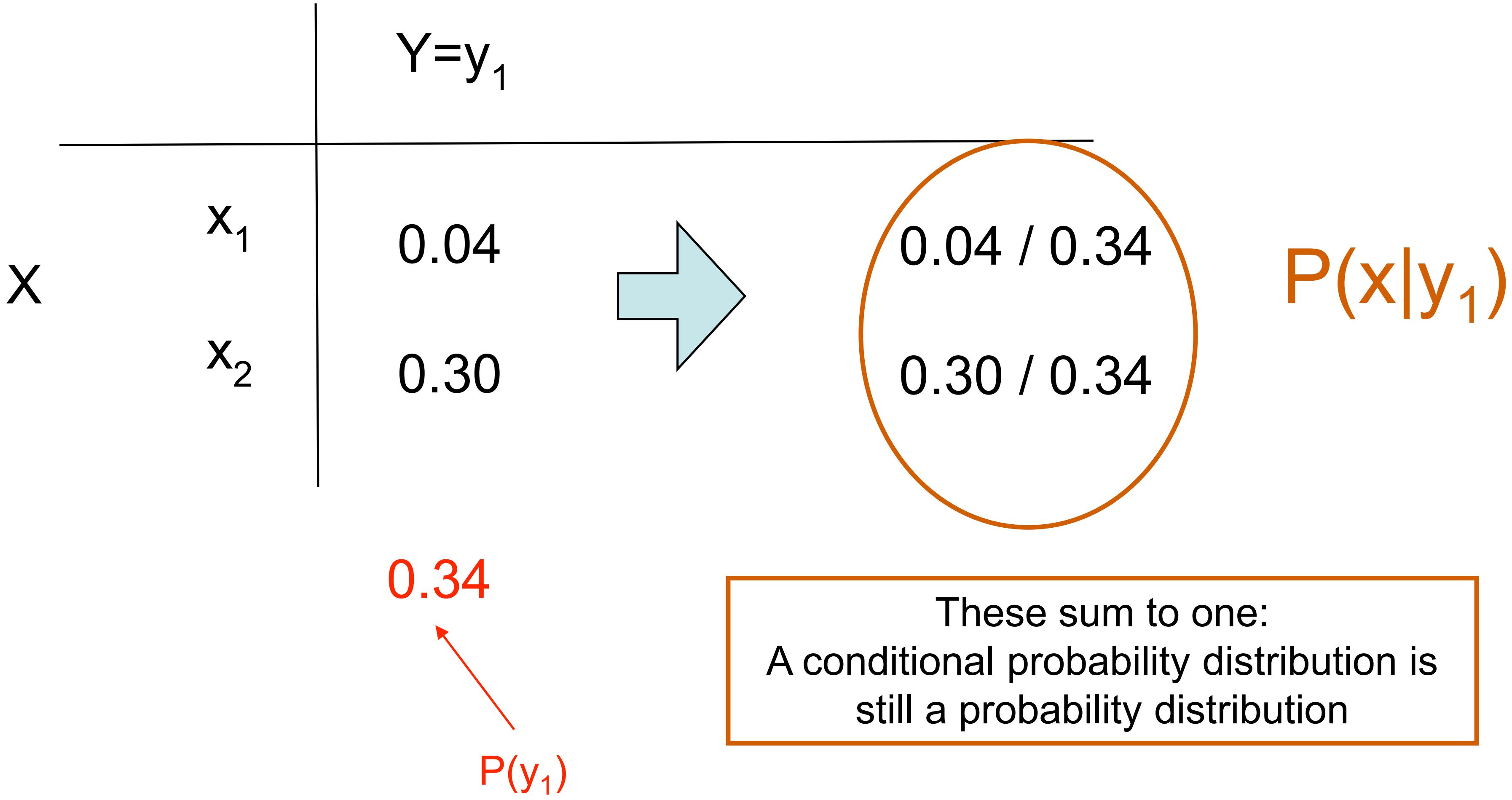
		y_1	y_2
X	x_1	0.04	
	x_2	0.30	
		0.34	

Censored!

P(y_1)

P(x|y_1)=?

Useful Discrete Distributions



		$Y=y_1$
X	x_1	0.04
	x_2	0.30
		0.34

$P(y_1)$

0.04 / 0.34

0.30 / 0.34

$P(x|y_1)$

These sum to one:
A conditional probability distribution is still a probability distribution

Useful Discrete Distributions

Question: How many flips until we observe a success?

Geometric *Distribution on number of independent draws of* $X \sim \mathrm{Bernoulli}(\pi)$ *until success:*

$$p(Y = n) = (1 - \pi)^{n-1}\pi \qquad \mathbf{E}[Y] = \frac{1}{\pi}$$

E.g. for fair coin $\pi = 1/2$ takes two flips on avg.

e.g. there must be n-1 failures (tails) before a success (heads).

Question: How many more flips of we have already seen k failures?

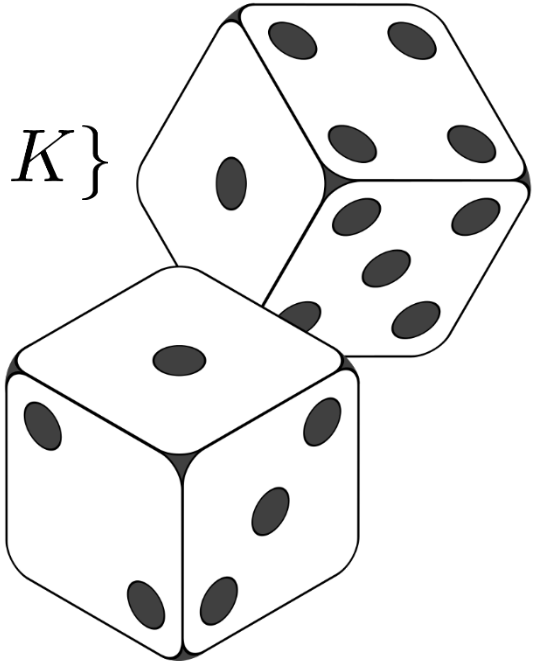
$$p(Y = n + k \mid Y > k) = \frac{p(Y{=}n{+}k, Y{>}k)}{p(Y{>}k)} = \frac{p(Y{=}n{+}k)}{p(Y{>}k)}$$

$$= \frac{(1{-}\pi)^{n+k-1}\pi}{\sum_{i=k}^{\infty}(1{-}\pi)^{i}\pi} = \frac{(1{-}\pi)^{n+k-1}\pi}{(1{-}\pi)^{k}} = (1 - \pi)^{n-1}\pi = p(Y = n)$$

For $0 < x < 1, \sum_{i=k}^{\infty} x^i = x^k/(1 - x)$

Corollary: $p(Y > k) = (1 - \pi)^{k-1}$

Useful Discrete Distributions



Categorical *Distribution on integer-valued RV* $X \in \{1, \dots, K\}$

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

with parameter $p(X = k) = \pi_k$ *and Kronecker delta:*

$$\mathbf{I}(X = k) = \begin{cases} 1, & \text{If } X = k \\ 0, & \text{Otherwise} \end{cases}$$

Can also represent X as *one-hot* binary vector,

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

This representation is special case of the **multinomial distribution**

Useful Discrete Distributions

What if we count outcomes of *N* independent categorical RVs?

Multinomial *Distribution on K-vector* $X \in \{0, N\}^K$ *of counts of N repeated trials* $\sum_{k=1}^K X_k = N$ *with PMF:*

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

Number of ways to partition N objects into K groups:

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

Leading term ensures PMF is properly normalized:

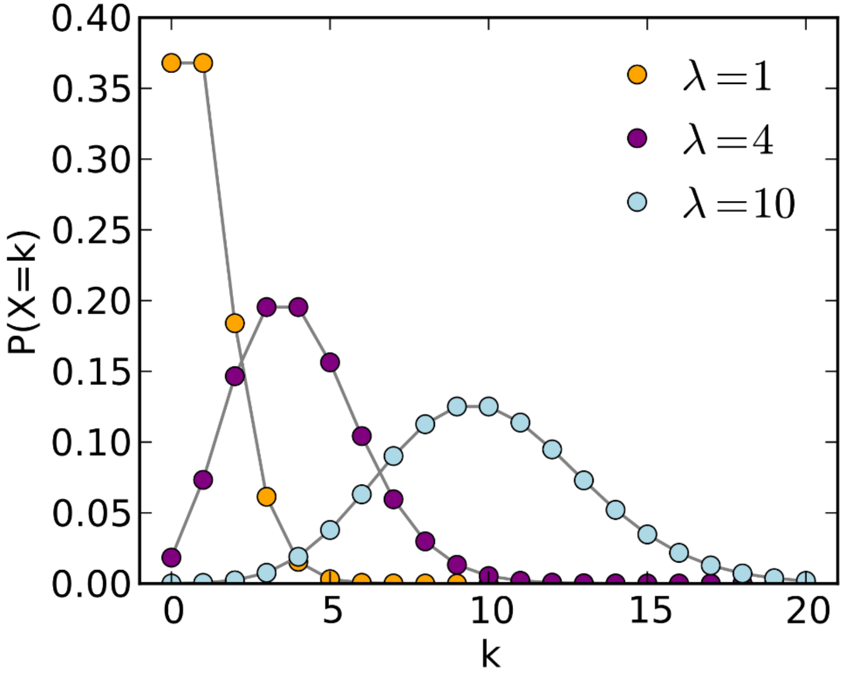
$$\sum_{x_1} \sum_{x_2} \ldots \sum_{x_K} p(x_1, x_2, \ldots, x_K) = 1$$

Useful Discrete Distributions

*A **Poisson** RV X with <u>rate</u> parameter λ has the following distribution:*

Mean and variance both scale with parameter

$$p(X = k) = \frac{e^{-\lambda}\lambda^k}{k!} \qquad \mathbf{E}[X] = \mathbf{Var}[X] = \lambda$$



Represents number of times an *event* occurs in an interval of time or space.

***Ex.** Probability of overflow floods in 100 years,*

$$p(k \text{overflow floods in 100 yrs}) = \frac{e^{-1}1^k}{k!}$$

Avg. 1 overflow flood every 100 years, makes setting rate parameter easy.

***Lemma (additive closure)** The sum of a finite number of Poisson RVs is a Poisson RV.*

$$X \sim \text{Poisson}(\lambda_1), \quad Y \sim \text{Poisson}(\lambda_2), \quad X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$$

Useful Discrete Distributions

Theorem *Let* $X \sim \text{Binomial}(n, \pi(n))$ *where* $\pi(n)$ *is a function of n and* $\lim_{n\to\infty} n \cdot \pi(n) = \lambda$ *for some constant* λ*. Then for any fixed k:*

$$\lim_{n\to\infty} \text{Binomial}(X \mid n, \pi(n)) = \text{Poisson}(X \mid \lambda)$$

Proof Sketch Use Taylor expansion of e^x and $(1-\pi)^k \geq (1-\pi k)$ to upper and lower bound Binomial probability as a function of n:

$$\underbrace{\frac{e^{\pi n}((n-k+1)\pi)^k}{k!}(1-\pi^2 n)}_{\textit{\textbf{LB(n)}}} \leq \text{Binomial}(X = k \mid n, \pi) \leq \underbrace{\frac{e^{-\pi n}(n\pi)^k}{k!}\frac{1}{1-\pi k}}_{\textit{\textbf{UB(n)}}}$$

As $n \to \infty$ it must be that $\pi(n) \to 0$ so that $\lim_{n\to\infty} n \cdot \pi(n) = \lambda$ is constant. Then $1/(1-\pi k) \to 1$ and $1-\pi^2 n \to 1$. The difference $[(n-k+1)\pi] - n\pi$ approaches 0. Therefore:

$$\lim_{n\to\infty} \text{LB}(n) = \frac{e^{-\lambda}\lambda^k}{k!} \qquad \text{and} \qquad \lim_{n\to\infty} \text{UB}(n) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Bounds converge so result holds.

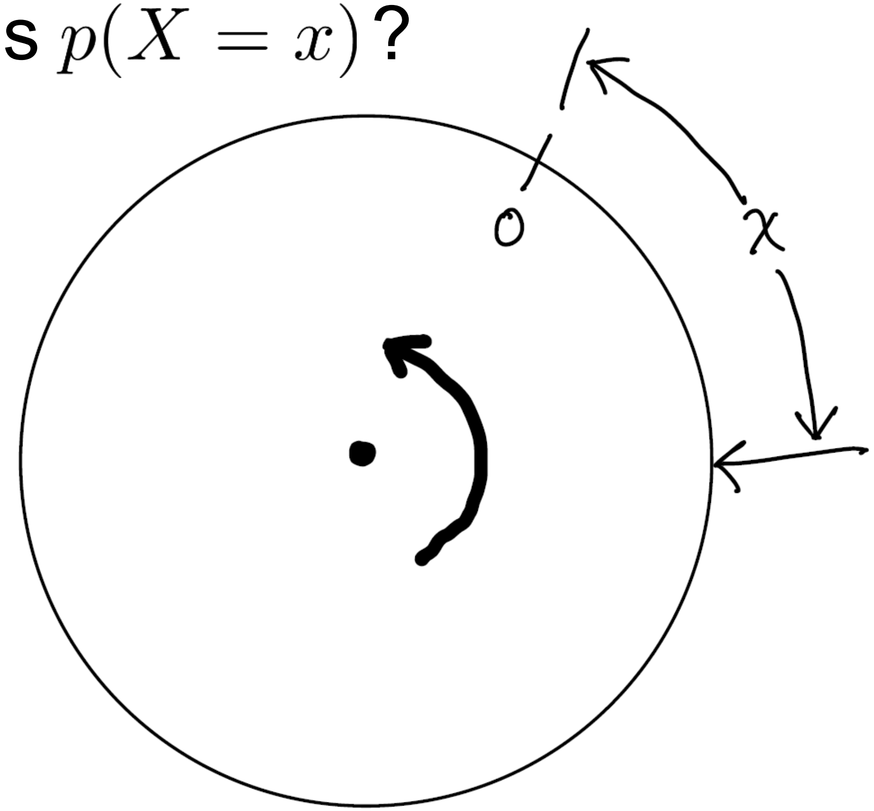
Outline

- Moments of (discrete) random variables
- Some useful discrete distributions
- Continuous probability

Continuous Probability

Experiment Spin continuous wheel and measure X displacement from 0

Question Assuming uniform probability, what is $p(X = x)$?



First, recall axioms of probability…

1. For any event E, $0 \leq P(E) \leq 1$
2. $P(\Omega) = 1 \text{ and } P(\emptyset) = 0$
3. For any *finite* or *countably infinite* sequence of <u>pairwise mutually disjoint events</u> $E_1, E_2, E_3, \ldots$

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

Sample space Ω is all points (real numbers) in $[0, 1)$

Continuous Probability

- Let $p(X = x) = \pi$ be the probability of any single outcome
- Let $S(k)$ be set of any k *distinct* points in $[0, 1)$ then,

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

- Since $0 < P(x \in S(k)) < 1$ by axioms of probability, $k\pi < 1$ for any k
- Therefore: $\pi = 0$ and $P(x \in S(k)) = p(X = x) = 0$

What does this mean?

- Let E be event that $x \in S(k)$
- In infinite sample space, an event may be **possible** but have zero "probability"
- Since $P(\bar{E}) = 1 - P(E) = 1$ events may have "probability" 1 but **not certain**

Assign probability to intervals, not individual values

Continuous Probability

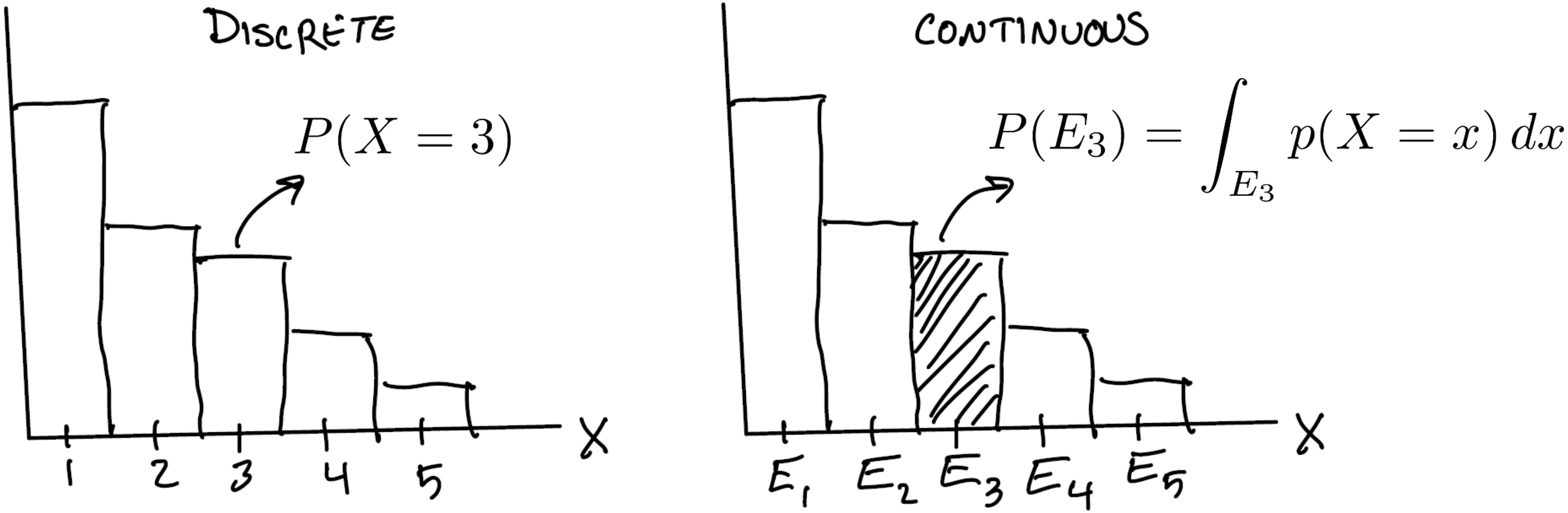
We could just accept this oddity…

Good Earth®

In mathematics you don't understand things. You just get used to them.
Johann von Neumann
(1903 - 1957)

…or we could try to convince ourselves that it is sensible.

Continuous Probability



- 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 $P(X = 0) = P(x \leq X < x) = 0$
- But may have nonzero *probability density* $p(X = x)$

Continuous Probability

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

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

Different ways to represent probability of interval, CDF is just a convention.

- Can easily measure probability of closed intervals,

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

- If *X* is *absolutely continuous* (i.e. differentiable) then,

Fundamental Theorem of Calculus

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

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

- Typically use shorthand P for CDF and p for PDF instead of F and f

Continuous Probability

Most definitions for discrete RVs hold, replacing PMF with PDF/CDF…

Two RVs X & Y are **independent** if and only if,

$p(x,y) = p(x)p(y)$ or $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$

Conditionally independent given Z iff,

Shorthand: $P(x) = P(X \leq x)$

$p(x,y \mid z) = p(x \mid z)p(y \mid z)$ or $P(x,y \mid z) = P(x \mid z)P(y \mid z)$

Probability chain rule,

$p(x,y) = p(x)p(y \mid x)$ and $P(x,y) = P(x)P(y \mid x)$

Continuous Probability

…and by replacing summation with integration…

Law of Total Probability for continuous distributions,

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

Expectation of a continuous random variable,

$$\mathbf{E}[X] = \int_{\mathcal{X}} x \cdot p(x)\, dx$$

Covariance of two continuous random variables X & Y,

$$\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \int_{\mathcal{X}} \int_{\mathcal{Y}} (x - \mathbf{E}[X])(y - \mathbf{E}[Y]) p(x, y)\, dx dy$$

Continuous Probability

***Caution** Some technical subtleties arise in continuous spaces…*

For **discrete** RVs X & Y, the conditional

P(Y=y)=0 means impossible

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

is **undefined** when P(Y=y) = 0 … no problem.

For **continuous** RVs we have,

$$P(X \leq x \mid Y = y) = \frac{P(X \leq x, Y = y)}{P(Y = y)}$$

but numerator and denominator are 0/0.

**P(Y=y)=0 means improbable,
but not impossible**

Continuous Probability

Defining the conditional distribution as a limit fixes this…

$$P(X \leq x \mid Y = y) = \lim_{\delta \to 0} P(X \leq x \mid y \leq Y \leq y + \delta)$$

$$= \lim_{\delta \to 0} \frac{P(X \leq x, y \leq Y \leq y + \delta)}{P(y \leq Y \leq y + \delta)}$$

$$= \lim_{\delta \to 0} \frac{P(X \leq x, Y \leq y + \delta) - P(X \leq x, Y \leq y)}{P(Y \leq y + \delta) - P(Y \leq y)}$$

Definition The conditional PDF is given by,

$$p(x \mid y) = \frac{p(x, y)}{p(y)}$$

$$= \int_{-\infty}^{x} \lim_{\delta \to 0} \frac{\frac{\partial}{\partial x} P(u, y + \delta) - \frac{\partial}{\partial x} P(u, y)}{P(y + \delta) - P(y)} \, du$$

(Fundamental theorem of calculus)

(Assume interchange limit / integral)

$$= \int_{-\infty}^{x} \lim_{\delta \to 0} \frac{\left(\frac{\partial}{\partial x} P(u, y + \delta) - \frac{\partial}{\partial x} P(u, y)\right) / \delta}{\left(P(y + \delta) - P(y)\right) / \delta} \, du$$

(Multiply by $\frac{\delta}{\delta} = 1$)

$$= \int_{-\infty}^{x} \frac{\frac{\partial^2}{\partial x \partial y} P(u, y)}{\frac{\partial}{\partial y} P(y)} \, du \quad = \int_{-\infty}^{x} \frac{p(u, y)}{p(y)} \, du$$

(Definition of partial derivative)

(Definition PDF)

Useful Continuous Distributions

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

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

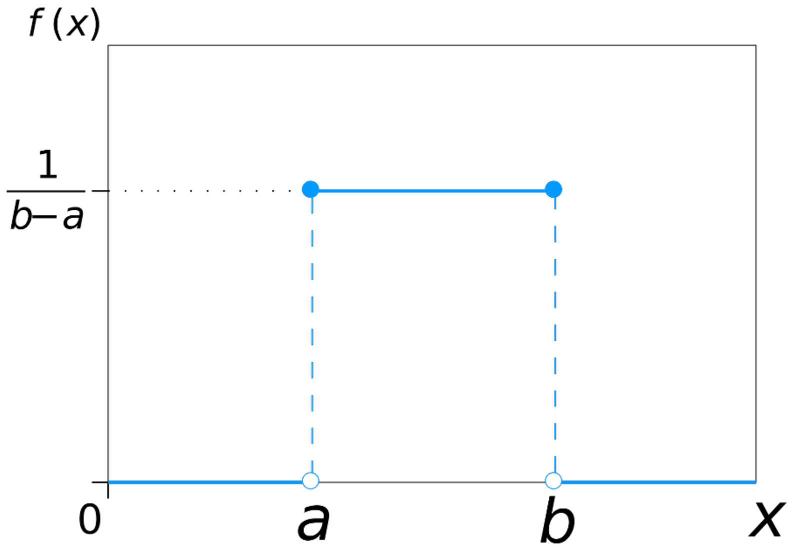
Say that $X \sim U(a, b)$ whose moments are,

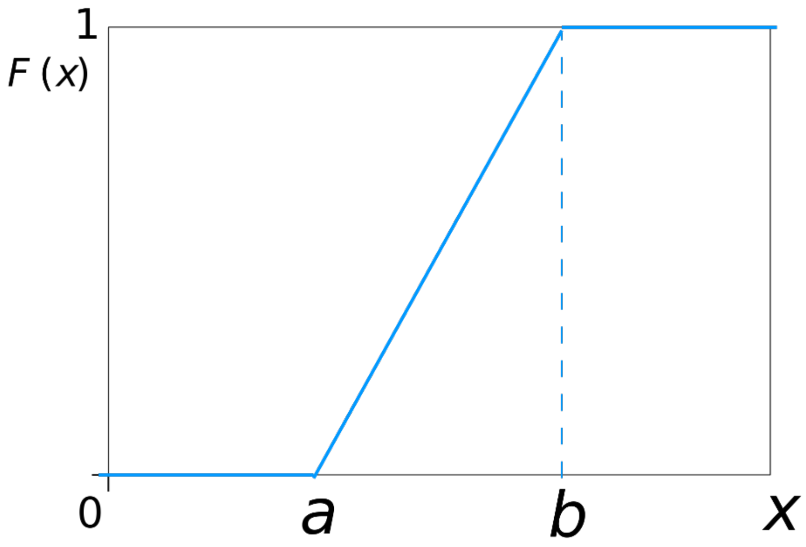
$$\mathbf{E}[X] = \frac{b+a}{2} \qquad \mathbf{Var}[X] = \frac{(b-a)^2}{12}$$

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

$$P(X \le x \mid X \le \tfrac{1}{2}) = U(0, \tfrac{1}{2})$$

Holds generally: Uniform closed under conditioning





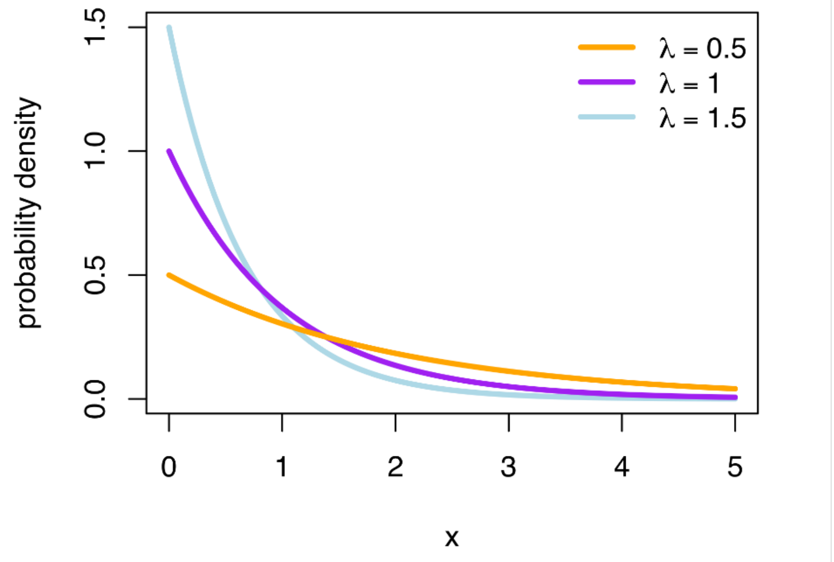
Useful Continuous Distributions

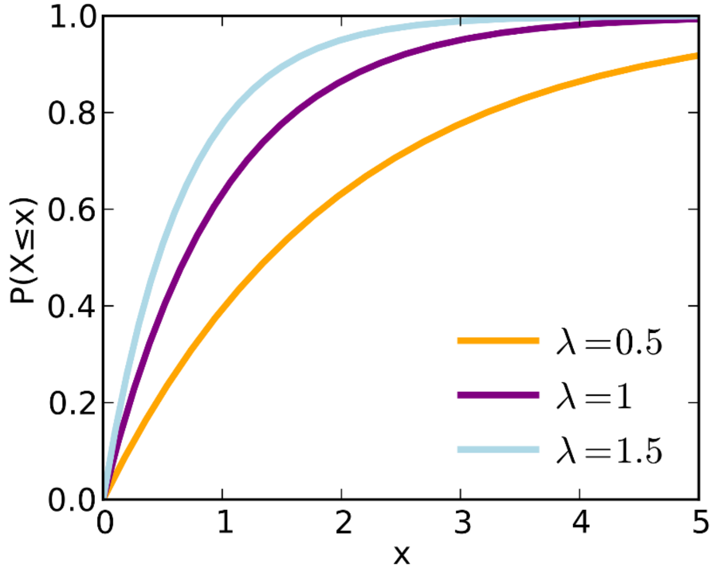
Exponential distribution with scale λ,

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

for X>0. Moments given by,

$$\mathbf{E}[X] = \frac{1}{\lambda} \qquad \mathbf{Var}[X] = \frac{2}{\lambda^2}$$





Useful properties

- **Closed under conditioning** If $X \sim \text{Exponential}(\lambda)$ then,

$$P(X \geq s + t \mid X \geq s) = P(X \geq s) = e^{-\lambda s}$$

- **Minimum** Let $X_1, X_2, \ldots, X_N$ be i.i.d. exponentially distributed with scale parameters $\lambda_1, \lambda_2, \ldots, \lambda_N$ then,

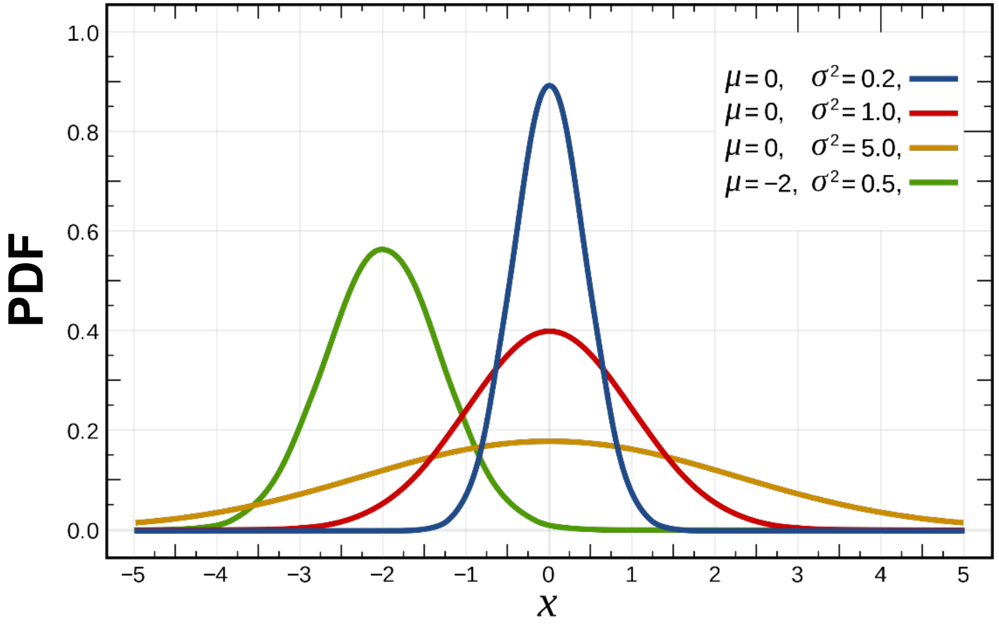
$$P(\min(X_1, X_2, \ldots, X_N)) = \text{Exponential}(\textstyle\sum_i \lambda_i)$$

Useful Continuous Distributions

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

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

We say $X \sim \mathcal{N}(\mu, \sigma^2)$.



Useful Properties

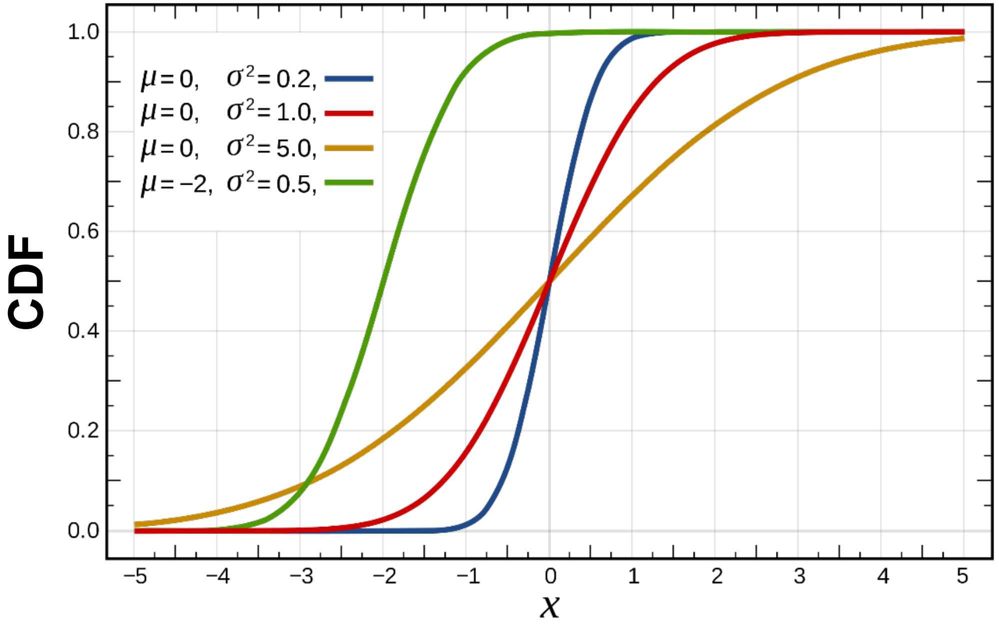
- Closed under additivity:

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

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

- Closed under linear functions (a and b constant):

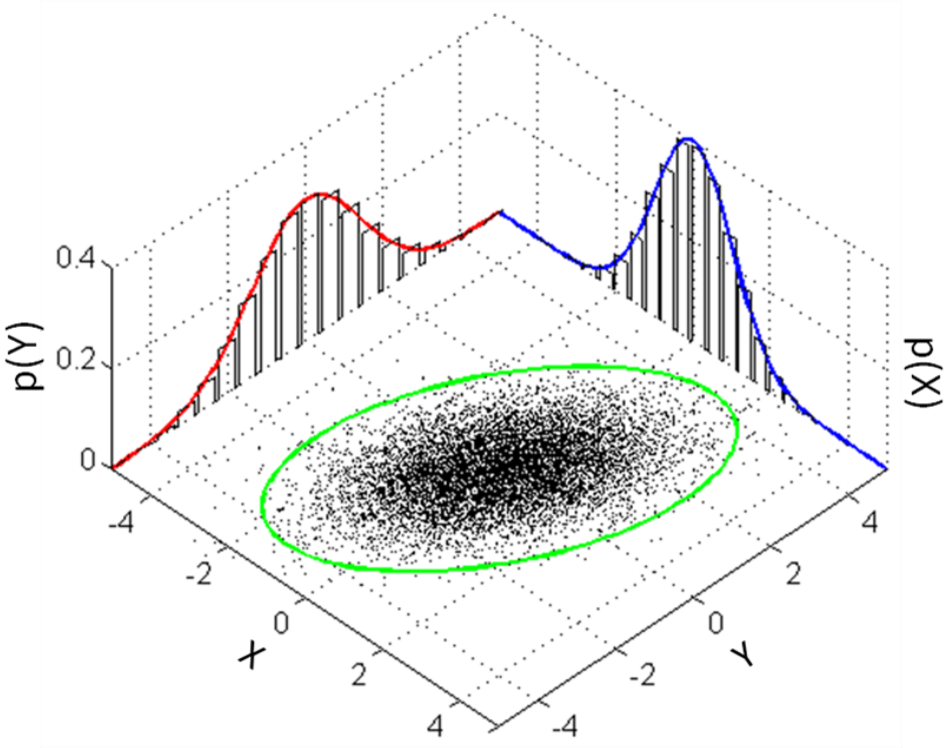
$$aX + b \sim \mathcal{N}(a\mu_x + b, a^2\sigma_x^2)$$



Useful Continuous Distributions

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

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



Moments given by parameters directly.

Useful Properties

- Closed under additivity (same as univariate case)
- Closed under linear functions,

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

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

- Closed under conditioning and marginalization

Will discuss Gaussians a lot more when we cover exponential families