

Basics of Probability

Northwestern EECS 349
Doug Downey

Events

- ▶ **Event space** Ω

- ▶ E.g. for dice, $\Omega = \{1, 2, 3, 4, 5, 6\}$

- ▶ **Set of measurable events** $\mathcal{S} \subseteq 2^\Omega$

- ▶ E.g.,

- ▶ $\alpha = \text{event we roll an even number} = \{2, 4, 6\} \in \mathcal{S}$

- ▶ \mathcal{S} must:

- ▶ Contain the empty event \emptyset and the trivial event Ω

- ▶ Be closed under union & complement

- $\alpha, \beta \in \mathcal{S} \rightarrow \alpha \cup \beta \in \mathcal{S}$ and $\alpha \in \mathcal{S} \rightarrow \Omega - \alpha \in \mathcal{S}$



Probability Distributions

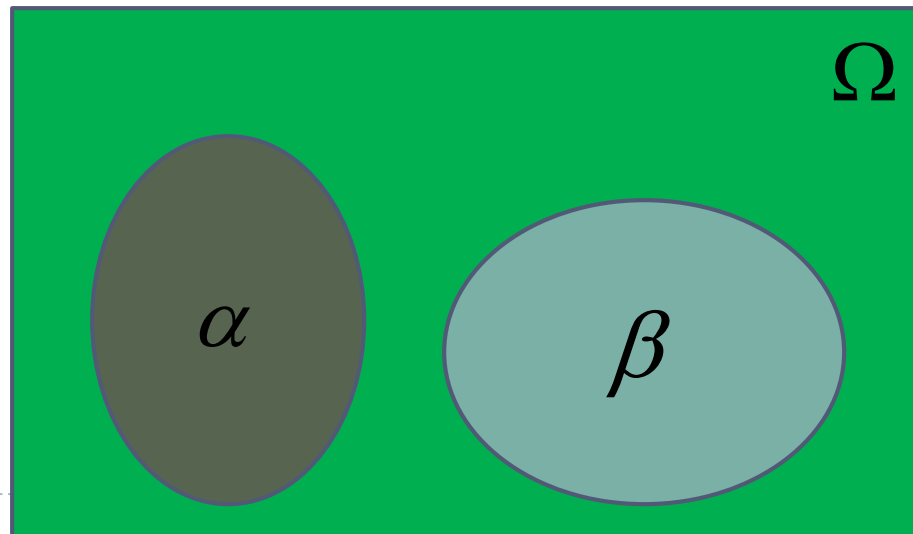
- ▶ A **probability distribution** P over (Ω, \mathcal{S}) is a mapping from \mathcal{S} to real values such that:

1. $P(\alpha) \geq 0 \quad \forall \alpha \in \mathcal{S}$

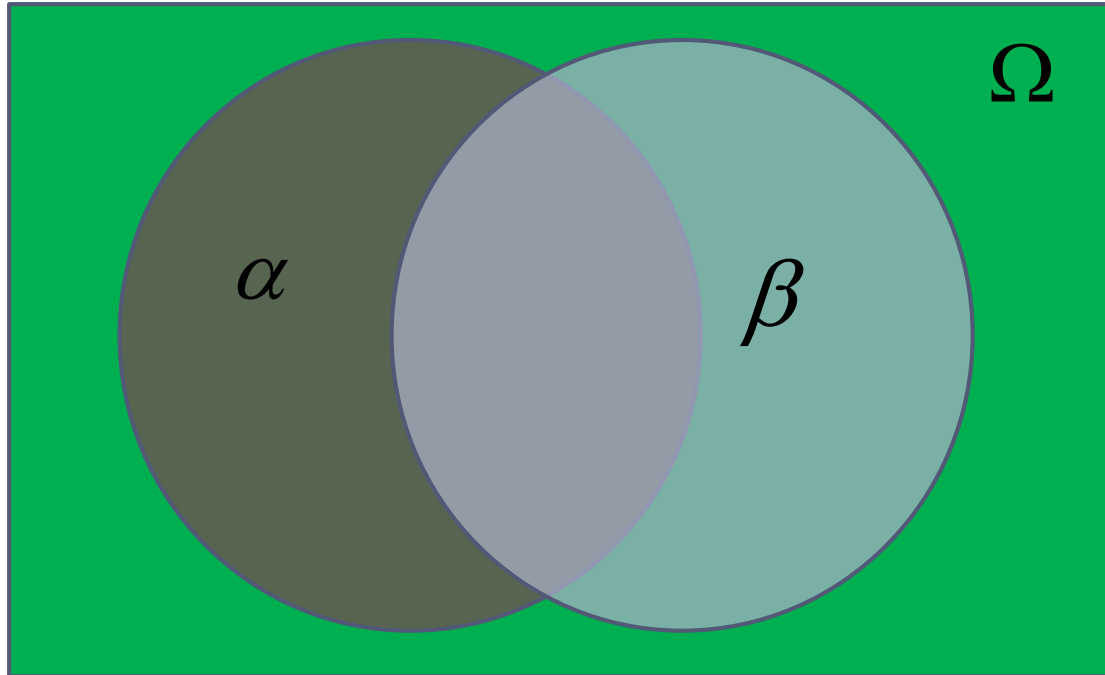
2. $P(\Omega) = 1$

3. $\alpha, \beta \in \mathcal{S} \wedge \alpha \cap \beta = \emptyset \rightarrow P(\alpha \cup \beta) = P(\alpha) + P(\beta)$

Sidenote – 1st and 3rd axioms ensure P is a *measure*



Probability Distributions



Can visualize probability as fraction of area




Probability: Interpretations & Motivation

- ▶ Interpretations: Frequentist vs. Bayesian
- ▶ **Why** use probability for subjective beliefs?
 - ▶ Beliefs that violate the axioms can lead to bad decisions *regardless* of the outcome [de Finetti, 1931]
 - ▶ Example: $P(A) = 0.6$, $P(\text{not } A) = 0.8$?
 - ▶ Example: $P(A) > P(B)$ and $P(B) > P(A)$?



Random Variables

- ▶ A **random variable** is a function from Ω to a value
 - ▶ A *partition* of the event space Ω
 - ▶ A short-hand for referring to *attributes* of events
- ▶ **Examples**
 - ▶ $\Omega = \{1, 2, 3, 4, 5, 6\}$
DieRollEven $\in \{\text{true}, \text{false}\}$  = Val(**DieRollEven**)
 - ▶ $\Omega = \{\text{all possible hmwk/exam grade combinations}\}$
FinalGrade $\in \{a, b, c\}$



Joint Distributions

Grade	Interest	Course load	P(G, I, C)
a	high	full-time	0.10
a	high	part-time	0.08
a	low	full-time	0.03
a	low	part-time	0.04
b	high	full-time	0.07
b	high	part-time	0.02
b	low	full-time	0.12
b	low	part-time	0.16
c	high	full-time	0.01
c	high	part-time	0.02
c	low	full-time	0.20
c	low	part-time	0.15



Conditioning!

Grade	Interest	Course load	P(G, I, C)
a	high	full-time	0.10
a	high	part-time	0.09
a	low	full-time	0.03
a	low	part-time	0.04
b	high	full-time	0.07
b	high	part-time	0.02
b	low	full-time	0.12
b	low	part-time	0.16
c	high	full-time	0.01
c	high	part-time	0.02
c	low	full-time	0.20
c	low	part-time	0.15



Conditioning!

Grade	Interest	Course load	P(G, I, C)
a	high	full-time	0.10 / 0.53
a	low	full-time	0.03 / 0.53
b	high	full-time	0.07 / 0.53
b	low	full-time	0.12 / 0.53
c	high	full-time	0.01 / 0.53
c	low	full-time	0.20 / 0.53

0.53



Conditioning!

Grade	Interest	Course load	$P(G, I C=f)$
a	high	full-time	0.19
a	low	full-time	0.06
b	high	full-time	0.13
b	low	full-time	0.23
c	high	full-time	0.02
c	low	full-time	0.38
			<hr/> 1.0



Conditional Probability

- ▶ $P(\text{Grade} = A \mid \text{Interest} = \text{High}) = 0.6$
 - ▶ the probability of getting an A given **only** *Interest* = High, and nothing else.
 - ▶ If we know *Motivation* = High or *OtherInterests* = Many, the probability of an A changes even given high *Interest*
- ▶ **Formal Definition:**
 - ▶ $P(\alpha \mid \beta) = P(\alpha, \beta) / P(\beta)$
 - ▶ When $P(\beta) > 0$



Conditional Probability

- ▶ **Also:**

- ▶ $P(A \mid B, C) = P(A, B, C) / P(B, C)$

- ▶ **More generally:**

- ▶ $P(\mathbf{A} \mid \mathbf{B}) = P(\mathbf{A}, \mathbf{B}) / P(\mathbf{B})$

- ▶ (Boldface indicates vectors of variables)

- ▶ $P(\textit{Grade} = A \mid \textit{Grade} = A, \textit{Interest} = \textit{high}) ?$



Marginalization

Grade	Interest	Course load	P(G, I, C)
a	high	full-time	0.10
a	high	part-time	0.08
a	low	full-time	0.03
a	low	part-time	0.04
b	high	full-time	0.07
b	high	part-time	0.02
b	low	full-time	0.12
b	low	part-time	0.16
c	high	full-time	0.01
c	high	part-time	0.02
c	low	full-time	0.20
c	low	part-time	0.15



Marginalization

Grade	Interest	Course load	P(G, I, C)
a	high	*	0.10
a	high	*	0.08
a	low	*	0.03
a	low	*	0.04
b	high	*	0.07
b	high	*	0.02
b	low	*	0.12
b	low	*	0.16
c	high	*	0.01
c	high	*	0.02
c	low	*	0.20
c	low	*	0.15



Marginalization

Grade	Interest	Course load	P(G, I)
a	high	*	0.18
a	low	*	0.07
b	high	*	0.09
b	low	*	0.28
c	high	*	0.03
c	low	*	0.35



Marginalization

Grade	Interest	P(G, I)
a	high	0.18
a	low	0.07
b	high	0.09
b	low	0.28
c	high	0.03
c	low	0.35
		<hr/> 1.0



Marginalization

$$P(X) = \sum_{y \in \text{Val}(Y)} P(X, Y = y)$$



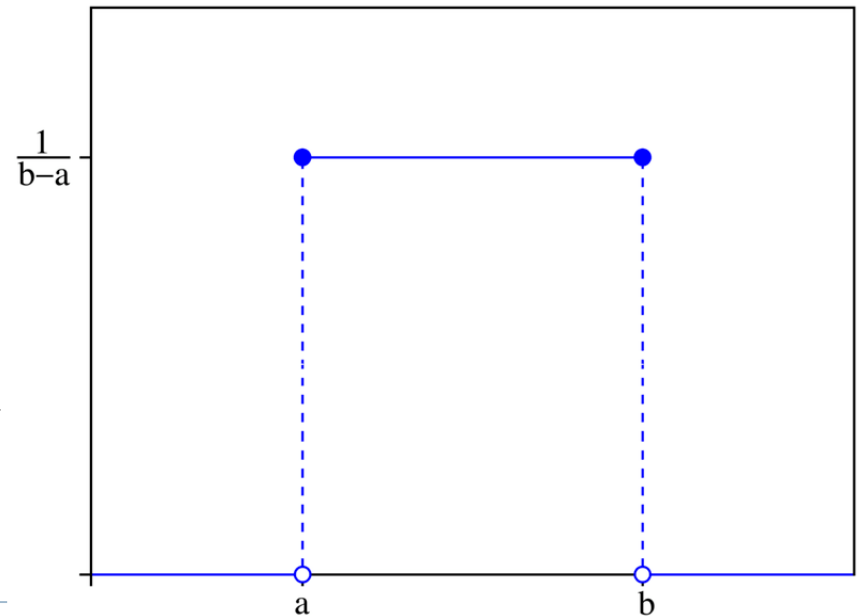
Continuous Random Variables

- ▶ For continuous r.v. X , specify a *density* $p(x)$, such that:

E.g.,

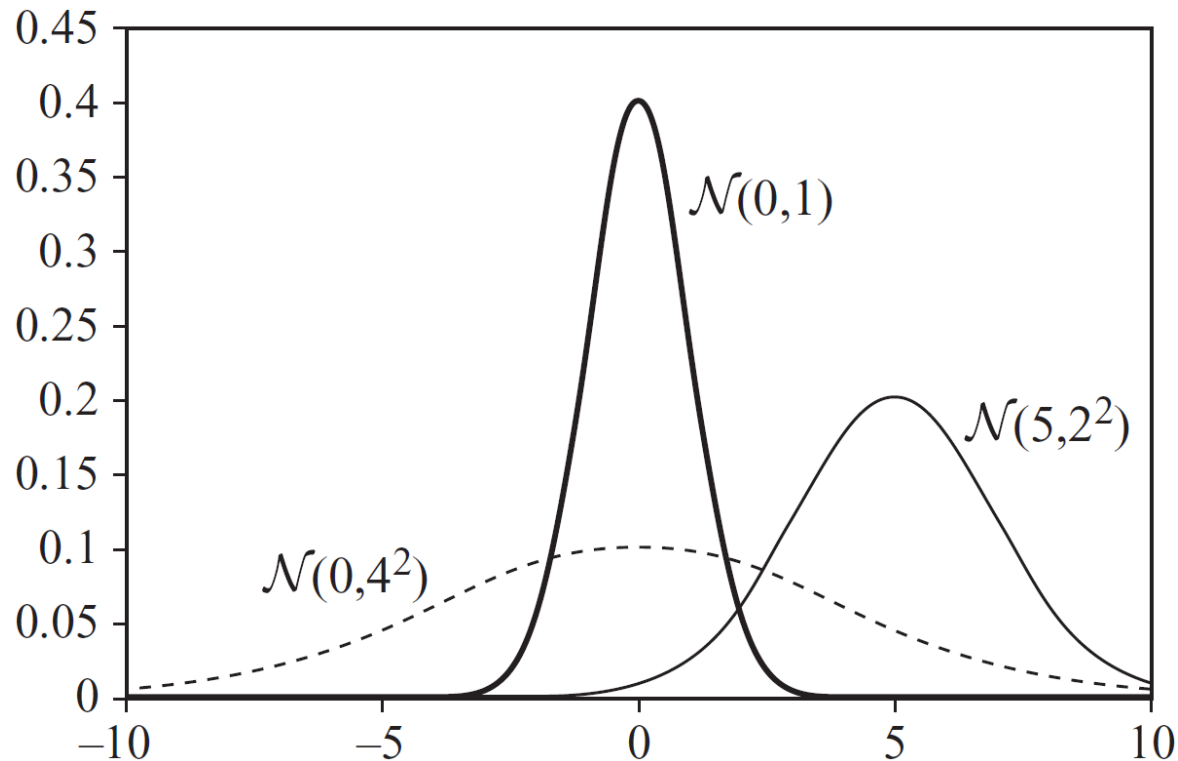
$$P(r \leq X \leq s) = \int_{x=r}^s p(x) dx$$

$$p(x) = \begin{cases} \frac{1}{b-a} & b \geq x \geq a \\ 0 & \text{otherwise} \end{cases}$$



Gaussian Density

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



Joint Distribution

		Interest	
		low	high
Grade	a	0.07	0.18
	b	0.28	0.09
	c	0.35	0.03

Joint Distribution specified with $2*3 - 1 = 5$ values



Conditional Probability

		Interest	
		low	high
Grade	a	0.07	0.18
	b	0.28	0.09
	c	0.35	0.03

$P(\text{Grade} = a \mid \text{Interest} = \text{high}) ?$

$P(\text{Grade} = a, \text{Interest} = \text{high}) = 0.18$

$P(\text{Interest} = \text{high}) = 0.18 + 0.09 + 0.03 = 0.30$

$\Rightarrow P(\text{Grade} = a \mid \text{Interest} = \text{high}) = 0.18 / 0.30 = \mathbf{0.6}$



Conditional Probability

		Interest	
		low	high
Grade	a	0.07	0.18
	b	0.28	0.09
	c	0.35	0.03

$P(\text{Interest} \mid \text{Grade} = a)?$

Interest	
low	high
0.28	0.72



Conditional Probability

		Interest	
		low	high
Grade	a	0.07	0.18
	b	0.28	0.09
	c	0.35	0.03

$P(\text{Interest} \mid \text{Grade})?$

Actually three separate distributions, one for each *Grade* value

(has three independent parameters total)



Chain Rule

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i \mid X_{i-1} = x_{i-1}, \dots, X_1 = x_1)$$

- ▶ E.g., $P(\text{Grade}=\text{b}, \text{Int.} = \text{high})$
 $= P(\text{Grade}=\text{b} \mid \text{Int.}=\text{high})P(\text{Int.} = \text{high})$
- ▶ Can be used for distributions...
 - ▶ $P(A, B) = P(A \mid B)P(B)$



Handy Rules for Cond. Probability (1 of 2)

- ▶ $P(A \mid B = b)$ is a single distribution, like $P(A)$
- ▶ $P(A \mid B)$ is *not* a single distribution
 - ▶ a set of $|\text{Val}(B)|$ distributions



Handy Rules for Cond. Probability (2 of 2)

- ▶ Any statement true for arbitrary distributions is also true if you condition on a new r.v.
 - ▶ $P(A, B) = P(A | B)P(B)$? (chain rule)
Then also $P(A, B | C) = P(A | B, C) P(B | C)$
- ▶ Likewise, any statement true for arbitrary distributions is also true if you replace an r.v. with two/more new r.v.s
 - ▶ $P(A | B) = P(A, B) / P(B)$? (def. of cond. Prob)
 - ▶ $P(A | C, D) = P(A, C, D) / P(C, D)$ or $P(\mathbf{A} | \mathbf{B}) = P(\mathbf{A}, \mathbf{B}) / P(\mathbf{B})$



Independence

- ▶ $P(\text{Rain} \mid \text{Cloudy}) \neq P(\text{Rain})$
 - ▶ But: $P(\text{FairDie}=6 \mid \text{PreviousRoll}=6) = P(\text{FairDie}=6)$
- ▶ We say A and B are **independent** iff

$$P(A \mid B) = P(A)$$

- ▶ Logically equivalent to $P(A, B) = P(A) * P(B)$
- ▶ Denoted $A \perp B$



Conditional Independence (1 of 2)

- ▶ A and B are **conditionally independent** given C *iff*

$$P(A \mid B, C) = P(A \mid C)$$

- ▶ Equivalent to $P(A, B \mid C) = P(A \mid C) P(B \mid C)$

- ▶ Denoted $(A \perp B \mid C)$



Conditional Independence (2 of 2)

▶ **Example: university admissions**

- ▶ $Val(GetIntoX) = \{yes, no, wait\}$
- ▶ $Val(Application) = \{good, bad\}$

$3*3*2*2 = 36$ Parameters

$P(GetIntoNU \mid GetIntoUIUC, GetIntoStanford, Application)$

=

$P(GetIntoNU \mid Application)$

$2*2 = 4$ Parameters



Properties of Conditional Independence

- ▶ **Decomposition**

- ▶ $(X \perp Y, W \mid Z) \Rightarrow (X \perp Y \mid Z)$

- ▶ **Weak Union**

- ▶ $(X \perp Y, W \mid Z) \Rightarrow (X \perp Y \mid Z, W)$

- ▶ **Contraction**

- ▶ $(X \perp W \mid Z, Y) \& (X \perp Y \mid Z) \Rightarrow (X \perp Y, W \mid Z)$



Expectation

- ▶ Discrete

$$E_P[X] = \sum_x x P(x)$$

- ▶ Continuous

$$E_P[X] = \int x p(x) dx$$

- ▶ E.g., $E[\text{FairDie}] = 3.5$



Expectation is Linear

$$\begin{aligned} \boxed{E_P[X + Y]} &= \sum_{x,y} (x + y)P(x, y) \\ &= \sum_{x,y} x P(x, y) + \sum_{x,y} y P(x, y) \\ &= \sum_x x \sum_y P(x, y) + \sum_y y \sum_x P(x, y) \\ &= \sum_x x P(x) + \sum_y y P(y) = \boxed{E_P[X] + E_P[Y]} \end{aligned}$$



What have we learned?

- ▶ **Probability** – a calculus for dealing with uncertainty
 - ▶ Built from small set of axioms (ignore at your peril)
- ▶ **Joint Distribution $P(A, B, C, \dots)$**
 - ▶ Specifies probability of all combinations of r.v.s
- ▶ **Conditional Probability $P(A | B)$**
 - ▶ Specifies probability of $A=a$ given $B=b$
- ▶ **Conditional Independence**
 - ▶ Can radically reduce number of model parameters
- ▶ **Expectation**
- ▶ **Next time: Bayes' Rule, Statistical Estimation**

