Basics of Probability

Northwestern EECS 349
Doug Downey

Events

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

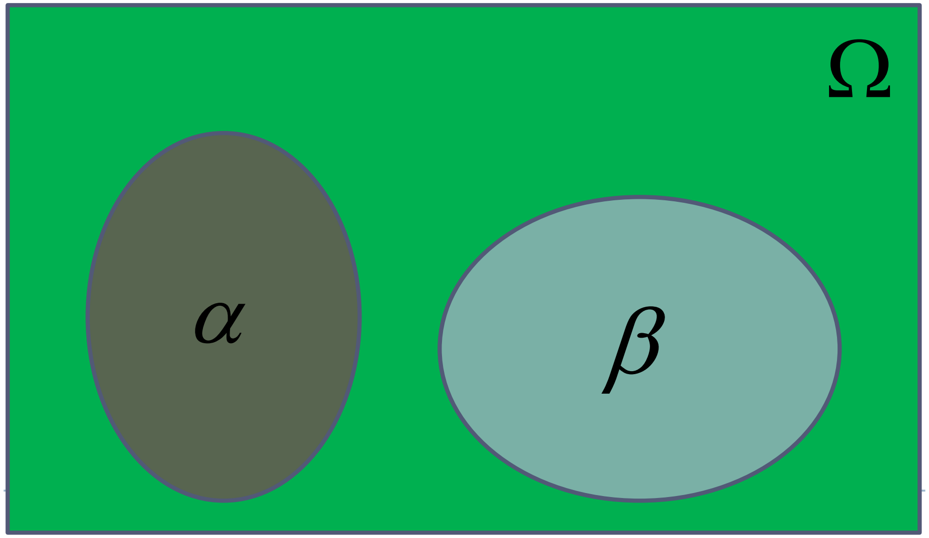
- Set of measurable events $S \subseteq 2^{\Omega}$
  - E.g.,
    α = event we roll an even number = $\{2, 4, 6\} \in S$
  - S must:
    - Contain the empty event $\emptyset$ and the trivial event Ω
    - Be closed under union & complement
      - $\alpha, \beta \in S \rightarrow \alpha \cup \beta \in S$ and $\alpha \in S \rightarrow \Omega - \alpha \in S$

Probability Distributions

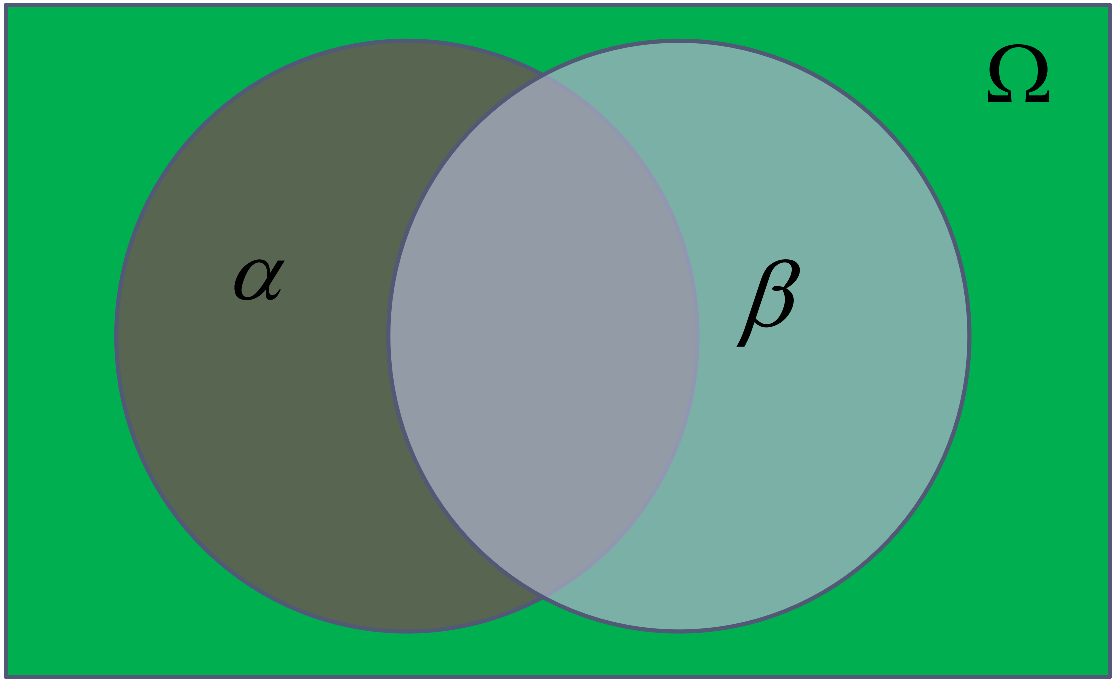
- A probability distribution P over (Ω, S) is a mapping from S to real values such that:

1. $P(\alpha) \geq 0 \quad \forall \alpha \in S$
2. $P(\Omega) = 1$
3. $\alpha, \beta \in S \ \wedge \ \alpha \cap \beta = \emptyset \ \rightarrow \ P(\alpha \cup \beta) = P(\alpha) + P(\beta)$

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(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
  - Ω = {1, 2, 3, 4, 5, 6}
    DieRollEven $\in$ {true, false} → = Val(***DieRollEven***)
  - Ω = {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~~	~~0.08~~
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	/ 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

1.0

Conditional Probability

- P(*Grade* = A | *Interest* = 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:
  - $P(\boldsymbol{A} \mid \boldsymbol{B}) = P(\boldsymbol{A}, \boldsymbol{B}) / P(\boldsymbol{B})$
  - (Boldface indicates vectors of variables)
- P(*Grade* = A | *Grade* = A, *Interest* = 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

1.0

Marginalization

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

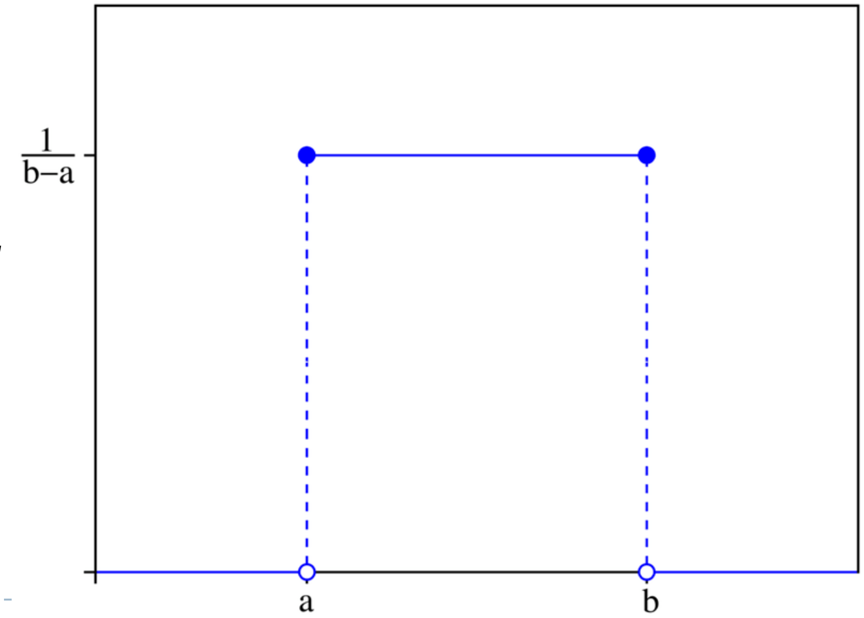
Continuous Random Variables

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

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

E.g.,

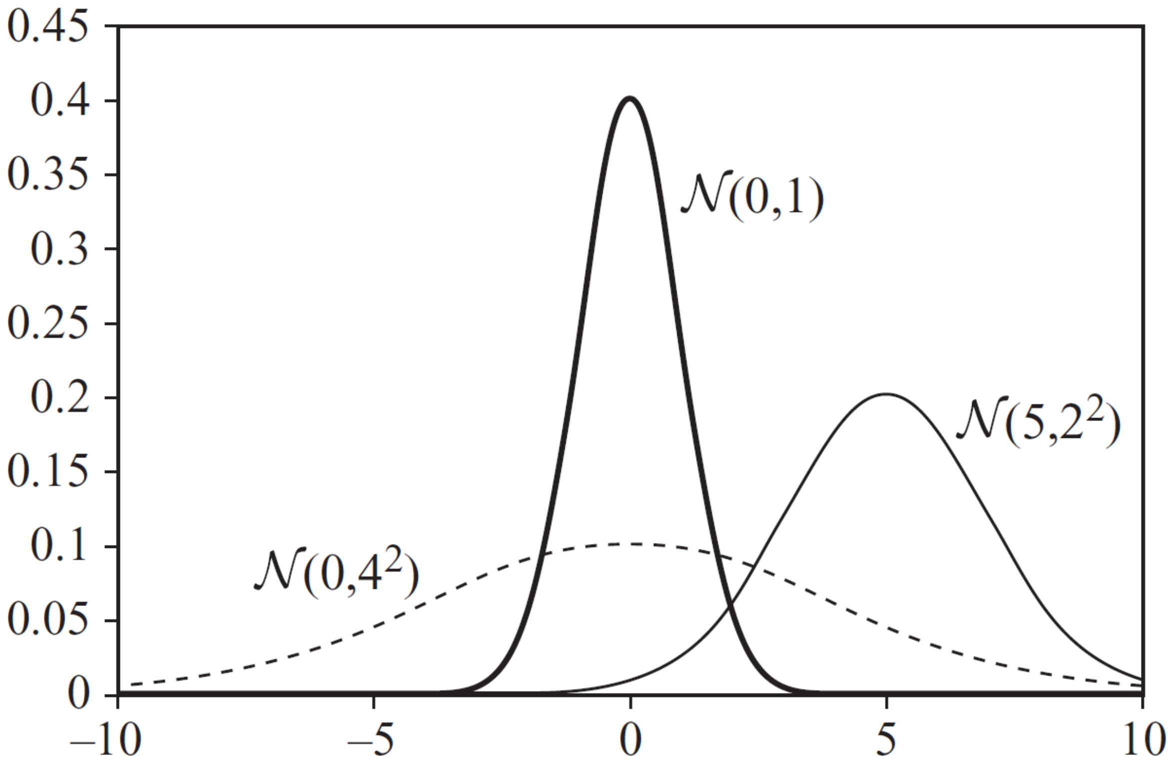
$$p(x) = \begin{cases} \dfrac{1}{b-a} & b \ge x \ge a \\ 0 & \text{otherwise} \end{cases}$$



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(*Grade* = a | *Interest* = high) ?

P(*Grade* = a, *Interest* = high) = 0.18

P(*Interest* = high) = 0.18+0.09+0.03 = 0.30

=> P(*Grade* = a | *Interest* = high) = 0.18/0.30 = **0.6**

Conditional Probability

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

P(*Interest* | *Grade* = a)?

Interest	
low	high
0.28	0.72

Conditional Probability

		Interest	
		low	high
Grade	a	0.28	0.72
	b	0.76	0.24
	c	0.92	0.08

P(*Interest* | *Grade*)?

Actually three separate distributions, one for each *Grade* value
(has three independent parameters total)

Chain Rule

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

- E.g., P(*Grade*=b, *Int.* = high)
  = P(*Grade*=b | *Int.*= high)P(*Int.* = high)
- Can be used for distributions…
  - P(*A*, *B*) = P(*A* | *B*)P(*B*)

Handy Rules for Cond. Probability (1 of 2)

- P(A | B = b) is a single distribution, like P(A)
- P(A | B) is *not* a single distribution
  - a *set* of |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(***A*** | ***B***) = P(***A***, ***B***) / P(***B***)

Independence

- P(*Rain* | *Cloudy*) ≠ P(*Rain*)
  - But: P(*FairDie*=6 | *PreviousRoll*=6) = P(*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* | *GetIntoUIUC*, *GetIntoStanford*, *Application*)

=

P(*GetIntoNU* | *Application*)

2*2= **4** Parameters

Properties of Conditional Independence

- Decomposition
  - $(\boldsymbol{X} \perp \boldsymbol{Y}, \boldsymbol{W} \mid \boldsymbol{Z}) \Rightarrow (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$
- Weak Union
  - $(\boldsymbol{X} \perp \boldsymbol{Y}, \boldsymbol{W} \mid \boldsymbol{Z}) \Rightarrow (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{W})$
- Contraction
  - $(\boldsymbol{X} \perp \boldsymbol{W} \mid \boldsymbol{Z}, \boldsymbol{Y})$ & $(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}) \Rightarrow (\boldsymbol{X} \perp \boldsymbol{Y}, \boldsymbol{W} \mid \boldsymbol{Z})$

Expectation

- Discrete

$$\boldsymbol{E_P[X] = \sum_x x\, P(x)}$$

- Continuous

$$\boldsymbol{E_P[X] = \int x\, p(x)\, dx}$$

- E.g., ***E***[FairDie]=3.5

Expectation is Linear

$$\boxed{E_P[X+Y]} = \sum_{x,y} (x+y) P(x,y)$$

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

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

$$= \sum_{x} x\, P(x) + \sum_{y} y\, P(y) = \boxed{E_P[X] + E_P[Y]}$$

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*, …)
  - 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