Recitation 2: Probability

Colin White, Kenny Marino

January 23, 2018

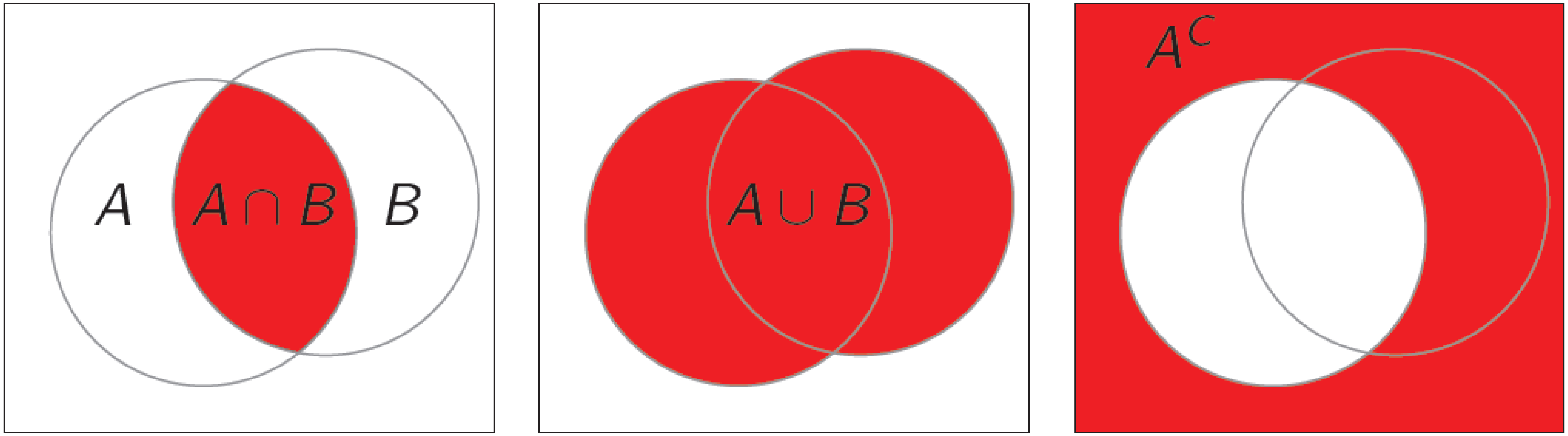
Outline

- Facts about sets
- Definitions and facts about probability
- Random Variables and Joint Distributions
- Characteristics of distributions (mean, variance, entropy)
- Any other questions

Set Basics

A *set* is a collection of *elements*

- **Intersection:** $A \cap B = \{x : x \in A \text{ and } x \in B\}$
- **Union:** $A \cup B = \{x : x \in A \text{ or } x \in B\}$
- **Complement:** $A^{\mathrm{C}} = \{x : x \notin A\}$



Disjointness, Partitions

- Sets $A_1, A_2, \ldots$ are **pairwise disjoint** or **mutually exclusive** if for all $i \neq j$, $A_i \cap A_j = \emptyset$.
- Sets $A_1, A_2, \ldots$ form a **partition** of a set S if they are pairwise disjoint and if $\bigcup_i A_i = S$

Useful facts about partitions:

$$B \cap S = B \cap \left(\bigcup_i A_i \right)$$

$$= \bigcup_i (B \cap A_i) \qquad \text{by the distributive property}$$

- $B \cap A_i$ are also pairwise disjoint

Probability Definitions

- Sample space Ω: set of possible outcomes
- Event space F: collection of subsets
- Probability measure P: assigns probabilities to events
- Probability space (Ω, F, P): set of sample space, event space, and probability measure

Example, rolling a die:

- $\Omega = \{1,2,3,4,5,6\}$
- $F = \{\{1\}, \{2\}, \ldots, \{1,2\}, \ldots, \{1,2,3\}, \ldots \{1,2,3,4,5,6\}, \emptyset\}$
- $P(\{1\}) = \frac{1}{6}$, $P(\{2,4,6\}) = \frac{1}{2}$, etc

Probability Axioms

Kolmogorov conditions for a probability space (Ω, F, P):

- $P(A) \geq 0$ for all $A \in F$
- $P(\Omega) = 1$
- $P(\bigcup_i A_i) = \sum_i P(A_i)$ where $\{A_i\}_i \in F$ are pairwise disjoint

These imply the following:

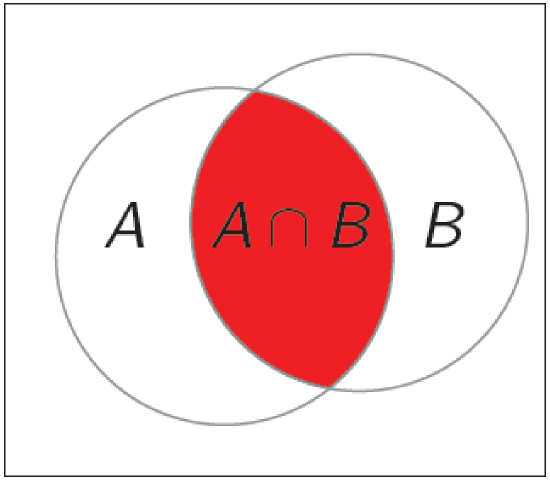
- $P(A^C) = 1 - P(A)$
- $P(A) \leq 1$
- $P(\emptyset) = 0$

Law of Total Probability

$$B = B \cap \Omega = B \cap (A \cup A^C) = (B \cap A) \cup (B \cap A^C)$$

So $\quad P(B) = P(B \cap A) + P(B \cap A^C)$

Called "law of total probability"



$$\begin{aligned} P(A \cup B) &= P\big(A \cup (B \cap A^C)\big) \\ &= P(A) + P(B \cap A^C) \\ &= P(A) + P(B) - P(B \cap A) \\ &\leq P(A) + P(B) \end{aligned}$$

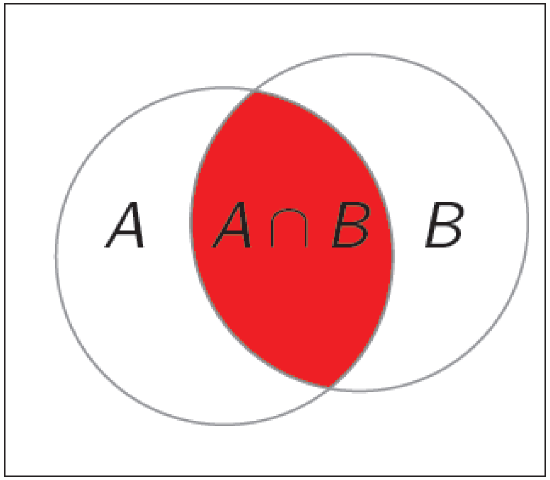
A similar proof for the union bound

Conditional Probabilities

The **conditional probability of $\boldsymbol{A}$ given $\boldsymbol{B}$**:

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

I.e., treat B as the entire sample space, and then find the probability of A.



This implies $P(A|B)P(B) = P(A \cap B)$

"chain rule for probabilities"

Given a partition $A_1, A_2, \ldots$ of Ω,

$$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i)P(A_i)$$

Conditional Probability Example

Given a die, $\Omega = \{1,2,3,4,5,6\}$, $F = 2^{\Omega}$, $P(\{i\}) = 1/6$,

$A = \{1,2,3,4\}$, i.e., the roll is < 5,

$B = \{1,3,5\}$, i.e., the roll is odd.

- $P(A) = 2/3$
- $P(B) = 1/2$
- $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{1,3\})}{P(B)} = \frac{2}{3}$
- $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(\{1,3\})}{P(A)} = \frac{1}{2}$
- Note these quantities are not the same!

Bayes' Rule

Using the chain rule,

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

Rearranging gives us **Bayes' rule:**

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

If B_1, B_2, ... is a partition of Ω, we have

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_i P(A|B_i)P(B_i)}$$

(from Bayes' rule + Law of Total Probability)

Independence

A, B are **independent** if $P(A \cap B) = P(A)P(B)$

When $P(A) > 0$, we can also write this as $P(B|A) = P(B)$

i.e. rolling two dice, etc

A, B are **conditionally independent given $\boldsymbol{C}$** when

$$P(A \cap B|C) = P(A|C)P(B|C).$$

When $P(A) > 0$, we can write $P(B|A, C) = P(B|C)$

i.e., the weather tomorrow is independent of the weather yesterday, given the weather today.

Random Variables

A **random variable** is a function $X: \Omega \to \mathbb{R}^d$,

i.e.

- Roll n dice, X= sum of the numbers
- Indicators of events: $X(\omega) = 1_\omega$, e.g., the indicator of a coin toss coming up heads.
- Throw a dart at a dartboard, $X \in \mathbb{R}^2$ are the coordinates where the dart lands.

Distributions

- By considering random variables, we can think of probability measures as functions on the real numbers
- The probability measure associated with the random variable is characterized by its **cumulative distribution function (CDF):** $F_X(x) = P(X \leq x)$. We write $X \sim F_X$
- If two random variables have the same CDF, we call them **identically distributed.**

Discrete Distributions

- If X only has a countable number of values, then we can characterize it using a **probability mass function (PMF)** which describes the probability of each value $f_X(x) = P(X = x)$.
- We have $\sum_X f_X(x) = 1$ (law of total probability)
- Example: Bernoulli distribution $X \in \{0,1\}$, $f_X(x) = \theta^x(1-\theta)^{1-x}$
- In general, $f_X(x_i) = \theta_i$, where $\sum_i \theta_i = 1$, $\theta_i \geq 0$.
- General model for binary outcomes

Continuous Distributions

- When the CDF is continuous, we can look at the derivative $f_X(x) = \frac{d}{dx} F_X(x)$.
- This is called the **probability density function (PDF).**
- We can compute the probability of an interval (a, b) with $P(a < X < b) = \int_a^b f_X(x)dx$.
- Note the probability of any specific point c, $P(X = c) = 0$
- E.g. Uniform distribution, $f_X(x) = \frac{1}{b-a} * 1_{(a,b)}(x)$
- E.g. Gaussian distribution, $f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp(\frac{(x-\mu)^2}{2\sigma^2})$

Multiple Random Variables

- We can also consider multiple functions from the same sample space, e.g., $X(\omega) = 1_A(\omega), Y(\omega) = 1_B(\omega)$:

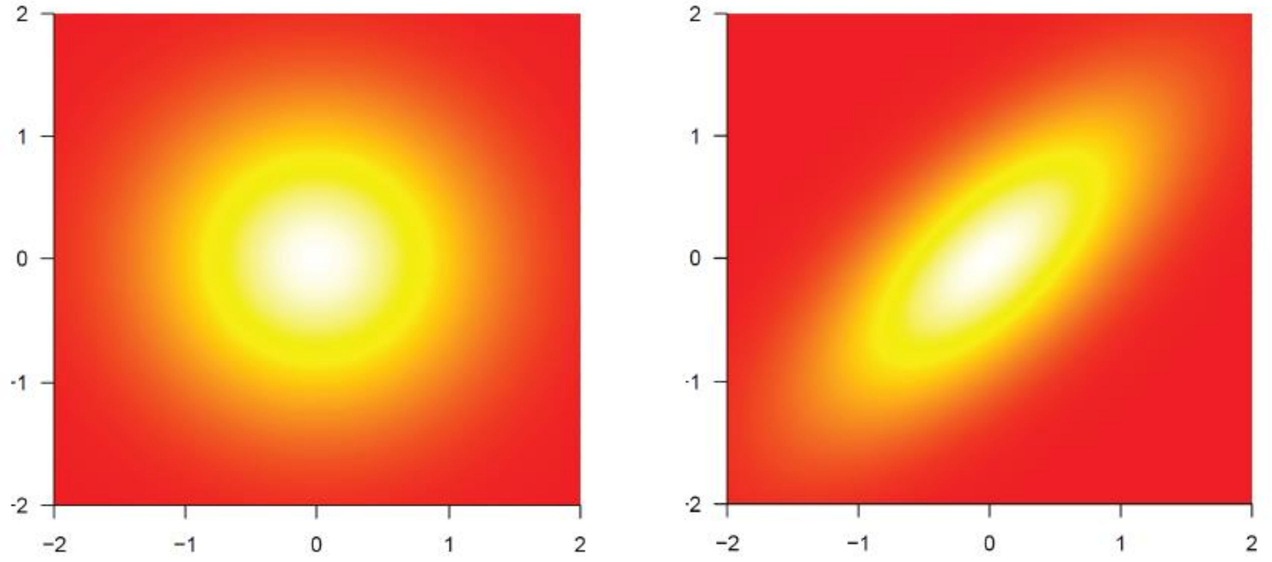
- We can represent the **joint distribution** as a table:

	$\boldsymbol{X = 0}$	$\boldsymbol{X = 1}$
$Y = 0$	.25	.15
$Y = 1$	.35	.25

We write the joint PMF or PDF as $f_{X,Y}(x, y)$

Multiple Random Variables

- If $f_{X,Y}(x, y) = f_X(x) f_Y(y)$, then the two random variables are **independent**
- If the two RVs are independent and identically distributed, we denote this as "i.i.d"



Joint Distributions

- **Marginalizing:** $f_X(x) = \int_y \; f_{X,Y}(x, y)\, dy$. (Similar to the law of total probability)
- **Conditioning:** $f_{X|Y}(x, y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{\int_X \; f_{X,Y}(x,y)\, dx}$.

Mean of a Distribution

- **Expectation** or **mean** of a distribution:

  $E(X) = \sum_X x f_X(x)$ if X is discrete

  $\int_{-\infty}^{\infty} x f_X(x)\, dx$ if X is continuous

- Linearity of Expectation:

$$E(aX + bY + c) = aE(X) + bE(Y) + c$$

- $E(X * Y) = E(X)E(Y)$ is only true when $f_{X,Y} = f_X f_Y$
- $E\big(E(X)\big) = E(X)$

Variance of a Distribution

- **Variance** of a distribution: $Var(X) = E(X - EX)^2$

  how "spread out" is the distribution?

- $E(X - EX)^2$ $$= E(X^2 - 2XE(X) + (EX)^2)$$
  $$= E(X^2) - 2E(X)E(X) + (EX)^2$$
  $$= E(X^2) - (EX)^2$$

What is the variance of a coin toss?

Example of mean/variance

Given $X_1, \ldots, X_n$ i.i.d, $EX_i = \mu$, and $Var(X_i) = \sigma^2$.

What is the expectation and variance of $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$?

$$E(\overline{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} * n * \mu = \mu$$

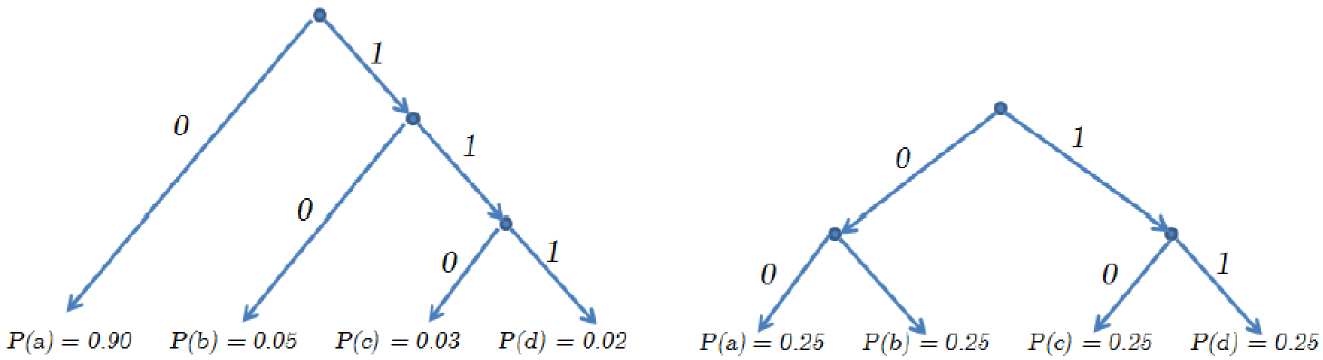
$$Var(\overline{X}_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} * n * \sigma^2 = \frac{\sigma^2}{n}$$

Entropy of a Distribution

Entropy is a measure of uniformity in a distribution

$$H(X) = -\sum_X f_X(x) \log f_X(x)$$

Think about the expected number of bits used to send labeled points



Entropy is the expected depth of the tree (expected number of bits)

Law of Large Numbers

Recall the example, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. What happens when $n \to \infty$?

- **Weak law of large numbers:**

$$\lim_{n\to\infty} P\left(\left|\overline{X}_n - \mu\right| < \varepsilon\right) = 1$$

I.e., given any ε , there exists an n such that $\left|\overline{X}_n - \mu\right| < \varepsilon$

- **Strong law of large numbers:**

$$P\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1$$

I.e., the mean converges to the expectation as n increases

Central Limit Theorem

The distribution of $\overline{X}_n$ starts to look like a Gaussian distribution

$$\lim_{n\to\infty} F_{\overline{X}_n}(x) = \phi\left(\frac{x-\mu}{\sqrt{n}\sigma}\right)$$