STA 35C: Statistical Data Science III

Lecture 2: Probability Review

Dogyoon Song

Spring 2025, UC Davis

Agenda[1]

- Probability basics
- Conditional probability
- Bayes' theorem
- Random variables
- Joint, marginal, and conditional distributions

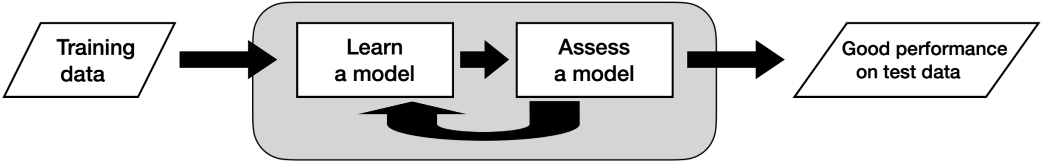
[1] Most of today's topics were covered in STA 35A; see Lectures 13–16

Motivation: Statistical learning recap

Recall the goals of statistical learning:

- Predict Y given X (learn a function $f : X \to Y$)
- Identify patterns in data X

Standard workflow:



Key challenge:

- We aim for good predictions or insights on ***new, unseen* data**
- How should we assess a model, given that training data $\neq$ test data?

Motivation: Why probability?

Probabilistic tools and viewpoints offer a formal way to manage and quantify uncertainty

In particular,

- **Issue/Need:** Training data $\neq$ test data
  - *Remedy:* Assume training and test data are *randomly drawn* from same distribution
- **Issue/Need:** Uncertainty in prediction
  - *Remedy:* Model Y as a *random variable*, and predict Y conditioned on X
- **Issue/Need:** Choosing among many models
  - *Remedy:* Update our belief about the "best model" based on observed data

We will discuss these aspects in more detail throughout the course

Today: We review probability concepts

Probability in everyday examples

- Coin toss
  - Possible outcomes are Head or Tail; each has probability 0.5
- Die roll
  - Possible outcomes $\{1, 2, \ldots, 6\}$; each has probability $1/6$
- Y chromosomes in the US childbirths[2]
  - About 51.2% of births are to babies with Y chromosomes, and 48.8% do not
  - The probability of having a baby with a Y chromosome is 0.512
- Commute time
  - An average commute might take 20 minutes, but it varies with traffic, weather, etc.
- Subjective probability
  - You may personally estimate the likelihood of a stock price rising or falling, based on your own analysis or expert opinions
  - This kind of probability reflects beliefs rather than strict long-run frequencies

[2] Source: CDC National Vital Statistics Reports, Births: Final Data for 2023

Formalizing probability: Sample space and events

- **Sample space**: the set of all possible outcomes, often denoted by Ω
  - e.g., $\{H, T\}$, $\{1, 2, 3, 4, 5, 6\}$
- **Event:** a subset of Ω
  - e.g., $\emptyset$, $\{H\}$, $\{T\}$ $\{H, T\}$, $\{6\}$, $\{1, 2\}$, $\{2, 4, 6\}$
- **Probability**[3]**:** a map P that assigns a number in $[0, 1]$ to each event such that
  - $P(\Omega) = 1$;
  - For disjoint events $A_1, A_2, \ldots$, $P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$
    - Simply put, if $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$
    - Example: $A_1 = \{1, 2\}$, $A_2 = \{6\}$

[3]A formal, mathematically rigorous definition of probability measure is beyond the scope of STA 35C

Basic properties of probability

All the following properties can be derived from the two axioms:

- $P(A^c) = 1 - P(A)$ where $A^c = \Omega \setminus A$
- $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(B \setminus A) = P(B) - P(A \cap B)$
- If $A \subseteq B$, then $P(A) \leq P(B)$

It is often useful to visualize and verify these using a Venn diagram

A quick exercise: a die roll example

Setup:

- $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $P(\{1\}) = P(\{2\}) = \cdots = P(\{6\}) = 1/6$
- $A = \{2, 3, 5\}$ (prime faces)
- $B = \{2, 4, 6\}$ (even faces)

Questions:

- Draw a Venn diagram to visualize Ω, A and B
- Identify $A \cup B$, $A \cap B$ and $A \setminus B$ on the Venn diagram
- Compute $P(A \cup B)$, $P(A \cap B)$, and $P(A \setminus B)$

Conditional probability

Probability an event will occur given that another event has occurred

The conditional probability of A given B is

$$P(A|B) = \begin{cases} \frac{P(A\cap B)}{P(B)} & \text{if } P(B) > 0, \\ 0 & \text{if } P(B) = 0. \end{cases}$$

Example: Compare $P(A)$ vs $P(A|B)$ in the die roll example on the previous slide

Some rules for conditional probability:

- (*Multiplication rule*) $P(A \cap B) = P(B)P(A|B)$
- (*Law of total probability*) $P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$

*Note: Marginal probability refers to "unconditional" probability

Independence

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

- Recall that $P(A \cap B) = P(B)P(A|B)$
- Thus, if A and B are independent, then $P(A|B) = P(A)$
- That is, knowing the outcome of B provides no useful information about the outcome of A and vice versa

Example: Flipping a coin and rolling a die

- Knowing the coin was heads does not help determine the outcome of a die roll

Counter-example: Seeing someone with an umbrella and the day being rainy are not independent

- If we see someone with an umbrella, it is more likely to be a rainy day

Bayes' theorem

Often, we know $P(B|A)$ when what we really want is $P(A|B)$

- A: cause, B: effect
- A: having cancer, B: positive mammogram screening result
- A: "good" prediction function, B: observed data

Assuming that we know (1) marginal probabilities of A and (2) conditional probabilities of "A (cause) $\to$ B (effect)," we want to *"update our belief"* about the cause, A, conditioning on observed effect B

Bayes' theorem states that

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

Bayes' theorem: Examples

Example: Let A=cancer and B=positive screening result

- Suppose $P(A) = 0.01$
- $P(B|A) = 0.8$ (true positive)
- $P(B|A^c) = 0.1$ (false positive; positive screening though a person does not have cancer)

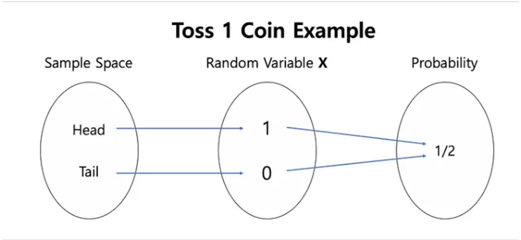
What is $P(A|B)$? How does observing B affect our "belief" on A?

Food for thought: Let A=a model (or a set of models) and B=observed data

Example: A coin gambler's bet on a fair coin $(\mathrm{Bern}(1/2))$ vs biased $(\mathrm{Bern}(1))$

Random variables

- A **random variable**[4] $X : \Omega \to \mathbb{R}$ maps a possible outcome to a number
  - Instead of enumerating all outcomes, we can track a number



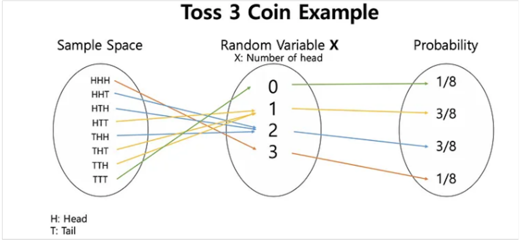


Figure: Random variable example: outcome of tossing a coin[5]

[4] Again, a mathematically rigorous definition is beyond the scope of STA 35C

[5] Source: https://medium.com/jun94-devpblog/prob-stats-1-random-variable-483c45242b3c

Distribution of a random variable

How do we describe the probability that a random variable takes certain values?

- When X is a discrete random variable, its **probability mass function** (PMF) $p_X : \mathbb{R} \to [0, 1]$ satisfies
$$p_X(x) = P(X = x)$$
- When X is a continuous random variable, its **probability density function** (PDF) $f_X : \mathbb{R} \to \mathbb{R}_+$ satisfies
$$P[a \leq X \leq b] = \int_a^b f_X(x)\, dx$$
- In either case, its **cumulative distribution function** (CDF) $F_X : \mathbb{R} \to [0, 1]$ is defined by
$$F_X(x) := P(X \leq x)$$

Examples: Bernoulli, uniform, normal (=Gaussian), ...

Expectation, variance, and covariance

Distributions can be complex; we might want summaries of location, spread, etc.

- The **expected value** of a random variable X is the average outcome you can expect:
  - Discrete: $\mathbb{E}[X] = \sum_x x \cdot p_X(x)$
  - Continuous: $\mathbb{E}[X] = \int x \cdot f_X(x)\, dx$
- The **variance** of a random variable X is the "spread" around the mean:
  - $\mathrm{Var}(X) = \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right]$
  - Alternative formula: $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- The **covariance** between X and Y measures their "joint variability" around means:
  - $\mathrm{Cov}(X, Y) = \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)\left(Y - \mathbb{E}[Y]\right)\right]$
  - Correlation coefficient: $\rho(X, Y) := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \in [-1, 1]$

Examples: (1) symmetric distribution on $\{-1, 0, 1\}$; (2) 2x2 contingency table

Some properties of expectation and variance

Expectation: Let X, Y be random variables and $a, b \in \mathbb{R}$.

- $\mathbb{E}[a] = a$
- $\mathbb{E}[bX] = b \cdot \mathbb{E}[X]$
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

Variance: Let X, Y be random variables and $a, b \in \mathbb{R}$.

- $\mathrm{Var}(a) = 0$
- $\mathrm{Var}(bX) = b^2 \cdot \mathrm{Var}(X)$
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$

Example: a mixture of two Gaussians

Distribution of multiple random variables

Let X, Y be discrete random variables

Joint distribution:

- The joint PMF of X and Y satisfies $p_{X,Y}(x,y) = P(X = x \ \& \ Y = y)$
- The joint CDF of X and Y is defined by $F_{X,Y}(x,y) = P(X \leq x \ \& \ Y \leq y)$

Marginal distribution:

- The marginal PMF of X satisfies

$$p_X(x) = \sum_{y'} p_{X,Y}(x,y')$$

Conditional distribution:

- The conditional PMF of X given Y satisfies

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_{X,Y}(x,y)}{\sum_{x'} p_{X,Y}(x',y)}$$

Question: Can you write marginal PDF and conditional PDF using joint PDF similarly?