

Section 4: Basics in probability theory

STA 141A – Fundamentals of Statistical Data Science

Instructor: Akira Horiguchi

Fall Quarter 2025 (Sep 24 – Dec 12)

MWF, 9:00 AM – 9:50 AM, TLC 1215

University of California, Davis

- 1 Probability measure and random variables
- 2 PMF/PDF
- 3 Some distributions
- 4 Expected value
- 5 Variance and covariance
- 6 Conditional probability and independence

The prerequisite for this class is either STA 108 (regression) or STA 106 (ANOVA), so I expect you have already learned everything in this slide deck.

- If you need a refresher on probability, you can refer to this free textbook:
<https://www.probabilitycourse.com/>

Probability measure and random variables

Probability is a way to quantify randomness and/or uncertainty.

- e.g., coin flips, dice rolls, stocks, weather.
- Rules of probability should be intuitive and self-consistent.
- Self-consistent: the rules shouldn't lead to contradictions.
- Thus these rules must be constructed in a certain way.
- Suppose we want to assign a probability to each event in a set of possible events.
- We would like, at the very least:
 1. each probability to be a value between 0 and 1 (inclusive)
 2. the probability assigned to the full set of events to be 1
 3. the probability assigned to the empty set to be 0
- We need more restrictions to ensure self-consistency.

The following definition will lead to intuitive and self-consistent rules of probability.

Definition 1: Probability measure $P(\cdot)$

For a nonempty set Ω , the set function $P: \Omega \rightarrow [0, 1]$ is a *probability measure*, if

- $P(\Omega) = 1$,
- for any pairwise disjoint sets $A_1, A_2, \dots \subseteq \Omega$ (i.e. $A_i \cap A_j = \emptyset$ for all i, j with $i \neq j$), holds:

$$P\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} P(A_i). \quad (1)$$

This definition fulfills the three properties from the previous slide:

- $P(\Omega) = 1$: the probability of the biggest possible set is equal to 1.
- Property (1) allows us to add probabilities of disjoint sets.
 - ▶ Disjoint means having no shared elements.
 - ▶ (Property (1) is called the *countable additivity* property.)

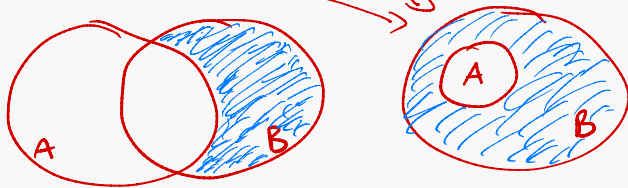
Probability measure - Properties

Definition 1 implies the following additional properties:

Properties of $P(\cdot)$

With \emptyset being the empty set, with some sets $A, B \subset \Omega$, and with $A^c = \Omega \setminus A$ denoting the complement of A , holds,

- i) $P(\emptyset) = 0$; $1 = P(\Omega) = P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset) = 1 + P(\emptyset)$
- ii) $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$; \rightarrow follows from countable additivity
- iii) $P(A^c) = 1 - P(A)$; $1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$
- iv) $P(B \setminus A) = P(B) - P(A)$ if $A \subseteq B$; $P(B) = P(A \cup (B \setminus A)) = P(A) + P(B \setminus A)$
- v) $P(A) \leq P(B)$ if $A \subseteq B$.



Probability measures allow us to characterize the "randomness" of events.

- But we are often interested in more than just probabilities. For example:
 - ▶ the number of heads from three (independent) flips of some coin
 - ▶ the sum of the faces after throwing two dice
 - ▶ the lifetime of a battery
- We call each of these a *random variable* because they take on different values based on random events.
- The probability that a random variable is a certain value will depend on the probabilities of individual events.

PMF/PDF

When doing probability calculations, rather than use probability measures (which are functions of sets), it is often easier to describe a probability distribution using functions of single variables

1. PMF/PDF

The idea behind a PMF/PDF is to assign probabilities to the possible values of a random variable.

- The concept is different for discrete and continuous random variables.

PMF/PDF - discrete and continuous case

A random variable X is **discrete** if its range is finite or countably infinite.

■ Examples:

1. number of heads after two coin flips,
2. number of coin flips needed before a heads turns up.

e.g. subsets of set of integers

■ Here probabilities can be assigned to each realizable value. Examples:

1. For $\{0, 1, 2\}$ (finite), we can assign probabilities $1/4$, $1/2$, and $1/4$.
2. For \mathbb{N} (countably infinite), we can assign probabilities $(1/2)^k$ to each $k \in \mathbb{N}$.

■ The **probability mass function** (PMF) f_X of a discrete random variable X assigns probabilities to each realizable value of X . Examples:

1. $f_X(0) = 1/4$, $f_X(1) = 1/2$, and $f_X(2) = 1/4$.
2. $f_X(k) = (1/2)^k$ for each $k \in \mathbb{N}$.

The PMF at a , $f_X(a) := P(X = a)$, is "the probability that X equals a ."

■ The probability that X lies in a set A can be calculated by

$$\underline{P(X \in A)} = P\left(\bigcup_{a \in A} [X = a]\right) = \sum_{a \in A} P(X = a) = \sum_{a \in A} \underline{f_X(a)} \quad (2)$$

\downarrow countable additivity

► E.g. for example 2, what is the probability that $X < 3$?

$$P(X < 3) = P(X \in \{1, 2, 3\}) = \sum_{k=1}^2 f_X(k) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$$

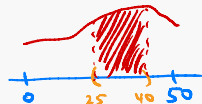
PMF/PDF - discrete and continuous case

e.g. \mathbb{R} , intervals of \mathbb{R}

A random variable X is continuous if its range is uncountably infinite.

- Examples: lifetime of a person, time it takes you to finish the first midterm exam
- For any value in the range of a continuous random variable X , the probability that X is that value must be zero. Why?
 - If uncountably many values are assigned positive probability, the sum of those values would then be infinity!
- For a continuous random variable X , at any value a we have $P(X = a) = 0$.
- The probability density function (PDF) f_X of a continuous random variable X describes how likely it is for X to lie in a set A of values:

$$P(X \in A) = \int_A f_X(s) ds.$$



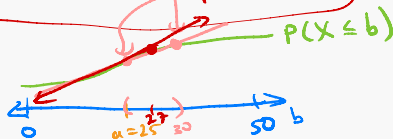
(3)

- Letting $A = (a, a + h]$, we can think of the PDF f_X at a value a as

$$\lim_{h \rightarrow 0^+} \frac{P(X \in A)}{h} = \lim_{h \rightarrow 0^+} \frac{P(a < X \leq a + h)}{h}$$

$$= \lim_{h \rightarrow 0^+} \frac{P(X \leq a+h) - P(X \leq a)}{h}$$

Want $f_X(a)$



From the properties of probability measures, it follows that any PMF f_X of a discrete random variable X must satisfy both

1. $f_X(a) \geq 0$ for all a , and
2. $\sum_{\text{all } a} f_X(a) = 1$.

Similarly, it follows that any PDF f_X of a continuous random variable X must satisfy both

1. $f_X(a) \geq 0$ for all a , and
2. $\int_{\text{all } a} f_X(a) da = 1$.

Some distributions

Discrete case - Uniform distribution

A random variable X with values in a finite set M is *uniformly* distributed if each element in M has the same probability:

$$P(X = k) = \frac{1}{\#M} \quad \text{for all } k \in M$$

"is distributed as"

- Such distributions occur when all possible outcomes are equally likely.
- We write $X \sim U(M)$ or $X \sim \underline{Unif(M)}$.
- Nine random draws in R:

```
sample(c(1,2,3,4,5,6), size=9, replace=TRUE)
```

possible outcomes
of a six-sided die

Discrete case - Bernoulli distribution

A random variable X is **Bernoulli** distributed with parameter $p \in (0, 1)$, if $P(X = 1) = p$ and $P(X = 0) = 1 - p$.

- For when our random experiment has only two possible outcomes ("success" and "failure").
- Example: flip a coin with probability p of heads ("success"). Is it heads?
- We write $X \sim \text{Ber}_p$ or $X \sim \text{Bern}(p)$.
- Nine random draws in R: `rbinom(n=9, size=1, prob=1/3)`

parameter p

End of 10/15
lecture

Discrete case - Binomial distribution

A random variable X is *Binomial* distributed with parameters $n \in \mathbb{N}$ and $p \in (0, 1)$ if

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad \text{for all } k = 0, \dots, n.$$

- We think of n as the number of experiments and p the success probability. In the above equation, k is the number of successes.
- For measuring the probability of the number of successes of n independent Bernoulli experiments with parameter p .
- Example: flip a coin n times, each flip with probability p of heads ("success"). How many heads?
- We write $X \sim \text{Bin}_{n,p}$ or $X \sim \text{Bin}(n, p)$.
- A random draw in R: `rbinom(n=3, size=1, prob=0.25) |> sum()`

A random variable X is **uniformly** distributed on an interval $M = (a, b)$, with $b > a$, if the PDF has the form

$$f_X(x) = \frac{1}{b-a} \quad \text{for all } x \in (a, b).$$

- Such distributions occur when all (uncountably many) possible outcomes are equally likely.
- The interval M can also instead be $[a, b)$, or $(a, b]$, or $[a, b]$.
- Here we also write $X \sim U(M)$ or $X \sim \text{Unif}(M)$.
- Nine random draws in (3, 5) in R: `runif(n=9, min=3, max=5)`

Continuous case - Normal distribution

A random variable X is **normally** distributed with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, if the PDF has the form

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for all } x \in \mathbb{R}.$$

- This distribution appears often in this class, in future classes, and in life!
- We write $X \sim N(\mu, \sigma^2)$. We also call it **Gaussian** distributed.
- Thereby, $E(X) = \mu$ (location parameter), and $\text{Var}(X) = \sigma^2$ (squared scale).
- If $X \sim N(0, 1)$, the distribution of X is said to be **standard normal**.
- Nine random draws in R: `rnorm(n=9, mean=2, sd=1)`

PDF of $X \sim N(0, 1)$, $Y \sim N(2, 1)$, $Z \sim N(0, 3)$

Expected value

Expected value - Introduction

The expected value of a random variable is the weighted average of all of its **values**, where the **weights** are the probabilities that these values occur.

Definition 2: Expected value $E(\cdot)$

Let X be a random variable. Then, the **expected value** of X is in the discrete case and in the continuous case (given the PDF f_X) is defined as

$$E(X) = \sum_{\text{all } k} P(X = k) \cdot k \quad \text{resp.} \quad E(X) = \int_{\text{all } s} f_X(s) \cdot s \, ds. \quad (4)$$

- The expected value of a random variable sometimes does not exist if, for example, the random variable is continuous and the weights are "large" for large values of the random variable (e.g. $E(X) = \int_1^\infty \frac{1}{s^2} \cdot s \, ds = \infty$).

Expected value - Calculating expected value by hand

Calculate $E(X)$ with PDF $f_Y(a) = \frac{3}{7}a^2$ where $a \in [1, 2]$

Properties of $E(\cdot)$

Let $c \in \mathbb{R}$ be a constant, and let X, Y be random variables for which their expected values $E(X)$ and $E(Y)$ exists. Then, the following rules hold.

- i) $E(c) = c$;
- ii) $E(cX) = cE(X)$;
- iii) $E(X + Y) = E(X) + E(Y)$.

Example with $c = 2, E(X) = 1, E(Y) = 5$

Variance and covariance

Heuristics

Variance - Definition and properties

The variance of a random variable is the expected squared deviation of its values to its expected value.

Definition 3: Variance $\text{Var}(\cdot)$

Let X be a random variable with $E(X^2) < \infty$. Then the **variance** of X is defined as

$$\text{Var}(X) := E[\{X - E(X)\}^2]. \quad (5)$$

Think of $\text{Var}(X)$ as “how much X varies about its mean.” We can deduce:

- $\text{Var}(X) \geq 0$.
- $\text{Var}(X) = 0 \Rightarrow X$ is constant.
- The variance of X can also be calculated as

$$\text{Var}(X) = E(X^2) - (E(X))^2. \quad (6)$$

Properties of $\text{Var}(\cdot)$

Let $c \in \mathbb{R}$ be a constant, and let X be a random variable with $E(X^2) < \infty$. Then

- i) $\text{Var}(c) = 0$;
- ii) $\text{Var}(X + c) = \text{Var}(X)$;
- iii) $\text{Var}(cX) = c^2 \text{Var}(X)$;

Recall intuition: $\text{Var}(X)$ is “how much X varies about its mean.”

Example with $c = 5$, $\text{Var}(X) = 1$, $\text{Var}(Y) = 2$.

Expected value and variance help characterize the distribution of a single random variable X .

Now suppose we want to characterize the relationship between two random variables X and Y .

- A complete characterization requires assigning probabilities to every possible pair of values that (X, Y) could be.
- Simpler characterizations are the *covariance* and *correlation* of X and Y .

Heuristics

Definition 4: Covariance $\text{Cov}(\cdot, \cdot)$

Let X, Y be random variables with $E(X^2), E(Y^2) < \infty$. Then the **covariance** between X and Y is defined as

$$\text{Cov}(X, Y) := E((X - E(X))(Y - E(Y))). \quad (7)$$

- The covariance between X and Y can also be calculated as

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y). \quad (8)$$

- We say X and Y are **uncorrelated** if $\text{Cov}(X, Y) = 0$. Then X and Y have no linear relationship, and $E(XY) = E(X)E(Y)$.
- $\text{Cov}(X, Y) > 0$ indicate a positive linear relationship between X and Y .
- $\text{Cov}(X, Y) < 0$ indicate a negative linear relationship between X and Y .
- Covariance is symmetric: $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.

Definition 5: Correlation coefficient $\rho(\cdot, \cdot)$

Let X, Y be random variables with $E(X^2), E(Y^2) < \infty$. Then, the **correlation coefficient** between X and Y is defined as, provided $\text{Var}(X) > 0$ and $\text{Var}(Y) > 0$,

$$\rho(X, Y) := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} \in [-1, 1]. \quad (9)$$

- $\rho(X, Y) = 0 \Rightarrow$ between X and Y is no linear relationship.
- $\rho(X, Y) = -1$ (1) \Rightarrow all values of X and Y lie on a line with negative (positive) slope.
- If $\rho(X, Y)$ is close to -1 (1), there is a strong negative (positive) linear relationship between X and Y .

Properties of $\text{Var}(\cdot)$ and $\text{Cov}(\cdot, \cdot)$

Let $c \in \mathbb{R}$ be a constant, and let X, Y, Z be random variables with $E(X^2) < \infty$, $E(Y^2) < \infty$, and $E(Z^2) < \infty$. Then

- iv) $\text{Var}(X) = \text{Cov}(X, X)$
- v) $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$
- vi) $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- vii) $\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$ and $\text{Cov}(cX, Z) = c\text{Cov}(X, Z)$

(Property vii says $\text{Cov}(\cdot, \cdot)$ is linear in its first argument. Because $\text{Cov}(\cdot, \cdot)$ is symmetric, it is also linear in its second argument. Thus we call it *bilinear*.)

Example with $c = 5$, $\text{Var}(X) = 1$, $\text{Var}(Y) = 2$, $\text{Cov}(X, Y) = 1/3$.

Conditional probability and independence

Heuristics

Definition and properties

An **event** is a subset of the sample space Ω .

Definition 6: Conditional probability

For events $A, B \subseteq \Omega$, the **conditional probability** of A given B is defined by

$$P(A|B) = \begin{cases} \frac{P(A \cap B)}{P(B)}, & \text{if } P(B) > 0, \\ 0, & \text{if } P(B) = 0. \end{cases} \quad (10)$$

- Events A and B are called **independent** if

$$P(A \cap B) = P(A)P(B). \quad (11)$$

Here knowing B provides no information about A , and vice versa.

- Equivalently, events A and B are independent if $P(A|B) = P(A)$.
- Random variables X and Y are called **independent** if for all sets A, B holds,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B). \quad (12)$$

- Independent random variables are uncorrelated.
- But uncorrelated random variables are not necessarily independent!