STA 141A – Fundamentals of Statistical Data Science

Department of Statistics; University of California, Davis

Instructor: Dr. Akira Horiguchi (ahoriguchi@ucdavis.edu)
A01 TA: Zhentao Li (ztlli@ucdavis.edu)
A02 TA: Zijie Tian (zijtian@ucdavis.edu)
A03 TA: Lingyou Pang (lyopang@ucdavis.edu)

Section 4: Basics in probability theory

Spring 2025 (Mar 31 – Jun 05), MWF, 01:10 PM – 02:00 PM, Young 198

Section 4: Basics in probability theory

1. Section 4: Basics in probability theory
   - Section 4.1: Probability measure and random variables
   - Section 4.2: PMF/PDF and CDF
   - Section 4.3: Some distributions
   - Section 4.4: Expected value
   - Section 4.5: Variance and covariance
   - Section 4.6: Conditional probability and independence

The prereq 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/`

Section 4: Basics in probability theory

Section 4.1: 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: Probabilty measure $P(\cdot)$

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

"Omega"

A_i & A_j do not overlap

- $P(\Omega) = 1$,
- for any pairwise disjoints 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:

union

empty set

$$P\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) = \sum_{i\in\mathbb{N}} P(A_i). \tag{1}$$

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.)

Section 4.1 – 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,

"A complement"

i) $P(\emptyset) = 0$;

ii) $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$;

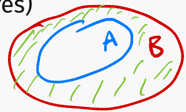
iii) $P(A^c) = 1 - P(A)$;

iv) $P(B \setminus A) = P(B) - P(A)$ if $A \subseteq B$;

v) $P(A) \leq P(B)$ if $A \subseteq B$.

(Pictures)

iv)



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.

Section 4: Basics in probability theory

Section 4.2: PMF/PDF and CDF

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
2. CDF

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.

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.
- 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 $k \in \mathbb{N}$.

  Here $f_X(a)$ is "the probability that X equals a."
- The probability $P(X \in A)$ that X lies in a set A can be calculated by

$$P(X \in A) = \sum_{a \in A} f_X(a), \qquad \text{with} \qquad f_X(a) := P(X = a). \tag{2}$$

- It is common to plot the PMF.

Section 4.2 – PMF/PDF - discrete and continuous case

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

- Examples: the lifetime of a battery, the lifetime of a person, the 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:

$$P(X \in A) = \int_A \overbrace{f_X(s)}^{\text{PDF}} \mathrm{d}s. \tag{3}$$

- It is common to plot the PDF.

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

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

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

1. $f_X(x) \geq 0$ for all x, and
2. $\int_{\text{all } x} f_X(x)\,\mathrm{d}x = 1$.

Section 4.2 – CDF

The *cumulative distribution function* (CDF) of a random variable X is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(a) := P(X \leq a), \quad a \in \mathbb{R}. \tag{4}$$

This is "the probability that X is less than or equal to a."

- Definition holds regardless of whether X is continuous or discrete.
- In the discrete case – recall Eq. (2) – holds for any $a \in \mathbb{R}$,

$$F_X(a) = \sum_{s \leq a} f_X(s) \, .$$

- In the continuous case – recall Eq. (3) – holds for any $a \in \mathbb{R}$,

$$F_X(a) = \int_{-\infty}^{a} f_X(s) \, \mathrm{d}s.$$

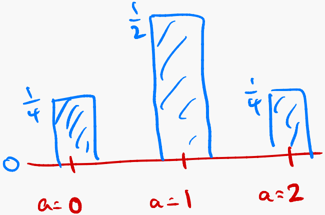
- From the definition in Eq. (4) come the following properties:
  1. The function F_X is (right-continuous) and $\underbrace{\text{monotonically increasing,}}_{\text{nondecreasing}}$
  2. $\lim_{a \to -\infty} F_X(a) = 0$,
  3. $\lim_{a \to \infty} F_X(a) = 1$.
- For any $a, b \in \mathbb{R}$ with $b > a$ holds,

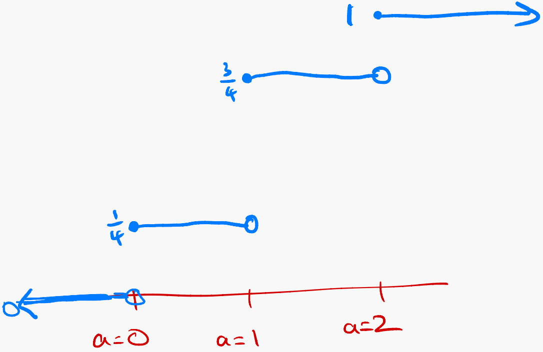
$$P(a < X \leq b) = \overbrace{F_X(b)}^{\text{CDF at b}} - \overbrace{F_X(a)}^{\text{CDF at a}} \, .$$

Discrete random variables

PMF $P(X=a)$

CDF $P(X \le a)$

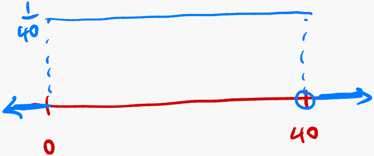


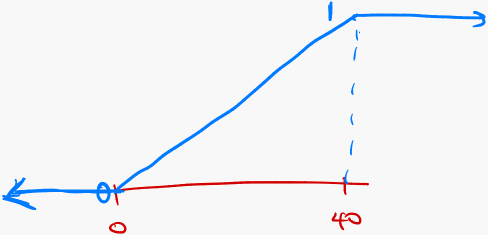


Continuous random variables

PDF $f_X(a)$

CDF $P(X \le a) = \int_{-\infty}^{a} f_X(s)\,ds$





Section 4: Basics in probability theory

Section 4.3: Some distributions

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

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

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

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

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 $X \sim Ber_p$ or $X \sim Bern(p)$.
- Nine random draws in R: `rbinom(n=9, size=1, prob=1/3)`

Section 4.3 – Discrete case - Binomial distr.

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

$\frac{n!}{k!(n-k)!}$

HTT
THT
TTH

$\binom{n}{k} = \binom{3}{1} = \frac{3!}{1!2!} = 3$ outcomes with 1 success

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

2^3 total outcomes

- 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 $X \sim Bin_{n,p}$ or $X \sim Bin(n,p)$.
- 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

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for all } x \in (a, b), \\ 0, & \text{otherwise} \end{cases} \qquad \int_a^b \frac{1}{b-a}\,dx = 1$$

- 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 $X \sim Unif(M)$.
- Nine random draws in $(3, 5)$ in R: `runif(n=9, min=3, max=5)`

Section 4.3 – Continuous case - Normal distr.

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

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

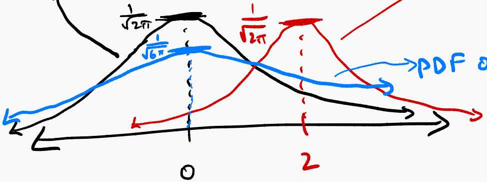
"mu" "sigma squared"

- 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 $Var(X) = \sigma^2$ (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)` $N(2, 1^2)$

PDF and CDF of $X \sim N(\underset{\mu}{0}, \underset{\sigma^2}{1})$, $Y \sim N(2, 1)$, $Z \sim N(0, 3)$

PDF of X: $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x)^2}$

PDF of Y: $f_Y(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2}$



PDF of Z

$f_Z(x) = \frac{1}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sqrt{3}}\right)^2}$

$f_Z(0) = \frac{1}{\sqrt{6\pi}}$

Section 4: Basics in probability theory

Section 4.4: Expected value

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

$$\underbrace{E(X) = \sum_{\text{all } k} \overbrace{P(X = k)}^{\text{weight}} \cdot \overbrace{k}^{\text{value}}}_{\text{discrete}} \quad \text{resp.} \quad \underbrace{E(X) = \int_{\text{all } s} \overbrace{f_X(s)}^{\text{weight}} \cdot \overbrace{s}^{\text{value}}\, \mathrm{d}s}_{\text{continuous}}\,. \tag{5}$$

- 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. $E(X) = \int_1^\infty \frac{1}{s^2} \cdot s \mathrm{d}s = \infty$).

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

$$E(X) = \int_1^2 f_X(a) \; a \; da$$

$$= \int_1^2 \frac{3}{7} a^2 \; a \; da$$

$$= \frac{3}{7} \int_1^2 a^3 \, da$$

$$= \frac{3}{7} \; \frac{a^4}{4} \Big|_{a=1}^{a=2}$$

$$= \frac{3}{7} \left[\frac{2^4}{4} - \frac{1^4}{4} \right]$$

$$= \frac{3}{7} \; \frac{15}{4} = \boxed{\frac{45}{28}}$$

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

i) $E(2) = 2$

ii) $E(2X) = 2E(X) = 2$

iii) $E(X+Y) = E(X) + E(Y) = 1 + 5 = 6$

$EX + EY$

Section 4: Basics in probability theory

Section 4.5: Variance and covariance

Heuristics "How much values vary about their mean"

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

Definition 3: Variance $Var(\cdot)$

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

$$Var(X) := E[\underbrace{\{X - E(X)\}^2}_{\{X - EX\}^2}]. \tag{6}$$

Think of $Var(X)$ as "how much X varies about its mean." We can deduce:

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

$$Var(X) = E(X^2) - (E(X))^2. \tag{7}$$

Properties of $Var(\cdot)$

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

i) $Var(c) = 0$;
ii) $Var(X + c) = Var(X)$;
iii) $Var(cX) = c^2 Var(X)$;

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

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

i) $Var(5) = 0$

ii) $Var(X+5) = Var(X) = 1$

iii) $Var(5X) = 25\, Var(X) = 25$

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 $Cov(\cdot,\cdot)$

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

$$Cov(X,Y) := E\{\underbrace{[X - E(X)]}_{\text{"centered X"}}\overbrace{[Y - E(Y)]}^{\text{"centered Y"}}\}. \tag{8}$$

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

$$\underline{Cov(X,Y)} = E(XY) - E(X)E(Y)\,. \tag{9}$$

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

$X \sim N(0,1)$ $\quad$ $\rho(X,Y) = -1$ $\quad$ $X \sim \text{Unif}(0,1)$ $\quad$ $\rho(X,Y) = 1$

$Y = -X$ $\quad$ $Y = 2X$

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 $Var(X) > 0$ and $Var(Y) > 0$,

$$\rho(X, Y) := \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} \in [-1, 1] . \tag{10}$$

- $\rho(X, Y) = 0 \Rightarrow$ 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 $Var(\cdot)$ and $Cov(\cdot,\cdot)$

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) $Var(X) = Cov(X, X)$

v) $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X, Y)$

vi) $Cov(X, Y) = Cov(Y, X)$

vii) $Cov(\underline{X+Y}, Z) = Cov(\underline{X}, Z) + Cov(\underline{Y}, Z)$ and $Cov(\underline{c}X, Z) = \underline{c}Cov(X, Z)$

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

a bilinear function

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

$$Var(X+Y) = Var(X) + Var(Y) + 2\cdot Cov(X,Y) = 1 + 2 + 2\cdot\tfrac{1}{3} = 3\tfrac{2}{3} = \tfrac{11}{3}$$

$$\begin{aligned} Var(2X+Y) &\overset{v)}{=} Var(2X) + Var(Y) + 2\cdot Cov(2X,Y) \\ &= 4Var(X) + 2 + 4\,Cov(X,Y) \\ &= 4\cdot 1 + 2 + 4\cdot\tfrac{1}{3} = 6 + \tfrac{4}{3} = \tfrac{22}{3} \end{aligned}$$

$\xrightarrow{\text{more}}$ examples

Section 4: Basics in probability theory

Section 4.6: Conditional probability and independence

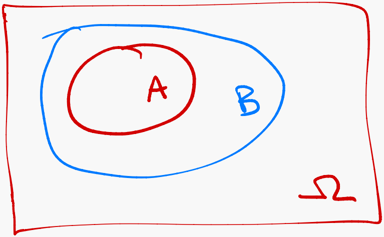
$$\mathrm{Var}(X-2Y) = \mathrm{Var}(X) + \underbrace{\mathrm{Var}(-2Y)}_{(-2)^2\,\mathrm{Var}(Y)} + \underbrace{\mathrm{Cov}(X,-2Y)}_{-2\,\mathrm{Cov}(X,Y)}$$

$$= 1 + 4\cdot 2 - 2\cdot 1/3 = 9 - \tfrac{2}{3} = \boxed{25/3}$$

$$\mathrm{Cov}(X+Y,\ Z+Y) = \underbrace{\mathrm{Cov}(X,\ Z+Y)}_{\mathrm{Cov}(X,Z) + \mathrm{Cov}(X,Y)} + \underbrace{\mathrm{Cov}(Y,\ Z+Y)}_{\mathrm{Cov}(Y,Z) + \underbrace{\mathrm{Cov}(Y,Y)}_{\mathrm{Var}(Y)}}$$

Section 4.6 – Conditional probability - Introduction

Heuristics



$X \in \{0, 1, 2\}$

$P(X=0) = \frac{1}{4}$

$P(X=1) = \frac{1}{2}$

$P(X=2) = \frac{1}{4}$

$$P(X=0 \mid X<2) = P(X=0 \mid X=0 \text{ or } X=1)$$

$$= \frac{P(X=0)}{P(X=0 \text{ or } X=1)} = \frac{1/4}{\frac{1}{4}+\frac{1}{2}} = \frac{1}{3}$$

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

Definition 6: Conditional probability

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

$$P(A|B) = \begin{cases} \frac{P(A\cap B)}{P(B)}, & \text{if } P(B) > 0, \\ 0, & \text{if } P(B) = 0. \end{cases} \tag{11}$$

- Events A and B are called *independent* if

$$P(A \cap B) = P(A)P(B). \tag{12}$$

  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,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B). \tag{13}$$

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