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

Overview

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

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 measure - Motivation

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.

Probability measure - Definition

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

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

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

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.)

Definition 1 implies the following additional properties:

Properties of $P(\cdot)$

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

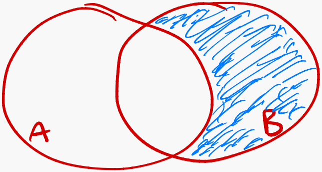
i) $P(\varnothing) = 0$; $1 = P(\Omega) = P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset) = 1 + P(\emptyset)$

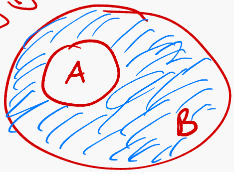
ii) $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \varnothing$; → follows from countable additivity

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 $A \subseteq B$.





Random variables - Notion

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

Motivation

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

PMF/PDF - concept

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.

e.g. subsets of set of integers

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

  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

$$P(X \in A) = P\Big(\bigcup_{a \in A} [X = a]\Big) = \sum_{a \in A} P(X = a) = \sum_{a \in A} f_X(a) \tag{2}$$

countable additivity

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

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

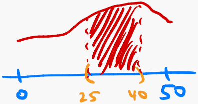
e.g. $\mathbb{R}$, intervals of $\mathbb{R}$

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

$[0,\infty)$ $[0,50]$

- 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:

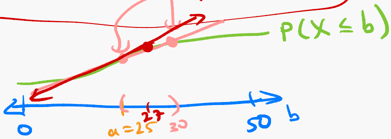
$$P(X \in A) = \int_A f_X(s)\mathrm{d}s. \tag{3}$$



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

$$\lim_{h\to 0^+} \frac{P(X \in A)}{h} = \lim_{h\to 0^+} \frac{P(a < X \le a + h)}{h} = \lim_{h\to 0^+} \frac{P(X \le a+h) - P(X \le a)}{h}$$

Want $f_X(a)$



PMF/PDF - discrete and continuous case

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. $\int_{\text{all } a} f_X(a)\,\mathrm{d}a = 1$.

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"

- We write $X \sim U(M)$ or $X \sim Unif(M)$.
- Such distributions occur when all possible outcomes are equally likely.
- 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 $P(X = 0) = 1 - p$.

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

parameter p

End of 10/15 lecture

Discrete case - Binomial distribution

\# ways we can get k successes and n-k failures

prob of k successes

prob of n-k failures

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

$P(X=0) = (1-p)^2$

$P(X=2) = p^2$

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

$P(X=1) = p(1-p) + (1-p)p \longrightarrow \binom{2}{1} = 2$

- We write $X \sim Bin(n, p)$.
- For measuring the probability of the number of successes of n *independent* Bernoulli experiments with “success probability” parameter p.
- Example: how many heads in n independent coin flips, each flip with probability p of heads ("success")?
- A random draw in R: `rbinom(n=3, size=1, prob=0.25) |> sum()`

n p

e.g. 0 1 1

0 0 0

Continuous case - Uniform distribution

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

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

- Here we also write $X \sim U(M)$ or $X \sim Unif(M)$.
- 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 $[a, b]$.
- Nine random draws in $(3, 5)$ in R: `runif(n=9, min=3, max=5)`

(a,b) $\qquad\qquad\qquad\qquad$ a $\qquad$ b

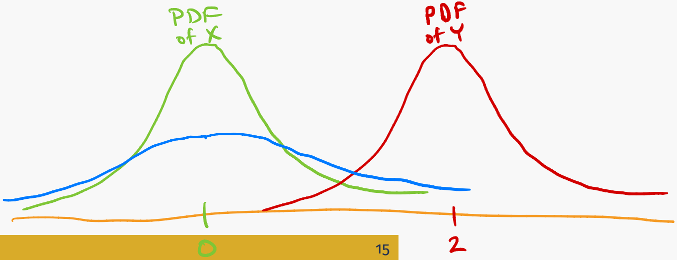
A random variable X is *normally* distributed with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, if the PDF has the form

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

- We write $X \sim N(\mu, \sigma^2)$. We also call it *Gaussian* distributed.
- This distribution appears often in this class, in future classes, and in life!
- It can be shown that $E(X) = \mu$ (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)`

  μ σ

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

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

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

Calculate the expected value of a random variable with PDF $f_X(a) = \frac{3}{7}a^2$ **where** $a \in [1, 2]$

$$E(X) = \int_{-\infty}^{\infty} f_X(a) \cdot a \, da$$

$$= \cancel{\int_{-\infty}^{1} 0 \cdot a \, da} + \int_{1}^{2} \frac{3}{7} a^2 \cdot a \, da + \cancel{\int_{2}^{\infty} 0 \cdot a \, da}$$

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

$$= \frac{3}{7} \left[\frac{a^4}{4}\right]_1^2$$

$$= \frac{3}{28} (2^4 - 1^4)$$

$$= \frac{45}{28}$$

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

$E(cX) = cE(X) = 2 \cdot 1 = 2$

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

$E(X+c) = E(X) + E(c) = E(X) + c = 1+2 = 3$

$E(cY + X) = E(cY) + E(X) = cE(Y) + E(X) = 2 \cdot 5 + 1 = 11$

Variance and covariance

Heuristics

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

$E[X - E(x)]$
$= E[x] - E[E(x)]$
$= E[x] - E(x) = 0$

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

$Var(cX) = E[\{cX - E(cX)\}^2]$

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

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

Properties of $Var(\cdot)$

Let $c \in \mathbb{R}$ be a constant, and let X be a random variable with $\underline{E(X^2) < \infty}$. 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 $c = 5$, $Var(X) = 1$, $Var(Y) = 2$.

$Var(X+c) = Var(X) = 1$

$Var(cY) = c^2 Var(Y) = 5^2 \cdot 2 = 50$

End of 10/17 lecture

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

Covariance - Definition and properties

$$Var(X) = E[(X-EX)^2] = E[(X-EX)(X-EX)] = Cov(X,X)$$

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

$$Cov(X,Y) := E\left[\overbrace{(X - E(X))}^{X \text{ centered}}\overbrace{(Y - E(Y))}^{Y \text{ centered}}\right]. \tag{7}$$

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

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

- We say X and Y are *uncorrelated* if $Cov(X,Y) = 0$. Then X and Y have no linear relationship, and $E(XY) = E(X)E(Y)$.
- $Cov(X,Y) > 0$ 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: $Cov(X,Y) = Cov(Y,X)$.

$\sqrt{Var(X)}$ = standard deviation of 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 $Var(X) > 0$ and $Var(Y) > 0$,

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

Covariance between two r.v.s w/ unit variance $\leftarrow = Cov\left(\frac{X}{\sqrt{Var(X)}}, \frac{Y}{\sqrt{Var(Y)}}\right)$

- $\rho(X, Y) = 0 \Rightarrow$ 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) $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$ and $Cov(cX, Z) = cCov(X, Z)$

(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*.)

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

$$Var(cX + Y) \overset{v)}{=} Var(cX) + Var(Y) + 2Cov(cX, Y)$$
$$= c^2 Var(X) + Var(Y) + 2c\,Cov(X, Y)$$
$$= 25 + 2 + \frac{10}{3} = \boxed{\frac{91}{3}}$$

Conditional probability and independence

Heuristics

$\boxed{\text{Independent} \neq \text{disjoint!}}$

If A and B are disjoint, then knowing B occurred tells us that A did not occur.

$\Downarrow$

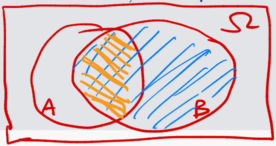
Knowing B occurred tells us a lot of information about A.

Definition and properties

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

Definition 6: Conditional probability

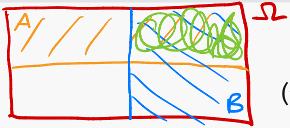
For events $A, B \subseteq \Omega$, the *conditional probability* of A given B is defined by



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

- Events A and B are called *independent* if

If A and B are disjoint, then knowing B occurred tells us that A did not occur → independent ≠



$$P(A \cap B) = P(A)P(B). \tag{11}$$

  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,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B). \tag{12}$$

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

Conditional probability and independence

Bayes' rule

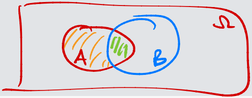
From the definition of conditional probability, we know for any two events A and B that

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

Dividing by $P(A)$ (assuming it is not zero), we get *Bayes' rule*:

Theorem 2: Bayes' theorem

Let $\Omega \neq \varnothing$. For any events $A, B \subseteq \Omega$ with $P(A) \neq 0$ holds, $P(B)$ = prior information about B



$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \tag{13}$$

Often $P(A)$ is unknown and difficult to deduce; can use the *law of total probability* (14).

- Because the sets $A \cap B$ and $A \cap B^c$ partition the set A, we can write $P(A)$ as

$$P(A) = P\big((A\cap B)\cup(A\cap B^c)\big) = P(A\cap B) + P(A\cap B^c) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

- More generally, for any partition $\{B_1, B_2, \ldots\}$ of Ω, we can write $P(A)$ as

Then $A\cap B_1$, $A \cap B_2$, ... partition A

$$P(A) = \sum_{j=1}^{\infty} P(A|B_j)P(B_j). \tag{14}$$

Example: False positive paradox

$P(D) \approx \frac{1}{10000} = 0.0001$

A certain disease affects about 1 out of 10,000 people. There is a test to check whether the person has the disease. In particular, we know that

- the probability that the test result is positive, given that the person does not have the disease, is 2%; $P(+|D^c) = 0.02$
- the probability that the test result is negative, given that the person has the disease, is 1%. $P(+^c|D) = 0.01 \Rightarrow P(+|D) = 1 - 0.01 = 0.99$

Suppose a random person gets tested for the disease and the test result is positive. What is the probability that the person has the disease?

"+" – positive test result

"D" – person has disease

Bayes Thm: $P(D|+) = \frac{P(+|D)\,P(D)}{P(+)} = \frac{0.00099}{0.020097} \approx 0.005$

$\downarrow$

50x base rate

$$P(+) = P(+|D)\,P(D) + P(+|D^c)\,P(D^c)$$

$$= 0.99 \cdot 0.0001 + 0.02 \cdot (1 - 0.0001)$$

$$= 0.020097$$

Example: False positive paradox

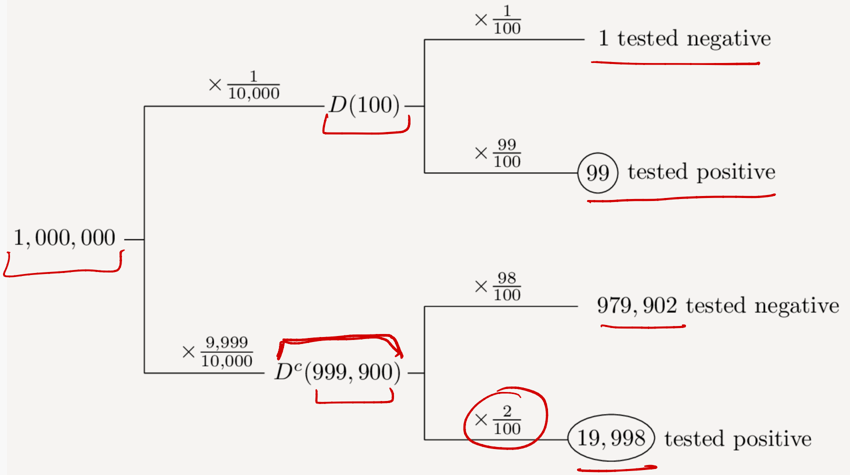


Fig.1.25 - Tree diagram for Example 1.26.

Figure: From textbook Pishro-Nik (2014).

Bayesian paradigm

Bayes' rule enables *Bayesian statistics* (STA 145).

- *Bayesian* interpretation: probability expresses a degree of belief in an event. Use *Bayes' rule* to update degree of belief based on observed data.
- *Frequentist* interpretation: probability is the long-run relative frequency of an event after many trials.
- Don't need to know for this course. More intuition here `https://www.youtube.com/watch?v=9wCnvr7Xw4E`