Section 5: Basics in probability theory

STA 141A – Fundamentals of Statistical Data Science (UC Davis, Spring 2026)

Instructor: Akira Horiguchi

Overview

Probability measure

Random variables, PMF/PDF

Expected value and variance

- Some distributions

Relationship between events or random variables

- Covariance and correlation
- 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 the free textbook: Introduction to probability, statistics, and random processes by H. Pishro-Nik (2014).

Probability measure

Random variables, PMF/PDF

Expected value and variance

Some distributions

Relationship between events or random variables

Covariance and correlation

Conditional probability and independence

Bayes' rule

Probability measure - Motivation

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

- e.g., coin flips, dice rolls, stocks, weather.
- Suppose we have possible events (sets) that we want to assign probabilities to.
- Let's mathematically define "probability" so that
  1. each probability is a value between 0 and 1 (inclusive),
  2. the probability assigned to the empty set is 0,
  3. the probability assigned to the full set of events is 1.
- We need more restrictions to ensure *self-consistency* (our definition shouldn't lead to contradictions).

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 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:

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

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

This definition fulfills the three properties from the previous slide (prove this):

1.
2.
3.

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 $A^c = \Omega \backslash A$ denoting the complement of A, holds,

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 $A \subseteq B$.

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 your phone 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.

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. Idea is to assign probabilities to the possible values of a random variable.
   - The concept is different for discrete and continuous random variables.
2. CDF. Idea is to assign probabilities to ranges of the possible values of a random variable.
   - Nice because it works for all types of random variables, but not covered in this class.

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

- Example 1: let X be the number of heads after two coin flips.
  - Can assign probabilities to each realizable value.
    E.g., For $\{0, 1, 2\}$ (finite), we can assign probabilities $1/4$, $1/2$, and $1/4$.
  - Can compute the probability that $X < 2$:
- Example 2: let X be the number of coin flips needed before a heads turns up.
  - Can assign probabilities to each realizable value.
    E.g., for $\mathbb{N}$ (countably infinite), we can assign probabilities $(1/2)^k$ to each $k \in \mathbb{N}$.
  - Can compute the probability that $X > 2$:

General terminology

- The *probability mass function* (PMF) f_X of a discrete random variable X specifies a probability for each realizable value of X.
  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

$$P(X \in A) = \qquad = \sum_{a \in A} f_X(a) \quad (2)$$

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:

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

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. $\sum_{\text{all } a} f_X(a) = 1$.

...any PDF f_X of a continuous random variable X must satisfy both

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

Motivation

Sometimes we want to do summarize a probability distribution with a handful of scalar numbers. We would at least like to capture:

- Central tendency (e.g., mean, median, mode).
  - Mode is nice because it can be computed for a non-numeric distribution.
- Spread (e.g., variance or standard deviation).

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

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

Example - Calculating expected value by hand

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

.

Expected value can be infinite

Don't worry about this for this class, but...

- The expected value of a random variable sometimes does not exist (i.e., is infinity) if, for example, the random variable is continuous and the weights are "large" for large values of the random variable (e.g. $E(X) = \int_1^\infty \frac{1}{s^2} \cdot s \, ds = \infty$).
- This is why you might see written explicitly "assume that the random variable has finite expected value" or "assume that $E(X) < \infty$."

Expected value - Calculation tools

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 - Definition and properties

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

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

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

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

Variance - Calculation tools

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

.

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:

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

- We write $X \sim U(M)$ or $X \sim Unif(M)$.
- Such distributions occur when all possible outcomes are equally likely.
- The set M might be non-numeric, in which case we cannot compute the expected value.
- If M contains only scalar numbers, then we can easily show:
  - $E(X) = \sum_{k \in M} k \frac{1}{\#M} = \frac{1}{\#M} \sum_{k \in M} k.$
- Nine random draws in R: `sample(c(1,2,3,4,5,6), size=9, replace=TRUE)`

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?
- Can easily show:
  - $E(X) \ = \ 0 \cdot P(X = 0) + 1 \cdot P(X = 1) \ = \ 0 + 1 \cdot p \ = \ p;$
  - $E(X^2) \ = \ 0 \cdot P(X = 0) + 1 \cdot P(X = 1) \ = \ 0 + 1 \cdot p \ = \ p;$
  - $Var(X) \ = \ E(X^2) - [E(X)]^2 \ = \ p - p^2 \ = \ p(1 - p).$
- Nine random draws in R: `rbinom(n=9, size=1, prob=1/3)`

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 $[a, b]$.
- Can easily show:
  - $E(X) = \int_a^b t \frac{1}{b-a} \, \mathrm{d}t = \frac{a+b}{2}$;
  - $E(X^2) = \int_a^b t^2 \frac{1}{b-a} \, \mathrm{d}t = \frac{1}{b-a} \frac{b^3-a^3}{3} = \cdots$ (simplify by factoring out $(b-a)$ from $b^3 - a^3$);
  - $Var(X) = E(X^2) - [E(X)]^2 = \cdots = \frac{(b-a)^2}{12}$.
- 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

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

- 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!
- One can (less easily) do the math to show that:
  - $E(X) = \int_{-\infty}^{\infty} t\, f_X(t)\, \mathrm{d}t = \cdots = \mu$ (location parameter);
  - $E(X^2) = \int_{-\infty}^{\infty} t^2\, f_X(t)\, \mathrm{d}t = \cdots = \mu^2 + \sigma^2$;
  - $Var(X) = E(X^2) - [E(X)]^2 = \sigma^2$ (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 $X \sim N(0,1), Y \sim N(2,1), Z \sim N(0,3)$

Covariance and correlation - Motivation

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.

Covariance - Definition and properties

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

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

- 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: $Cov(X,Y) = Cov(Y,X)$.

Correlation coefficient

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 $Var(Y) > 0$,

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

- $\rho(X,Y) = 0 \Rightarrow$ 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.

Variance and covariance - More calculation tools

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

Another exposition

The YouTube channel StatQuest (made by a UNC prof) has some great expositions:

- Expected Values, Main Ideas!!! (13:39)
- Calculating the Mean, Variance and Standard Deviation, Clearly Explained!!! (14:21)
- Covariance, Clearly Explained!!! (22:22)
- Pearson's Correlation, Clearly Explained!!! (19:12)

Conditional probability and independence: events

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

Definition 6: Conditional probability of events

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

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

  Here knowing B provides no information about A, and vice versa.
- Equivalently, events A and B are independent if $P(A|B) \;=\; P(A)$.

Conditional probability and independence: random variables

- Random variables X and Y are called *independent* if for all sets A and 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!
- StatQuest: Conditional Probabilities, Clearly Explained!!! (10:55)

Bayes' rule: Definition

For any two events A and B, the definition of conditional probability tells us 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' rule

Let $\Omega \neq \emptyset$. For any events $A, B \subseteq \Omega$ with $P(A) \neq 0$ holds,

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

Often $P(A)$ in (13) is unknown and difficult to deduce; can use the *law of total probability* (14).

Bayes' rule: LOTP

Law of total probability (LOTP)

For $A \subseteq \Omega$ and any partition $\{B_1, B_2, B_3, \dots\}$ of the set Ω, we can write $P(A)$ as

$$\begin{aligned} P(A) &= P\left(\bigcup_{i=1}^{\infty} [A \cap B_i]\right) \\ &= \sum_{i=1}^{\infty} P(A \cap B_i) \\ &= \sum_{i=1}^{\infty} P(A|B_i)\, P(B_i). \end{aligned} \tag{14}$$

Special case: $P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$.

Bayes' rule: application

Example: False positive paradox

A certain disease affects about 1 out of 10,000 people. There is a test to check whether the person has a certain 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%;
- the probability that the test result is negative, given that the person has the disease, is 1%.

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?

.

Bayes' rule: application

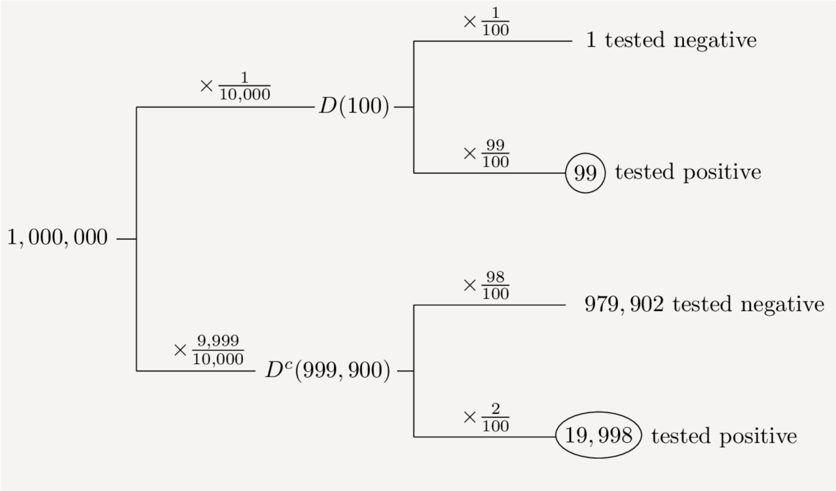

Fig.1.25 - Tree diagram for Example 1.26.

Figure 1: From textbook Pishro-Nik (2014). Example: False positive paradox.

Bayesian paradigm

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

- *Frequentist* interpretation: probability is the long-run relative frequency of an event after many trials.
- *Bayesian* interpretation: probability expresses a degree of belief in an event.
  Use *Bayes' rule* to update degree of belief based on observed data.
- Helpful for e.g. hierarchical modeling: how can we appropriately share information about similar but not identically distributed groups?
- Don't need to know for this course. More intuition on StatQuest video: Bayes' Theorem, Clearly Explained!!!!