Probability Background Review

(STAT 32950-24620 FINM 34700 Spring 2026)

Random variables, conditional probability, conditional expectation

Contents

1 Probability, sample space, independence

Probability is used to quantify uncertain events.

1.1 Sample space and events

- Sample space is the set of possible outcomes of an experiment.

  Other terms for sample outcomes: sample realizations, sample points or sample elements.

- Notations:

  Sample spaces are denoted as $\Omega = \{\omega\}$ or $\mathcal{S} = \{s\}$, where ω, s denote possible outcomes.

  For example,

  - If the experiment consists of tossing a coin once, $\Omega = \{\text{Head, Tail}\} = \{H, T\}$.
  - If the experiment is to toss a coin twice, then $\Omega = \{HH, HT, TH, TT\}$.

- Events are subsets of sample space, denoted as A, B, C, D, E, etc.

  For example, in the experiment of tossing a coin twice, $A = \{\text{first toss is a head}\} = \{HH, HT\}$ is an event consisting of two possible outcomes.

It is not uncommon for choosing sample space to be larger than necessary.

For example, the sample space for possible temperates in Fahrenheit in Chicago area may be chosen as $\Omega = \{t;\ t \in (-\infty, \infty) = \mathbb{R}\}$, where t means $t\,^\circ F$, t degrees in Fahrenheit.

Venn diagram: Since events are sets, Venn diagrams are often used to illustrate event operations.

1.2 Operations of events

The following are common event operations and notations.

- Complement $A^c = \{\text{not } A\} = \{\omega : \omega \notin A\}$
- Union $A \cup B = \{A \text{ or } B\} = \{\omega : \omega \in A \text{ or } \omega \in B\}$

  Note that here "or" means "or/and", either or both.
- Intersection $A \cap B = AB = \{A \text{ and } B\} = \{\omega : \omega \in A \text{ and } \omega \in B\}$

  Note that here "and" means "simultaneously".
- Set difference or set minus $A \setminus B = \{A \text{ but not } B\} = A \cap B^c$, also denoted as $A - B$.
- Set inclusion $A \subset B$ (any element of A is in B), thus if $A \subset B$ then $\omega \in A \ \Rightarrow\ \omega \in B$.
- Null event $\emptyset$ (impossible event, always false).
- True event Ω (deterministic event, always true).
- Disjoint events $A \cap B = \emptyset$ (A and B are mutually exclusive).
- **De Morgan's laws**

$$(A \cup B)^c = A^c \cap B^c, \qquad (A \cap B)^c = A^c \cup B^c$$

- Partition of the sample space

  $\{A_i, i = 1, 2, \cdots\}$ is a partition of Ω if $\Omega = A_1 \cup A_2 \cup \cdots = \bigcup_{i=1}^{\infty} A_i$, where A_i's are disjoint.

  The disjointness means pairwise mutually exclusive, i.e. $A_i \cap A_j = \emptyset$ for $i \neq j,\ i, j = 1, 2, \cdots$.

- Monotone increasing events $A_1 \subset A_2 \subset \cdots$.

  Define $\lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i = A$, expressively denoted as $A_n \nearrow A$ (or simply $A_n \to A$).

- Monotone decreasing events $A_1 \supset A_2 \supset \cdots$.

  Define $\lim_{n\to\infty} A_n = \bigcap_{i=1}^{\infty} A_i = A$, denoted as $A_n \searrow A$ (or $A_n \to A$).

- Limits of sequence of events.

$$\{\text{Infinitely many } A_n \text{ occur}\} = \overline{\lim_{n\to\infty}} A_n = \limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{j=n}^{\infty} A_j$$

$$\{\text{All but finitely many } A_n \text{ occur}\} = \underline{\lim_{n\to\infty}} A_n = \liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{j=n}^{\infty} A_j$$

- Indicator function of A:

$$\mathbf{1}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A; \\ 0, & \text{if } \omega \notin A. \end{cases}$$

  Often the simpler notation $\mathbf{1}_A$ or I_A are used, when the context is clear.

1.3 Probability of events

Definition

$\mathbb{P}$ (or simply P) is a probability function (*a.k.a.* probability distribution, probability measure) on a sample space Ω if

- $\mathbb{P}(A) \geq 0$,
- $\mathbb{P}(\Omega) = 1$,
- $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$, if A_i's are disjoint.

where A, A_i's are (measurable) events $\subset \Omega$.

Properties

- $\mathbb{P}(\emptyset) = 0$
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
- $A \subset B \ \Rightarrow\ \mathbb{P}(A) \leq \mathbb{P}(B)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB)$.
- Similarly, for n events,

$$\mathbb{P}(A_1 \cup A_2 \cdots \cup A_n) = \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i A_j) + \sum_{i<j<k} \mathbb{P}(A_i A_j A_k) - \cdots + (-1)^{n-1} \mathbb{P}(A_1 A_2 \cdots A_n)$$

- Boole's inequality

$$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) \ \leq\ \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

- Continuity

$$A_n \subset A_{n+1},\ n \geq n_o \quad or \quad A_{n+1} \subset A_n,\ n \geq n_o \quad \Longrightarrow \quad \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(\lim_{n\to\infty} A_n)$$

- Borel-Cantelli Lemmas

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \ \Longrightarrow\ \mathbb{P}(\text{infinitely many } A_n \text{ occur}) = 0$$

$$A_n\text{'s are independent}, \ \sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty \ \Longrightarrow\ \mathbb{P}(\text{infinitely many } A_n \text{ occur}) = 1$$

Remarks

- Frequency interpretation of probability:

  If the experiment is repeated over and over again, then the proportion of time that event A occurs will be $\mathbb{P}(A)$.

- Measure theory concerns on the definition of probability:

  In general, it is mathematically impossible to assign probability to all subsets of a general sample space Ω. Measure theory deals with this problem. Usually we restrict to a collection of "good" (measurable) subsets of Ω, denoted as $\mathcal{B}$, such that

  - $\emptyset \in \mathcal{B}$,
  - $A \in \mathcal{B}$ then $A^c \in \mathcal{B}$,
  - $A_1, A_2, \cdots \in \mathcal{B}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$.

  Such $\mathcal{B}$ is called a σ-field. Only to the subsets in $\mathcal{B}$ we can assign probability $\mathbb{P}$.

  The triplet $(\Omega, \mathcal{B}, \mathbb{P})$ is called a probability space.

  However, almost all probability events that will ever concern us belong to such a $\mathcal{B}$.

- Implicit assumptions on common sample space and probability

  Often events and their probability are used without mentioning their sample space. Implicitly, a common sample space and its probability measure is assumed.

  It is important to clearly identify when events and probabilities in discussion are not from the sample space, or the sample sample space but different probability measures.

1.4 Independence

Independence is one the most important concepts in probability and statistics.

Definitions

Two events A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$$

A common notation for independence is

$$A \perp\!\!\!\perp B$$

A set of events $\{A_i, i \in K\}$ are independent if for and finite subset $J \subset K$,

$$\mathbb{P}\left(\bigcap_{i\in J} A_i\right) = \prod_{i\in J} \mathbb{P}(A_i)$$

The last expression is the products of all $\mathbb{P}(A_i), i \in J$. The sets J, K are called index sets.

When do we have independence?

- By verifying that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$ holds for the given sample space Ω and the corresponding probability measure $\mathbb{P}$, for any $A, B \subset \Omega$.
- Or, by assuming that the independence condition holds, as often done in practice.

Independent vs disjoint

"Independent events" and "disjoint events" are not the same.

In fact, disjoint events of positive probability are not independent, since for $A \cap B = \emptyset$,

$$0 = \mathbb{P}(\emptyset) = \mathbb{P}(AB) \neq \mathbb{P}(A)\mathbb{P}(B) > 0.$$

1.5 Conditional probability

If $P(B) > 0$, the conditional probability of A given B is defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}.$$

Consequently, the multiplication rule follows:

$$\mathbb{P}(AB) = \mathbb{P}(A|B)\,\mathbb{P}(B)$$

Remarks

- A is independent of B if and only if $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$, thus, if and only if $\mathbb{P}(A|B) = \mathbb{P}(A)$.
- Independence or dependence is a symmetric relationship.

$$\mathbb{P}(A|B) = \mathbb{P}(A) \quad \Longrightarrow \quad \mathbb{P}(B|A) = \mathbb{P}(B)$$

- In general,

$$\mathbb{P}(A|B) \neq \mathbb{P}(B|A), \qquad \text{since} \qquad \frac{\mathbb{P}(AB)}{\mathbb{P}(B)} \neq \frac{\mathbb{P}(AB)}{\mathbb{P}(A)}.$$

  whenever $\mathbb{P}(A) \neq \mathbb{P}(B)$.

- The "left side" of conditional probability $\mathbb{P}(\bullet|B)$ is again a probability.
  - $\mathbb{P}(A|B) \geq 0$,
  - $\mathbb{P}(\Omega|B) = 1$,
  - $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i|B) = \sum_{i=1}^{\infty} \mathbb{P}(A_i|B)$ if A_i's are disjoint.

- One can define further conditional probability built on a conditional probability. Let

$$\mathbb{P}^*(\bullet) = \mathbb{P}(\bullet|B), \qquad \mathbb{P}(B) > 0.$$

  If $\mathbb{P}(C|B) > 0$, then

$$\mathbb{P}^*(A|C) = \frac{\mathbb{P}^*(AC)}{\mathbb{P}^*(C)}\ .$$

Proof. Given B, the probability of A given C is the probability of A given $\{B$ and $C\}$.

$$\begin{aligned}\mathbb{P}^*(A|C) &= \mathbb{P}(A|C|B) = \mathbb{P}(A|CB) = \frac{\mathbb{P}(ACB)}{\mathbb{P}(CB)} \\ &= \frac{\mathbb{P}(AC|B)\mathbb{P}(B)}{\mathbb{P}(C|B)\mathbb{P}(B)} = \frac{\mathbb{P}(AC|B)}{\mathbb{P}(C|B)} \\ &= \frac{\mathbb{P}^*(AC)}{\mathbb{P}^*(C)}\end{aligned}$$

□

- The "right side" of conditional probability $\mathbb{P}(A|\bullet)$ is not a probability. Generally,

$$\mathbb{P}(A|B \cup C) \neq \mathbb{P}(A|B) + \mathbb{P}(A|C)$$

since

$$\frac{\mathbb{P}((AB) \cup (AC))}{\mathbb{P}(B \cup C)} \neq \frac{\mathbb{P}(AB)}{\mathbb{P}(B)} + \frac{\mathbb{P}(AC)}{\mathbb{P}(C)}.$$

1.6 Bayes' Theorem

The law of total probability

If $\{A_1, \cdots, A_k\}$ is a partition of the sample space, then for any event B,

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i).$$

Proof. By definition, $\bigcup_{i=1}^{\infty} A_i = \Omega$, $A_i \cap A_j = \emptyset$ for $i \neq j$. Then $BA_i \cap BA_j = \emptyset$ for $i \neq j$.

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(B \cap \Omega) \\ &= \mathbb{P}\left(B \cap \left(\bigcup_{i=1}^{k} A_i\right)\right) = \mathbb{P}\left(\bigcup_{i=1}^{k} BA_i\right) \\ &= \sum_{i=1}^{k} \mathbb{P}(BA_i) = \sum_{i=1}^{k} \mathbb{P}(B|A_i)P(A_i).\end{aligned}$$

□

Bayes' Theorem

Let $\{A_1, \cdots, A_k\}$ be a partition of the sample space, $\mathbb{P}(A_i) > 0$ for $i = 1, \cdots, k$.
If $\mathbb{P}(B) > 0$, then for any $j = 1, \cdots, k$,

$$\mathbb{P}(A_j|B) = \frac{P(B|A_j)\mathbb{P}(A_j)}{\sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}.$$

Proof.

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(A_jB)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}$$

by the definition of conditional probability and the law of total probability.

□

Notes: In the Bayesian context, as in the example below,

- $\mathbb{P}(A)$ is the *prior* probability of A,
- $\mathbb{P}(A|B)$ is the *posterior* probability of A.

Example (on conditional probability)

A disease (e.g. caused by coronavirus) is present in one percent of people in a population. Denote D as the event that the disease is present in a typical individual in the population. A medical test is developed to identify the presence of the disease, with test results as positive or negative. The probabilities are:

Test outcome	D (present)	D^c (absent)
$+$	$\mathbb{P}(D \text{ and } +) = .009$	$\mathbb{P}(D^c \text{ and } +) = .036$
$-$	$\mathbb{P}(D \text{ and } -) = .001$	$\mathbb{P}(D^c \text{ and } -) = .954$

For example, 95.4% of the population do not have the disease and have test result negative.
Is the test a good one? There are a few relevant measures.

- The *prevalence* of the disease is

$$\mathbb{P}(D) = \mathbb{P}(D \text{ and } +) + \mathbb{P}(D \text{ and } -) = .009 + .001 = 1.0\%.$$

Therefore,

$$\mathbb{P}(D^c) = 1 - \mathbb{P}(D) = .990 = 99.0\%.$$

- The *sensitivity* of the test seems good:

$$\mathbb{P}(\ +\ |D) = \frac{\mathbb{P}(+ \bigcap D)}{\mathbb{P}(D)} = \frac{.009}{.01} = .90 = 90.0\%$$

- The *specificity* of the test looks fine:

$$\mathbb{P}(\ -\ |D^c) = \frac{\mathbb{P}(- \bigcap D^c)}{\mathbb{P}(D^c)} = \frac{.954}{.99} \approx 0.964 = 96.4\%$$

- The probability of *Type II error*, or *false negative*, is

$$\mathbb{P}(-|D) = \frac{\mathbb{P}(-\ and\ D)}{\mathbb{P}(D)} = \frac{.001}{.01} = 10.0\%$$

- The probability of *Type I error*, or *false positive*, is

$$\mathbb{P}(+|D^c) = \frac{\mathbb{P}(+\ and\ D^c)}{\mathbb{P}(D^c)} = \frac{.036}{.99} = 3.64\%$$

- If the test is positive, what is the chance that you do have the disease? This is the rate of *positive predictivity* $\mathbb{P}(D|\ +\ )$. Let's practice using Bayes' formula and write out every term.

$$\begin{aligned}\mathbb{P}(D|\ +\ ) &= \frac{\mathbb{P}(+|D)\mathbb{P}(D)}{\mathbb{P}(+|D)\mathbb{P}(D) + \mathbb{P}(+|D^c)\mathbb{P}(D^c)} \\ &= \frac{\mathbb{P}(+|D)\mathbb{P}(D)}{\mathbb{P}(+|D)\mathbb{P}(D) + (1 - \mathbb{P}(-|D^c))\mathbb{P}(D^c)} \\ &= \frac{.9 \times .01}{.9 \times .01 + (1 - .964) \times .99} \approx .20 = 20\%\end{aligned}$$

One should not get upset right away when the test result is positive: only 20% of the tested-positive individuals actually have the disease. In other words, the proportion of alarms that are in error, $P(D^c|\ +\ )$, is quite high:

$$\mathbb{P}(D^c|\ +\ ) = 1 - \mathbb{P}(D|\ +\ ) \approx 80\%.$$

- On the other hand, the *negative predictivity* is reassuring:

$$\mathbb{P}(D^c|\ -\ ) = \frac{.954}{.001 + .954} = 99.9\%$$

Remarks

- $\mathbb{P}(D)$ is our prior knowledge of the disease D.
- $\mathbb{P}(D|+)$ is our posterior knowledge of the disease D, after the event "test positive" already happened. The conditional probability answers the following question: what is the probability of an individual having the disease D, knowing that the test result is positive?
- The seemingly paradoxical numbers in the example are largely due to the small probability of the event D.

2 Random Variables

In scientific studies, often the numerical data produced by mechanisms or processes with uncertainty factors are measured or observed. We relate data to the original sample spaces by random variables.

Definition

A random variable is a function from a sample space (the domain of the random variable) to the real line (the range of the random variable) that assigns a real number to each outcome.

$$\begin{aligned} X:\ \Omega &\to \mathbb{R} \\ \omega &\to X(\omega) = x \in \mathbb{R} \end{aligned}$$

Remarks

We often work directly with the range of random variables. However, the sample space is always in the background.

The probability distribution of a random variable is induced by the probability distribution in the sample space:

$$\mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}\{\omega \in \Omega:\ X(\omega) \in A\}$$

For example,

$$\mathbb{P}\{X \in [a,b]\} = \mathbb{P}\{\omega \in \Omega:\ X(\omega) \in [a,b]\}$$
$$\mathbb{P}(X = x) = \mathbb{P}\{\omega \in \Omega:\ X(\omega) = x\}$$

A more specific example: Tossing a die with six labeled faces. The sample space is

$$\Omega = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}\} = \{\omega:\ \omega = 1,2,3,4,5,6\}$$

Define a random variable

$$X(\omega) = \begin{cases} 1, & \text{if the outcome } \omega \text{ is even;} \\ 0, & \text{if the outcome } \omega \text{ is odd.} \end{cases}$$

Then the probability of events in the sample space induces the distribution of X.

$$\mathbb{P}\{X=1\} = \mathbb{P}\{\{2\} \cup \{4\} \cup \{6\}\} = \mathbb{P}\{\{\omega=2\} \cup \{\omega=4\} \cup \{\omega=6\}\}$$
$$\mathbb{P}\{X=0\} = \mathbb{P}\{\{1\} \cup \{3\} \cup \{5\}\} = \mathbb{P}\{\{\omega=1\} \cup \{\omega=3\} \cup \{\omega=5\}\}$$

Common notations

$X, Y, Z, \cdots$ denote random variables.

$x, y, z, \cdots$ denote particular values of the random variables.

$p, \lambda, \mu, \sigma, \alpha, \beta, \cdots$ denote parameters.

Cumulative distribution function (CDF)

The cumulative distribution function (CDF or cdf) of a random variable X is defined as

$$F(x) = F_X(x) = \mathbb{P}\{X \le x\}.$$

Properties:

- A CDF is a probability: $F:\ \mathbb{R} \to [0,1]$.

- A CDF is non-decreasing: $F(x_1) \le F(x_2)$ if $x_1 < x_2$.
- A CDF is right continuous: $F(x^+) = F(x)$, where $F(x^+) = \lim_{y \to x^+} F(y) = \lim_{y \to x, y > x} F(y)$.
- A CDF is defined for all real numbers.
- There is always

$$\lim_{x \to -\infty} F(x) = 0, \qquad \lim_{x \to \infty} F(x) = 1$$

This is because, based on the non-decreasing property,

$$\begin{aligned} \mathbb{P}\{X < \infty\} = 1 &\implies F(x) \to 1 \quad as \quad x \to \infty, \\ \mathbb{P}\{X < -\infty\} = 0 &\implies F(x) \to 0 \quad as \quad x \to -\infty, \end{aligned}$$

- A CDF contains all the information about the probability distribution of the random variable. In other words,

$$\begin{aligned} \text{If} \quad & F_X(t) = F_Y(t), \quad \text{for all } t \in \mathbb{R}, \\ \text{then} \quad & \mathbb{P}(X \in (a,b)) = \mathbb{P}(Y \in (a,b)), \quad \text{for all } a, b \in \mathbb{R}. \end{aligned}$$

- Note that $F_X = F_Y$ does not imply $X \equiv Y$.
  Example: X is the outcome value of tossing a die, $Y = 7 - X$.
- Given a function $F: \mathbb{R} \to [0,1]$, F can be viewed as a CDF with respect to some probability $\mathbb{P}$ if

$$F \text{ is right-continuous, non-decreasing, } \lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to \infty} F(x) = 1.$$

- Notations: F is conventionally used as CDF. $X \sim F$ means X has distribution F.

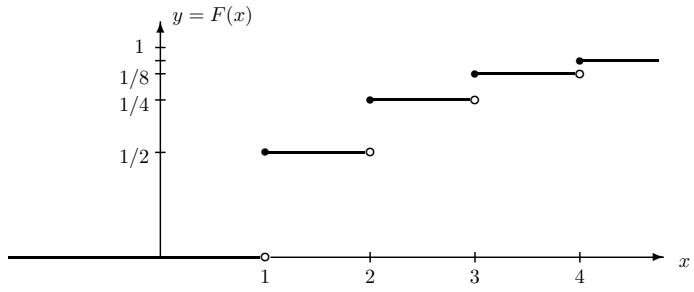
Example (Discrete probability distribution)

Define X as the number of coins tossed until the first head appears. If $p = \mathbb{P}\{\text{head}\} = \frac{1}{2}$, then

$$\mathbb{P}(X = k) = (1-p)^{k-1}p, \quad k = 1, 2, \cdots,$$

This probability distribution is called a geometric distribution with parameter $p = 1/2$.

Cumulative distribution functions (CDFs) of discrete random variables are step functions, as illustrated in the CDF of geometric distribution in the following graph (up to $x < 5$).



Remarks

- we can evaluate a cumulative distribution function at any point.
  In the example of geometric distribution, $F(1.2) = 1/2$, even though $\mathbb{P}(X = 1.2) = 0$.
- The above $F(x)$ is differentiable everywhere except at $x = 1, 2, \cdots$.
- The probability at a single point $\mathbb{P}(X = k)$ can be recovered from CDF by

$$\mathbb{P}(X = k) = \mathbb{P}(X \le k) - \mathbb{P}(X < k) = F(k) - F(k^-)$$

  where

$$F(k^-) = \lim_{x \to k^-} F(x).$$

  At $k = 1, 2, \cdots$, the probability at the single point for this random variable can be simplified to

$$\mathbb{P}(X = k) = F(k) - F(k-1).$$

- For convenience, sometimes the discontinuous points of a cumulative function are connected by vertical lines in the graph.

2.1 Discrete random variables

Definition

A random variable is discrete if it takes finite or countably many values.

$$X : \ \Omega \to \{x_k; \ k = 1, 2, \cdots\} \subset \mathbb{R}.$$

Often the possible values of a discrete random variable are chosen as non-negative integers. Then it becomes

$$X : \ \Omega \to \{0, 1, 2, \cdots\} = \mathbb{N} \subset \mathbb{R}.$$

Probability functions

For a discrete random variable X, the set of values

$$p_k = p(x_k) = \mathbb{P}\{X = x_k\}, \qquad with \quad \sum_{x_k} \mathbb{P}\{X = x_k\} = 1$$

is the *probability mass function* of X. Its CDF can be written as

$$F(x) = \mathbb{P}\{X \le x\} = \sum_{x_k \le x} \mathbb{P}\{X = x_k\}$$

The CDF F is a right continuous step function, as depicted in the example of geometric distribution.

In particular, if $x_1 < x_2 < x_3 < \cdots$,

$$\mathbb{P}\{X = x_k\} = \mathbb{P}\{X \le x_k\} - \mathbb{P}\{X \le x_{k-1}\} = F(x_k) - F(x_{k-1}), \quad k = 1, 2, \cdots .$$

Important discrete random variables

Discrete uniform distribution

A random variable X is of uniform distribution on $\{x_1, \cdots, x_N\}$ if

$$\mathbb{P}(X = x_k) = \begin{cases} \frac{1}{N}, & k = 1, 2, \cdots, N, \\ 0, & otherwise. \end{cases}$$

Point mass distribution

A random variable X is of point mass distribution at $x_o \in \mathbb{R}$ if

$$\mathbb{P}(X = x) = \begin{cases} 1, & x = x_o, \\ 0, & x \ne x_o. \end{cases}$$

The cumulative function is

$$F(x) = \begin{cases} 0, & x < x_o, \\ 1, & x \ge x_o. \end{cases}$$

Bernoulli distribution

Many experiments, especially pilot studies, often consist of just two possible outcomes, called success vs. failure, or true vs. false, or accept vs. reject.

A random variable X is Bernoulli distribution with parameter $p \in [0, 1]$ if

$$X = \begin{cases} 1 \text{ (i.e., success)}, & \text{with probability } p \ ; \\ 0 \text{ (i.e., failure)}, & \text{with probability } 1 - p \ . \end{cases}$$

Remarks. Do not confuse Bernoulli distribution with a point mass distribution.

For a Bernoulli random variable, the probability mass function is

$$\mathbb{P}(X = x) = \begin{cases} p, & if \ x = 1; \\ 1 - p, & if \ x = 0; \\ 0, & otherwise. \end{cases}$$

The cumulative distribution function of the Bernoulli random variable is

$$F(x) = \begin{cases} 0, & x < 0; \\ 1 - p, & 0 \le x < 1; \\ 1, & x \ge 1. \end{cases}$$

A toss of a fair coin corresponds to Bernoulli distribution with $p = 1/2$.

Binomial distribution

A random variable is of Binomial distribution with parameters $n = 1, 2, \cdots$ and $p \in [0, 1]$, if

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 1, 2, \cdots, n.$$

Notation: $X \sim Bin(n, p)$.

Remarks

- The notation
$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}.$$
- For non-negative integers, the factorial function
$$k! = k \times (k-1) \times \cdots \times 2 \times 1, \qquad 1! = 1, \quad 0! = 1.$$
- A Binomial random variable $X \sim Bin(n,p)$ is the sum of n independent Bernoulli random variables with common parameter p.
- The Mode of binomial distribution
$\mathbb{P}(X=k)$ increases for $k \le [(n+1)p]$ and decreases for $k \ge [(n+1)p]$, because
$$\frac{\mathbb{P}(X=k)}{\mathbb{P}(X=k-1)} = \frac{\binom{n}{k}p^k(1-p)^{n-k}}{\binom{n}{k-1}p^{k-}(1-p)^{n-k+1}} = \frac{(n-k+1)}{k}\frac{p}{1-p} \begin{cases} >1, & k<(n+1)p \\ =1, & k=(n+1)p \\ <1, & k>(n+1)p \end{cases}$$
- An example (applications of binomial random variables)

  A type of airplane engine has probability p to be operative throughout a flight. Assume that an airplane will make a successful flight if at least fifty percent of its engines remain operative. For what values of p is a four-engine plane preferable to a two-engine plane?

  It is reasonable to assume that the performance of engines are independent of each other. Consequently, during a flight,
$$\begin{aligned} \text{Number of operative engines on a 4-engine plane} &\sim Bin(4,p), \\ \text{Number of operative engines on a 2-engine plane} &\sim Bin(2,p). \end{aligned}$$

  For a four-engine plane,
$$\begin{aligned} & \mathbb{P}(\text{a 4-engine plane makes a successful flight}) \\ = \ & \mathbb{P}(\text{at least 2 engines are operative}) \\ = \ & \mathbb{P}(\text{4 engines operative}) + \mathbb{P}(\text{3 engines operative}) + \mathbb{P}(\text{2 engines operative}) \\ = \ & \binom{4}{4}p^4 + \binom{4}{3}p^3(1-p) + \binom{4}{2}p^2(1-p)^2 \\ = \ & p^4 + 4p^3(1-p) + 6p^2(1-p)^2 \end{aligned}$$

  While for a two-engine plane,
$$\begin{aligned} & \mathbb{P}(\text{a 2-engine plane makes a successful flight}) \\ = \ & \mathbb{P}(\text{at least 1 engine is operative}) \\ = \ & \mathbb{P}(\text{2 engines operative}) + \mathbb{P}(\text{1 engine operative}) \\ = \ & \binom{2}{2}p^2 + \binom{2}{1}p(1-p) = p^2 + 2p(1-p) \end{aligned}$$

  A four-engine plane is preferable if
$$\mathbb{P}(\text{4-engine plane makes a successful flight}) > \mathbb{P}(\text{2-engine plane makes a successful flight})$$

That is, when
$$p^4 + 4p^3(1-p) + 6p^2(1-p)^2 \ > \ p^2 + 2p(1-p).$$
Equivalently, when
$$g(p) = 3p^4 - 8p^3 + 7p^2 - 2p > 0.$$

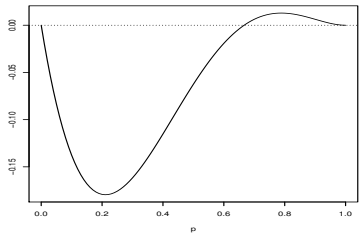


Figure 1: Plot of $g(p)$

The graph indicates that, when the reliability of engines is high ($p > 67\%$), the four-engine plane is preferable ($g(p) > 0$). In fact,
$$g(p) = 3p^4 - 8p^3 + 7p^2 - 2p = p(p-1)^2(3p-2).$$
Consequently,
$$g(p) > 0 \quad when \quad p > \frac{2}{3} \approx 67\%.$$
Are there other conclusions you can get from the graph?

Geometric distribution

A random variable X is of geometric distribution with parameter $p \in [0,1]$ if
$$\mathbb{P}(X=k) = (1-p)^{k-1}p, \quad k = 1, 2, \cdots$$

Example: The probability for you to win a chess game against your friend is p. Let X be the number of games needed to play until you win a game, then X is of geometric distribution with parameter p.

Negative binomial distribution

A random variable X is of negative binomial with parameter (r,p) if
$$\mathbb{P}(X=k) = \binom{k-1}{r-1}p^r(1-p)^{k-r}, \qquad k = r, r+1, r+2, \cdots$$

Example: If the probability for you to win a chess game against your friend is p, X is the number of games needed to play until you win r games, then X is of negative binomial with parameters (r,p).

Hypergeometric distribution

A random variable X is of hypergeometric distribution with parameters (M, N, K) if

$$\mathbb{P}(X = k|N, M, K) = \frac{\binom{M}{k}\binom{N-M}{K-k}}{\binom{N}{K}}, \qquad \max\{0, K-(N-M)\} \le k \le \min\{K, M\}.$$

Hypergeometric distribution is often described by the urn-ball model:
There are N balls in an urn, M are blue and $N-M$ are red. Randomly select K balls from the urn. The number of blue balls selected is of hypergeometric distribution.

Poisson distribution

A random variable $X \sim \text{Poi}(\lambda)$ is of Poisson distribution with parameter $\lambda \in (0, \infty)$ with

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}, \quad k = 0, 1, 2, \cdots$$

Poisson distribution is often used as a model for counts of rare events, such as radioactive decay or traffic accidents.

Remarks.

- We can check that $\sum_k \mathbb{P}(X = k) = 1$ holds for each of the probability distributions.

  The following are a few examples.
  - For geometric distribution,

    $$\sum_k \mathbb{P}(X = k) = \sum_{k=1}^{\infty}(1-p)^{k-1}p = \frac{1}{1-(1-p)}p = 1,$$

    where we used the expression for sums of the series

    $$a + a^2 + a^3 + \cdots = \frac{a}{1-a} \qquad for \quad |a| < 1.$$

  - For binomial distribution,

    $$\sum_k P(X = k) = \sum_{k=1}^{n}\binom{n}{k}p^k(1-p)^{n-k} = (p + (1-p))^n = 1,$$

    where we used

    $$(a+b)^n = \sum_{k=1}^{n}\binom{n}{k}a^k b^{n-k}, \qquad a, b \in \mathbb{R}.$$

  - For Poisson distribution,

    $$\sum_k P(X = k) = \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda} = e^{\lambda}\, e^{-\lambda} = 1,$$

    where we applied the Taylor expansion formula

    $$e^x = \sum_{k=0}^{\infty}\frac{x^k}{k!} \qquad \text{for} \quad x \in \mathbb{R}.$$

- A sample space can be constructed as outcomes of an experiment for each of the above distributions.

2.2 Continuous random variables

For a continuous random variable,

$$\mathbb{P}(X = x) = 0 \qquad \text{for any} \quad x \in \mathbb{R}.$$

So we have to work with cumulative distribution functions and their derivatives, the density functions.

Definition.

A random variable is continuous if there is a **probability density function** (PDF) $f(x) = f_X(x)$ such that,

- $f(x) \ge 0$ for all $x \in \mathbb{R}$,
- $\int_{-\infty}^{\infty} f(x)dx = 1$,
- $\mathbb{P}\{a < X < b\} = \int_a^b f(x)dx$.

Properties.

- The relationship between PDF $f(x)$ and CDF $F(x)$.

  From the definition of cumulative function,

  $$\mathbb{P}\{a < X \le b\} = \mathbb{P}\{X \le b\} - \mathbb{P}\{X \le a\} = F(b) - F(a).$$

  From the definition of derivatives in calculus, when $F'(x)$ exists,

  $$\frac{\mathbb{P}\{x < X \le x + \Delta x\}}{\Delta x} = \frac{F(x+\Delta x) - F(x)}{\Delta x} \longrightarrow F'(x), \quad as\ \Delta x \to 0.$$

  The above fact and

  $$\mathbb{P}\{a < X \le b\} = \mathbb{P}\{a < X < b\} = F(b) - F(a) = \int_a^b f(x)dx$$

  imply that, the density function

  $$f(x) = F'(x) \qquad \text{whenever } F'(x) \text{ exists},$$

  and

  $$\mathbb{P}\{X \le t\} = F(t) = \int_{-\infty}^{t} f(x)dx.$$

- The density function $f(x) \ne \mathbb{P}(X = x)$.
- $f(x)$ is not a probability. It is possible that $f(x) > 1$. In fact $f(x)$ can be unbounded. For example,

  $$f(x) = \begin{cases} \dfrac{1}{2\sqrt{x}}, & 0 < x < 1, \\ 0, & otherwise. \end{cases}$$

  Then $f(x) > 0$, $\int_{-\infty}^{\infty} f(x)dx = \int_0^1 \frac{1}{2\sqrt{x}}dx = 1$. But $f(x)$ is unbounded near $x = 0$.

Important continuous distributions.

Uniform distribution on interval (a, b): For $X \sim U(a, b)$,

$$f(x) = \begin{cases} 0, & x < a; \\ \dfrac{1}{b-a}, & a \le x \le b; \\ 0, & x > b. \end{cases} \qquad F(x) = \begin{cases} 0, & x < a; \\ \dfrac{x-a}{b-a}, & a \le x \le b; \\ 1, & x > b. \end{cases}$$

Exponential distribution with rate parameter $\lambda > 0$: For $X \sim Exp(\lambda)$,

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0; \\ 0, & x < 0. \end{cases} \qquad F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \ge 0; \\ 0, & x < 0. \end{cases}$$

Another common parametrization for exponential distribution is to use the mean value $\beta = 1/\lambda$ as the parameter. The density function then becomes

$$f(x) = \begin{cases} \dfrac{1}{\beta}\, e^{-x/\beta}, & x \ge 0; \\ 0, & x < 0. \end{cases}$$

The definition of the parameter used has to be checked from the context.

Exponential distribution is frequently used to model lifespans and arrival times between rare events.

Gamma distribution with parameters $n > 0,\ \lambda > 0$: For $X \sim Gamma(n, \lambda)$,

$$f(x) = \begin{cases} \dfrac{(\lambda x)^{n-1}}{\Gamma(n)} \lambda e^{-\lambda x}, & x \ge 0; \\ 0, & x < 0, \end{cases}$$

where the parameter n is not restricted to integers.

The Gamma function $\Gamma(n)$ is defined as

$$\Gamma(n) = \int_0^\infty x^{n-1} e^{-x} dx.$$

Thus $\Gamma(1) = 1$. Integration by parts gives the property

$$\Gamma(n) = (n-1)\,\Gamma(n-1).$$

Therefore Gamma function has a simpler form when n is a positive integer:

$$\Gamma(n) = (n-1)! \quad \text{for} \quad n = 1, 2, \cdots$$

commonly we define $0! = 1$.

Note that $Exp(\lambda) \equiv Gamma(1, \lambda)$. Gamma density function in the above form hints the relationship between exponential distribution and Gamma distribution: the sum of independent, identically distributed (*i.i.d.*) exponential random variables has a Gamma distribution.

Since $n > 0$ is not restricted to integers, Gamma distribution parameters are often denoted as (α, λ) instead of (n, λ). Another parametrization for Gamma distribution is $X \sim Gamma(\alpha, \beta)$, with $\alpha = n,\ \beta = 1/\lambda$. The Gamma density function then becomes

$$f(x) = \begin{cases} \dfrac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}, & x \ge 0; \\ 0, & x < 0. \end{cases}$$

Again, the definition of the parameters used has to be checked from the context.

Beta distribution with parameters $\alpha > 0,\ \beta > 0$: For $X \sim Beta(\alpha, \beta)$,

$$f(x) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & x \in (0,1); \\ 0, & otherwise. \end{cases}$$

Notice that $Gamma(1, 1)$ is the uniform distribution on $(0, 1)$.

A derivation of Beta distribution.

In the following we give an interesting derivation of the property

$$\int_0^1 f(x) dx = 1$$

for the density function of Beta distribution.

Let $X_1, \cdots, X_{n+m}, Y$ be *i.i.d.* $\sim Unif(0, 1)$.
Denote $X_{(1)}, \cdots, X_{(n+m)}$ as the order statistics of the X_i's,

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n+m)}.$$

Define

$$A = \{X_{(1)} \le \cdots \le X_{(m)} \le Y \le X_{(m+1)} \le \cdots \le X_{(n+m)}\}$$

Then by the *i.i.d.* nature of $X_1, \cdots, X_{n+m}, Y$, the probability that Y falls between two adjacent order statistics $X_{(i)} \le Y \le X_{(i+1)}$ should be equal for any $i = 0, 1, 2, \cdots, n$ (with $X_{(0)} = -\infty,\ X_{(n+m+1)} = \infty$). Therefore,

$$\mathbb{P}(A) = \frac{1}{m+n+1}$$

For given $Y = y,\ y \in [0, 1]$, each X_i has the same probability to fall to the left of y, thus

$$\text{the number of } X_i \le y \ \sim \ Bin(n+m, y)$$

In particular,

$$\mathbb{P}(A|Y = y) = \mathbb{P}(m\ X_i's \le y) = \binom{n+m}{m} y^m (1-y)^n, \quad \forall y \in [0, 1].$$

Combining with the previous result yields

$$\frac{1}{m+n+1} = \mathbb{P}(A) = \int_0^1 \mathbb{P}(A|Y = y) dy = \int_0^1 \binom{n+m}{m} y^m (1-y)^n dy$$

Thus

$$(n+m+1) \int_0^1 \binom{n+m}{m} y^m (1-y)^n dy = 1$$

Note that

$$(n+m+1) \int_0^1 \binom{n+m}{m} y^m (1-y)^n dy = \int_0^1 \frac{(n+m+1)!}{n!\ m!} y^m (1-y)^n dy$$

therefore, let $= \alpha - 1, n = \beta - 1$, we have

$$\int_0^1 \frac{(\alpha+\beta-1)!}{(\alpha-1)!\ (\beta-1)!} x^{\alpha-1}(1-x)^{\beta-1} dx = \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} dx = 1$$

The integral

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$$

is call the Beta function.

Note that there are several ways to write the Beta function,

$$B(n,m)=\frac{\Gamma(n)\Gamma(m)}{\Gamma(n+m)}=\frac{(n-1)!(m-1)!}{(n+m-1)(n+m-2)!}=\frac{1}{(n+m-1)\binom{n+m-2}{n-1}}$$

Normal (Gaussian) distribution with parameters $\mu,\sigma>0$:

A random variable $X\sim N(\mu,\sigma^2)$ is of normal distribution with parameters (μ,σ^2) if X has density function

$$f(x)=\frac{1}{\sqrt{2\pi}\ \sigma}\ e^{-(x-\mu)^2/2\sigma^2},\qquad x\in\mathbb{R}$$

and CDF

$$F(x)=\frac{1}{\sqrt{2\pi}\ \sigma}\int_\infty^x e^{-(t-\mu)^2/2\sigma^2}dt\qquad x\in\mathbb{R}.$$

- Normal distribution is extremely important in probability and statistics.
- Many natural phenomena approximate Normal distributions.
- The Central Limit Theorem says that sums of random variables can be approximated by normal distributions, under moderate regularity conditions.
- The case $\mu=0,\ \sigma=1$ is called *standard normal*.
- Conventionally, standard normal random variable is denoted by Z, standard normal density function and cumulative function are denoted by φ and Φ respectively.

$$Z\sim N(0,1),\quad \varphi(z)=\frac{1}{\sqrt{2\pi}}\ e^{-z^2/2},\quad \Phi(z)=\frac{1}{\sqrt{2\pi}}\int_\infty^z e^{-x^2/2}dx.$$

- On the web and in most statistical texts, a table of values of $\Phi(z)=P(Z\le z)$ is available. For a general Normal random variable $X\sim N(\mu,\sigma)$, the values of its CDF $F(x)$ can be obtained from a table of the standard normal by the following relation:

$$X\sim N(\mu,\sigma)\quad\Longrightarrow\quad\frac{X-\mu}{\sigma}=Z\sim N(0,1).$$

2.3 Joint distributions

We consider joint probability distribution functions for two random variables

$$(X,Y):\quad \Omega\to\mathbb{R}\times\mathbb{R}=\mathbb{R}^2$$

or n random variables $X_1,X_2,\cdots,X_n$, each $\Omega\to\mathbb{R}$.

The joint CDF of two random variables X and Y is defined as

$$F(x,y)=F_{X,Y}(x,y)=P\{X\le x,Y\le y\}=\mathbb{P}\{X\le x\ \textbf{and}\ Y\le y\}.$$

Similarly, for n random variables,

$$F(x_1,\cdots,x_n)=F_{X_1,\cdots,X_n}(x_1,\cdots,x_n)=\mathbb{P}\{X_1\le x_1,\cdots,X_n\le x_n\}.$$

2.3.1 Discrete joint distributions

- The joint probability (mass) function of X and Y is given by

$$p(x,y)=\mathbb{P}\{X=x,Y=y\}=\mathbb{P}\{X=x\ \textbf{and}\ Y=y\}$$

with

$$\sum_{x,y}p(x,y)=\sum_x\sum_y p(x,y)=\sum_y\sum_x p(x,y)=1.$$

The joint distribution of $X_1,X_2,\cdots,X_n$ is given by

$$p(x_1,x_2,\ldots,x_n)==\sum_x\sum_y p(x,y)P\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}$$

with

$$\sum_{x_1,\cdots,x_n}p(x_1,x_2,\ldots,x_n)=1.$$

- Marginal distributions (from joint distributions).

$$\mathbb{P}\{X=x\}=\sum_y\mathbb{P}\{X=x,Y=y\}\qquad\text{— marginal probability mass function for } X.$$
$$\mathbb{P}\{Y=y\}=\sum_x\mathbb{P}\{X=x,Y=y\}\qquad\text{— marginal probability mass function for } Y.$$

Marginal distributions for n dimensional case can be similarly defined.

For example, if the random variables are X_i: $\Omega\to\mathbb{N}=\{0,1,2,\cdots\}$ for $i=1,\cdots,n$, then

$$\begin{aligned}&\mathbb{P}\{X_1=7,X_2\in\mathbb{N},\cdots,X_n\in\mathbb{N}\}\\ =\ &\mathbb{P}\left\{\bigcup_{k_2,\cdots,k_n\in\mathbb{N}}(X_1=7,X_2=k_2,\cdots,X_n=k_n)\right\}\\ =\ &\sum_{k_2,\cdots,k_n\in\mathbb{N}}\mathbb{P}\{\ X_1=7,X_2=k_2,\cdots,X_n=k_n\ \}=\mathbb{P}\{X_1=7\}\end{aligned}$$

which is the marginal probability mass function for X_1 evaluated at $X_1=7$.

$$\begin{aligned}&P\{X_1=7,X_2=11,X_3\in\mathbb{N},\cdots,X_n\in\mathbb{N}\}\\ =\ &\sum_{k_3,\cdots,k_n\in\mathbb{N}}P\{\ X_1=7,X_2=11,X_3=k_3,\cdots,X_n=k_n\ \}\\ =\ &P\{X_1=7,X_2=11\}\end{aligned}$$

which is the marginal probability mass function for (X_1,X_2) evaluated at $(7,11)$.

- Sum of two discrete random variables

$$P\{X+Y=z\}=\sum_x P\{X=x,Y=z-y\}=\sum_y P\{X=z-y,Y=y\}$$

2.3.2 Continuous joint distributions

Definitions.

$f(x,y) = f_{X,Y}(x,y)$ is a joint probability density function for random variables X and Y if

- $f(x,y) \geq 0$,
- $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$,
- $\mathbb{P}((X,Y) \in A) = \displaystyle\iint_A f(x,y)dxdy$ for $A \subset \mathbb{R}\times\mathbb{R}$.

Similarly, for joint density $f(x_1,\cdots,x_n) = f_{X_1,\ldots,X_n}(x_1,\cdots,x_n)$,

$$f(x_1,\cdots,x_n) \geq 0, \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1,\cdots,x_n)\ dx_1dx_2\cdots dx_n = 1.$$

Properties.

- The relation between density and cumulative function.

  $$f(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F(x,y)$$

  at all differentiable points of $F(x,y)$, which implies

  $$f(x,y) \approx \frac{P\{x \leq X \leq x+\Delta x,\ y \leq Y \leq y+\Delta y\}}{\Delta x\Delta y}$$

  for small Δx and Δy.

  Similarly,

  $$f_{X_1,\ldots,X_n}(x_1,\cdots,x_n) = \frac{\partial}{\partial x_1}\cdots\frac{\partial}{\partial x_n}F_{X_1,\ldots,X_n}(x_1,\cdots,x_n),$$

  and

  $$f_{X_1,\ldots,X_n}(x_1,\cdots,x_n) \approx \frac{P\{x_1 \leq X_1 \leq x_1+\Delta x_1,\cdots,x_n \leq X_n \leq x_n+\Delta x_n\}}{\Delta x_1\cdots\Delta x_n}.$$

- Marginal distributions.

  $$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \quad \text{— marginal density function for } X,$$

  $$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx \quad \text{— marginal density function for } Y.$$

  Marginal distribution for higher dimensions are defined similarly.

- Probability on a subset.

  $$P\{X \in (a,b), Y \in (c,d)\} = \int_a^b\int_c^d f_{X,Y}(x,y)dxdy.$$

$$P\{X+Y \leq z\} = \int_{-\infty}^{z}\left(\int_{-\infty}^{z-y} f_{X,Y}(x,y)dx\right)dy = \int_{-\infty}^{z}\left(\int_{-\infty}^{z-x} f_{X,Y}(x,y)dy\right)dx.$$

2.3.3 Independence

Definition.

Two random variables X and Y are independent if

$$\mathbb{P}\{X \in A, Y \in B\} = \mathbb{P}\{X \in A\}\mathbb{P}\{X \in B\}, \qquad \textit{for all } A, B.$$

Equivalently, X and Y are independent if

$$F_{X,Y}(a,b) = \mathbb{P}(X \leq a, Y \leq b) = \mathbb{P}(X \leq a)\mathbb{P}(Y \leq b) = F_X(a)F_Y(b)$$

for all a and b. Consequently, X and Y are independent if

$$P(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$

for discrete random variables, and

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

for continuous random variables.

Similarly, $X_1, X_2, \cdots, X_n$ are independent if

$$F_{X_1,\ldots,X_n}(x_1,x_2,\cdots,x_n) \quad = \quad F_{X_1}(x_1)f_{X_2}(x_2)\cdots F_{X_n}(x_n) \qquad \textit{for all } x_i \in \mathbb{R},$$

which means

$$\mathbb{P}(X_1 = x_1,\ldots,X_n = x_n) = \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2)\cdots\mathbb{P}(X_n = x_n) \qquad \textit{for all } x_i \in \mathbb{R}$$

for the discrete case, and

$$f_{X_1,\ldots,X_n}(x_1,x_2,\cdots,x_n) \quad = \quad f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n) \qquad \textit{for all } x_i \in \mathbb{R}$$

for the continuous case.

Sums of independent random variables and convolution

If X and Y are independent and discrete,

$$\mathbb{P}\{X+Y=z\} = \sum_x \mathbb{P}\{X=x\}\mathbb{P}\{Y=z-y\} = \sum_y \mathbb{P}\{Y=y\}\mathbb{P}\{X=z-y\}$$

If X and Y are independent and continuous,

$$\begin{aligned}\mathbb{P}\{X+Y \leq z\} &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z-y} f_X(x)dx\right)f_Y(y)dy = \int_{-\infty}^{\infty} F_X(z-y)f_Y(y)dy = F_X * F_Y \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z-x} f_Y(y)dy\right)f_X(x)dx = \int_{-\infty}^{\infty} F_Y(z-x)f_X(x)dx = F_Y * F_X\end{aligned}$$

where the last expressions are using the notation called *convolution*. The density of the sum has the expression

$$\frac{d}{dz}\mathbb{P}\{X+Y \leq z\} = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy = f_X * f_Y(z) = \int_{-\infty}^{\infty} f_Y(z-x)f_X(x)dx = f_Y * f_X(z)$$

2.3.4 Important multivariate distributions

Multinomial — the multivariate version of binomial distribution.

Consider discrete random variables

$$X_i \in \{0, 1, \cdots, n\}, \qquad i = 1, 2, \cdots, k, \qquad \sum_{i=1}^{k} X_k = n,$$

where parameter n is a fixed integer. Let

$$X = (X_1, \cdots, X_k), \quad p = (p_1, \cdots, p_k), \qquad p_i \geq 0, \quad \sum_{i=1}^{k} p_i = 1.$$

We say that

$$X \sim Multinomial(n, p)$$

if the probability function is

$$\mathbb{P}(X_1 = n_1, \cdots, X_k = n_k) = \binom{n}{n_1 \cdots n_k} p_1^{n_1} \cdots p_k^{n_k} = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}$$

where $\sum_{i=1}^{k} n_i = n$.

Properties

- $\mathbb{E}(X_i) = np_i,\ V(X_i) = np_i(1-p_i),\ Cov(X_i, X_j) = -np_ip_j$
- *Multinomial and binomial.*
  If $k = 2$, then
  $$X \sim Multinomial(n, p)$$
  is equivalent to
  $$X_1 \sim Bin(n, p_1), \quad X_2 \sim Bin(n, p_2), \qquad X_1 + X_2 = n.$$

  Proof.

  Let $n_1 + n_2 = n$. Since $X_1 + X_2 = n,\ p_1 + p_2 = 1$,

  $$\begin{aligned} & \mathbb{P}(X_1 = n_1, X_2 = n_2) = \frac{n!}{n_1! n_2!} p_1^{n_1} p_2^{n_2} \\ = & \begin{cases} \dfrac{n!}{n_1!(n-n_1)!} p_1^{n_1} p_2^{n-n_1} = \dbinom{n}{n_1} p_1^{n_1} (1-p_1)^{n-n_1} = \mathbb{P}(X_1 = n_1) \\ \dfrac{n!}{(n-n_2)! n_2!} p_1^{n-n_2} p_2^{n_2} = \dbinom{n}{n_2} p_2^{n_2} (1-p_2)^{n-n_2} = \mathbb{P}(X_2 = n_2). \end{cases} \end{aligned}$$

- *Combinations of elements of multinomial random variables.*
  If
  $$X = (X_1, \cdots, X_{k-1}, X_k) \sim Multinomial(n, p), \quad p = (p_1, \cdots, p_{k-1}, p_k),$$
  then
  $$X' = (X_1, \cdots, X_{k-2}, X^*) \sim Multinomial(n, p'),$$

where

$$X^* = X_{k-1} + X_k, \quad p' = (p_1, \cdots, p_{k-2}, p_{k-1} + p_k).$$

Proof. Let $n = n_1 + \cdots + n_{k-2} + n^*$.

$$\begin{aligned} & \mathbb{P}(X_1 = n_1, \cdots, X_{k-2} = n_{k-2}, X^* = n^*) \\ = & \sum_{m=0}^{n^*} \mathbb{P}(X_1 = n_1, \cdots, X_{k-2} = n_{k-2}, X_{k-1} = n^* - m, X_k = m) \\ = & \sum_{m=0}^{n^*} \frac{n!}{n_1! \cdots n_{k-2}!\ (n^*-m)!\ m!} p_1^{n_1} \cdots p_{k-2}^{n_{k-2}} p_{k-1}^{n^*-m} p_k^m \\ = & \frac{n!}{n_1! \cdots n_{k-2}!\ n^*!} p_1^{n_1} \cdots p_{k-2}^{n_{k-2}} (p_{k-1} + p_k)^{n^*} \\ & \times \sum_{m=0}^{n^*} \frac{n^*!}{(n^*-m)!\ m!} \left(\frac{p_{k-1}}{p_{k-1}+p_k}\right)^{n^*-m} \left(\frac{p_k}{p_{k-1}+p_k}\right)^m \\ = & \frac{n!}{n_1! \cdots n_{k-2}!\ n^*!} p_1^{n_1} \cdots p_{k-2}^{n_{k-2}} (p_{k-1} + p_k)^{n^*} \end{aligned}$$

The above result of combining $X_{k-1} + X_k$ can be applied to any combination of $X_i + X_j$, since the order of X_i's does not matter in multinomial distribution.

- *Marginal distributions of multinomial random variable.*

  If $X \sim Multinomial(n, p)$, then the marginal for X_i is

  $$X_i \sim Bin(n, p_i) \qquad for\ i = 1, 2, \cdots, k.$$

  Proof. Fix any $i \in \{1, 2, \cdots, k\}$. Let

  $$Y_1 = X_i, \quad Y_2 = X_1 + \cdots + X_{i-1} + X_{i+1} + \cdots + X_k.$$

  Then

  $$\mathbb{P}(Y_1 = n_i) \quad \text{if and only if} \quad \mathbb{P}(Y_2 = n - n_i).$$

  Therefore,

  $$\mathbb{P}(X_i = n_i) = \mathbb{P}(Y_1 = n_i) = \mathbb{P}(Y_1 = n_i, Y_2 = n - n_i).$$

  Apply the combination property of multinomial (repeatedly) to

  $$X_1 + \cdots + X_{i-1} + X_{i+1} + \cdots + X_k,$$

  we have

  $$Y = (Y_1, Y_2) = (X_i, X_1 + \cdots + X_{i-1} + X_{i+1} + \cdots + X_k) \sim Multinomial(n, p),$$

  with

  $$Y_1 + Y_2 = n, \quad p = (p_i, p_1 + \cdots + p_{i-1} + p_{i+1} + \cdots, p_k) = (p_i, 1 - p_i).$$

  Therefore Y is multinomial with $k = 2$, and the marginal for $Y_1 = X_i$ is

  $$Y_1 = X_i \sim Bin(n, p_i).$$

Multivariate Normal.

X and Y are bivariate normal if the joint density function is

$$\begin{aligned} f(x,y) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \\ &\quad \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}. \end{aligned}$$

Consequently (after some derivations), the marginal distributions are

$$X \sim N(\mu_x, \sigma_x^2), \quad Y \sim N(\mu_y, \sigma_y^2).$$

If X and Y are independent, then $\rho = 0$, and

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}.$$

$X_1, \cdots, X_k$ are multivariate normal if the joint density function is

$$f(x_1, \cdots, x_k) = f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\},$$

where vector and matrix notations are used. μ and $\mathbf{x}$ both are vectors of length k, $\mathbf{x}^T$ is the transpose of $\mathbf{x}$:

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}, \qquad \mathbf{x}^T = (x_1, \cdots, x_k).$$

Here a Σ is $k \times k$ symmetric, positive definite matrix (which implies $x^T\Sigma x > 0$ for any nonzero vector $x \in \mathbb{R}$) $|\Sigma| = det(\Sigma)$ is the determinant of the matrix σ. Σ^{-1} denotes the inverse matrix of Σ, and $x^T y = x_1y_1 + \cdots + x_ky_k$ is the inner product of vectors x and y.

When $k = 2$, we have the bivariate normal again, with

$$x = x_1, \quad y = x_2, \quad \mu_x = \mu_1, \quad \mu_y = \mu_2, \quad \sigma_x = \sigma_1, \quad \sigma_y = \sigma_2$$

Using the matrix notation and $\sigma_{12} = \rho\sigma_1\sigma_2$ to define ρ, we have

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}, \qquad |\Sigma| = \sigma_x^2\sigma_y^2(1-\rho^2), \qquad \Sigma^{-1} = \frac{1}{|\Sigma|}\begin{bmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{bmatrix}.$$

where the last item is from the inverse matrix expression for 2×2 invertible matrices.

2.4 Functions of Random Variables

Y is a function or a transformation of X if $Y = g(X)$ for some function of X.

- $Y = g(X)$, X discrete. Then

$$\mathbb{P}(Y = y) = \mathbb{P}(g(X) = y).$$

If g is invertible, $X = g^{-1}(Y)$, then

$$\mathbb{P}(Y = y) = \mathbb{P}(g(X) = y) = \mathbb{P}(X = g^{-1}(y)).$$

- $Y = g(X)$, X continuous. Then

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \le y) = \mathbb{P}(g(X) \le y) = \int_{\{x:g(x)\le y\}} f_X(x)dx, \\ f_Y(y) &= F_Y'(y). \end{aligned}$$

If g is invertible, $h(y) = g^{-1}(y)$ exists, then

$$f_Y(y) = f_X(h(y))|h'(y)|, \quad F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt = \int_{-\infty}^{y} f_X(h(t))|h'(t)|dt$$

- $Y_1 = g_1(X_1, X_2),\ Y_2 = g_2(X_1, X_3),\ X_1, X_2$ discrete. Then

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2) = \mathbb{P}(g_1(X_1, X_2) = y_1, g_2(X_1, X_2) = y_2).$$

- $Y_1 = g_1(X_1, X_2),\ Y_2 = g_2(X_1, X_3),\ X_1, X_2$ continuous. Then

$$\begin{aligned} F_{Y_1,Y_2}(y_1, y_2) &= \mathbb{P}(Y_1 \le y_1, Y_2 \le y_2) \\ &= \mathbb{P}(g_1(X_1, X_2) \le y_1, g_2(X_1, X_2) \le y_2) \\ &= \iint_{\{(x_1,x_2):g_1(x_1,x_2)\le y_1, g_2(x_1,x_2)\le y_2\}} f_{X_1,X_2}(x_1, x_2)dx_1dx_2 \end{aligned}$$

At the points that F has continuous second derivatives,

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{\partial}{\partial y_1}\frac{\partial}{\partial y_2}F_{Y_1,Y_2}(y_1, y_2)$$

If there exist functions h_1, h_2 such that $X_1 = h_1(Y_1, Y_2),\ X_2 = h_2(Y_1, Y_2)$, and the Jacobian

$$J(x_1, x_2) = det\begin{bmatrix} \dfrac{\partial g_1}{\partial x_1} & \dfrac{\partial g_1}{\partial x_2} \\ \dfrac{\partial g_2}{\partial x_1} & \dfrac{\partial g_2}{\partial x_2} \end{bmatrix} = \frac{\partial g_1}{\partial x_1}\frac{\partial g_2}{\partial x_2} - \frac{\partial g_1}{\partial x_2}\frac{\partial g_2}{\partial x_1} \neq 0,$$

exists, then the joint density of Y_1, Y_2 can be expressed in terms of the joint density of X_1, X_2 and the absolute value of Jacobian.

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(h_1(y_1, y_2), h_2(y_1, y_2))|J(h_1(y_1, y_2), h_2(y_1, y_2))|^{-1}.$$

Ofter monotone or smoothness conditions hold piecewise with respect to individual variables, then all properties need to be adjusted to apply piecewise or locally.

2.5 Relations among distributions

Bernoulli and binomial.

$X_i, i = 1, 2, \cdots, n$ are independent Bernoulli with parameter p, then the sum

$$S_n = \sum_{i=1}^{n} X_i \sim Bin(n, p).$$

Since $Bin(1, p)$ is Bernoulli with parameter p, the above also implies that the sum of two independent binomial random variables with the same parameter is also binomial.

$$X \sim Bin(n, p),\ Y \sim Bin(m, p),\ X \perp\!\!\!\perp Y \quad \Longrightarrow \quad X + Y \sim Bin(n + m, p)$$

Proof. By the (convolution) formula for sums of independent random variables,

$$\begin{aligned}\mathbb{P}(X+Y=j) &= \sum_k \mathbb{P}(X=k)\,\mathbb{P}(Y=j-k)\\ &= \sum_k \binom{n}{k}p^k(1-p)^{n-k}\binom{m}{j-k}p^{j-k}(1-p)^{m-(j-k)}\\ &= \sum_k \binom{n}{k}\binom{m}{j-k}p^j(1-p)^{n+m-j} \;=\; \binom{n+m}{j}p^j(1-p)^{n-j}\end{aligned}$$

In the last equality we used the formula

$$\sum_k \binom{n}{k}\binom{m}{j-k} = \binom{n+m}{j}$$

where the sum is over all possible k (consider choosing j kids from n girls and m boys). Therefore $X+Y \sim Bin(n+m,p)$.

□

Bernoulli and geometric.

$X_i, i=1,2,\cdots$ are a sequence of independent Bernoulli with parameter p, then

$$Y = \min\{k : X_1 = \cdots = X_{k-1} = 0, X_k = 1\} \sim Geometric(p).$$

Binomial and hypergeometric.

$X \sim Bin(n,p), Y \sim Bin(m,p)$ are independent, then the conditional distribution of X or Y given the sum is of hypergeometric distribution.

$$X \mid X+Y=j \quad \sim Hypergeometric(n, n+m, j)$$

$$\mathbb{P}(X_1 = k \mid X_1+X_2=j) = \frac{\binom{n}{k}\binom{m}{j-k}}{\binom{n+m}{j}}, \qquad \max(0, j-m) \le k \le \min(j,n).$$

Poisson and binomial.

Binomial as conditioned Poisson.

$X_i \sim Poi(\lambda_i),\ i=1,2$ are independent, then

$$X_1 \mid X_1+X_2=n \quad \sim Bin\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$$

Poisson as a limit of Binomial.

$$Bin(n,p_n) \to Poi(\lambda) \qquad if \quad np_n \to \lambda \ \ as\ n\to\infty$$

in the sense that, if $X_n \sim Bin(n,p_n), X \sim Poi(\lambda)$, then

$$\mathbb{P}(X_n = k) \to \mathbb{P}(X=k) \qquad \forall k,\ as\ n\to\infty.$$

Proof. Write $\lambda_n = np_n$.

$$\begin{aligned}\mathbb{P}(X_n=k) &= \binom{n}{k}p_n^k(1-p_n)^{n-k} = \frac{n!}{(n-k)!\ k!}\ p_n^k\ (1-p_n)^{n-k}\\ &= \frac{n(n-1)(n-2)\cdots(n-k-1)}{k!}p_n^k(1-p_n)^{n-k}\\ &= \frac{n(n-1)(n-2)\cdots(n-k-1)}{n^k}\frac{n^k}{k!}\left(\frac{\lambda_n}{n}\right)^k\left(1-\frac{\lambda_n}{n}\right)^{n-k}\\ &= \frac{n(n-1)(n-2)\cdots(n-k+1)}{n^k}\ \frac{(\lambda_n)^k}{k!}\left(1-\frac{\lambda_n}{n}\right)^n\left(1-\frac{\lambda_n}{n}\right)^{-k}\\ &= 1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\frac{(\lambda_n)^k}{k!}\left(1-\frac{\lambda_n}{n}\right)^n\left(1-\frac{\lambda_n}{n}\right)^{-k}.\end{aligned}$$

As $n\to\infty,\ \lambda_n = np_n \to \lambda$, so $p_n \to 0$, and

$$1-\frac{1}{n}\to 1, \qquad \left(1-\frac{\lambda_n}{n}\right)^k = (1-p_n)^k \to 1, \qquad \left(1-\frac{\lambda_n}{n}\right)^n \to e^{-\lambda}.$$

Therefore as $n\to\infty$,

$$\mathbb{P}(X_n=k) \quad\longrightarrow\quad 1\times 1\cdots\times 1\times\frac{\lambda^k}{k!}\times e^{-\lambda}\times 1 = \frac{\lambda^k}{k!}e^{-\lambda} = \mathbb{P}(X=k).$$

□

Exponential and Geometric.

Exponential distribution is the continuous cousin of geometric distribution. If $X \sim Exp(\lambda)$, then $Y = [X] \sim Geometri(p)$ with $p = 1-e^{\lambda}$. where $[x]$ denotes the largest integer $\le x$.

On the other hand, exponential distribution is a type of limit of geometric distribution, which can be loosely expressed as

$$Geometric\left(\frac{\lambda}{n}\right) \quad \text{“}\to\text{”} \quad Exp(\lambda) \qquad as\ n\to\infty$$

in the sense that for $Y_n \sim Geometric(\lambda/n), Y \sim Exp(\lambda)$, as $n\to\infty$,

$$\mathbb{P}(Y_n/n \le t) \to \mathbb{P}(Y \le t)$$

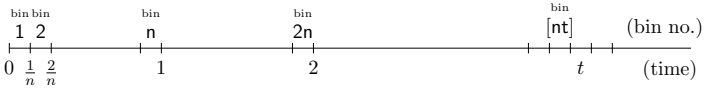
Proof. For each n, partition $[0,\infty)$ into bins of width $\frac{1}{n}$. Let

$$X_i \;=\; \begin{cases}1, & Success\ in\ bin\ i,\ with\ probability\ p_n = \dfrac{\lambda}{n},\\ 0, & otherwise.\end{cases}$$

$$Y_n \;=\; \min\{k : X_1 = \cdots = X_{k-1} = 0, X_k = 1\}$$

Then

$$Y_n \sim Geometric(p_n), \qquad p_n = \frac{\lambda}{n}$$



Let

$$Y \sim Exp(\lambda)$$

Then

$$\begin{aligned}\mathbb{P}(Y_n/n \le t) = \mathbb{P}\{Y_n \le [nt]\} &= \sum_{k=1}^{[nt]} \mathbb{P}\{Y_n = k\} = \sum_{k=1}^{[nt]} p_n(1-p_n)^{k-1} \\ &= p_n \frac{1-(1-p_n)^{[nt]}}{1-(1-p_n)} = 1-(1-p_n)^{[nt]} \\ &= 1 - \left(1-\frac{\lambda}{n}\right)^{[nt]} = 1 - \left[\left(1-\frac{\lambda}{n}\right)^{\frac{n}{\lambda}}\right]^{\lambda\frac{[nt]}{n}} \\ &\longrightarrow \quad 1-e^{-\lambda t} = \mathbb{P}\{Y \le t\} \qquad \text{as} \;\; n \to \infty,\end{aligned}$$

where we apply the limits

$$\left(1-\frac{\lambda}{n}\right)^{\frac{n}{\lambda}} \to e^{-1} \quad \text{and} \quad \frac{[nt]}{n} \to t \qquad \text{as} \;\; n \to \infty.$$

□

2.6 Expectation, variance and covariance

Definitions.

Discrete: $\mathbb{E}(X) = \sum_{k=0}^{\infty} k\,\mathbb{P}\{X=k\}$

Continuous: $\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx$

Variance: $Var(X) = \mathbb{E}(X-\mathbb{E}(X))^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

Examples.

- Bernoulli λ: $\mathbb{E}(X) = p,\ Var(X) = p(1-p)$
- Geometric p: $\mathbb{E}(X) = \frac{1}{p},\ Var(X) = \frac{1-p}{p^2}$

  (Heuristics: Out of 100 trials, we should expect about $100p$ success and $100(1-p)$ failures. So the average number of trials per success is $100/100p = 1/p$, which is the expected value of the geometric distribution.)
- Binomial (n,p): $\mathbb{E}(X) = np,\ Var(X) = np(1-p)$
- Poisson λ: $\mathbb{E}(X) = \lambda = Var(X)$
- Uniform (a,b): $\mathbb{E}(X) = \frac{a+b}{2},\ Var(X) = \frac{(b-a)^2}{12}$
- Exponential λ: $\mathbb{E}(X) = \frac{1}{\lambda},\ Var(X) = \frac{1}{\lambda^2}$
- Gamma (n,λ): $\mathbb{E}(X) = \frac{n}{\lambda},\ Var(X) = \frac{n}{\lambda^2}$

- Beta (α,β): $\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta},\ Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- Normal (μ,σ^2): $\mathbb{E}(X) = \mu,\ Var(X) = \sigma^2$

Expectation of functions of random variables.

- Discrete: $\mathbb{E}(g(X)) = \sum_{k=0}^{\infty} g(k)\mathbb{P}\{X=k\}$
- Continuous: $\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$

Moments of random variables. $g(x) = x^m$ in the above.

Linearity. $\mathbb{E}(aX+b) = a\mathbb{E}(X) + b$

Expectation, variance and covariance for several variables.

- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
- $\mathbb{E}(aX+bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$
- $\mathbb{E}(\sum_i a_i X_i) = \sum_i a_i \mathbb{E}(X_i)$
- $Cov(X,Y) = \mathbb{E}(X-\mathbb{E}(X))(Y-\mathbb{E}(Y)) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
- $Cov(X,Y) = 0$ if X and Y are independent (or if and only if X and Y are uncorrelated).
- $Var(\sum_i X_i) = \sum_i Var(X_i)$ if X_i's are independent. Otherwise,

$$\begin{aligned}Var\left(\sum_i X_i\right) &= \mathbb{E}\left(\sum_i X_i - \mathbb{E}\left(\sum_i X_i\right)\right)^2 = \mathbb{E}\left(\sum_i X_i\right)^2 - \left(\sum_i \mathbb{E}(X_i)\right)^2 \\ &= \sum_i \sum_j \mathbb{E}(X_i X_j) - \sum_i \sum_j \mathbb{E}(X_i)E(X_j) \\ &= \sum_i \mathbb{E}(X_i^2) + \sum_{i\ne j} \mathbb{E}(X_i X_j) - \sum_i \mathbb{E}(X_i)^2 - \sum_{i\ne j} \mathbb{E}(X_i)E(X_j) \\ &= \sum_i \left(\mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2\right) + \sum_{i\ne j} \left(\mathbb{E}(X_i X_j) - \mathbb{E}(X_i)E(X_j)\right) \\ &= \sum_i Var(X_i) + \sum_{i\ne j} Cov(X_i, X_j) = \sum_i Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)\end{aligned}$$

- $Cov(\sum_i X_i, \sum_j Y_j) = \sum_i \sum_j Cov(X_i, Y_i)$

An example X = Number of people getting their own car keys when drawing randomly from the key box that contains everyone's keys (n total). $\mathbb{E}(X)$ =?

Let

$$X_i = \begin{cases} 1, & \text{person } i \text{ gets key } i, \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\mathbb{E}(X_i) = 1 \times \mathbb{P}\{X_i = 1\} + 0 \times \mathbb{P}\{X_i = 0\} = \mathbb{P}\{X_i = 1\} = \frac{1}{n}$$

In fact, $X_i \sim$ *Bernoulli* $\left(\frac{1}{n}\right)$,

$$\mathbb{E}(X_i) = \frac{1}{n}, \qquad Var(X_i) = \frac{1}{n}\left(1 - \frac{1}{n}\right)$$

By the linearity of expectation,

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(X_i) = \sum_{i=1}^{n} \frac{1}{n} = n \times \frac{1}{n} = 1$$

Notice that X_i's are not independent,

$$\begin{aligned}\mathbb{E}(X_i X_j) &= \mathbb{P}\{X_i X_j = 1\} = \mathbb{P}\{X_i = 1, X_j = 1\} \\ &= \mathbb{P}\{X_i = 1\}\mathbb{P}\{X_j = 1 | X_i = 1\} = \frac{1}{n} \times \frac{1}{n-1} \neq \mathbb{E}(X_i)\mathbb{E}(X_j)\end{aligned}$$

Thus pairwise covariance is non-zero,

$$Cov(X_i X_j) = \mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j) = \frac{1}{n(n-1)} - \left(\frac{1}{n}\right)^2 = \frac{1}{n^2(n-1)}$$

Hence

$$Var(X) = \sum_{i=1}^{n} Var(X_i) + \sum_{i \neq j} Cov(X_i X_j) = n\frac{1}{n}\left(1 - \frac{1}{n}\right) + n(n-1)\frac{1}{n^2(n-1)} = 1.$$

2.7 Moment generating functions

The moment generating function (MGF) of a random variable X

$$\phi_X(t) = \mathbb{E}\left(e^{tX}\right)$$

typically exists for some range of t. Notice that $\phi_X(-t)$ is the Laplace transform.

- Moment generating property:

$$\begin{aligned}\phi'_X(0) &= \mathbb{E}\left(Xe^{tX}\right)\Big|_{t=0} = \mathbb{E}(X), \\ \phi''_X(0) &= \mathbb{E}\left(X^2 e^{tX}\right)\Big|_{t=0} = \mathbb{E}(X^2), \\ &\vdots \\ \phi^{(n)}_X(0) &= \mathbb{E}\left(X^n e^{tX}\right)\Big|_{t=0} = \mathbb{E}(X^n).\end{aligned}$$

- $Y = aX + b$, then

$$\phi_Y(t) = \phi_{aX+b}(t) = \mathbb{E}\left(e^{t(aX+b)}\right) = e^{bt}\phi_X(at).$$

- If X, Y are independent, then

$$\phi_{X+Y}(t) = \mathbb{E}\left(e^{t(X+Y)}\right) = \mathbb{E}\left(e^{tX}e^{tY}\right) = \mathbb{E}\left(e^{tX}\right)\mathbb{E}\left(e^{tY}\right) = \phi_X(t)\phi_Y(t).$$

- If X_i's are independent, then

$$\phi_{X_1+X_2+\cdots+X_n}(t) = \phi_{X_1}(t)\phi_{X_2}(t)\cdots\phi_{X_n}(t).$$

- Joint moment generation functions.

  For any n random variables $X_1, X_2, \cdots, X_n$ (not necessarily mutually independent),

$$\phi_{(X_1,X_2,\cdots,X_n)}(t_1, t_2, \cdots, t_n) = \begin{cases} \sum_{x_1,\cdots,x_n} e^{t_1x_1+\cdots+t_nx_n}\ p(x_1, \cdots, x_n) \\ \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{t_1x_1+\cdots+t_nx_n} f(x_1, \cdots, x_n)dx_1\cdots dx_n \end{cases}$$

- Marginal moment generating function:

  Example: $\phi_{X_2}(t) = \phi_{(X_1,X_2,\cdots,X_n)}(0, t, 0, \cdots, 0)$.

2.7.1 MGF for some discrete distributions

- $X \sim$ Bernoulli(p).

$$\phi(t) = \mathbb{E}\left(e^{tX}\right) = e^{t\times 1}P\{X = 1\} + e^{t\times 0}P\{X = 0\} = e^t p + e^0(1-p) = pe^t + (1-p).$$

- $X \sim$ Binomial(n, p). Since $X = \sum_{i=1}^{n} X_i$, $X_i \sim$ Bernoulli(p), X_i's are indepdent,

$$\phi(t) = \mathbb{E}(e^{tX}) = \mathbb{E}\left(e^{t\sum_{i=1}^{n} X_i}\right) = \phi_{X_1}(t)\cdots\phi_{X_n}(t) = \left(pe^t + (1-p)\right)^n.$$

- $X \sim$ Poisson(λ).

$$\phi(t) = \sum_{k=0}^{\infty} e^{tk}\frac{\lambda^k}{k!}e^{-\lambda} = \sum_{k=0}^{\infty}\frac{\left(\lambda e^t\right)^k}{k!}e^{-\lambda} = e^{\lambda(e^t-1)}$$

  If X_1 and X_2 are independent, it is easy to prove using MGF that

$$X_1 \sim Poi(\lambda_1),\ X_2 \sim Poi(\lambda_2) \implies X_1 + X_2 \sim Poi(\lambda_1 + \lambda_2),$$

  because

$$\phi_{X_1+X_2}(t) = \phi_{X_1}(t)\phi_{X_2}(t) = e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}.$$

- $X = (X_1, \cdots, X_k) \sim$ multinomial $(n; p_1, \cdots, p_k), \sum_{i=1}^{k} p_i = 1$.

  Moment generation function for multivariate variables $X_1, \cdots, X_k$ is defined as

$$\phi_{X_1,\cdots,X_k}(t_1, \cdots, t_k) = E[e^{t_1X_1+\cdots+t_kX_k}]$$

  Using multinomial probability, the moment generating function can be derived as

$$\phi_{X_1,\cdots,X_k}(t_1, \cdots, t_k) = \left(\sum_{i=1}^{k} p_i e^{t_i}\right)^n = \left(\sum_{i=1}^{k}\phi_{X_i}(t_i)\right)^n, \qquad \phi_{X_i}(t) = p_i e^t + (1-p_i),$$

  which takes the form of the product of the moment generating functions of the n multinomial observation. Notice that the components X_i's are dependent. In particular, $X_k = n - (X_1 + \cdots + X_{k-1})$. Therefore the moment generation function for multinomial can be on the first $k-1$ components, thus has the form

$$\phi_{X_1,\cdots,X_{k-1}}(t_1, \cdots; t_{k-1}) = E\left(e^{t_1X_1+\cdots+t_{k-1}X_{k-1}}\right) = \left(\sum_{i=1}^{k-1} p_i e^{t_i} + 1 - p_1 - \cdots - p_{k-1}\right)^n = \left(\sum_{i=1}^{k-1} p_i e^{t_i} + p_k\right)^n$$

2.7.2 MGF for some continuous distributions

- $X \sim Exp(\lambda)$.

$$\begin{aligned}\phi(t) &= \mathbb{E}(e^{tX}) = \int_0^\infty e^{tx}\lambda e^{-\lambda x}dx \\ &= \lambda\int_0^\infty e^{(t-\lambda)x}dx = \frac{\lambda}{t-\lambda}e^{-\lambda x}\Big|_0^\infty = \frac{\lambda}{\lambda - t}. \qquad for \quad t<\lambda.\end{aligned}$$

for $t \geq \lambda$, the integral diverges, the MGF does not exist.

- $X \sim Gamma(\alpha, \lambda)$.

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}, \qquad x \geq 0.$$

If $\alpha = n$ is a positive integer, then $X = X_1 + \cdots + X_n$, for $X_i, i.i.d. \sim Exp(\lambda)$.

$$\phi(t) = \mathbb{E}(e^{tX}) = \mathbb{E}\left(e^{t\sum_{i=1}^n X_i}\right) = \phi_{X_1}(t)\cdots\phi_{X_n}(t) = \left(\frac{\lambda}{\lambda - t}\right)^n, \quad for \quad t<\lambda.$$

Continuity of $\phi(t)$ in α yields the general result

$$\phi(t) = \left(\frac{\lambda}{\lambda - t}\right)^\alpha, \qquad for \quad t<\lambda.$$

- $X \sim$ normal distribution.

  - Standard normal $Z \sim N(0,1)$.

$$\begin{aligned}\phi(t) &= \int_{-\infty}^\infty e^{tx}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty e^{-\frac{1}{2}(x^2-2xt+t^2)}e^{t^2/2}dx \\ &= e^{t^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty e^{-\frac{1}{2}(x-t)^2}dx \\ &= e^{t^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty e^{-\frac{1}{2}(x^*)^2}dx^* \qquad (x^* = x-t) \\ &= e^{t^2/2}\end{aligned}$$

  - Univariate normal $X \sim N(\mu, \sigma^2)$. Then $X = \mu + \sigma Z, Z \sim N(0,1)$.

$$\phi(t) = e^{\mu t}\phi_Z(\sigma t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

For $a \neq 0, aX$ has MGF

$$\phi_{aX}(t) = \phi_X(at) = e^{(a\mu)t + \frac{1}{2}(a\sigma)^2 t^2}$$

  - Multivariate normal.

$Z_1, Z_2, \cdots, Z_n,\ i.i.d. \sim N(0,1)$,

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} a_{11}Z_1 + a_{12}Z_2 + \cdots + a_{1n}Z_n \\ a_{21}Z_1 + a_{22}Z_2 + \cdots + a_{2n}Z_n \\ \vdots \\ a_{n1}Z_1 + a_{n2}Z_2 + \cdots + a_{nn}Z_n \end{pmatrix} = A_{n\times n}Z_{n\times 1}, \quad Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix}$$

X is multivariate normal. The MGF is

$$\begin{aligned}\phi_{X_1,X_2,\cdots,X_n}(t_1,t_2,\cdots,t_n) &= \mathbb{E}(e^{t_1X_1+\cdots+t_nX_n}) \\ &= \mathbb{E}(e^{t^TX}) = \mathbb{E}(e^{t^T(AZ)}) = \mathbb{E}(e^{(t^TA)Z}) \\ &= \mathbb{E}\left(e^{(t_1a_{11}+t_2a_{21}+\cdots+t_na_{n1})Z_1+\cdots+(t_1a_{1n}+t_2a_{2n}+\cdots+t_na_{nn})Z_n}\right) \\ &= \phi_{Z_1}(t_1a_{11}+\cdots+t_na_{n1})\cdots\phi_{Z_n}(t_1a_{1n}+\cdots+t_na_{nn}) \\ &= \phi_{Z_1}(t^Ta_{\cdot 1})\phi_{Z_2}(t^Ta_{\cdot 2})\cdots\phi_{Z_n}(t^Ta_{\cdot n}) \\ &= e^{(t^Ta_{\cdot 1})^2/2+(t^Ta_{\cdot 2})^2/2+\cdots+(t^Ta_{\cdot n})^2/2} \\ &= e^{(t^Ta_{\cdot 1}a_{\cdot 1}^Tt+\cdots+t^Ta_{\cdot n}a_{\cdot n}^Tt)/2} \\ &= e^{t^TAA^Tt/2} = e^{t^TA^TAt/2},\end{aligned}$$

where $t^T = (t_1, t_2, \cdots, t_n)$, and $a_{\cdot k} = \begin{pmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{nk} \end{pmatrix}$ denotes the kth column of matrix A. The covariance matrix of the X_i's is A^TA.

$$\begin{aligned}(A^TA)_{ij} &= \sum_{k=1}^n a_{ik}a_{kj} = \sum_{k=1}^n \mathbb{E}(a_{ik}a_{kj}Z_k^2) = \mathbb{E}\left(\sum_{k=1}^n a_{ik}a_{kj}Z_k^2\right) \\ &= \mathbb{E}\left((a_{i1}Z_1 + \cdots + a_{in}Z_n)(a_{1j}Z_1 + \cdots + a_{nj}Z_n)\right) \\ &= Cov(X_i, X_j) = Cov(X_j, X_i) = (A^TA)_{ji}\end{aligned}$$

Notice that the covariance matrix A^TA (positive definite and symmetric) is all the information needed to get the MGF of $(X_1, \cdots, X_n)$.

2.7.3 Uniqueness of MGF

There is a one-to-one correspondence between the moment generating function and the probability distribution of the random variable.

$$\phi_X(t) \equiv \phi_Y(t) \quad \Longrightarrow \quad X, Y \text{ have the same distribution.}$$

An application. $X_i,\ i = 1, \cdots, n$ are $i.i.d. \sim N(\mu, \sigma^2)$. As $n \to \infty$,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \longrightarrow N(0,1) \qquad (in\ distribution)$$

is usually proved by the Central Limit Theorem (CLT).

The above can also be proved using MGF and its uniqueness. Let $\phi(t) = \phi_{X_i}(t)$.

We consider the case $\mu = 0$ (the easier proof).

$$\begin{aligned}\mathbb{E}\left(e^{t\frac{X_1+\cdots+X_n}{\sigma\sqrt{n}}}\right) &= \left[\mathbb{E}\left(e^{t\frac{X}{\sigma\sqrt{n}}}\right)\right]^n = \phi\left(\frac{t}{\sigma\sqrt{n}}\right)^n \\ &= \left(1 + \frac{t^2}{2n} + \frac{c_{t,n}}{n^2}\right)^n \\ &\longrightarrow \lim_{n\to\infty}\left(1 + \frac{t^2}{2n}\right)^n = e^{t^2/2},\end{aligned}$$

which is the MGF of $N(0,1)$. In the above, a useful Taylor expansion is applied:

$$\begin{aligned}\phi(s) &= \sum_{j=0}^{\infty} \frac{s^j \phi^{(j)}(0)}{j!} \\ &= \phi(0) + s\phi'(0) + \frac{s^2\phi''(0)}{2} + \frac{s^3\phi^{(3)}(0)}{3!} + \frac{s^4\phi^{(4)}(0)}{4!} + \cdots \\ &= 1 + s\mu + \frac{s^2(\sigma^2+\mu^2)}{2} + \frac{s^3\phi^{(3)}(0)}{3!} + \frac{s^4\phi^{(4)}(0)}{4!} + \cdots\end{aligned}$$

When $\mu = 0$ (consequently $\phi^{(3)}(0) = 0$ for normal distribution), apply Taylor expansion at $s = \frac{t}{\sigma\sqrt{n}}$ gives the following:

$$\phi\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 + \frac{t^2(\sigma^2)}{2\sigma^2 n} + \frac{t^4\phi^{(4)}(0)}{24\sigma^4 n^2} + \cdots = 1 + \frac{t^2}{2n} + \frac{c_{t,n}}{n^2},$$

where the constant $c_{t,n}$ depending on t, n, however $c_{t,n} \le c_t < \infty$, thus $\frac{t^2}{2n} + \frac{c_{t,n}}{n^2} \approx \frac{t^2}{2n}$ for large n.

2.8 Some useful inequalities

- **Markov inequality**

  $X > 0$. For any $a > 0$,

  $$\mathbb{P}\{X > a\} \le \frac{\mathbb{E}(X)}{a}.$$

  Proof (for the case $X = 0, 1, \cdots, p_k = P(X = k)$)

  $$a\mathbb{P}\{X > a\} = a\sum_{k>a} p_k \le \sum_{k>a} k p_k \le \sum_k k p_k = \mathbb{E}(X).$$

- **Chebyshev's inequality**

  $\mu = \mathbb{E}(X), \sigma^2 = Var(X)$. For any $\varepsilon > 0$,

  $$\mathbb{P}\{|X - \mu| > \varepsilon\} \le \frac{\sigma^2}{\varepsilon^2}.$$

  Proof

  $$\mathbb{P}\{|X-\mu| > \varepsilon\} = \mathbb{P}\{|X-\mu|^2 > \varepsilon^2\} = \mathbb{P}\{X^* > a\} \le \frac{\mathbb{E}(X^*)}{a} = \frac{\sigma^2}{\varepsilon^2},$$

  where Markov inequality is applied to $X^* = |X-\mu|^2 > 0$ and $a = \varepsilon^2 > 0$.

- For $a > 0$,

  $$\mathbb{P}\{e^{tX} > a\} \le \frac{\phi_X(t)}{a}.$$

  Proof Applying Markov's inequality to $X^* = e^{tX} > 0$.

- For X with MGF $\phi(t)$ and $a > \mu = \mathbb{E}(X)$,

  $$\mathbb{P}\{X > a\} \le \phi_a(t) = \mathbb{E}(e^{t(X-a)}). \qquad \text{(shifted MGF)}$$

  Proof Applying the above inequality to $X^* = e^{tX}$ and $a^* = e^{ta}$,

  $$\mathbb{P}\{X > a\} = \mathbb{P}\{e^{tX} > e^{ta}\} \le \frac{\phi(t)}{e^{ta}} = e^{-at}\mathbb{E}\left(e^{tX}\right) = \mathbb{E}(e^{t(X-a)}) = \phi_a(t).$$

- $X_1, \cdots, X_n$ are i.i.d. with mean μ and MGF $\phi(t)$, then for $a > \mu$, $t > 0$,

  $$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n X_i > a\right\} \le \phi_a(t)^n.$$

  Proof

  $$\begin{aligned}\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n X_i > a\right\} &= \mathbb{P}\left\{\sum_{i=1}^n X_i > na\right\} = \mathbb{P}\left\{e^{t\sum_{i=1}^n X_i} > e^{tna}\right\} \\ &\le \frac{\phi_{\sum_{i=1}^n X_i}(t)}{e^{tna}} = \left(\phi(t)e^{-at}\right)^n = \phi_a(t)^n.\end{aligned}$$

  Remark: For fixed a and $t > 0$ near 0, $\phi_a(t)^n \to 0$ very quickly as $n \to \infty$, since $\phi_a(0) = 1$ and $\phi_a(t) < 1$ for small t near 0:

  $$\phi_a'(0) = \frac{d}{dt}\left(\phi(t)e^{-at}\right)\Big|_{t=0} = \left(\phi'(t)e^{-at} - a\phi(t)e^{-at}\right)\Big|_{t=0} = \mu - a < 0.$$

2.9 Limiting Properties

Strong Law of Large Numbers.

Let X_i be *i.i.d.* with $\mu = \mathbb{E}(X_i)$. Then

$$\frac{X_1 + \cdots + X_n}{n} \to \mu \quad \text{with probability 1, as } n \to \infty.$$

Central Limit Theorem.

Let X_i be *i.i.d.* with $mu = (X_i), \sigma^2 = Var(X_i)$. Then

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \to N(0,1) \quad as\ n \to \infty.$$

3 Conditional probability and conditional expectation

3.1 Basic conditioning

To study random variables with dependent structures, such as random variables $X_1, \cdots, X_t, \cdots$, in a stochastic process, we need to know the relationship among the X_t's, especially between the adjacent ones, say, X_t and X_{t+1}.

Recall that from the definition of conditional probability, we have

$$\mathbb{P}(AB) = \mathbb{P}(A)\,\mathbb{P}(B|A).$$

This product rule can be generalized. Apply the above equation twice,

$$\mathbb{P}(ABC) = \mathbb{P}(AB)\,\mathbb{P}(C|AB) = \mathbb{P}(A)\,\mathbb{P}(B|A)\,\mathbb{P}(C|BA).$$

Apply it again and again,

$$\mathbb{P}(A_1, A_2, \cdots, A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2|A_1)\cdots\mathbb{P}(A_n|A_{n-1}\cdots A_1)$$

The above leads to the following useful facts.

- If $X_1, X_2, \cdots, X_n$ are discrete random variables, then their joint probability mass function is

$$\begin{aligned} p(x_1, \cdots, x_n) &= \mathbb{P}(X_1 = x_1, \cdots, X_n = x_n) \\ &= \mathbb{P}(X_1 = x_1)\,\mathbb{P}(X_2 = x_2|X_1 = x_1)\,\mathbb{P}(X_n = x_n|X_{n-1} = x_{n-1}, \cdots, X_1 = x_1) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i = x_i|X_{i-1} = x_{i-1}, \cdots, X_1 = x_1) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i = x_i| \text{ given the past of } i) \\ &= \prod_{i=1}^{n} p(x_i \mid past_i) \end{aligned}$$

- If $X_1, X_2, \cdots, X_n$ are continuous random variables, then their joint probability density function has similar properties,

$$\begin{aligned} f(x_1, \cdots, x_n) &= f(x_1)f(x_2|x_1)\cdots f(x_n|x_{n-1}, \cdots, x_1) \\ &= \prod_{i=1}^{n} f(x_i|x_{i-1}, \cdots, x_1) \\ &= \prod_{i=1}^{n} f(x_i \mid past_i) \end{aligned}$$

In almost all interesting stochastic processes, X_t's are not independent. We need to characterize the dependence structure, and the above factorization of the joint probability or density function is very useful.

3.2 Conditional probability

Discrete case

The joint probability function

$$F(x, y) = F_{X,Y}(x, y) = F(X = x, Y = y) = \mathbb{P}\{X \le x, Y \le y\}$$

Definition of conditional probability:

$$p_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x \text{ and } Y = y)}{\mathbb{P}(Y = y)} = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

when the denominator is not zero.

Example — an urn-ball model illustration

Total of N balls with K white balls. Let $p = K/N$ be the proportion of original white balls, $X_i,\ i = 1, \cdots, n$ be 1 if the ith draw is white, 0 otherwise.

- $X_1, X_2, \cdots, X_n$, n draws with replacement. $X_i \sim Bernoulli(p)$ are independent.

$$S_n = \sum_{i=1}^{n} X_i, \qquad \mathbb{E}(S_n) = np, \quad Var(S_n) = np(1-p).$$

$$\Longrightarrow \quad S_n \sim Bin(n, p)$$

- $X_1, X_2, \cdots, X_n$, n draws without replacement. Then X_i are not independent.
  For example, one realization of the probability distribution of $(X_1, X_2, X_3, \cdots, X_n)$ can be

$$\begin{aligned} & P\{X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 0, X_5 = 1, \cdots\} \\ = \ & P\{X_1 = 1\}P\{X_2 = 1|X_1 = 1\}P\{X_3 = 0|X_2 = 1, X_1 = 1\}\cdots \\ = \ & \frac{K}{N}\,\frac{K-1}{N-1}\,\frac{N-K}{N-2}\,\frac{N-K-1}{N-3}\,\frac{K-2}{N-4}\cdots \end{aligned}$$

  not the same as the independent case.

  Remarks This is an example of Exchangeability.
  Consider the probability distribution of a permutation of $(X_1, X_2, X_3, \cdots, X_n)$, say, $(X_1, X_3, X_2, \cdots, X_n)$, that is, when X_1 and X_3 are permuted, the joint probability is the same.

$$\begin{aligned} & P\{X_1 = 1, X_3 = 0, X_2 = 1, X_4 = 0, X_5 = 1, \cdots\} \\ = \ & \frac{K}{N}\,\frac{N-K}{N-1}\,\frac{K-1}{N-2}\,\frac{N-K-1}{N-3}\,\frac{K-2}{N-4}\cdots \end{aligned}$$

  No matter what the permutation is, X_i's are exchangeable (yet not independent): the joint distribution of $(X_1, X_2, \cdots, X_n)$ is the same as the joint distribution of $(X_{\pi_1}, X_{\pi_2}, \cdots, X_{\pi_n})$ for any permutation $\pi = (\pi_1, \pi_2, \cdots, \pi_n)$ of $(1, 2, \cdots, n)$.

The mean and variance of the partial sum $S_n = \sum_{i=1}^{n} X_i$ are

$$\mathbb{E}(S_n) = np, \quad Var(S_n) = np(1-p)\frac{N-n}{N-1}.$$

Proof. Th mean equation above is by linearity of the integration or sum. In the following we prove the variance equation (for the with-replacement sample).

$$\begin{aligned}\mathbb{E}(X_iX_j) &= 1\times\mathbb{P}\{X_i=1,X_j=1\}+0\times\mathbb{P}\{X_i=0 \text{ or } X_j=0\}\\ &= \mathbb{P}\{X_j=1|X_i=1\}\,\mathbb{P}\{X_i=1\}\\ &= \mathbb{P}\{X_2=1|X_1=1\}\,\mathbb{P}\{X_1=1\}\\ &= \frac{K-1}{N-1}\,\frac{K}{N}=\frac{K-1}{N-1}\,p.\end{aligned}$$

The probability $\mathbb{P}\{X_i=1,X_j=1\}$ can alternatively obtained by

$$\mathbb{P}\{X_i=1,X_j=1\}=\frac{\binom{2}{2}\binom{N-2}{K-2}}{\binom{N}{K}}=\frac{K(K-1)}{N(N-1)}$$

From $\mathbb{E}(X_i)=K/N=p$, the covariance

$$Cov(X_i,X_j) = \mathbb{E}(X_iX_j)-\mathbb{E}(X_i)\mathbb{E}(X_j)=p\frac{K-1}{N-1}-p^2=-\frac{p(1-p)}{N-1}$$

$$\begin{aligned}\Longrightarrow\quad Var(S_n) &= \sum_{i=1}^n Var(X_i)+2\sum_{i<j}Cov(X_i,X_j)\\ &= np(1-p)+2\binom{n}{2}\left(-\frac{p(1-p)}{N-1}\right)\\ &= np(1-p)+n(n-1)\frac{p(1-p)}{N-1}\\ &= np(1-p)\left(1-\frac{n-1}{N-1}\right)=np(1-p)\frac{N-n}{N-1}\end{aligned}$$

which is slightly smaller than the binomial variance for the independent without-replacement sample.

□

Continuous case

$f_{X,Y}(x,y)$ — joint density of the pair of random variables (X,Y).

Conditional density functions are defined as

$$f_{X|Y}(x|y)=\frac{f_{X,Y}(x,y)}{f_Y(y)},\qquad f_{Y|X}(y|x)=\frac{f_{X,Y}(x,y)}{f_X(x)}$$

when the denominators are not zero.

Bayes Theorem for random variables.

Let $p_{X|Y}$ stands for conditional probability mass function when X is discrete or density function when X is continuous for Y discrete or continuous, then

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

with

$$P_X(x)=\begin{cases}\sum_y p_{X|Y}(x|y)p_Y(y), & Y \text{ discrete};\\ \int p_{X|Y}(x|y)f_Y(y)dy, & Y \text{ continuous}.\end{cases}$$

Bayes Theorem also applies to the case when one variable is discrete and the other is continuous.

Examples of mixed cases

Y is a continuous random variable with density function $f_Y(y)$. X is a discrete random variable. When $Y=y$, X has the conditional probability $P_{X|Y}(X=x|y)$ that depends on y. Therefore the conditional density function for Y given $X=x$ is

$$f_{Y|X}(y|X=x)=\frac{\mathbb{P}_{X|Y}(X=x|y)f_Y(y)}{\int\mathbb{P}_{X|Y}(X=x|y)f_Y(y)dy}$$

Similarly, The conditional probability of $X=x$ given $Y=y$ is

$$\mathbb{P}_{X|Y}(X=x|y)=\frac{f_{Y|X}(y|X=x)\mathbb{P}(X=x)}{\sum_{x'}f_{Y|X}(y|X=x')\mathbb{P}(X=x')}$$

Proof. By definition,

$$\begin{aligned}\mathbb{P}_{X|Y}(X=x|y) &= \lim_{\Delta y\to0}\mathbb{P}(X=x|Y\in(y,y+\Delta y))\\ &= \lim_{\Delta y\to0}\frac{\mathbb{P}(X=x,Y\in(y,y+\Delta y))}{\mathbb{P}(Y\in(y,y+\Delta y))}\\ &= \lim_{\Delta y\to0}\frac{\mathbb{P}(Y\in(y,y+\Delta y)|X=x)\,\mathbb{P}(X=x)}{\sum_{x'}\mathbb{P}(Y\in(y,y+\Delta y)|X=x')\,\mathbb{P}(X=x')}\\ &= \frac{\lim_{\Delta y\to0}\frac{\mathbb{P}(Y\in(y,y+\Delta y)|X=x)}{\Delta y}\mathbb{P}(X=x)}{\sum_{x'}\lim_{\Delta y\to0}\frac{\mathbb{P}(Y\in(y,y+\Delta y)|X=x')}{\Delta y}\mathbb{P}(X=x')}\\ &= \frac{f_{Y|X}(y|X=x)\mathbb{P}(X=x)}{\sum_{x'}f_{Y|X}(y|X=x')\mathbb{P}(X=x')}\end{aligned}$$

where we have assumed that the limits are well defined, and the denominators are non-zero.

□

3.3 Conditional expectation and conditional variance

If we have no information about the future outcome X of a random experiment, our best guess of the value of X is $\mathbb{E}(X)$.

If we have complete knowledge of the outcome, then we know the exact value of X.

Conditional expectation is the best guess of the value of X when we have partial knowledge of the outcome, for example, the knowledge of the values of other random variables.

Conditional Expectation

When X, Y both are discrete or both continuous, conditional expectation of X conditioned on Y, $\mathbb{E}(X|Y)$, evaluated at $Y = y$ is defined as

$$\mathbb{E}(X|Y=y) = \begin{cases} \sum_x x\,\mathbb{P}(X=x|Y=y), & X \text{ discrete,} \\ \int_{-\infty}^{\infty} x f_{X|y}(x|y)dx, & X \text{ continuous,} \end{cases}$$

which is a function of y.

Hence, $\mathbb{E}(X|Y)$ is treated as a random function of Y.

One of the most important property of conditional expectation is the tower law of expectation:

$$\mathbb{E}[\mathbb{E}(X|Y)] = \mathbb{E}[X].$$

Proof. We will prove the case when X, Y both discrete or both continuous.

For X, Y discrete with finite expectations,

$$\begin{aligned} \mathbb{E}[\mathbb{E}(X|Y)] &= \sum_y \mathbb{E}(X|Y=y)P(Y=y) = \sum_y \left(\sum_x x\mathbb{P}(X=x|Y=y)\right)\mathbb{P}(Y=y) \\ &= \sum_x x\left(\sum_y \mathbb{P}(X=x|Y=y)\,\mathbb{P}(Y=y)\right) = \sum_x x\,\mathbb{P}(X=x) = \mathbb{E}[X] \end{aligned}$$

For X, Y continuous with finite expectations,

$$\begin{aligned} \mathbb{E}[\mathbb{E}(X|Y)] &= \int_{-\infty}^{\infty} \mathbb{E}(X|Y=y)f_Y(y)dy = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx\right) f_Y(y)dy \\ &= \int_{-\infty}^{\infty} x\left(\int_{-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)dy\right) = \int_{-\infty}^{\infty} x f_X(x) = \mathbb{E}[X] \end{aligned}$$

□

Characteristic properties of conditional expectation:

- $\mathbb{E}(X|Y)$ is a (Borel measurable) function of the random variable Y.
- If A is an event depending only on Y, then

$$\mathbb{E}(X \cdot \mathbf{1}_A] = \mathbb{E}[\mathbb{E}(X \cdot \mathbf{1}_A \mid Y)] = \mathbb{E}[\mathbb{E}(X|Y)\mathbf{1}_A]$$

Moreover, $\mathbb{E}(X|Y)$ is the unique random function of Y which has the above property for every A depending on Y only.

Conditional variance

The definition of conditional variance involves conditional expectations at two levels:

$$Var(X|Y=y) = \mathbb{E}\left[(X - \mathbb{E}[X|Y=y])^2 \mid Y=y\right].$$

As in the non-conditional case, a convenient form is

$$Var(X|Y=y) = \mathbb{E}[X^2|Y=y] - (\mathbb{E}[X|Y=y])^2.$$

The conditional variance formula (the law of total variance)

$$Var(X) = V(\mathbb{E}[X|Y]) + \mathbb{E}[Var(X|Y)].$$

Proof. From the above formula for $Var(X|Y=y)$, we obtain

$$\mathbb{E}[Var(X|Y)] = \mathbb{E}[\mathbb{E}(X^2|Y)] - \mathbb{E}[(\mathbb{E}[X|Y])^2].$$

Then

$$\begin{aligned} Var(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[\mathbb{E}(X^2|Y)] - \mathbb{E}[\mathbb{E}(X|Y)]^2 \\ &= \mathbb{E}[Var(X|Y)] + \mathbb{E}[\mathbb{E}(X|Y)]^2 - \mathbb{E}[\mathbb{E}(X|Y)]^2 \qquad \text{(use the equation above)} \\ &= \mathbb{E}[Var(X|Y)] + Var(\mathbb{E}[X|Y]) \qquad \text{(use the variance formula)} \end{aligned}$$

□

An Example. (Finding the expectation by first-step conditioning)

A mouse is trapped in a maze. It has equal probability to try any of the three tunnels, without aware of their destinies:

First tunnel will lead to the outside in t_1 minutes;

Second tunnel will lead back to the same site in t_2 minutes;

Third tunnel will lead back to the same site in t_3 minutes.

Let X be the time for the mouse to exit the maze,

$$Y = \begin{cases} 1, & \text{if the mouse exits in the first try;} \\ 0, & \text{if the mouse comes back to the same site after the first try.} \end{cases}$$

Assuming that the three doors looked so identical that the mouse would not know when door it has tried unsuccessfully. (The mouse's intelligence deserves more credits but this is just a hypothetically example.) That means, if, say, the mouse tried the second door first, then the expected time to exit the maze becomes

$$t_2 + \mathbb{E}(X)$$

Now we can calculate the expected time to exit.

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[\mathbb{E}(X|Y)] \\ &= \mathbb{P}\{Y=1\}\,\mathbb{E}(X|Y=1) + \mathbb{P}\{Y=0\}\,\mathbb{E}(X|Y=0) \\ &= \frac{1}{3}\times t_1 + \frac{2}{3}\left(\frac{1}{2}\times(t_2+\mathbb{E}(X)) + \frac{1}{2}\times(t_3+\mathbb{E}(X))\right) \\ &= \frac{t_1}{3} + \frac{t_2+t_3}{3} + \frac{2}{3}\mathbb{E}(X) \end{aligned}$$

Moving $\mathbb{E}(X)$ to the same side, we obtain

$$\mathbb{E}(X) \;=\; t_1 + t_2 + t_3 \;(minutes)$$

Note that in calculating $\mathbb{E}(X|Y=0)$, we have used the above mentioned assumption that the mouse will not learn from its experience: it always chooses any of the three tunnels with equal probability.

The conditioning may also be on variables different from the one used. For example, we may use

$$Y = \begin{cases} 1, & \text{if the mouse takes the fist tunnel in the first try;} \\ 2, & \text{if the mouse takes the second tunnel in the first try;} \\ 3, & \text{if the mouse takes the third tunnel in the first try.} \end{cases}$$

The conclusions will be the same.

An Example. (A recursive formula of expectation by conditioning on the previous step.)

Consider coin-toss with probability p of having Heads.

Find $\mathbb{E}(N_k)$, where N_k = the first time of a sequence of k Heads.

Start by conditioning on N_{k-1}. Notice that $N_{k-1}=m$ is the event that the first sequence of $k-1$ Heads occurred at the mth toss.

$$\begin{aligned} \mathbb{E}(N_k|N_{k-1}=m) &= p(m+1)+(1-p)(m+1+\mathbb{E}(N_k)) \\ &= m+1+(1-p)\mathbb{E}(N_k) \end{aligned}$$

Using the tower law of expectation,

$$\begin{aligned} \mathbb{E}(N_k) &= \mathbb{E}[\,\mathbb{E}(N_k|N_{k-1})] \\ &= \sum_{m=k-1}^{\infty} \mathbb{P}\{N_{k-1}=m\}\mathbb{E}(N_k|N_{k-1}=m) \\ &= \sum_{m=k-1}^{\infty} \mathbb{P}\{N_{k-1}=m\}(m+1+(1-p)\mathbb{E}(N_k)) \\ &= \sum_{m=k-1}^{\infty} m\mathbb{P}\{N_{k-1}=m\} + \sum_{m=k-1}^{\infty} \mathbb{P}\{N_{k-1}=m\}(1+(1-p)\mathbb{E}(N_k)) \\ &= \mathbb{E}(N_{k-1})+1+(1-p)\mathbb{E}(N_k). \end{aligned}$$

We obtain a recursive formula for the desired expectation:

$$\mathbb{E}(N_k) = \frac{1}{p}\left(\mathbb{E}(N_{k-1})+1\right)$$

We can derive the first few:

$$\begin{aligned} \mathbb{E}(N_1) &= \frac{1}{p} && (N_1 \sim \text{Geometric distribution}) \\ \mathbb{E}(N_2) &= \frac{1}{p^2}+\frac{1}{p} \\ \mathbb{E}(N_3) &= \frac{1}{p^3}+\frac{1}{p^2}+\frac{1}{p} \\ &\vdots \end{aligned}$$

The explicit formula can be derived by induction:

$$\mathbb{E}(N_k) = \sum_{i=1}^{k} \frac{1}{p^i}$$

An Example. (Probability of a discrete variable conditioned on a continuous variable.)

n people are invited to a party for a VIP. The party starts at noon. The arrival times of the guests are i.i.d. $\sim Exp(\lambda)$ (hours), the VIP may arrive any time T between noon and β O'clock p.m.

Let N be the number of guests arriving before the VIP. Then conditioned on the VIP arriving at $T=t$,

$$N|T=t \;\sim\; Bin(n,p_t), \qquad \mathbb{E}(N|T=t)=np_t$$

where

$$p_t = \mathbb{P}\{\text{a guest arrives before } t\} = \int_0^t \lambda e^{-\lambda s}ds = 1-e^{-\lambda t}.$$

It is reasonable to assume that $T\sim U(0,\beta)$. Therefore, the probability that k guests arrived before the VIP is

$$\begin{aligned} \mathbb{P}(N=k) &= \int_{-\infty}^{\infty} \mathbb{P}(N=k|T=t)f_T(t)dt \\ &= \int_0^{\beta} \binom{n}{k}(1-e^{-\lambda t})^k(e^{-\lambda t})^{n-k}\frac{1}{\beta}dt \\ &= \binom{n}{k}\frac{1}{\beta\lambda}\int_{e^{-\lambda\beta}}^{1}(1-y)^k y^{n-k-1}dy \end{aligned}$$

where variable substitution $y=e^{-\lambda t},\ dy=-\lambda e^{-\lambda t}dt$ is used in the last step.

Another interesting quantity to calculate is

$$\mathbb{E}(N)=\mathbb{E}[\mathbb{E}(N|T)] = \int_0^{\beta} n(1-e^{-\lambda t})\frac{1}{\beta}dt = n\left(1-\frac{1}{\lambda\beta}(1-e^{-\lambda\beta})\right).$$

In particular, if the arrival time of the VIP is any time before 1 p.m. ($\beta=1$), and the mean arrival time of the guests is 1 p.m. ($\lambda=1$), then only 37% of the guests are expected to be there when the VIP arrives.