Math 431 Intro to Probability

Chenxi Wu

Fall 2025

Contents

Instructor: Chenxi Wu (he/him)

Email: cwu367@wisc.edu

Lecture: 1-2:15pm Tu Th

Office Hours: 9-10 am Tuesday and Wednesday at Van Vleck 517, or by appointment.

Grades: 10% weekly HW, 2% Quiz on prerequisites, 5% weekly quizzes, $2 \times 25\%$ Midterms, 33% Final Exam.

Do as much of the exercises as possible, but make sure you understand the basic concepts first.

Why study probability:

- Foundation of statistics, Statistical Literacy
- Applications in other areas of mathematics
- Applications in science and engineering

Content colored in blue are materials that might help with understanding but will not be covered in the exam.

1 Probability Spaces

1.1 Definition and Basic Properties

1.1-1.4 of textbook

One way to formulate the concept of probability is the **Kolmogorov's Axioms**:

Definition 1.1. A **probability space** is a tuple $(\Omega, \mathcal{F}, P)$, where:

1. Ω is a set called the **sample space**, an element $\omega \in \Omega$ is called a **sample**.
2. $\mathcal{F}$ is a subset of the set of subsets of Ω. Elements of $\mathcal{F}$ are called **events**. F is further required to be a **σ-algebra**, which means that:
   (a) $\emptyset \in \mathcal{F}$
   (b) $A \in \mathcal{F} \implies A^c \in \mathcal{F}$
   (c) Let $\{A_n\}$ be a countable sequence of elements of $\mathcal{F}$, then $\bigcup_n A_n \in \mathcal{F}$
3. P is a function from $\mathcal{F}$ to $\mathcal{R}$, called the **probability**, or **probability measure** or **probability distribution**. It satisfies the following axioms:
   (i) For any $A \in \mathcal{F}$, $0 \leq P(A) \leq 1$
   (ii) $P(\emptyset) = 0$, $P(X) = 1$
   (iii) Let A_i be a sequence of pairwise disjoint ($i \neq j$ then $A_i \cap A_j = \emptyset$) events, then $P(\bigcup_i A_i) = \sum_i P(A_i)$.

Remark 1.2.

1. Elements of Ω denotes "possible outcomes of an experiment" or "possible states of the world". The set Ω and its elements are usually unimportant.
2. Elements of $\mathcal{F}$ denotes events whose probability we care about. Let $A \in \mathcal{F}$, $\omega \in A$ means "at state ω the event A can be said to have happened". The assumptions on $\mathcal{F}$ means:
   (a) There is an event that would never happen.
   (b) If A is an event, "A does not happen" is an event as well.
   (c) If we have a countable sequence of events $\{A_n\}$, then "at least one of the A_i happens" is an event.
3. P is a function that assigns each event its probability. The axioms on P means:
   (i) The probability of an event must be a number between 0 and 1
   (ii) If an event never happens, its probability is 0. If it always happens, its probability is 1.

(iii) If there is a countable sequence of events, none of the two can happen at the same time, then the probability that at least one happens is the sum of the probabilities they happen.

Remark 1.3. In probability, we sometimes denote $A \cap B$ as AB.

Example 1.4. A fair coin flip can be represented as the probability space $(\Omega, \mathcal{F}, P)$, where

$$\Omega = \{\text{head}, \text{tail}\}$$

$$\mathcal{F} = \{\emptyset, \{\text{head}\}, \{\text{tail}\}, \{\text{head}, \text{tail}\}\}$$

$$P(\emptyset) = 0, P(\{\text{head}\}) = 1/2, P(\{\text{tail}\}) = 1/2, P(\Omega) = 1$$

Example 1.5. More generally, an experiment with N possible outcomes with equal probability can be represented by $(\Omega, \mathcal{F}, P)$, where Ω is a finite set of N elements, $\mathcal{F}$ is the set of subsets of Ω, and $P(A) = |A|/N$ where $|\cdot|$ is the cardinality of finite set A.

Example 1.6. Similar probability spaces can be written down to represent

1. Fair dice
2. Multiple coin flips
3. Multiple dice rolls.
4. Random sampling of one object among finitely many objects with equal chances
5. Random sampling of multiple objects among finitely many objects, with and without put back, with or without order

Example 1.7. Sometimes Ω need to be infinite sets, for example

1. Infinitely many coin flips:

$$\Omega = \{\text{head}, \text{tail}\}^{\mathbb{N}}, P(\{\omega = (\omega_i) : \omega_k = c_k, k = 1, \dots, m\}) = 2^{-m}$$

2. Random point on interval $[0, 1]$, with uniform distribution

$$\Omega = [0, 1], P((a, b)) = b - a$$

3. Random point on a disc (or other shape with finite and non-zero volume), with uniform distribution

Remark 1.8. Axiom (iii) can be used to calculate the probability of an event, after one decomposes it into simpler events. For example, for the model of infinitely coin flips, the probability of getting a head at the $2k$-th flip for some $k \in \mathbb{N}$, is $1/4 + 1/4^2 + \cdots = 1/3$.

Below are some basic properties of probability:

Theorem 1.9.

1. $P(A) + P(A^c) = 1$
2. (Monotonicity) $A \subseteq B$ then $P(A) \leq P(B)$
3. (Inclusion-Exclusion) A_i, $i = 1, \dots n$ are events, then

$$P(A_1 \cup \cdots \cup A_n) = \sum_{j=1}^{n} \left((-1)^{j-1} \sum_{1 \leq i_1 < \cdots < i_j \leq n} P(A_{i_1} \cap \cdots \cap A_{i_j}) \right)$$

Proof. 1. $1 = P(\Omega) = P(A) + P(A^c)$.

2. $P(B) = P(A) + P(B \cap A^c) \geq P(A)$.

3. Induction on n.

□

Theorem 1.10. Let A_i be a sequence of events, $A_i \subseteq A_j$ if $i < j$, then

$$P\left(\bigcup_{i=1}^{\infty} A_i \right) = \lim_{n \to \infty} A_n$$

Proof.

$$P\left(\bigcup_{i=1}^{\infty} A_i \right) = P(A_1) + \sum_{i=1}^{\infty} P(A_{i+1} \backslash A_i)$$

Here $A_{i+1} \backslash A_i = A_{i+1} \cap A_i^c$ are events. Hence

$$P\left(\bigcup_{i=1}^{\infty} A_i \right) = \lim_{n \to \infty} \left(P(A_1) + \sum_{i=1}^{n-1} P(A_{i+1} \backslash A_i) \right) = \lim_{n \to \infty} A_n$$

□

Example 1.11. N persons with distinct names drawing their own names without put back, the number of possible outcomes is $N!$, all with same probability due to symmetry. Let A be the event that at least one person get their own name, A_i be the event that the i-th person gets their own name, then $A = \bigcup_i A_i$, and by inclusion-exclusion,

$$P(A) = N \times \frac{1}{N} - \binom{N}{2} \frac{1}{N(N-1)} + \cdots = \sum_{i=1}^{N} \frac{(-1)^{i-1}}{i!}$$

Which, as $N \to \infty$, converges to $1 - e^{-1}$.

1.2 Random Variables

1.5 of textbook

Definition 1.12. A **random variable** X on the probability space $(\Omega, \mathcal{F}, P)$ is a real valued function on Ω such that $X^{-1}((a, \infty)) \in \mathcal{F}$ for all $a \in \mathbb{R}$.

Example 1.13.

1. Let $(\Omega, \mathcal{F}, P)$ be the probability space that represents throwing a fair dice, $\Omega = \{1, 2, 3, 4, 5, 6\}$, X is defined as $X(n) = n$, then X is a random variable.

2. Let $(\Omega, \mathcal{F}, P)$ be the probability space that represents picking a real number uniformly at random from interval $(0, 1)$, then $\Omega = (0, 1)$, and $X(x) = x$ is a random variable.

Definition 1.14. When there are at most countable $k_i \in \mathcal{R}$ such that $P_X(\{k_i\}) = \sum_i P_X(\{k_i\}) = 1$, we say that X is a **discrete random variable**. The probability distribution P_X now depends completely on the **probability mass function (p.m.f)** $p(k_i) = P(X^{-1}(\{k_i\}))$. When there is only one k_i we say that X is called **degenerate** or **almost surely constant**.

Example 1.15.

1. Consider infinitely many fair coin flips (see Example 1.7), let X be the number of flips needed to get a head, then the pmf is $p(n) = 2^{-n}$, $n \in \mathbb{Z}$, $n > 0$.

2. Consider picking a number from open interval $(0, 1)$ uniformly at random, $X(x) = \lfloor 1/x \rfloor$, then the pmf is $p(n) = \frac{1}{n(n+1)}$, $n \in \mathbb{Z}$, $n > 0$.

Definition 1.16. Given a random variable X on a probability space $(\Omega, \mathcal{F}, P)$, one can define another probability space $(\mathbb{R}, \mathcal{B}, P_X)$, called the **probability distribution** of X, as follows:

$$P_X(A) = P(X^{-1}(A))$$

Here $\mathcal{B}$ is the **Borel σ-algebra**, which is the smallest σ-algebra containing all the intervals.

Remark 1.17.

1. One can show that $(\mathbb{R}, \mathcal{B}_X, P_X)$ satisfies Definition 1.1. In other words, the **probability distribution** is well defined.

2. When X is discrete, $P_X(A) = \sum_{k_i \in A} p(k_i)$.

2 Conditional Probability and Independence

2.1 Conditional Probability

2.1-2.2 of textbook

Definition 2.1. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $B \in \mathcal{F}$, $P(B) > 0$, then for any event A, the **conditional probability of A given B** is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Example 2.2. To see why this definition make sense, let's consider the probability space in Example 1.5, i.e.

$$\Omega = \{\omega_1, \ldots, \omega_N\}$$

$$\mathcal{F} = 2^{\Omega}$$

$$P(A) = |A|/N$$

Now suppose $P(B) = k/N > 0$, and we already know that B happened, then each of the k experimental outcomes in B are equally likely to happen, while the remaining experimental outcomes would definitly not happen. Hence,

$$P(\{\omega\}|B) = \begin{cases} 1/k & \omega \in B \\ 0 & \omega \notin B \end{cases}$$

and given this information, the probability that another event A would happen should be

$$\sum_{\omega \in A} P(\{\omega\}|B) = \frac{|A \cap B|}{k} = \frac{|A \cap B|/N}{k/N} = \frac{P(A \cap B)}{P(B)}$$

Theorem 2.3. If $(\Omega, \mathcal{F}, P)$ is a probability space, $P(B) > 0$, then so is $(\Omega, \mathcal{F}, P(\cdot|B))$.

Proof. We only need to check conditions (i)-(iii) in Definition 1.1:

(i) By Theorem 1.9 Part 2, $P(A \cap B) \leq P(B)$, by (i) of Definition 1.1, $P(A \cap B) \geq 0$, hence

$$0 \leq \frac{P(A \cap B)}{P(B)} = P(A|B) \leq 1$$

(ii)

$$P(\emptyset|B) = \frac{P(\emptyset \cap B)}{P(B)} = 0$$

$$P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = 1$$

(iii) If $\{A_n\}$ is a countable sequence of mutually disjoint events, so is $\{A_n \cap B\}$. Hence

$$P(\bigcup_n A_n|B) = \frac{P((\bigcup_n A_n) \cap B)}{P(B)} = \frac{P(\bigcup_n (A_n \cap B))}{P(B)}$$
$$= \sum_n \frac{P(A_n \cap B)}{P(B)} = \sum_n P(A_n|B)$$

□

Definition 2.4. Let $(\Omega, \mathcal{F}, P)$ be a probability space, a **Partition** is a finite set of pairwise disjoint events $B_1, \ldots, B_n$ whose union is Ω.

Example 2.5. Let $B \in \mathcal{F}$, then $\{B, B^c\}$ is a partition.

The followings are some basic properties of conditional probability, the proofs are all very straightforward.

Proposition 2.6. Let $(\Omega, \mathcal{F}, P)$ be a probability space, then:

1. If $A_1, \ldots, A_n \in \mathcal{F}$, then

   $$P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \ldots P(A_n|A_1 \cap \cdots \cap A_{n-1})$$

2. If $B_1, \ldots, B_n$ is a partition, $A \in \mathcal{F}$, then

   $$P(A) = \sum_i P(A|B_i)P(B)$$

   In particular,

   $$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

3. (Bayer's formula) If $B_1, \ldots, B_n$ is a partition, $A \in \mathcal{F}$, $P(A) > 0$, then

   $$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_j P(A|B_j)P(B_j)}$$

   In particular,

   $$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

Proof.

1. This is done by repeatedly applying $P(A \cap B) = P(A)P(B|A)$.
2. $A = \bigcup_i (A \cap B_i)$, and when $i \neq j$,

   $$(A \cap B_i) \cap (A \cap B_j) \subseteq B_i \cap B_j = \emptyset$$

   Hence

   $$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$$

3.
$$P(B_k|A) = \frac{P(B_k \cap A)}{P(A)} = \frac{P(B_k)P(A|B_k)}{\sum_i P(A|B_i)P(B_i)}$$

□

Remark 2.7. In the proposition above, the conditional probability might not be well defined, but the identities are still valid if we use the convention that 0 times something undefined equals 0.

Example 2.8. If the prevalence of a disease in the population 0.001, a test has false positive probability (the probability that the test result is positive while there is no disease) and false negative probability (the probability that the test result is negative while there is disease) 0.01, and a person is tested positive, then the probability that they actually have the disease can be calculated as follows:

Let A be the event that the person has the disease, B be the event that the person get tested positive, then the assumption becomes $P(A) = 0.001$, $P(B|A) = 0.99$, $P(B^c|A^c) = 0.99$. Hence

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \\ &= \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.01 \times 0.999} \approx 0.09 \end{aligned}$$

Basically, if we are testing for something really rare, the false positive rate shouldn't be too high, otherwise the positive tests will mostly come from false positives and not true positives.

Example 2.9. If one draws 2 balls at random, from a box with 6 balls of identical shape, 3 colored in red and 3 colored in green. And suppose we know that at least one of the two balls are red. The probability that the other ball is also red, given this information, would be

$$\frac{P(\text{Getting 2 red balls})}{P(\text{Getting at least one red ball})} = \frac{\frac{\binom{3}{2}}{\binom{6}{2}}}{1 - \frac{\binom{3}{2}}{\binom{6}{2}}} = \frac{1}{4}$$

If the first ball drawn is red, the probability that the second ball drawn is also red is $\frac{2}{5}$.

2.2 Independence

2.3, 2.5 of textbook

Definition 2.10. Let $(\Omega, \mathcal{F}, P)$ be a probability space.

1. $A, B \in \mathcal{F}$, we say that they are **independent** if $P(A \cap B) = P(A)P(B)$.

2. We say a sequence (or set) of events A_i are **mutually independent**, if for any $i_1 < \cdots < i_k$, $P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \dots P(A_{i_k})$.

Example 2.11.

1. Consider a roll of a symmetrical cubical dice, with numbers $1, \dots, 6$ on its faces. Then the event of getting an even number, and the event of getting a number smaller than 3, are independent.

2. In example 1.7 Part 1, the events "the n-th flip gets a head" are all mutually independent.

3. Let Ω be a square whose vertices are $(0,0), (1,0), (0,1), (1,1)$. Pick a point $p = (x, y)$ uniformly at random from S, then $x < 1/2$ and $1/3 < y < 2/3$ are independent.

4. Let D be the unit disc, which is a disc on $\mathbb{R}^2$ centered at origin and has radius 1. Pick a point $p = (x, y)$ uniformly at random from D. Then

   - $x > 0$ and $y > 0$ are independent.
   - $x > 0.75$ and $y > 0.75$ are not independent.

The followings are some basic properties of independence:

Theorem 2.12. Suppose A is an event, $P(A) = 0$ or 1

1. For any event B, A and B are independent.

2. For any set of events $\{B_n\}$ which are mutually independent, $\{A\} \cup \{B_n\}$ is a set of events that are mutually independent.

Proof. For Part 1, if $P(A) = 0$,

$$0 \leq P(A \cap B) \leq P(A) = 0$$

so

$$P(A \cap B) = 0 = P(A)P(B)$$

If $P(A) = 1$, then

$$P(A \cap B) = P(B) - P(A^c \cap B)$$

$$0 \leq P(A^c \cap B) \leq P(A^c) = 1 - P(A) = 0$$

Hence

$$P(A \cap B) = P(B) = P(A)P(B)$$

The proof for Part 2 is analogous. □

Theorem 2.13. If A and B are independent, so are A^c and B, A and B^c, A^c and B^c. If $\{A_i\}$ is a set of events that are mutually independent, let B_i be either A_i or A_i^c, then $\{B_i\}$ is a set of mutually independent events.

Proof. $P(A^c \cap B) = P(B) - P(A \cap B) = P(B)(1 - P(A)) = P(A^c)P(B)$. The remaining cases are analogous. □

Theorem 2.13 shows that independence is preserved by taking complements. Similarly, it is also preserved by taking countable disjoint unions:

Theorem 2.14. Let I be a non empty set, for each $i \in I$, let $\{A_{ij}\}$ be a set of disjoint events, such that for any choice of j_i, $\{A_{ij_i} : i \in I\}$ is a set of mutually independent events. Then $\{\bigcup_j A_{ij}\}$ is a set of mutually independent events.

Note that any set constructed from $A_1, \ldots, A_n \in 2^\Omega$ by finite union, finite intersection and complement can always be written as a finite disjoint union of the intersections of the various A_i and their complements. Hence, by Theorems 2.13, 2.14 and Definition 2.10 Part 2, we have

Theorem 2.15. Let $\{A_i\}$ be a set of mutually independent events. Let $\{B_j\}$ be a set of events, each B_j is constructed from elements in $\{A_i\}$ by complement, finite intersection and finite union, and no A_i appears in the expression of two different B_js. Then B_js are mutually independent.

Example 2.16. If an event A is independent from itself, then $P(A) = P(A \cap A) = P(A)^2$, hence $P(A) = 0$ or 1.

Definition 2.17. Let X_i be a sequence of random variables. We say X_i are mutually independent if for any Borel sets B_i in $\mathbb{R}$, the events $X_i \in B_i$ (which means $\{\omega \in \Omega : X_i(\omega) \in B_i\}$) are mutually independent. Or, equivalently, for any $x_i \in \mathbb{R}$, the events $X_i \leq x_i$ are mutually independent.

Remark 2.18. When all X_i are discrete random variables, they are mutually independent iff for any distinct $i_1, \ldots, i_k$, any real numbers $x_1, \ldots, x_k$, we have

$$P(X_{i_1} = x_1, \ldots, X_{i_k} = x_k) = \prod_{j=1}^{k} P(X_{i_j} = x_j)$$

Definition 2.19. Let $(\Omega, \mathcal{F}, P)$ be a probability space, $A_i \in \mathcal{F}$, $B \in \mathcal{F}$, $P(B) > 0$. We say A_i are mutually **conditionally independent** given B, if they are independent in $(\Omega, \mathcal{F}, P(\cdot|B))$.

Example 2.20. In Example 2.8, suppose we perform two consecutive tests, the results are independent conditioning on whether the patient has or does not have the disease. Let B_i, $i = 1, 2$, be the event that the i-th test is positive. Then if someone get tested positive twice, the probability that the person actually has the disease is

$$P(A|B_1 \cap B_2) = \frac{P(A)P(B_1|A)P(B_2|A)}{P(A)P(B_1|A)P(B_2|A) + P(A^c)P(B_1|A^c)P(B_2|A^c)}$$

$$= \frac{0.001 \times 0.99^2}{0.001 \times 0.99^2 + 0.999 \times 0.01^2} \approx 0.91$$

Also,

$$P(B_2) = P(B_1) = P(A)P(B_1|A) + P(A^c)P(B_1|A^c) = 0.01098$$

$$P(B_1 \cap B_2) = P(A)P(B_1|A)P(B_2|A) + P(A^c)P(B_1|A^c)P(B_2|A^c) = 0.00108$$

So without conditioning B_1 and B_2 are not independent.

Remark 2.21. As seen in the example above, conditioning can turn independence into dependence or dependence into independence.

2.3 Distributions from independent tests

2.4, 2.5 of textbook

The followings are some important discrete probability distributions (recall Definition 1.16):

1. **Bernoulli distribution**: It has pmf

   $$p(1) = p, p(0) = 1 - p$$

   where $p \in [0, 1]$ is a parameter. "X has Bernoulli distribution of parameter p" is denoted as $X \sim Ber(p)$

2. **Binomial distribution**: The sum of n random variables with distribution $Ber(p)$. The pmf is

   $$p(i) = \binom{n}{i} p^i (1-p)^{n-i}, i = 0, 1, \ldots, n$$

   "X has Binomial distribution of parameters n, p" is denoted as $X \sim Bin(n, p)$ (or $Binom(n, p)$).

3. **Geometric distribution**: It has pmf

   $$p(i) = (1-p)^{i-1} p, i = 1, 2, 3, \ldots$$

   "X has Geometric distribution of parameter p" is denoted as $X \sim Geom(p)$.

4. **Hypergeometric distribution**: Pick n balls at random from N_A red balls and $N - N_A$ blue balls, the number of red balls satisfies this distribution. The pmf is

   $$p(i) = \frac{\binom{N_A}{k}\binom{N-N_A}{n-k}}{\binom{N}{n}}, i = 0, \ldots, n$$

   "X has hypergeometric distribution of parameter N, N_A, n" is denoted as $X \sim Hypergeom(N, N_A, n)$.

Example 2.22.

1. Flip a coin once, the number of heads one gets is $Ber(1/2)$.
2. Flip a coin N times, the number of heads one get is $Bin(N, 1/2)$.
3. Flip a coin till we get head, the number of flips is $Geom(1/2)$.

Example 2.23. Suppose two person takes turns rolling a dice, the first person getting 4 wins. Then by the pmf of $Geom(1/6)$, the probability that the game ends at the k-th roll is

$$\left(\frac{5}{6}\right)^{k-1}\frac{1}{6}$$

hence the probability that the first player wins is

$$\sum_{n=0}^{\infty}\left(\frac{5}{6}\right)^{2n+1-1}\frac{1}{6}=\frac{6}{11}$$

If we change the rule, so that the first player wins if they get a 4 and the second player wins if they get 1 or 2, then the probability that the game ends after $2n+1$ rolls, where n is a non negative integer, equals

$$\left(\frac{5}{6}\right)^{n}\left(\frac{2}{3}\right)^{n}\frac{1}{6}$$

and the probability that the first player wins now becomes

$$\sum_{n=0}^{\infty}\left(\frac{5}{6}\right)^{n}\left(\frac{2}{3}\right)^{n}\frac{1}{6}=\frac{3}{8}$$

2.4 Some Classic Examples in Probability

Example 2.24. (Birthday Problem) Let p persons each come up with a natural number uniformly at random from 1 to n. The probability that none of them pick the same number is

$$\frac{n-1}{n}\cdot\frac{n-2}{n}\cdots\frac{n-p+1}{n}=\frac{n!}{n^p p!}$$

Example 2.25. (Monty Hall Problem) Person A is given three identical boxes, one with 1000 dollars, the other two empty. After A pick a box to open, B, **who knows where the money is**, opens one of the other boxes showing that it is empty. Should A keep the original choice or should A change?

1. Because the boxes are identical and A has no idea which has the money, the probability that the initial choice is correct is 1/3. Let C be the event that A made the right choice from the beginning.
2. If C happens, and A decides to change, then A gets nothing.

3. If C does not happen, and A decides to change, then A gets the money.

4. So the strategy of changing gets A money at probability 2/3.

If B **does not know where the money is**, and just opens one of the other boxes at random, which happens to be empty. Then:

1. The probability that the box B picked is empty is $1/3 \times 1 + 2/3 \times 1/2 = 2/3$.

2. The conditional probability that changing the box will get A the money is

$$\frac{P(\text{The box B picked is empty, and A gets the money})}{P(\text{The box B picked is empty})}$$

$$= \frac{P(\text{The box B picked is empty, and the initial box A picked is empty})}{P(\text{The box B picked is empty})}$$

$$= \frac{2/3 \times 1/2}{2/3} = 1/2$$

3 Random Variables

3.1 cdf, pmf and pdf

3.1, 3.2 of textbook

Definition 3.1. Let X be a random variable, the **cumulative distribution function** (c.d.f) is $F(s) = P(X \leq s)$.

Remark 3.2. By measure theory, the cdf uniquely determines the probability distribution of a real valued random variable.

Example 3.3. Suppose X is discrete with pmf $p(x_i) = P(X = x_1)$ (see Definition 1.14), then the cdf of X is $F(s) = \sum_{x_i \leq s} p(x_i)$.

Definition 3.4. If there is a real valued function f such that $F(s) = \int_{-\infty}^{s} f(t)dt$, we say that X is a **continuous random variable**, the function f is called the **probability density function** (p.d.f).

Example 3.5. Let x be a point picked uniformly at random from some interval $[a, b]$, $X = x$. Then the cdf of X is

$$F(s) = \begin{cases} 0 & s \leq a \\ \frac{x-a}{b-a} & a < s < b \\ 1 & s \geq b \end{cases}$$

$$= \int_{-\infty}^{s} \frac{\chi_{[a,b]}(t)}{b-a} dt$$

Here $\chi_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$, called the **characteristic function**. Hence X is a continuous random variable. We call the probability distribution of X the **Uniform Distribution**, denoted as $X \sim Unif[a, b]$.

From the definition and properties of probability spaces and random variables, we can deduce some elementary properties of the cdf:

Theorem 3.6. Let X be a random variable, F its cdf.

1. $s < s'$ implies $F(s) \leq F(s')$.
2. $\lim_{s \to -\infty} F(s) = 0$, $\lim_{s \to \infty} F(s) = 1$.
3. F is right continuous, i.e. $F(s) = \lim_{t \to s^+} F(t)$.

Proof.

1. $$F(s') = P(X \leq s') = P(X \leq s) + P(s < X \leq s') \geq P(X \leq s)$$

2. By Part 1 above, the limits exist. By Theorem 1.10,

$$\lim_{s\to\infty} F(s) = \lim_{n\to\infty} P(X \le n) = P(\Omega) = 1$$

$$\lim_{s\to-\infty} F(s) = 1 - \lim_{n\to\infty} P(X > -n) = 1 - P(\Omega) = 0$$

3. By Theorem 1.10,

$$1 - F(s) = P(X > s) = \lim_{n\to\infty} P\left(X > s + \frac{1}{n}\right)$$

Hence for any $\epsilon > 0$, there is some N such that if $n > N$,

$$P\left(X > s + \frac{1}{n}\right) > P(X > s) - \epsilon$$

Hence

$$F(s) \le F(s + 1/(N+1)) < F(s) + \epsilon$$

So for any $t \in (s, s + 1/(N+1))$, $|F(t) - F(s)| < \epsilon$, hence $F(s) = \lim_{t\to s^+} F(t)$.

□

The following is another consequence of Theorem 1.10:

Theorem 3.7. Let F be the cdf of random variable X, then $P(X < s) = \lim_{t\to s^-} F(t)$.

When the random variable X is continuous, we can get various properties on its cdf and pdf via properties of integrals we learned in calculus (definition, mean value theorem, fundamental theorem of calculus etc). For example, the follows can be shown by Theorem 3.6, 3.7 and properties of integrals:

Theorem 3.8. Let X be a continuous random variable, with cdf F and pdf f. Then

1. F is continuous.
2. $P(X = s) = P(X \le s) - P(X < s) = F(s) - \lim_{t\to s^-} F(t) = 0$
3. If f is continuous at a, $P(a < X < a + \epsilon) = \int_a^{a+\epsilon} f(t)dt = \epsilon f(a) + o(\epsilon)$.
4. If f_1 and f differs on only finitely many points, f_1 is also a pdf of X.
5. f can not be negative on an interval with positive length. Actually, f can always be chosen to be non negative.
6. $\int_{-\infty}^{\infty} f(t)dt = 1$.

Remark 3.9.

1. If a function F satisfies the conclusions of Theorem 3.6, it can be the cdf of a random variable.
2. If f is a non negative integrable function on $\mathbb{R}$, and $\int_{-\infty}^{\infty} f(t)dt = 1$ then f can be the pdf of some continuous random variable.
3. If p is a non negative function on a countable set $A \subseteq \mathbb{R}$, and $\sum_{a\in A} p(a) = 1$, then p is the pmf of some discrete random variable.

For continuous random variables, one can recover pdf from cdf as below:

Theorem 3.10. Furthermore, if the cdf F of some random variable X is continuous and **piecewise differentiable**, i.e. there are a_n, $a_{n-1} < a_n$ for all n, such that F is differentiable on the open interval (a_{n-1}, a_n) for all n, then X is a continuous random variable, with pdf $f = F'$.

Proof. This is an immediate consequence of the fundamental theorem of calculus. □

Example 3.11. Let X be a random variable with cdf $F(s) = \frac{1}{\pi}\arctan(x) + \frac{1}{2}$, then it is a continuous random variable with pdf $f(s) = \frac{1}{\pi(1+s^2)}$.

For some discrete random variables one can also recover pmf from cdf as follows:

Theorem 3.12. Let F be the cdf of a random variable X. If F is piecewise constant, i.e. there are a_n, $a_{n-1} < a_n$ for all n, such that F is constant on the open interval (a_{n-1}, a_n) for all n, then X is a discrete random variable, and the pmf is

$$p(a_n) = F(a_n) - \lim_{b\to a_n^-} F(b)$$

Remark 3.13.

- Not all discrete random variables have piecewise constant cdf. For example, let $\{q_i\}$, $i = 1, 2, \dots$ be a sequence going through all rational numbers without repetition, and let $p(q_i) = 2^{-i}$.
- Some random variables are neither discrete nor continuous. For example, if we flip a fair coin, and when we get tail pick a real number a uniformly at random from $[0, 1]$, and let

$$X = \begin{cases} 0 & \text{Got Head} \\ a & \text{Got Tail} \end{cases}$$

  Or, if we do infinitely many coin flips, let $X_i = 0$ if the i-th flip got tail, $X_i = 1$ if the i-th flip got head, and let $X = \sum_i 4^{-i} X_i$.

Example 3.14. Let X be a random variable with cdf F, $b, a > 0$ are real numbers, then $aX + b$ has cdf $s \mapsto F((s-b)/a)$. If X is continuous with pdf f, then the pdf of $aX + b$ is $s \mapsto \frac{f((s-b)/a)}{a}$. What if $a < 0$?

3.2 Expectation and Variance

3.3, 3.4 of textbook

Definition 3.15. Let X be a discrete random variable taking values at countably many x_i, i.e. $\sum_i P(X = x_i) = 1$, and suppose

$$\sum_i |x_i| P(X = x_i) < \infty$$

Then the **Expectation** of X is defined as

$$E(X) = \sum_i x_i P(X = x_i)$$

Definition 3.16.

1. Let X be a non negative random variable, the **Expectation** of X is defined as
$$E(X) = \sup_{X_1 \leq X, X_1 \text{ is discrete}} E(X_1)$$
where $E(X_1)$ is defined as in Definition 3.15.

2. Let X be a general random variable, the **Expectation** of X is defined as
$$E(X) = E(\max\{X, 0\}) - E(\max\{-X, 0\})$$
Where $E(\max\{X, 0\})$ and $E(\max\{-X, 0\})$ are defined as in Part 1 above.

The following follows from the definition of integration:

Theorem 3.17. If X is a continuous random variable with pdf f, and

$$\int_{-\infty}^{\infty} |t| f(t) dt < \infty$$

then

$$E(X) = \int_{-\infty}^{\infty} t f(t) dt$$

Example 3.18.

1. If $X \sim Ber(p)$, $E(X) = p$.
2. If $X \sim Geom(p)$, $E(X) = 1/p$.
3. If $X \sim Unif[a, b]$, $E(X) = \frac{a+b}{2}$.
4. Some random variables may not have expectations which are real numbers. For example,

(a) if X is a discrete random variable, with pmf $p(2^n) = 2^{-n}$, $n = 1, 2, \ldots$, we say $E(X) = \infty$

(b) if X is a continuous random variable with pdf $f(x) = \frac{1}{2(1+|x|)^2}$, we say $E(X)$ does not exist.

Remark 3.19. Let A be an event, the **Indicator random variable** I_A is defined as

$$I_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

Then $I_A \sim Ber(P(A))$, and $E(I_A) = P(A)$.

Remark 3.20. $E(X)$ is sometimes also denoted as $E[X]$. If both $E(\max\{X, 0\})$ and $E(\max\{-X, 0\})$ are infinite we say $E(X)$ is **undefined**, if only the former, or only the latter is infinite, we say the expectation is **positive infinity** or **negative infinity**, respectively.

Theorem 3.21.

1. If $X \geq 0$, $E(X) \geq 0$.
2. If $P(X = c) = 1$, $E(X) = c$.
3. If $E(X)$ exists, $a \in \mathbb{R}$, then $E(aX) = aE(X)$.
4. If $E(X)$ and $E(Y)$ exists, $E(X + Y) = E(X) + E(Y)$
5. If $X \leq Y$ then $E(X) \leq E(Y)$.

When X is discrete this follows from Definition 3.15. For example, to prove part 4, if X takes one of the x_i with probability 1, Y takes one of the y_i with probability 1, then

$$E(X + Y) = \sum_{i,j} (x_i + y_j) P(X = x_i, Y = y_j)$$

$$= \sum_i \left(x_i \sum_j P(X = x_i, Y = y_j) \right) + \sum_j \left(y_j \sum_i P(X = x_i, Y = y_j) \right)$$

$$= \sum_i x_i P(X = x_i) + \sum_j y_j P(Y = y_j) = E(X) + E(Y)$$

The proof for general X is by Definition 3.16.

Example 3.22. If $X \sim Bin(n, p)$ then $E(X) = np$.

Theorem 3.23. Let $g : \mathbb{R} \to \mathbb{R}$ be a function where the preimage of Borel sets are Borel sets (called Borel measurable), X is a random variable.

1. If X is discrete with pmf p, $g(X)$ is discrete, $E(g(X)) = \sum_i g(x_i)p(x_i)$.

2. If X is continuous with pdf f, $E(g(X)) = \int_{-\infty}^{\infty} g(t)f(t)dt$.

Part 1 follows from Definition 3.15, and Part 2 follows from the definition of Lebesgue integrals.

Definition 3.24. Let X be a random variable, $n > 0$ a positive integer. The n**th moment** of X is $E(X^n)$. When $E(X)$ exists, we define the **variance** of X as $Var(X) = E((X - E(X))^2)$.

Theorem 3.25. If X is a random variable, $E(X)$ and $Var(X)$ both exists, then $Var(X) = E(X^2) - (E(X))^2$

Proof.

$$X^2 = (X - E(X))^2 + 2E(X)X - (E(X))^2$$

Hence, by Theorem 3.21,

$$E(X^2) = Var(X) + (E(X))^2$$

The conclusion follows. □

Example 3.26. Consider a continuous random variable X with pdf

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geq 0 \end{cases}$$

Then $E(X^k) = k!\lambda^{-k}$, hence $Var(X) = 2\lambda^{-2} - (\lambda^{-1})^2 = \lambda^{-2}$. Such a random variable is said to satisfy the **Exponential Distribution**, denoted as $X \sim Exp(\lambda)$.

Theorem 3.27. If X is a random variable, $E(X)$ and $Var(X)$ both exists, then $Var(X) = 0$ iff X is degenerate, i.e. $P(X = E(X)) = 1$.

Proof.

- If X is degenerate, it is a discrete random variable, and its variance can be calculated via Theorem 3.23 Part 1.
- If X is not degenerate,

$$0 < P(X \neq E(X)) = P\left(\bigcup_{n=1}^{\infty} |X - E(X)| > \frac{1}{n}\right)$$

$$= \lim_{n\to\infty} P(|X - E(X)| > 1/n)$$

So there is some $N > 0$ such that $P(|X - E(X)| > 1/N) > 0$. Now let

$$Y = \begin{cases} N^{-2} & |X - E(X)| > 1/N \\ 0 & \text{otherwise} \end{cases}$$

then $(X - E(X))^2 \geq Y$, $Var((X - E(X))^2) \geq E(Y) > 0$.

□

Another immediate consequence of Theorem 3.21 is the following:

Theorem 3.28. Let X be a random variable with expectation and variance. Then $E(aX+b) = aE(X)+b$, $Var(aX+b) = a^2 Var(X)$.

Example 3.29. Let $X \sim Bin(n,p)$, then $Var(X) = np(1-p)$. To show this, let $X = \sum_{i=1}^{n} X_i$ where X_i are mutually independent random variables of distribution $Ber(p)$. Then

$$Var(X) = E\left(\left(\sum_{i=1}^{n} X_i\right)^2\right) - (E(X))^2$$

$$= \sum_{i=1}^{n} E(X_i^2) + 2\sum_{1\le i<j\le n} E(X_iX_j) - (np)^2$$

$$= np + n(n-1)p^2 - n^2p^2 = np(1-p)$$

Example 3.30. Let P be a point picked uniformly at random from the unit disc, and let X be the distance between P and the origin. Then the pdf is

$$f(s) = \begin{cases} 0 & s \le 0 \text{ or } s \ge 1 \\ 2s & 0 < s < 1 \end{cases}$$

The expectation is $2/3$ and variance is $1/18$.

3.3 Gaussian Distribution

3.5 of textbook

By multivariable calculus,

$$\int_{-\infty}^{\infty} e^{-t^2}dt = \left(\int_{\mathbb{R}^2} e^{-x^2-y^2}dxdy\right)^{1/2} = \left(\int_0^{\infty} 2\pi r e^{-r^2}dr\right)^{1/2}$$

$$= \sqrt{\pi}\left(\int_0^{\infty} e^{-s}ds\right)^2 = \sqrt{\pi}$$

Hence

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}dt = 1$$

Definition 3.31. A random variable X with pdf

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

is called a variable with **standard normal distribution** or **standard Gaussian distribution**, denoted as $X \sim \mathcal{N}(0,1)$.

Remark 3.32. If $X \sim \mathcal{N}(0,1)$, $E(X) = 0$, $Var(X) = 1$.

Definition 3.33. A random variable X is said to satisfy **the normal distribution with mean μ and variance σ^2**, if it has pdf

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

denoted as $X \sim \mathcal{N}(\mu, \sigma^2)$.

If $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$P(X \leq s) = \int_{-\infty}^{s} \frac{dx}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

If $a > 0$, b are real numbers, then

$$P(aX + b \leq s) = P\left(X \leq \frac{s-b}{a}\right) = \int_{-\infty}^{\frac{s-b}{a}} \frac{dx}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

$$= \int_{-\infty}^{s} \frac{dy}{\sqrt{2\pi\sigma^2 a^2}} e^{-(y-a\mu-b)^2/(2\sigma^2 a^2)}$$

So $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$. When $a < 0$ we can calculate this similarly, hence

Theorem 3.34. If a, b are real numbers, $a \neq 0$, $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

4 Normal and Poisson approximation of binomial distribution

4.1 Normal Approximation

4.1 of textbook

If $X \sim Bin(n,p)$, then by Examples 3.22 and 3.29, $E(X) = np$, $Var(X) = np(1-p)$, hence $\frac{X-np}{\sqrt{np(1-p)}}$ has expectation 0 and variance 1. Furthermore, we have the following theorem which would be proved later in the semester:

Theorem 4.1 (Binomial Central Limit Theorem)**.** Let $X \sim Bin(n,p)$, then as $n \to \infty$, the cdf of $\frac{X-np}{\sqrt{np(1-p)}}$ converges to the cdf of $\mathcal{N}(0,1)$.

Remark 4.2. One can further show that, via e.g. Berry-Esseen theorem, the error is bounded by $\frac{C}{\sqrt{np(1-p)}}$, where $C < 1$. Generally in practice when $np(1-p) > 10$ we can do normal approximation.

Remark 4.3. To write down the cdf of $Bin(n,p)$, we usually make use of the *continuity correction*:

$$F(s) \approx \Phi\left(\frac{\lfloor s \rfloor + \mathbf{1/2} - np}{\sqrt{np(1-p)}}\right)$$

Where Φ is the cdf of $\mathcal{N}(0,1)$, and $\lfloor x \rfloor$ is the largest integer no more than x.

An interactive plot of the normal approximation of binomial distribution, as well as an illustration of the continuity correction, can be found at `https://wuchenxi.github.io/binomclt.html`. The source code can be found in Appendix D.

4.2 Applications

Sections 4.2, 4.3 of textbook

1. Let $n \to \infty$ in Theorem 4.1, we get:

   Theorem 4.4 (Law of Large Numbers)**.** If $S_n \sim Bin(n,p)$, then for any $\epsilon > 0$, as $n \to \infty$, $P(|S_n/n - p| > \epsilon) \to 0$.

2. Theorem 4.1 can be used to provide the confidence interval of p. Here

   Definition 4.5. Let X be a random variable, such that its probability distribution $X(\theta)$ has an unknown parameter θ.

   (1) A **point estimate** is a function $\hat{\theta}$ from the range of X to $\mathbb{R}$.

(2) When X is discrete, the **maximal likelihood estimate** (MLE) is

$$\hat{\theta}(x) = \arg\max P(X_\theta = x) \text{ where } X_\theta \sim X(\theta)$$

(3) An **interval estimate** is a function from the range of X to open intervals in $\mathbb{R}$.

(4) We say an interval estimate $x \mapsto (a(x), b(x))$ is a **confidence interval** with *p-value* p, if for any θ, if $X \sim X(\theta)$, then

$$P(a(X) < \theta < b(X)) \geq 1 - p$$

In practice we often let $p = 0.05$.

Now let $X \sim Bin(n, \theta)$, where n is fixed, then

(a) $\hat{\theta}(x) = x/n$ is the MLE.

(b) Let $(x/n - r, x/n + r)$ be a confidence interval, then, when $n >> 1$, by normal approximation, for any θ, if $X \sim Bin(n, \theta)$, then

$$p > P(\theta \geq X/n + r \text{ or } \theta \leq X/n - r) \approx 2\Phi\left(\frac{-rn}{\sqrt{n\theta(1-\theta)}}\right)$$

$$\Phi\left(\frac{-rn}{\sqrt{n\theta(1-\theta)}}\right) \leq 2\Phi(-2r\sqrt{n})$$

When $p = 0.05$, $r \geq \frac{0.98}{\sqrt{n}}$.

3. When $N, N_A \to \infty$, $N_A/N \to p$, $0 < p < 1$, then hypergeometric distribution with parameters (N, N_A, n) converges to binomial distribution with parameters (n, p). This shows that we can use binomial distributions, and their normal approximations, to analyze data from opinion polls.

Remark 4.6. The cdf of $\mathcal{N}(0, 1)$ is not an elementary function, but it can be calculated numerically and is implemented by most programming languages. For example in Python standard library:

```
import math
def gaussian_cdf(t):
    return (1+math.erf(t/2**0.5))/2
print(gaussian_cdf(-1.96))
```

or in C standard library

```
#include <cstdio>
#include <cmath>
double cdf(double x){return (1+erf(x/sqrt(2)))/2;}
int main(){printf("%g\n", cdf(-1.96));return 0;}
```

Remark 4.7. In exams we will also provide tables for the values of this cdf. Note that in the tables we only have positive numbers, but due to symmetry, we have $\Phi(-s) = 1 - \Phi(s)$.

4.3 Poisson Distribution and Exponential distribution

Section 4.4 and 4.5 of textbook

Suppose we have a data center with a large amount of machines. To model the number of malfunctions within a day, we can divide the day into hours, minutes, seconds, milliseconds, etc, and find the number of hours, minutes, seconds, milliseconds, etc with a malfunction. As the size of the time unit gets smaller, it is increasingly unlikely that within that unit there are more than one malfunctions; and as the size of the unit decreases the probability that there is a malfunction within that time unit goes down as well and goes down approximately linearly. Hence, the number of malfunctions in a day can be described by the limiting distribution of $Bin(n, \lambda/n)$ as $n \to \infty$, here n is the number of time units in a day. The limiting pmf would now be

$$p(k) = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}$$

Definition 4.8. The discrete random variable with pmf $p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$, where $k = 0, 1, 2, \ldots$, is called the **Poisson distribution** with parameter λ. If X satisfies Poisson distribution with parameter λ, we denote it as $X \sim Poisson(\lambda)$.

By calculation, we know that

Proposition 4.9. If $X \sim Poisson(\lambda)$, $E(X) = Var(X) = \lambda$.

Proof.

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

$$Var(X) = E(X^2) - (E(X))^2 = \sum_{k=0}^{\infty} \frac{k^2\lambda^k e^{-\lambda}}{k!} - \lambda^2$$

$$= \sum_{k=2}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-2)!} + \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-1)!} - \lambda^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

□

Remark 4.10. Because the cdf of $Bin(n, p)$ and $Poisson(\lambda)$ are both of the form $F(s) = \sum_{0\le k\le s} p(s)$, the cdf of $Bin(n, \lambda/n)$ converges to the cdf of $Poisson(\lambda)$ as well.

An interactive visualization of the Poisson approximation of binomial distribution can be found at `https://wuchenxi.github.io/poisson.html`, source code see Appendix F.

If we want to model the time when the next malfunction happens starting at a specific time t_0, we can do as follows: divide the time into segments of $1/n$ days, when $n >> 1$, the probability that a malfunction happens in each segment is approximately λ/n. Then the number of time segment needed to get the first

malfunction satisfies $Geom(\lambda/n)$. Now we show that if $X_n \sim Geom(\lambda/n)$, then as $n \to \infty$, $X_n/n \to Y$ where $Y \sim Exp(\lambda)$ in distribution:

The cdf of X_n/n is

$$F_n(s) = \begin{cases} 0 & s < 0 \\ \frac{1}{n} \cdot \sum_{k=0}^{\lfloor sn \rfloor} \left(\frac{\lambda}{n}\right) \left(1 - \frac{\lambda}{n}\right)^k & s \geq 0 \end{cases}$$

And

$$\lim_{n\to\infty} \cdot \sum_{k=0}^{\lfloor sn \rfloor} \left(\frac{\lambda}{n}\right) \left(1 - \frac{\lambda}{n}\right)^k = \lim_{n\to\infty} \frac{\lambda}{n} \frac{1 - \left(1 - \frac{\lambda}{n}\right)^{\lfloor sn \rfloor + 1}}{1 - \left(1 - \frac{\lambda}{n}\right)} = 1 - e^{-\lambda s}$$

So the limiting cdf is the cdf of $Exp(\lambda)$.

Remark 4.11. If instead of the time of the first malfunction we want the times of all malfunctions, the result is called the **Poisson point process**.

A key property of exponential distributions is that they are **memoryless**:

Theorem 4.12. If $X \sim Exp(\lambda)$, then for any $s, t \geq 0$,

$$P(X > s + t | X > s) = P(X > t)$$

Proof.

$$P(X > s + t | X > s) = \frac{P(X > s + t \text{ and } X > s)}{P(X > s)} = \frac{P(X > s + t)}{P(X > s)}$$

$$= \frac{\int_{s+t}^{\infty} \lambda e^{-\lambda x} dx}{\int_{s}^{\infty} \lambda e^{-\lambda x} dx} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \int_{t}^{-\infty} \lambda e^{-\lambda x} dx = P(X > t)$$

□

Example 4.13. Suppose the accidents in a factory happens with a constant possibility regardless of time and they all happen independently, and if the average time between two consecutive accidents is one day, what's probability that there will be more than 2 accidents in a day? How about 2 days?

Answer: The number of accidents in a day satisfies $Poisson(\lambda)$ and the time between two consecutive accidents satisfies $Exp(\lambda)$, here λ is defined as

$$\lambda = \lim_{N\to\infty} NP(\text{there is an accident in a time period of } 1/N \text{ day})$$

so $\lambda = 1$, the probability is

$$\sum_{i=3}^{\infty} \frac{\lambda^i}{i!} e^{-1} = 1 - \frac{5}{2e}$$

Similarly, the number of accidents in 2 days satisfies $Poisson(2\lambda)$, and the calculation can be done similarly.

Remark 4.14. We say a sequence of random variables A_n **converges in distribution** to a random variable B, if the cdf of A_n converges to the cdf of B. Hence,

1. The normal approximation says that if $X_n \sim Bin(n,p)$, $Y_n = \frac{X_n - np}{\sqrt{n}}$, then as $n \to \infty$, $Y_n \to \mathcal{N}(0, p(1-p))$ in distribution.
2. The Poisson approximation says that if $X_n \sim Bin(n, \lambda/n)$, then $X_n \to Poisson(\lambda)$ in distribution (see Remark 4.10).

Remark 4.15. Let T_i be mutually independent random variables of distribution $Exp(\lambda)$, then the Poisson point process of parameter λ is $\{\sum_{i=1}^{n} T_i\}$, and the largest number n such that $\sum_{i=1}^{n} T_i < C$ is $Poisson(C\lambda)$.

5 Moment Generating Function

Sections 5.1 and 5.2 of textbook

Definition 5.1. Let X be a random variable. The **moment generating function** (MGF) of X is defined as

$$M_X(t) = E(e^{tX})$$

Remark 5.2. Let X be a random variable.

1. If X is discrete with pmf $p(a_i) = p_i$, then $M_X(t) = \sum_i p_i e^{ta_i}$.
2. If X is continuous with pdf f, then $M_X(t) = \int_{-\infty}^{\infty} e^{ts} f(s) ds$.

Remark 5.3. The moment generating function, when exists around a neighborhood of 0, encodes all the moment:

$$M_X(t) = \sum_{i=0}^{\infty} \frac{t^i E(X^i)}{i!}$$

The following Theorem is key to the proof of central limit theorem. The proof is beyond the scope of this course.

Theorem 5.4. Let X and Y be two random variables. Suppose $M_X = M_Y$ on a neighborhood of 0, then the cdf of X and Y are the same, which implies that they have the same distribution (we call them **equal in distribution**).

Remark 5.5. A way to prove the theorem above is as follows (assume knowledge of complex analysis): let $t \in \mathbb{C}$, under the assumption $E(e^{tX})$ add $E(e^{tY})$ both exists and are analytic on a uniform neighborhood of the imaginary axis. Since they are identical on an interval they are identical. Now recover cdf by inverse Fourier transform.

Example 5.6. The followings are the calculation of MGF of some common distributions:

1. $X \sim Bin(n, p)$, then

$$M_X(t) = \sum_{i=0}^{n} \binom{n}{i} e^{ti} p^i (1-p)^{n-i} = (1 - p + pe^t)^n$$

2. $X \sim Geom(p)$, then

$$M_X(t) = \sum_{i=1}^{\infty} e^{ti} p(1-p)^{i-1} = \frac{pe^t}{1 - (1-p)e^t}$$

3. $X \sim Poisson(\lambda)$, then

$$M_X(t) = \sum_{i=0}^{\infty} \frac{e^{ti}\lambda^i}{i!} e^{-\lambda} = e^{e^t\lambda - \lambda}$$

4. $X \sim Exp(\lambda)$, then

$$M_X(t) = \int_0^{\infty} \lambda e^{ts} e^{-\lambda s} ds = \frac{\lambda}{\lambda - t}$$

5. $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$M_X(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{ts} e^{-(s-\mu)^2/2\sigma^2} ds = e^{\mu t + \sigma^2 t^2/2}$$

Example 5.7. Let X be a continuous random variable with pdf f. Then e^{tX} is the composition of X and a smooth function $g(\cdot) = e^{t\cdot}$.

1. When $t > 0$, g is increasing, hence the cdf of e^{tX} is

$$F(s) = P(e^{tX} \leq s) = P(X \leq \log(s)/t) = \int_{-\infty}^{\log(s)/t} f(r)dr$$

when $s > 0$ and $F(s) = 0$ when $s \leq 0$. Hence the pdf is

$$f(s) = \frac{d}{ds} \int_{-\infty}^{\log(s)/t} f(r)dr = \frac{f(\log(s)/t)}{ts} = \frac{f(y)}{g'(y)}$$

where $s > 0$, $g(y) = t$.

2. When $t < 0$, g is decreasing, the cdf of e^{tX} is

$$F(s) = P(e^{tX} \leq s) = P(X \geq \log(s)/t) = \int_{\log(s)/t}^{\infty} f(r)dr$$

when $s > 0$ and $F(s) = 0$ when $s \leq 0$. Hence the pdf is

$$f(s) = \frac{d}{ds} \int_{\log(s)/t}^{\infty} f(r)dr = -\frac{f(\log(s)/t)}{ts} = \frac{f(y)}{|g'(y)|}$$

where $s > 0$, $g(y) = t$.

By similar computation as above, we have:

Theorem 5.8. Let X be a discrete random variable with pmf p, $g : \mathbb{R} \to \mathbb{R}$, then $g(X)$ is also a discrete random variable with pmf $q(s) = \sum_{g(a)=s} p(a)$.

Theorem 5.9. Let X be a continuous random variable with pdf f, $g : \mathbb{R} \to \mathbb{R}$ differentiable with finitely many critical points, then $g(X)$ is a continuous random variable, and its pdf equals

$$f_{g(X)}(s) = \sum_{g(a)=s} \frac{f(a)}{|g'(a)|}$$

If g is one-to-one with inverse function g^{-1}, then

$$f_{g(X)}(s) = \frac{f(g^{-1}(s))}{|g'(g^{-1}(s))|}$$

Here the sum of an empty set of numbers is assumed to be 0.

Example 5.10. Let X be a random variable with distribution $Unif(0, 2)$, $Y = \sin(X)$, then the pdf of Y is

$$f_Y(s) = \begin{cases} 0 & s > 1 \text{ or } s < 0 \\ \frac{1}{\sqrt{1-s^2}} & \sin(2) \leq s \leq 1 \\ \frac{1}{2\sqrt{1-s^2}} & 0 \leq s < \sin(2) \end{cases}$$

Example 5.11. Let $X \sim Unif(0, 1)$, F a real valued function on $\mathbb{R}$ satisfying the three properties in Theorem 3.6, let

$$g(x) = \inf\{s : F(s) \geq x\} = \min\{s : F(s) \geq x\}$$

Then the cdf of $g(X)$ is then

$$F_1(s) = P(g(X) \leq s) = P(X \leq F(s)) = F(s)$$

Here, the second equality is because F is non decreasing and right continuous.

1. If F is the cdf of some discrete random variable Y with pmf $p(a_i) = p_i$, then $g(x) = \inf\{a_i : \sum_{a_j \leq a_i} p_j \geq x\}$.
2. If F is the cdf of some continuous random variable Y with positive pdf f, then $g(x) = F^{-1}(x)$, and by Theorem 5.9 the pdf of $g(X)$ is indeed f.

6 Joint Distributions

6.1 Definition and Examples

Section 6.1 and 6.2 of textbook

Recall that if X defined on a probability space $(\Omega, \mathcal{F}, P)$ is a random variable, the probability distribution of X is the probability space

$$(\mathbb{R}, \mathcal{B}, A \mapsto P(X \in A))$$

where $\mathcal{B}$ is the Borel σ-algebra.

Definition 6.1. Let $X_1, \ldots, X_n$ be random variables. The **joint distribution** is the probability space

$$(\mathbb{R}^n, \mathcal{B}, A \mapsto P((X_1, \ldots, X_n) \in A)$$

where $\mathcal{B}$ is the Borel σ algebra. The distribution of each X_i is called the **marginal distribution**.

Remark 6.2. Just like probability distribution of random variables are determined by their cdfs, joint distributions are determined by the joint cdf

$$F(a_1, \ldots, a_N) = P(X_1 \leq a_1, \ldots, X_n \leq a_n)$$

Definition 6.3. Let $X_1, \ldots, X_n$ be discrete random variables, $A_i \subseteq \mathbb{R}$ countable sets such that $P(X_i \in A_i) = 1$, then the **joint pmf** is defined as $p(a_1, \ldots, a_n) = P(X_1 = a_1, \ldots, X_n = a_n)$, where $a_i \in A_i$.

The following facts are straightforward:

Theorem 6.4. Let $X_1, \ldots, X_n$ be discrete random variables, $A_i \subseteq \mathbb{R}$ countable sets such that $P(X_i \in A_i) = 1$, and p be the joint pmf. Then

1. The probability distribution is now
$$P((X_1, \ldots, X_n) \in B) = \sum_{(a_1, \ldots, a_n) \in B, a_i \in A_i \text{ for all } i} p(a_1, \ldots, a_n)$$
2. The marginal distributions are
$$P(X_j \in B) = \sum_{a_j \in B \cap A_j} \left(\sum_{a_i \in A_i \text{ for all } i \neq j} p(a_1, \ldots, a_n) \right)$$
3. The pmf of X_j, which is also called the **marginal pmf**, equals
$$p_{X_j}(a_j) = \sum_{a_i \in A_i \text{ for all } i \neq j} p(a_1, \ldots, a_n)$$

4. Let $g : \mathbb{R}^n \to \mathbb{R}$ be a function, then

$$E(g(X_1, \ldots, X_n)) = \sum_{a_i \in A_i \text{ for all } i} g(a_1, \ldots, a_n) p(a_1, \ldots, a_n)$$

Definition 6.5. Consider an experiment with r outcomes, with probabilities $p_1, \ldots, p_r$, and $p_1 + \cdots + p_r = 1$. Now carry out experiment n times, the number of each of the r outcomes are denoted as $X_1, \ldots, X_r$, then their joint distribution is called the **multinomial distribution**, denoted as

$$(X_1, \ldots, X_n) \sim Multi(n, r, p_1, \ldots, p_r)$$

Where the joint pmf is

$$p(k_1, \ldots, k_r) = \begin{cases} \frac{n!}{k_1! \ldots k_r!} p_1^{k_1} \ldots p_r^{k_r} & k_1 + \cdots + k_r = n, k_i \in \mathbb{Z}_{\geq 0} \text{ for all } i \\ 0 & \text{otherwise} \end{cases}$$

Definition 6.6. We say $X_1, \ldots, X_n$ are called **jointly continuous**, if there is a function f on $\mathbb{R}^n$, called the **joint density function**, such that

$$P((X_1, \ldots, X_n) \in B) = \int_B f(s_1, \ldots, s_n) ds_1 ds_2 \ldots ds_n$$

Definition 6.7. Let $A \subseteq \mathbb{R}^n$ be a region with finite volume. We say $X_1, \ldots, X_n$ are **uniformly distributed on** A (see Example 1.7 Part 3) if they are jointly continuous with density function $f(s_1, \ldots, s_n) = \frac{\chi_A(s_1, \ldots, s_n)}{Vol(A)}$, where

$$\chi_A(s_1, \ldots, s_n) = \begin{cases} 1 & (s_1, \ldots, s_n) \in A \\ 0 & \text{otherwise} \end{cases}$$

Remark 6.8. X and Y both being continuous does not imply that they are jointly continuous. For example, if $X \sim Unif(0, 1)$ and $Y = X$.

Example 6.9. Let Ω be the triangle with vertices $(0, 0)$, $(1, 0)$, $(0, 1)$, $(X, Y) \sim Unif(\Omega)$. Then the cdf of X is

$$F(s) = \begin{cases} 0 & s \leq 0 \\ 1 & s \geq 1 \\ 1 - (1 - s)^2 & 0 < s < 1 \end{cases}$$

and the corresponding pdf is

$$f(s) = \begin{cases} 0 & s \leq 0 \text{ or } s \geq 1 \\ 2(1 - s) & 0 < s < 1 \end{cases}$$

Theorem 6.10. Let $X_1, \ldots, X_n$ be jointly continuous with joint density function f. Then each X_i is continuous, with pdf (called **marginal density function**)

$$f_{X_i}(s) = \int_{\mathbb{R}^{n-1}} f(s_1, \ldots, s_{i-1}, s, s_{i+1}, \ldots, s_n) ds_1 \ldots ds_{i-1} ds_{i+1} \ldots ds_n$$

Proof. By Fubini,

$$F_{X_i}(s) = P(X_i \le s)$$

$$= \int_{-\infty}^{s} dt \int_{\mathbb{R}^{n-1}} f(s_1, \dots, s_{i-1}, t, s_{i+1}, \dots, s_n) ds_1 \dots ds_{i-1} ds_{i+1} \dots ds_n$$

Take $\frac{d}{ds}$ we get the desired formula. □

This is a generalization of Theorem 3.23:

Theorem 6.11. Let g be a measurable function, $X_1, \dots, X_n$ be jointly continuous with joint density function f, then

$$E[g(X_1, \dots, X_n)] = \int_{\mathbb{R}^n} g(s_1, \dots, s_n) f(s_1, \dots, s_n) ds_1 \dots ds_n$$

Example 6.12. Let X, Y be uniformly distributed on the unit disc. $Z = \max\{X, Y\}$ Then

1. The marginal pdfs are

$$f_X(s) = f_Y(s) = \begin{cases} 0 & s > 1 \text{ or } s < -1 \\ \frac{2}{\pi}\sqrt{1 - s^2} & -1 \le s \le 1 \end{cases}$$

2.

$$E[Z] = \frac{1}{\pi} \left(\int_{x^2+y^2<1, x>y} x dx dy + \int_{x^2+y^2<1, y>x} y dx dy \right)$$

$$= \frac{1}{\pi} \int_0^1 r dr \left(\int_{-3\pi/4}^{\pi/4} r\cos(\theta) d\theta + \int_{\pi/4}^{5\pi/4} r\sin(\theta) d\theta \right) = \frac{2\sqrt{2}}{3\pi}$$

Remark 6.13. The analogy of Remark 3.9 for multiple random variables is also true.

1. Let $A \subseteq \mathbb{R}^n$ be a countable subset, $p : A \to \mathbb{R}$ a non negative function, $\sum_{a \in A} p(a) = 1$. Then there are discrete random variables $X_1, \dots, X_n$ with joint pmf p.

   To show this, consider probability space $(A, 2^A, B \mapsto \sum_{b \in B} p(b))$, and X_i are defined as $a = (a_1, \dots, a_n) \mapsto a_i$.

2. Let f be a non negative integrable function on $\mathbb{R}^n$, then $\int_{\mathbb{R}^n} f dx_1 \dots dx_n = 1$ iff there are random variables $X_1, \dots, X_n$ which are jointly continuous with joint pdf f.

   To show the "only if" part, consider probability space

   $$(\mathbb{R}^n, \mathcal{B}, B \mapsto \int_B f dx_1 \dots dx_n)$$

   and X_i are defined as $(x_1, \dots, x_n) \mapsto x_i$. The "if" part follows from the fact that $P(\Omega) = 1$.

6.2 Independence

Section 6.3 of textbook

Remark 2.18 can be restated as

Theorem 6.14. Let $X_1, \ldots, X_n$ be discrete, with pmf $p_1, \ldots, p_n$ respectively, then they are mutually independent iff their joint pmf is

$$p(a_1, \ldots, a_n) = p_1(a_1) \ldots p_n(a_n)$$

Similarly, for continuous random variables, we have:

Theorem 6.15. Let $X_1, \ldots, X_n$ be continuous random variables with pdf $f_1, \ldots, f_n$ respectively. Then they are mutually independent iff they are jointly continuous with joint pdf

$$f(s_1, \ldots, s_n) = f_1(s_1) \ldots f_n(s_n)$$

Proof. The "if" part follows from Fubini's theorem. To show the "only if", note that by measure theory, a Borel measure on $\mathbb{R}^n$ is completely determined by its value on sets of the form $A_1 \times \cdots \times A_n$. □

An immediate consequence of Theorem 2.13 is:

Theorem 6.16. Let X_i, $i \in I$ be a set of mutually independent random variables, Y_j are obtained by composing the various X_i with some Borel measurable functions, such that no X_i appears in the expression of two distinct Y_js, then the Y_js are mutually independent as well.

Example 6.17. Let X, Y be random variables with pdf f and g and cdf F and G respectively. $Z = \min\{X, Y\}$

1. The joint distribution function is $f_{X,Y}(s, s') = f(s)g(s')$.
2. The cdf of Z is

$$F_Z(s) = 1 - P(X > s, Y > s) = 1 - \int_s^\infty f(r)dr \int_s^\infty g(r)dr$$

   and the pdf is

$$f_Z(s) = \frac{d}{ds} F_Z(s) = f(s) \int_s^\infty g(r)dr + g(s) \int_s^\infty f(r)dr$$

$$= f(s)(1 - G(s)) + g(s)(1 - F(s))$$

3. Let

$$I = I_{X \geq Y} = \begin{cases} 1 & X \geq Y \\ 0 & X < Y \end{cases}$$

the pmf is

$$P(I=1)=\int_{s\geq t} f(s)g(t)dsdt=\int_{-\infty}^{\infty} f(s)G(s)ds$$

$$P(I=0)=\int_{s<t} f(s)g(t)dsdt=\int_{-\infty}^{\infty} F(t)g(t)dt$$

4. From definition 2.17, Z and I are independent iff

$$P(Z\in B)P(I=1)=P(Z\in B, I=1)$$

which means that

$$\int_{s\in B}\left(f(s)(1-G(s))+g(s)(1-F(s))\right)ds\cdot\int_{-\infty}^{\infty} f(s)G(s)ds=$$

$$\int_{t\in B, s\geq t} f(s)g(t)dsdt=\int_{t\in B} g(t)(1-F(t))dt$$

which is equivalent to the fact that there is a constant C independent of s, such that

$$f(s)(1-G(s))=Cg(s)(1-F(s))$$

So, for example, if F and G are both exponential distributions, this would be true.

Example 6.18. If X and Y are independent, and both are standard normal, let $X=r\cos(\theta)$, $Y=r\sin(\theta)$, where $r\geq 0$, $\theta\in[0,2\pi)$, then r and θ are independent and $r^2\sim Exp(1/2)$, $\theta\sim Unif[0,1)$. To show this one can make use of Definition 2.17 as in the previous example. This gives a way to generate normal distribution from uniform distribution without calculating inverse of the pdf of normal distribution, and is called the **Box-Muller algorithm**.

6.3 Sums and convolutions

Section 7.1 of textbook

The following follows from Theorem 6.14 and 6.15 and Fubini:

Theorem 6.19. Let X and Y be two independent random variables, $Z=X+Y$.

1. If both X and Y are discrete, with pmf p_X and p_Y respectively, then so is Z. The pmf of Z is

$$p(x)=\sum_a p_X(a)p_Y(x-b)$$

2. If X and Y are both continuous with pdf f_X and f_Y, then Z is continuous with pdf
$$f_Z(s) = \int_{-\infty}^{\infty} f_X(t) f_Y(s-t) dt$$
which is also called the convolution of f_X and f_Y, denoted as $f_X * f_Y$.

Example 6.20. X and Y are independent, $Z = X + Y$.

1. $X \sim Bin(n,p)$, $Y \sim Bin(m,p)$, then $Z \sim Bin(m+n,p)$.
2. $X \sim Poisson(\lambda)$, $Y \sim Poisson(\mu)$, then $Z \sim Poisson(\lambda+\mu)$.
3. $X \sim Geom(p)$, $Y \sim Geom(p)$, Z has pmf $p_Z(n) = (n-1)p^2(1-p)^{n-2}$.
4. $X \sim \mathcal{N}(\mu, \sigma^2)$, $Y \sim \mathcal{N}(\mu', \sigma'^2)$, then $Z \sim (\mu+\mu', \sigma^2+\sigma'^2)$. [1]
5. $X \sim Exp(\lambda)$, $Y \sim Exp(\lambda)$, Z has pdf
$$f_Z(s) = \begin{cases} 0 & s \le 0 \\ \lambda^2 s e^{-\lambda s} & s > 0 \end{cases}$$

Definition 6.21. The sum of n mutually independent random variables of distribution $Geom(p)$ is said to satisfy the (n,p)**-negative binomial distribution**, denoted as $Z \sim Negbin(n,p)$. Its pmf is
$$p(k) = \binom{k-1}{n-1} p^n (1-p)^{k-n}$$
where $k \ge n$ are integers. To see this, when $n = 1$ this is the same pmf as $Geom(p)$. If X and Y are independent and with pmf
$$p_X(k) = \binom{k-1}{n-2} p^{n-1}(1-p)^{k-n+1}$$

[1]
$$\begin{aligned} f_Z(s) = (f_X * f_Y)(s) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma'^2}} e^{-\frac{(s-t-\mu')^2}{2\sigma'^2}} dt \\ &= \frac{1}{2\pi\sigma\sigma'} \int_{-\infty}^{\infty} e^{-\frac{\sigma'^2(t-\mu)^2 + \sigma^2(s-t-\mu')^2}{2\sigma^2\sigma'^2}} dt \\ &= \frac{1}{2\pi\sigma\sigma'} \int_{-\infty}^{\infty} e^{-\frac{\left(t - \frac{\sigma'^2\mu + \sigma^2(s-\mu')}{\sigma^2+\sigma'^2}\right)^2 - \frac{(\sigma'^2\mu+\sigma^2(s-\mu'))^2}{(\sigma^2+\sigma'^2)^2} + \frac{\sigma'^2\mu^2+\sigma^2(s-\mu')^2}{\sigma^2+\sigma'^2}}{\frac{2\sigma^2\sigma'^2}{\sigma^2+\sigma'^2}}} dt \\ &= \frac{1}{2\pi\sigma\sigma'} \cdot \sqrt{2\pi \frac{\sigma^2\sigma'^2}{\sigma^2+\sigma'^2}} e^{-\frac{(\sigma'^2\mu^2+\sigma^2(s-\mu')^2)(\sigma^2+\sigma'^2) - (\sigma'^2\mu+\sigma^2(s-\mu'))^2}{2\sigma^2\sigma'^2(\sigma^2+\sigma'^2)}} \\ &= \frac{1}{\sqrt{2\pi(\sigma^2+\sigma'^2)}} e^{-\frac{(s-\mu-\mu')^2}{2(\sigma^2+\sigma'^2)}} \end{aligned}$$

where $k \geq n-1$, and

$$p_Y(k) = p(1-p)^{k-1}$$

where $k \geq 1$, then the pmf of $Z = X + Y$ is

$$p_Z(k) = \sum_{n-1 \leq k' \leq k-1} \left(\binom{k'-1}{n-2} p^{n-1}(1-p)^{k'-n+1} \cdot p(1-p)^{k-k'-1} \right)$$

$$= \binom{k-1}{n-1} p^n (1-p)^{k-n}$$

Definition 6.22. The sum of n mutually independent random variables of distribution $Exp(\lambda)$ is said to satisfy the (n, λ)**-Gamma distribution** (see Remark 4.15), denoted as $Z \sim Gamma(n, \lambda)$. Its pdf is

$$f(s) = \begin{cases} 0 & s \leq 0 \\ \frac{\lambda^n s^{n-1}}{(n-1)!} e^{-\lambda s} & s > 0 \end{cases}$$

6.4 Exchangability, iid

Section 7.2 of textbook

Definition 6.23. Let $X_1, \ldots, X_n$ be random variables. If for any bijection $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$, $X_1, \ldots, X_n$ and $X_{\sigma(1)}, \ldots, X_{\sigma n}$ have the same distribution, we say these random variables are **exchangable**.

These follow immediately from definition:

Theorem 6.24.

1. If $X_1 = X_2 = \cdots = X_n$, then they are exchangable.
2. If X_i are all discrete, then they are exchangable iff the joint pmf is a symmetric function.
3. If they are jointly continuous, they are exchangable iff the joint pdf is a symmetric function.

Theorem 6.25. If $X_1, \ldots, X_n$ are exchangable.

1. The marginal distributions of X_i are all identical.
2. Let $G : \mathbb{R}^n \to \mathbb{R}$ be any Borel measurable function, then for any permutation σ, $E(G(X_1, \ldots, X_n)) = E(G(X_{\sigma(1)}, \ldots, X_{\sigma(n)}))$.
3. Let $g : \mathbb{R} \to \mathbb{R}$ be any Borel measurable function, then $g(X_1), \ldots, g(X_n)$ are exchangable.

Example 6.26. Let N identical balls be each labeled a number, draw n balls without replacement from these balls, let X_i be the label of the ball obtained in the i-th draw. Then X_i are not mutually independent, but they are still exchangable.

As an application, we have:

Example 6.27. Suppose there are 10 red balls and 10 green balls, draw 10 from the 20 without replacement, then the probability that we get green at the 6th draw, red at the 9th, equals the probability that we get green at the first draw, red at the second, which equals $10/20 \times 10/19 = 5/19$. Similarly, the conditional probability of the 6th being green given that the 9th is red is $10/19$.

Example 6.28. If $X_1, \ldots, X_n$ are mutually independent with the same distribution (called **independent and identically distributed**, or **i.i.d.**, then they are exchangable.

As an application, we have:

Example 6.29. Let X, Y, Z be i.i.d. continuous random variables. Then by Theorem 6.15, they are jointly continuous, hence the probability that any two of the three are identical equals 0. Because they are exchangable,

$$P(X > Y, X > Z) = P(Y > X, Y > Z) = P(Z > X, Z > Y) = 1/3$$

Similarly, the probability that $X < Y < Z$ is $1/6$.

6.5 Indicator Method

Section 8.1 of textbook

By Remark 3.19, given any event A, its probability is the expectation of the indicator random variable I_A. Now we can make use of the properties of the expectation of random variables to calculate $P(A)$, or use $P(A)$ to calculate $E[I_A]$. This is called the **indicator method**.

Remark 6.30. Indicator method gives a simple proof for the inclusion-exclusion formula.

Example 6.31. Suppose there are 20 identical balls, 5 red 15 green. Draw 8 randomly without replacement, what's the expectation of the number of red balls drawn?

Answer: Let A_j be the event "the j-th ball drawn is red". Then the expectation of number of red balls drawn is

$$E\left(\sum_{i=1}^{8} I_{A_i}\right) = \sum_{i=1}^{8} E(I_{A_i}) = 8 \cdot 1/4 = 2$$

Example 6.32. Roll a dice 100 times, what's the expected number of sequences of consecutive 4s with length exactly 4?

Answer: Let A_j, $j = 1, \ldots, 97$, be the event that "there is a sequence of consecutive 4s with length exactly 4 beginning at the j-th roll". Then the required expectation is

$$E\left[\sum_{i=1}^{97} I_{A_i}\right] = 2 \cdot (5/6) \cdot (1/6)^4 + 95 \cdot (5/6)^2 \cdot (1/6)^4$$

Example 6.33. Suppose there are n individuals, whether or not any two of them know one another are independent and has probability $1/2$. Then the expected number of groups of k ppl where any two in the group know each other, would be, due to the same argument as the previous example, $\binom{n}{k}2^{-k}$.

6.6 Expectation of Products

Section 8.2 and 8.3 of textbook

Theorem 6.34. Let X_i, $i = 1, \ldots, n$, be mutually independent random variables, then $E[X_1 \ldots X_n] = E[X_1] \ldots E[X_n]$.

The case when X_i are discrete (or continuous) can be proved by computation. The general case can be shown via Definition 3.16.

Combining the above with Theorem 6.16, we have:

Theorem 6.35. If g_i are functions where preimages of Borel sets are Borel, X and Y be independent random variables, then

$$E[g_1(X_1) \ldots g_n(X_n)] = E[g_1(X_1)] \ldots E[g_n(X_n)]$$

As applications, we have the following:

Theorem 6.36. Let $X_1, \ldots, X_n$ be mutually independent random variables. $Z = \sum_i X_i$.

1. If their variances are $\sigma_1^2, \ldots, \sigma_n^2$ respectively, then the variance of Z is $\sum_i \sigma_i^2$.
2. If their MGFs are $M_1(t), \ldots, M_n(t)$ respectively, the MGF of Z is $\prod_i M_i(t)$.

Proof. Part 2 above follows immediately from Theorem 6.16. To show part 1,

$$Var(Z) = E((Z - E(Z))^2)$$

$$= E\left(\left(\sum_i X_i - \sum_i E(X_i)\right)^2\right) = E\left(\left(\sum_i (X_i - E(X_i))\right)^2\right)$$

$$= \sum_i E((X_i - E(X_i))^2) + \sum_{i \neq j} E((X_i - E(X_i))(X_j - E(X_j)))$$
$$= \sum_i \sigma_i^2 + \sum_{i \neq j} E(X_i - E(X_i))E(X_j - E(X_j)) = \sum_i \sigma_i^2$$

□

Example 6.37. Let $X_1, \ldots, X_n$ be iid random variables, each with expectation μ and variance σ^2.

1. Let $\overline{X} = \frac{1}{n}\sum_i X_i$, which we call the **sample mean**. Then $E(\overline{X}) = \mu$ (which we say "$\overline{X}$ is an **unbiased estimator** of μ"), $Var(\overline{X}) = \frac{\sigma^2}{n}$.

2. Let $s_n^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$ be called the **sample variance**. Then

$$E(s_n^2) = \frac{1}{n-1}\sum_i E\left(\left(\frac{n-1}{n}X_i - \sum_{j \neq i}\frac{1}{n}X_j\right)^2\right)$$
$$= \frac{n}{n-1}\left(\left(\frac{(n-1)^2}{n^2} + (n-1)\cdot\frac{1}{n^2}\right)E(X_i^2)\right.$$
$$\left. - \left(\frac{2\cdot(n-1)\cdot(n-1)}{n^2} + \frac{(n-1)(n-2)}{n^2}\right)E(X_i)^2\right) = \sigma^2$$

So s_n^2 is an unbiased estimator of σ^2.

Example 6.38. Parts 1, 2, and 4 of Example 6.20 follows from Part 2 of Theorem 6.36 and Theorem 5.4.

Example 6.39. Suppose an experiment has n equally likely outcomes. Let T_n be the number of independent experiments needed to get all n outcomes. Now let S_k be the number of experiments needed to go from getting $k-1$ outcomes to getting k outcomes, we have $T_n = \sum_{i=1}^n S_k$, the S_k are all independent, and $S_k \sim Geom\left(\frac{n-k+1}{n}\right)$.

So

$$E(T_n) = \sum_{k=1}^n E(S_k) = n\sum_{j=1}^n \frac{1}{j}$$

$$Var(T_n) = \sum_{k=1}^n \frac{(k-1)n^2}{n(n-k+1)^2} = n\sum_{j=1}^{n-1}\frac{n-j}{j^2}$$

6.7 Covariance and Correlation

Section 8.4 of the textbook

Definition 6.40. Let X and Y be two random variables on the same probability space. The **covariance** between them is

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

The **correlation** between them is

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

When $Cov(X,Y) > 0$ we say they are positively correlated. When $Cov(X,Y) < 0$ we say they are negatively correlated. When $Cov(X,Y) = 0$ we say they are uncorrelated.

The followings are some of their basic properties:

Theorem 6.41. Let X, X_i, Y, Y_i be random variables on the same probability space, $a_i, a, b_i, b \in \mathbb{R}$.

1. $Cov(X,Y) = E(XY) - E(X)E(Y)$
2. $Cov(X,Y) = Cov(Y,X)$
3. $Cov(a,X) = Cov(X,a) = 0$, $Cov(X,X) = Var(X)$
4. $Cov(\sum_i a_i X_i, \sum_i b_i Y_i) = \sum_{i,j} a_i b_j Cov(X_i, Y_j)$
5. $Var(\sum_i a_i X_i) = \sum_i a_i^2 Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$
6. $(Cov(X,Y))^2 \leq Var(X)Var(Y)$, and if $Var(X) \neq 0$, they are equal iff $Y = aX + b$ a. s. (aka. with probability 1) for some a, b.
7. $|Corr(X,Y)| \leq 1$. If $Var(X) \neq 0$, $Corr(X,Y) = 1$ iff $Y = aX + b$ a. s. for some $a > 0$
8. When $a > 0$, $Corr(aX + b, Y) = Corr(X,Y)$; when $a < 0$, $Corr(aX + b, Y) = -Corr(X,Y)$

Proof. Parts 2 and 3 follows from definition. Part 1 is due to

$$Cov(X,Y) = E((X - E(X))(Y - E(Y)))$$

$$= E(XY) - E(E(X)Y) - E(E(Y)X) - E(E(X)E(Y)) = E(XY) - E(X)E(Y)$$

Part 4 follows from Part 1 and the fact that expectation is linear. Part 5 follows from Part 3 and 4. Part 6 is due to the fact that for any $t \in \mathbb{R}$,

$$E((t(X - E(X)) - (Y - E(Y)))^2) \geq 0$$

Hence

$$t^2Var(X) - 2tCov(X,Y) + Var(Y) \geq 0$$

And the inquality follows. When equal sign is reached, there is some t such that

$$0 = E(t((X - E(X)) - (Y - E(Y)))^2) = Var(tX - Y)$$

Hence by Theorem 3.27, $Y = tX + C$ with probability 1. Part 7 follows from Part 6, Part 8 is due to Parts 2 and 3. □

Example 6.42. Let $X = I_A$, $Y = I_B$, then they are uncorrelated iff A and B are independent.

Example 6.43. Let $\Omega = \{(x,y) : |x| < 1, |y| < 1, |y| < |x|\}$, $X, Y \sim Unif(\Omega)$, then $Cov(X,Y) = Corr(X,Y) = 0$ but X and Y are not independent.

Example 6.44. Let $X \sim HyperGeom(n,m,k)$, then X is the sum of k random variables with $Ber(m/n)$ distribution. The covariance between them is $\frac{m(m-n)}{n^2(n-1)}$. So by Part 5 above, $Var(X) = \frac{km(n-m)}{n^2} - \frac{k(k-1)m(n-m)}{n^2(n-1)} = \frac{k(n-k)m(n-m)}{n^2(n-1)}$

Example 6.45. Suppose X, Y have joint pdf proportional to $e^{-(ax^2+2bxy+cy^2)}$, where $a > 0, ac > b^2$. Then by calculation, the variance of X and Y are $\frac{c}{2(ac-b^2)}$ and $\frac{a}{2(ac-b^2)}$ respectively, and $Cov(X,Y) = E(XY) = -\frac{b}{2(ac-b^2)}$ So the correlation between X and Y equals $-\frac{b}{\sqrt{ac}}$

7 LLN and CLT

7.1 Tail Bounds and Applications

Sections 9.1 and 9.2 of textbook

A generalization of Theorem 3.27 is the following **Chebyshev's inequality**:

Theorem 7.1. Let X be a random variable.

1. (Markov's inequality) If $X \geq 0$, $E[X] = C$, then for any $y \geq 0$, $P(X \geq y) \leq \frac{C}{y}$.
2. (Chebyshev's inequality) If X has finite first and second moment (i.e. $Var(X)$ is finite), then for any $y \geq 0$, $P(|X - E(X)| \geq y) \leq \frac{Var(X)}{y^2}$.

Proof. 1. Consider random variable $X_1(\omega) = \begin{cases} y & X(\omega) \geq y \\ 0 & X(\omega) < y \end{cases}$, then $X_1 \leq X$, hence $C = E(X) \geq E(X_1) = yP(X \geq y)$. Divide y on both sides we get the inequality.

2. Apply Part 1 to $(X - E(X))^2$.

□

An immediate application is the following (weak) law of large number:

Theorem 7.2. Let X_i be i.i.d. with finite finite expectation and variance. Then for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{\sum_{i=1}^n X_i}{n} - E(X_1)\right| > \epsilon\right) = 0$$

Proof. By Chebyshev's inequality (Theorem 7.1) and Theorem 6.36,

$$P\left(\left|\frac{\sum_{i=1}^n X_i}{n} - E(X_1)\right| > \epsilon\right) \leq \frac{Var(X_1)/n}{\epsilon^2}$$

which goes to 0 as $n \to \infty$. □

7.2 Central Limit Theorem

Section 9.3 of textbook

Theorem 7.3 (Central Limit Theorem)**.** Let X_i be i.i.d. with finite expectation and variance. Then for any $s \in \mathbb{R}$,

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^n X_i - nE(X_1)}{\sqrt{nVar(X)}} \leq s\right) = \Phi(s)$$

where Φ is the cdf of $\mathcal{N}(0,1)$.

Remark 7.4. WLOG assume that $E(X_i) = 0$ and $Var(X_i) = 1$. Consider the MGF of X_i, denoted as $M(t)$, then $M'(0) = 0$, $M''(0) = 1$. Now the MGF of $\frac{\sum_{i=1}^n X_i}{\sqrt{n}}$ is $M(t/\sqrt{n})^n$. Since $M(t) \approx 1 + t^2/2$, as $n \to \infty$ the limit we get approaches $e^{t^2/2}$ which is the MGF of $\mathcal{N}(0, 1)$.

The proof of the central limit theorem would follow if one consider t a complex number, i.e. replace MGF with "characteristic functions".

Remark 7.5. The X_i being uncorrelated and identical in distribution is insufficient for getting the conclusion of CLT. For example, let X_i be i.i.d. $Ber(1/2)$, $Y_i = X_1(2X_{i+1} - 1)$.

Example 7.6. The theorem above implies the normal approximation of binomial distribution (Theorem 4.1) and the Poisson distribution as $\lambda \to \infty$: if $Y_n \sim Poisson(n\lambda)$, then as $n \to \infty$, $Z_n = \frac{Y_n - n\lambda}{\sqrt{n\lambda}}$ converges in distribution to $\mathcal{N}(0, 1)$.

8 Conditional Distribution and Conditional Expectation

Section 10.1-10.3 of textbook

When X and Y are random variables, the conditional distribution of Y given $X = x$ can not always be defined as in Definition 2.1. The reason is that $P(X = x)$ may be 0. However, the **Disintegration theorem** in measure theory tells us that there is a almost everywhere uniquely defined family of probability spaces $(\mathbb{R}, \mathcal{B}, P_x)$ such that for any Borel set $A \subseteq \mathbb{R}^2$,

$$P((X, Y) \in A) = E[g_A(X)]$$

where

$$g_A(x) = P_x(\{y : (x, y) \in A\})$$

$P_x(\cdot)$ are called the **probability distribution**, denoted as $P(\cdot|X = x)$, and the expectation of Y in this conditional distribution is called the **conditional expectation**.

Remark 8.1.

1. One can define conditional probability distribution and conditional expectation for more than 2 random variables.
2. When X and Y are independent, the conditional distribution equals the marginal distribution.
3. When $Y = g(X)$, the conditional distribution of Y given $X = x$ is degenerate, with $P(Y = g(x)|X = x) = 1$.

8.1 Discrete Case

When X is discrete, we can use Definition 2.1. In particular, if both X and Y are discrete, $P(X = a) > 0$, then the **conditional pmf** of Y is

$$p_{Y|X}(y|a) = \frac{p(a, y)}{p_X(a)}$$

and the conditional expectation is:

$$E(g(Y)|X = a) = \sum_y g(y) p_{Y|X}(y|a)$$

Example 8.2. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli distribution with parameter $p = 1/2$. Let $Y = \sum_{i=1}^m X_i$, $X = \sum_{i=1}^n X_i$, where $m < n$. Then $X \sim Bin(n, 1/2)$.

1. The joint pmf is

$$p(x,y) = P\left(\sum_{i=1}^{m} X_i = x, \sum_{i=m+1}^{n} X_i = y - x\right) = \binom{m}{y} 2^{-m} \cdot \binom{n-m}{x-y} 2^{-n+m}$$

   where $0 \le y \le x \le n$, $y \le m$.

2. The conditional pmf is, for $0 \le y \le x \le n$, $y \le m$

$$p_{Y|X}(y|x) = \frac{\binom{m}{y}\binom{n-m}{x-y}}{\binom{n}{x}}$$

3. When $m = 1$,

$$P(X_1 = 1|X = x) = \frac{\binom{n-1}{x-1}}{\binom{n}{x}} = \frac{x}{n}$$

   So the conditional expectation of Y given X is

$$E[Y|X = x] = mE[X_1|X = x] = \frac{xm}{n}$$

8.2 Jointly Continuous Case

If X, Y are jointly continuous with joint pdf f, then the **conditional pdf** of Y is

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

and the **conditional expectation** is

$$E[g(Y)|X = x] = \int_{-\infty}^{\infty} \frac{g(y)f(x,y)}{f_X(x)} dy$$

Example 8.3. Suppose $X \sim Unif(0,1)$, and when $X = x$, $Y \sim Unif(0,x)$.

1. The joint pdf is now

$$f(x,y) = \begin{cases} 1/x & 0 < y < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

2. Suppose $Y = 1/2$, the conditional pdf of X is

$$f_{X|Y}(x|1/2) = \begin{cases} \frac{\frac{1}{x}}{\int_{1/2}^{1} \frac{1}{x} dx} & 1/2 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} \frac{1}{x \log(2)} & 1/2 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

3. $E(X|Y = 1/2) = \frac{1}{2\log(2)}$

A Midterm I Review

1. Probability spaces
   (a) Examples:
      i. Experiments with finitely many outcomes with equal probability
      ii. "uniformly at random" on geometric shapes
      iii. Infinite sequence of repeated experiments
   (b) Inclusion-Exclusion
2. Independence of events
3. Conditional Probability
   (a) Conditional Independence
   (b) Bayersian Theorem
4. Random Variables
   (a) CDF and Probability Distribution. Properties of CDF:
      i. Limits at $\pm\infty$
      ii. Non decreasing
      iii. Right continuity
   (b) Independence of random variables
   (c) Special kinds of random variables:
      i. Discrete random variables, PMF
      ii. Continuous random variables, PDF
      iii. Properties of PMF and PDF
         (i) PMF: Non-zero on countable set, non negative, sum equals 1
         (ii) PDF: Non negative, integrable, integral equals 1
      iv. Recovery of PMF and PDF from CDF
   (d) Expectation and Variance
      i. Expectation and Variance for discrete random variables
      ii. Expectation and Variance for continuous random variables
5. Normal Approximation of Binomial Distribution

The old midterm I and solution have been uploaded to this Overleaf project.

Practice Problems:

1. Consider the probability space corresponding to rolling a pair of fair dices, or when one picks a number uniformly at random from $[0, 1]$. Find three events A, B and C, such that they are pairwise independent but not mutually independent.

2. Show that if f_1 and f_2 are pdf of some random variables, then $\frac{f_1+f_2}{2}$ is also the pdf of some random variable.

3. Let X be a random variable. If X is independent with X^2, what do we know about X?

4. Flip a coin, if we get head, flip the coin again, and let $X = 1$ if we get head in the second flip, 0 if otherwise. If the first flip gets a tail, let X be a number chosen uniformly at random from interval $[0, 1]$. Now suppose we know that $X > 0.9$, what's the probability that we got head in the first flip?

5. Let X be a random variable such that $P(X = -1) + P(X = 0) + P(X = 1) = 1$. What's the largest possible expectation of X? What's the largest possible variance of X?

6. Let X and Y be two independent random variables, with distribution $Geom(1/2)$. Find the conditional distribution of $\max(X, Y)$ when $\min(X, Y) \geq 2$.

7. Let A_i, $i = 1, 2, 3$ be 3 real numbers chosen independently and uniformly at random from interval $[0, 1]$. Let X be the maximal number of points among A_i where the pairwise distance is no more than 0.1. Find the pmf and expectation of X.

8. Let X be a continuous random variable with pdf f. Find the cdf and pdf of $Y = |X|$.

Answer

1. For the 2 dices case, we can let A be rolling a 1 in the first dice, B be rolling a 2 in the second dice, and C be getting the same number on both dices. For the picking a point uniformly at random from $[0, 1]$ case, let A be the point being in $[0, 1/2]$, B be the point being in $[1/4, 3/4]$ and C be the point being in $[0, 1/4] \cup [1/2, 3/4]$.

2. Because $f_1 \geq 0$, $f_2 \geq 0$, so $\frac{f_1+f_2}{2} \geq 0$.

$$\int_{-\infty}^{\infty} \frac{f_1 + f_2}{2} dt = \frac{1}{2} \int_{-\infty}^{\infty} f_1 dt + \frac{1}{2} \int_{-\infty}^{\infty} f_2 dt = 1$$

3. For any $a \geq 0$, $-a \leq X \leq a$ and $X^2 \leq a^2$ are indentical events which are independent of one another, hence the cdf of X^2 takes value at only 0 or 1. Hence there is some $a \geq 0$ such that $P(X^2 = a^2) = 1$. If $a > 0$, $P(X = -a) = p$, $P(X = a) = 1 - p$, where $0 \leq p \leq 1$. If $a = 0$, $P(X = 0) = 1$.

4. Let B be the event where we get head in the first flip, and A be the event that we get $X > 0.9$. Then by Bayer's theorem,

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} = \frac{1/2 \times 1/2}{1/2 \times 1/2 + 0.1 \times 1/2} = \frac{5}{6}$$

5. Let $P(X = -1) = a$, $P(X = 0) = b$, $P(X = 1) = c$, then $0 \le a \le 1$, $0 \le b \le 1$, $0 \le c \le 1$, $a + b + c = 1$. $E(X) = -a + c$, so it is maximized when $c = 1$ and $a = b = 0$, and the maximum is 1. $Var(X) = E(X^2) - (E(X))^2 = (a+c) - (c-a)^2$ which is maximized when $b = 0$, $a = c = 1/2$, so the maximal possible variance is 1.

6. It is easy to see that when $\min(X, Y) \ge 2$, $\max(X, Y) \ge 2$.

$$P(\min(X, Y) \ge 2) = P(X \ge 2, Y \ge 2) = (P(X \ge 2))^2$$

$$= \left(\sum_{k=2}^{\infty} 2^{-k}\right)^2 = 1/4$$

For any integer $m \ge 2$,

$$P(\max(X, Y) = m, \min(X, Y) \ge 2)$$

$$= P(X = m, 2 \le Y \le m) + P(Y = m, 2 \le X \le m) - P(X = m, Y = m)$$

$$= 2 \times 2^{-m} \sum_{k=2}^{m} 2^{-k} - 2^{-m} \times 2^{-m} = 2^{1-m}(2^{-1} - 2^{-m}) - 2^{-2m} = 2^{-m} - 3 \times 2^{-2m}$$

So

$$P(\max(X, Y) = m | \min(X, Y) \ge 2) = 2^{2-m} - 3 \times 2^{2-2m}$$

7. Pick a point uniformly at random from the unit cube

$$\{(x, y, z) \in \mathbb{R}^3 : 0 \le x \le 1, 0 \le y \le 1, 0 \le z \le 1\}$$

Let the three coordinates be A_i, then these A_i are indeed mutually independent and all have uniform distribution on $[0, 1]$. X can take only 3 possible values, 1, 2 or 3.

$$P(X = 3) = \sum_i P(A_i \le A_j \le A_i + 0.1 \text{ for all } j \ne i)$$

$$= 3\left(\int_0^{0.9} 0.1^2 dt + \int_{0.9}^{1} (1-t)^2 dt\right) = 0.028$$

$$P(X = 1) = 3! P(A_2 > A_1 + 0.1, A_3 > A_2 + 0.1)$$

$$= 6\left(\int_0^{0.8} \frac{1}{2}(0.8 - t)^2 dt\right) = 0.512$$

$$P(X=2) = P(X \le 2) - P(X=3)$$
$$= 3P(A_1 - 0.1 \le A_2 \le A_1 + 0.1) - 3P(A_1 - 0.1 \le A_2 \le A_1 + 0.1,$$
$$A_1 - 0.1 \le A_3 \le A_1 + 0.1) + P(X=3) - P(X=3)$$
$$= 3\left(\int_0^{0.1}(t+0.1)dt + \int_{0.1}^{0.9} 0.2dt + \int_{0.9}^{1}(1.1-t)dt\right)$$
$$-3\left(\int_0^{0.1}(t+0.1)^2dt + \int_{0.1}^{0.9} 0.2^2dt + \int_{0.9}^{1}(1.1-t)^2dt\right)$$
$$= 0.57 - 0.11 = 0.46$$

8. The cdf of Y is

$$F(s) = P(|X| \le s) = \begin{cases} 0 & s < 0 \\ \int_{-s}^{s} f(t)dt & s \ge 0 \end{cases}$$

Hence the pdf of Y is

$$f(s) = \frac{d}{ds}F(s) = \begin{cases} 0 & s < 0 \\ f(s) + f(-s) & s \ge 0 \end{cases}$$

Remark A.1. There would not be complicated multiple integrals in the exams like in Problem 7 above. But the inclusion-exclusion idea we used to set up the integrals would be in the exam.

B Midterm II Review

1. Confidence Intervals
2. Exponential and Poisson Distibutions
3. Moment Generating Function and its applications (e.g. moments from mgf)
4. pmf and pdf of $g(X)$
5. Joint probability distributions
   (a) Joint pmf and joint pdf, relationship with independence
   (b) Multinomial and uniform distribution on regions
   (c) Sums of independent random variables
   (d) Exchangable random variables

Practice Problems:

1. Let $X_1, \ldots, X_6$ be i.i.d. normal with a variance of 4 and unknown expectation μ. Find C such that $\left[\frac{\sum_i X_i}{6} - C, \frac{\sum_i X_i}{6} + C\right]$ is a 95% ($p = 0.05$) confidence interval for μ.
2. Roll a dice 10 times, let X be the number of times getting 1, Y the number of times getting 2, Z the number of times getting 3 or above. What's the joint distribution of X, Y, Z? What's the marginal distribution of Y? What's the variance of $2Y - 1$?
3. Suppose the moment generating function of a random variable X is $M_X(t) = \frac{e^t + e^{-t}}{2}$. Find $E[X]$, $E[X^2]$, $Var(X)$ and the cdf of X.
4. Suppose $X \sim Poisson(1)$, $Y \sim Exp(1)$ are independent. What's the probability that $X < Y$?
5. Pick a point $P = (X, Y)$ uniformly at random from a disc of radius 1 centered at $(1, 1)$. Are X and Y independent?
6. Suppose $X \sim \mathcal{N}(0, 4)$, $Y \sim \mathcal{N}(0, 9)$ are independent. Find $t \in \mathbb{R}$ such that the variance of $tX + (1 - t)Y$ is minimized.

Answer

1. $C = -\sqrt{\frac{2}{3}}\Phi^{-1}(0.025)$ where Φ is the cdf of standard normal distribution.
2. The joint distribution is $Multi(10, 3, 1/6, 1/6, 2/3)$, and the marginal distribution of Y is $Bin(10, 1/6)$. $Var(2Y - 1) = 4Var(Y) = \frac{50}{9}$

3. X is discrete with pmf $p(1) = p(-1) = 1/2$. So cdf is $F(s) = \begin{cases} 0 & s < -1 \\ 1/2 & -1 \le s < 1 \\ 1 & s \ge 1 \end{cases}$,
   $E(X) = 0$, $Var(X) = E(X^2) = 1$.

4. $P(X < Y) = \sum_{k=0}^{\infty} P(X = k, Y > k) = \sum_{k=0}^{\infty} \frac{e^{-1}}{k!} \cdot e^{-k} = e^{e^{-1}-1}$.

5. No. One can verify by multiplying the marginal pdfs.

6. The variance of $tX + (1-t)Y$ equals $4t^2 + 9(1-t)^2$, so it is mininized when $t = 9/13$.

C Final Review

C.1 Common Distributions

Distribution	Expectation	Variance
$Ber(p)$	p	$p(1-p)$
$Bin(n,p)$	np	$np(1-p)$
$Geom(p)$ [(a)]	$1/p$	$(1-p)/p^2$
$HyperGeom(n,m,k)$	km/n	$k(n-k)m(n-m)/((n-1)n^2)$
$Negbin(n,p)$ [(b)]	n/p	$n(1-p)/p^2$
$Poisson(\lambda)$	λ	λ
$Unif(a,b)$	$(a+b)/2$	$(b-a)^2/12$
$Exp(\lambda)$	$1/\lambda$	$1/\lambda^2$
$\mathcal{N}(\mu,\sigma^2)$	μ	σ^2
$Multi(n,r,p_1,\ldots,p_r)$	$E[X_i]=np_i$	$Var(X_i)=np_i(1-p_i)$ $Cov(X_i,X_j)=-np_ip_j,\ i\neq j$ [(c)]

(a) $X \sim Geom(p)$, then

$$M_X(t) = E[e^{tX}] = \sum_{k=1}^{\infty} p(1-p)^{k-1}e^{kt} = \frac{pe^t}{1-(1-p)e^t}$$

$$M_X'(t) = \frac{pe^t}{(1-(1-p)e^t)^2}, E[X] = (M_X)'(0) = \frac{p}{p^2} = \frac{1}{p}$$

$$M_X''(t) = \frac{pe^t(1+(1-p)e^t)}{(1-(1-p)e^t)^3}, E[X^2] = M_X''(0) = \frac{2-p}{p^2}$$

$$Var(X) = E[X^2] - E[X]^2 = \frac{1-p}{p^2}$$

(b) $Negbin(n,p)$ is the sum of n i.i.d. $Geom(p)$.

(c) X_iX_j is the number of ordered pairs (k,l) such that $k \neq l$, and the k-th experiment gets outcome i, the l-th outcome j, hence $E[X_iX_j] = n(n-1)p_ip_j$,

$$Cov(X_i,X_j) = E[X_iX_j] - E[X_i]E[X_j] = n(n-1)p_ip_j - n^2p_ip_j = -np_ip_j$$

C.2 Topics

1. Probability Spaces
   (a) Inclusion-Exclusion
   (b) Independence of events
   (c) Conditional Probability
      i. Conditional Independence

        ii. Bayesian Theorem

2. Random Variables
    (a) Probability Distribution, CDF
    (b) Discrete RV, PMF
    (c) Continuous RV, PDF
    (d) Expectation
        i. Linearility
        ii. Indicator RV and Indicator Method
        iii. Variance
    (e) Transformations, MGF
3. Joint Distribution
    (a) Discrete Case, joint PMF
    (b) Jointly Continuous Case, Joint PDF
    (c) Marginal Distribution, marginal cdf/pmf/pdf
    (d) Conditional distribution
        i. Conditional expectation
    (e) Independece of RV
        i. pdf of sum and convolution
        ii. Expectation of product
        iii. MGF of sum
        iv. Variance of sum
    (f) Exchangability of RV
    (g) IID
    (h) Covariance and correlation
4. Tail bounds and CLT
    (a) Markov's inequality
        i. Chebyshev's inequality
        ii. Weak LLN
    (b) CLT

C.3 Practice Problems

1. Let X and Y be jointly continuous with joint pdf $f(x, y)$. Find the joint pdf of $X, kX + Y$ for $k \in \mathbb{R}$.

2. Let $X_1, \ldots, X_n$ be i.i.d. continuous random variables. Let Y be the number of j such that $X_j > X_1$.

   (a) Find the pmf of Y, $E[Y]$ and $Var(Y)$.

   (b) Let Z be the number of j such that $X_j > X_2$. Find $Cov(Y, Z)$.

3. Suppose there is a group of drivers, half of them gets on average one ticket per year, the other half one ticket every 2 years. Suppose someone in the group did not get any traffic ticket in 2025. What's the probability that the same person would not get a traffic ticket in 2026?

4. Roll a dice 100 times, let X be the sum of the points. Use Chebyshev's identity and CLT to estimate the probability that $X > 400$.

5. Let (X, Y) be chosen uniformly at random from the unit disc, find the expectation and variance of X and Y, $Cov(X, Y)$, and the conditional distribution of X given $Y = 0$.

6. Let X and Y be jointly continuous. Suppose the conditional expectation of Y is a constant that does not depend on the value of X. Is it true that X and Y are uncorrelated?

7. Let X and Y be i.i.d. $\mathcal{N}(0, 1)$.

   (a) For any real number s, find $P(X \leq s | X + Y < 0)$.

   (b) Find the conditional pdf of X given $X + Y < 0$.

Answer:

1. Let the new pdf be h, $F : (x, y) \mapsto (x, kx + y)$, then

$$\int_A h(s,t)dsdt = P((X, kX + Y) \in A) = P((X, Y) \in F^{-1}(A))$$

$$= \int_{F^{-1}(A)} f(x, y)dxdy$$

   By change of variable formula in multivariable calculus,

$$\int_A h(s,t)dsdt = \int_{F^{-1}(A)} h \circ F(x, y) |\det(F')| dxdy$$

   Here F' is the derivative, or the Jacobian matrix. Since $\det(F') = 1$, we have

$$f(x, y) = h \circ F(x, y) = h(x, kx + y)$$

   So $h(s, t) = f(s, t - ks)$.

2. Because they are continuous and iid, they are exchangable and jointly continuous, hence for any bijection σ from $\{1,\dots,n\}$ to itself, $P(X_{\sigma(1)} < \cdots < X_{\sigma(n)}) = \frac{1}{n!}$, hence for $0 \le k \le n-1$,

$$P(Y=k) = \frac{1}{n!} \cdot \binom{n-1}{k} \cdot k! \cdot (n-1-k)! = \frac{1}{n}$$

So $E[Y] = \frac{n-1}{2}$,

$$Var(Y) = E[Y^2] - E[Y]^2 = \frac{1}{n} \cdot \frac{(n-1)n(2n-1)}{6} - \frac{(n-1)^2}{4} = \frac{n^2-1}{12}$$

By exchangability, Z has the same distribution of Y, and the joint pmf can be similarly calculated as

$$p(i,j) = \frac{1}{n(n-1)} \text{ where } 0 \le i,j \le n-1, i \ne j$$

Hence

$$E(YZ) = \frac{2}{n(n-1)} \sum_{0 \le i < j \le n-1} ij = \frac{1}{n(n-1)} \left(\sum_{0 \le i,j \le n-1} ij - \sum_{i=0}^{n-1} i^2 \right)$$

$$= \frac{3n^2 - 7n + 2}{12}$$

$$Cov(Y,Z) = E(YZ) - E(Y)E(Z) = -\frac{n+1}{12}$$

3. Call the first kind bad drivers, the second kind good drivers. Then for bad drivers the number of ticket each year $\sim Poisson(1)$, for good drivers $\sim Poisson(1/2)$. Let the driver be a, then $P(a \text{ is good}) = 1/2$,

$$P(a \text{ is good} | a \text{ gets no ticket in 2025})$$

$$= \frac{e^{-1} \cdot \frac{1}{2}}{e^{-1} \cdot \frac{1}{2} + e^{-2} \cdot \frac{1}{2}} = \frac{e}{1+e}$$

$$P(a \text{ gets no ticket in 2026} | a \text{ gets no ticket in 2025})$$

$$= e^{-1} \cdot \frac{e}{1+e} + e^{-2} \cdot \frac{1}{1+e} = \frac{1+e^{-2}}{1+e}$$

4. $E[X] = 350$, $Var(X) = \frac{875}{3}$, by Chebyshev's identity,

$$P(X > 400) = \frac{1}{2} P(|X - 350| > 50) \le \frac{\frac{875}{3}}{2 \times 50^2} = \frac{7}{120}$$

By CLT,

$$P(X > 400) \approx 1 - \Phi\left(\frac{50}{\sqrt{\frac{875}{3}}} \right) \approx 1 - \Phi(2.928)$$

Where Φ is the cdf of $\mathcal{N}(0,1)$.

5. The marginal pdf of X and Y are

$$f(s) = \begin{cases} 0 & x < -1 \text{ or } x > 1 \\ \frac{\sqrt{1-s^2}}{\pi} & -1 \leq x \leq 1 \end{cases}$$

So their expectation is 0, variance is $\frac{1}{8}$.

$$Cov(X,Y) = \int_{x^2+y^2 \leq 1} \frac{xy}{\pi} dxdy = 0$$

When $Y = 0$, the conditional distribution of X is $Unif(-1,1)$.

6. Yes. Suppose $E[Y|X=x] = C = \frac{\int_{\mathbb{R}} yf(x,y)dy}{\int_{\mathbb{R}} f(x,y)dy}$

$$Cov(X,Y) = \int_{\mathbb{R}^2} xyf(x,y)dxdy - E[X]\int_{\mathbb{R}^2} yf(x,y)dxdy$$

$$\int_{\mathbb{R}^2} xyf(x,y)dxdy = \int_{\mathbb{R}} \left(xC \int_{\mathbb{R}} f(x,y)dy \right) dx = CE[X]$$

and

$$\int_{\mathbb{R}^2} yf(x,y)dxdy = \int_{\mathbb{R}} \left(C \int_{\mathbb{R}} f(x,y)dy \right) dy = C$$

So $Cov(X,Y) = 0$

7. The joint pdf is $f(x,y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}$.

   (a)

$$P(X \leq s | X+Y < 0) = \frac{P(X \leq s, X+Y < 0)}{P(X+Y<0)}$$

$$= \frac{1}{\pi} \int_{x \leq s, x+y<0} e^{-\frac{x^2+y^2}{2}} dxdy = 1 - \Phi^2(-s) = 2\Phi(s) - \Phi^2(s)$$

   where Φ is the cdf of $\mathcal{N}(0,1)$.

   (b) The conditional pdf is now $2f(s)(1-\Phi(s))$ where f is the pdf of $\mathcal{N}(0,1)$, and Φ its cdf.

D JS code for visualizing the normal approximation and continuity correction

```
<!DOCTYPE html>
<html>
<head>
<meta name="viewport" content="width=device-width,
    initial-scale=1.0">
<script src=https://cdnjs.cloudflare.com/ajax/libs/mathjs
    /14.8.1/math.min.js></script>
</head>
<h1>Normal Approximation of Binomial Distribution</h1>
<body>
<p>N: <input type="text" id="n" value="10"></p>
<p>p: <input type="text" id="p" value="0.5"></p>
<p><button type="button" onclick="plotCDF();">Plot CDF</
    button></p>
<canvas id="myCanvas" width="500" height="500"
style="border:1px solid #000000;">
</canvas>
<script>
  function plotCDF(){
      var n=parseInt(document.getElementById("n").value);
      var p=parseFloat(document.getElementById("p").value
          );
      var cvs=document.getElementById("myCanvas");
      var ctx=cvs.getContext("2d");
          ctx.clearRect(0, 0, cvs.width, cvs.height);
          ctx.beginPath();
      ctx.strokeStyle="#0000FF";
      function line(a, b, c, d){
         ctx.moveTo(a, b);
         ctx.lineTo(c, d);
         ctx.stroke();
      }
      var x=0;
      var y=500;
      for(let i=0;i<n;i++){
        var xn=x+500/n;
        var cp=math.combinations(n, i)*Math.pow(p, i)*
            Math.pow(1-p, n-i);
        var yn=y-500*cp;
        line(x, y, x, yn);
        line(x, yn, xn, yn);
        x=xn;
```

```
        y=yn;
      }
      line(500, y, 500, 0);
      ctx.strokeStyle="#FF0000";
      ctx.beginPath();
      var ny=[];
      for(let i=0;i<=500;i+=5){
          var k=i*n/500;
          var x=(k-n*p)/Math.pow(n*p*(1-p), 0.5);
          ny.push([i, 500*(0.5-0.5*math.erf(x/Math.pow(2,
              0.5)))]);
      }
      for(let i=0;i<100;i++){
        line(ny[i][0], ny[i][1], ny[i+1][0], ny[i+1][1]);
      }
      ctx.strokeStyle="#FF00FF";
      ctx.beginPath();
      var ny=[];
      for(let i=0;i<=500;i+=5){
          var k=i*n/500+0.5;
          var x=(k-n*p)/Math.pow(n*p*(1-p), 0.5);
          ny.push([i, 500*(0.5-0.5*math.erf(x/Math.pow(2,
              0.5)))]);
      }
      for(let i=0;i<100;i++){
        line(ny[i][0], ny[i][1], ny[i+1][0], ny[i+1][1]);
      }
  }
</script>
</body>
</html>
```

E Python code for calculating probabilies with normal approximation and continuity correction

```
import math
def cdf(t):
    return (1+math.erf(t/2**0.5))/2

def choose(n, k):
    r=1
    for j in range(k):
        r*=n-j
        r//=j+1
    return r

#binomial probability distribution
def binom(n, p, k1, k2):
    r=0
    for j in range(k1, k2+1):
        r+=choose(n, j)*(p**j)*((1-p)**(n-j))
    return r

#normal approximation without continuity correction
def approx(n, p, k1, k2):
    sigma=(n*p*(1-p))**0.5
    return cdf((k2-n*p)/sigma)-cdf((k1-1-n*p)/sigma)

#normal approximation with continuity correction
def approx_cc(n, p, k1, k2):
    sigma=(n*p*(1-p))**0.5
    return cdf((k2+0.5-n*p)/sigma)-cdf((k1-0.5-n*p)/sigma
        )

print(binom(100, 1/6, 20, 25))
print(approx(100, 1/6, 20, 25))
print(approx_cc(100, 1/6, 20, 25))
```

F JS code for visualizing Poisson approximation

```
<!DOCTYPE html>
<html>
<head>
<meta name="viewport" content="width=device-width,
    initial-scale=1.0">
<script src=https://cdnjs.cloudflare.com/ajax/libs/mathjs
    /14.8.1/math.min.js></script>
</head>
<h1>Poisson Approximation of Binomial Distribution</h1>
<body>
<p>N: <input type="text" id="n" value="10"></p>
<p>lambda: <input type="text" id="lambda" value="4"></p>
<p><button type="button" onclick="plotPMF();">Plot PMF</
    button></p>
<canvas id="myCanvas" width="500" height="500"
style="border:1px solid #000000;">
</canvas>
<script>
  function plotPMF(){
      var n=parseInt(document.getElementById("n").value);
      var lam=parseFloat(document.getElementById("lambda"
          ).value);
      var cvs=document.getElementById("myCanvas");
      var ctx=cvs.getContext("2d");
          ctx.clearRect(0, 0, cvs.width, cvs.height);
          ctx.beginPath();
      ctx.strokeStyle="#0000FF";
      function line(a, b, c, d){
         ctx.moveTo(a, b);
         ctx.lineTo(c, d);
         ctx.stroke();
      }
      for(let i=0;i<20;i++){
        var x=i*25+2;
        var cp=0;
        if(i<=n){
          cp=math.combinations(n, i)*Math.pow(lam/n, i)*
              Math.pow(1-lam/n, n-i);
        }
        line(x, 500, x, 500-500*cp);
      }
      ctx.strokeStyle="#FF0000";
      ctx.beginPath();
```

```
        for(let i=0;i<20;i++){
          var x=i*25+4;
          var cp=Math.pow(lam, i)/math.factorial(i)*Math.
              exp(-lam);
          line(x, 500, x, 500-500*cp);
        }
    }
</script>
</body>
</html>
```