Lecture Notes in Probability

Donlapark Ponnoprat

June 19, 2025

Contents

Chapter 1

Probability Fundamentals

The concept of probability has been a subject of debate among mathematicians and philosophers. Some view probabilities as long-run frequencies of events that can be repeated under identical conditions (frequentists), while others see them as a quantification of an individual's subjective degree of uncertainty (subjectivists). Regardless of the interpretation, the fundamental principles of combining probabilities can be understood by thinking about proportions.

1.1 Outcome Space and Events

Definition 1.1.1 (Outcome Space)**.** An experiment involving randomness results in one of several possible outcomes. The **outcome space**, denoted by Ω, is the set of all possible outcomes. For now, we assume Ω is a finite set.

Definition 1.1.2 (Outcome)**.** An **outcome**, denoted by ω, is an element of the outcome space Ω.

Definition 1.1.3 (Event)**.** An **event** is a subset of the outcome space Ω. Events are typically denoted by capital letters like A, B, C. The empty set $\emptyset$ and the entire space Ω are also considered events.

Example 1.1.4 (Permutations)**.** Consider shuffling three cards labeled a, b, c. The outcome space is the set of all possible permutations:

$$\Omega = \{abc, acb, bac, bca, cab, cba\}$$

Here are some examples of events:

- Event A: "a appears first". This corresponds to the subset $A = \{abc, acb\}$.
- Event B: "b and c are not next to each other". This corresponds to the subset $B = \{bac, cab\}$.

- Event C: "the letters are in alphabetical order". This corresponds to the subset $C = \{abc\}$.

Example 1.1.5 (Coin Tossing)**.** For three tosses of a coin, the outcome space is:

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

The event "no two consecutive tosses land the same way" is the subset $\{HTH, THT\}$.

1.2 Equally Likely Outcomes

A simple and powerful model for randomness is the assumption of equally likely outcomes. This model defines probabilities as proportions. Let Ω be a finite outcome space with $n = \#(\Omega)$ outcomes.

Definition 1.2.1 (Probability with Equally Likely Outcomes)**.** If all n outcomes in Ω are assumed to be equally likely, then the probability of an event $A \subseteq \Omega$ is defined as:

$$P(A) = \frac{\#(A)}{\#(\Omega)} = \frac{\#(A)}{n}$$

where $\#(A)$ is the number of outcomes in the event A.

Example 1.2.2 (Random Permutations)**.** Let Ω be the space of all permutations of the letters a, b, c. We have $\#(\Omega) = 6$. If we assume all permutations are equally likely, we can calculate the probabilities of the events defined earlier:

- $P(A) = P(\text{a appears first}) = \frac{\#\{abc, acb\}}{6} = \frac{2}{6} = \frac{1}{3}$.
- $P(C) = P(\text{the letters are in alphabetical order}) = \frac{\#\{abc\}}{6} = \frac{1}{6}$.

Example 1.2.3 (Random Number Generator)**.** A random number generator produces a pair of digits from 00 to 99, with all 100 pairs being equally likely.

- The probability that the pair consists of two different digits is calculated as follows: There are 10 choices for the first digit, and for each choice, there are 9 choices for the second digit. So, there are $10 \times 9 = 90$ such pairs. The probability is $\frac{90}{100} = 0.9$.
- The probability that the two digits are the same can be found by complementing the previous event: $1 - 0.9 = 0.1$. Alternatively, there are 10 pairs with identical digits (00, 11, ..., 99), so the probability is $\frac{10}{100} = 0.1$.

1.3 Collisions in Hashing

In computer science, a hash function assigns a hash value to each individual in a set. A collision occurs when two individuals are assigned the same hash value.

Let's assume there are N hash values and n individuals, and that all N^n possible assignments of values to individuals are equally likely. An assignment is a sequence $a_0, a_1, \ldots, a_{n-1}$ where individual i is assigned hash value a_i.

1.3.1 Probability of No Collisions

We want to find the probability that there are no collisions, assuming $n \leq N$. The number of ways to assign hash values with no collisions is the number of sequences $(a_0, a_1, \ldots, a_{n-1})$ where all the a_i are distinct.

- There are N choices for a_0.
- There are $N-1$ choices for a_1 (it must be different from a_0).
- There are $N-2$ choices for a_2 (different from a_0 and a_1).
- ...
- There are $N-(n-1)$ choices for a_{n-1}.

The total number of assignments with no collisions is $N(N-1)(N-2)\cdots(N-n+1)$. The probability of no collisions is:

$$P(\text{no collisions}) = \frac{N(N-1)(N-2)\cdots(N-n+1)}{N^n}$$

This can also be written as a product of fractions:

$$P(\text{no collisions}) = \prod_{i=0}^{n-1} \frac{N-i}{N} = \frac{N}{N} \cdot \frac{N-1}{N} \cdot \frac{N-2}{N} \cdots \frac{N-n+1}{N}$$

1.3.2 Probability of at Least One Collision

The event ”at least one collision” is the complement of the event ”no collisions”. Therefore, its probability is:

$$P(\text{at least one collision}) = 1 - P(\text{no collisions}) = 1 - \prod_{i=0}^{n-1} \frac{N-i}{N}$$

1.4 The Birthday Problem

The birthday problem is a classic application of collision probability. It asks for the probability that in a group of n people, at least two share a birthday.

1.4.1 Assumptions

We make the following simplifying assumptions:

- There are $N = 365$ days in a year.
- Each person is equally likely to be born on any of the 365 days, independently of the others.

These assumptions mean that all 365^n sequences of birthdays are equally likely.

1.4.2 The Chance of a Match

The problem is equivalent to the hashing problem with $N = 365$. A ”match” is a collision. The probability of no match (all birthdays are different) is:

$$P(\text{no match}) = \prod_{i=0}^{n-1} \frac{365 - i}{365}$$

The probability of at least one match is:

$$P(\text{at least one match}) = 1 - \prod_{i=0}^{n-1} \frac{365 - i}{365}$$

1.4.3 The Birthday ”Paradox”

The probability of a match increases sharply with n. With just 23 people, the probability of a match is greater than 50

1.5 An Exponential Approximation

To better understand the behavior of the collision probability, we can derive an approximation for $P(\text{no collision})$.

1. **Approximate the log of the probability**: It's easier to work with sums than products.

$$\log(P(\text{no collision})) = \log\left(\prod_{i=0}^{n-1} \frac{N-i}{N}\right) = \sum_{i=0}^{n-1} \log\left(1 - \frac{i}{N}\right)$$

2. **Use the Taylor approximation for log**: For small x, we have $\log(1 + x) \approx x$. For our case, this is $\log(1 - i/N) \approx -i/N$. This approximation is good when i/N is small.

$$\sum_{i=0}^{n-1} \log\left(1 - \frac{i}{N}\right) \approx \sum_{i=0}^{n-1}\left(-\frac{i}{N}\right) = -\frac{1}{N}\sum_{i=0}^{n-1} i$$

3. **Sum the series**: The sum of the first $n - 1$ integers is $\frac{(n-1)n}{2}$.

$$-\frac{1}{N}\sum_{i=0}^{n-1} i = -\frac{n(n-1)}{2N}$$

4. **Exponentiate to get the approximation for the probability**:

$$P(\text{no collision}) \approx e^{-\frac{n(n-1)}{2N}}$$

For large n, we can further approximate this as $e^{-\frac{n^2}{2N}}$.

So, the probability of at least one collision can be approximated by:

$$P(\text{at least one collision}) \approx 1 - e^{-\frac{n(n-1)}{2N}} \approx 1 - e^{-\frac{n^2}{2N}}$$

This approximation is very accurate, even for moderate values of n. It clearly shows why the probability of a match in the birthday problem grows quickly with n, as the term in the exponent is quadratic in n.

Chapter 2

Calculating Probabilities

2.1 Axioms of Probability

When dealing with situations where outcomes are not equally likely, a more general framework is needed. The foundation of modern probability theory is based on a set of axioms formulated by Andrey Kolmogorov.

We begin with an outcome space Ω, which for now we assume is finite. Probability is a function P defined on events (subsets of Ω). The axioms are:

1. **Non-negativity:** For any event A, $P(A) \geq 0$.
2. **Total Probability:** The probability of the entire outcome space is 1, i.e., $P(\Omega) = 1$.
3. **Additivity for Mutually Exclusive Events:** If events $A_1, A_2, \ldots$ are mutually exclusive (i.e., they don't intersect), then the probability of their union is the sum of their individual probabilities. For two events A and B with $A \cap B = \emptyset$, this means $P(A \cup B) = P(A) + P(B)$.

2.2 The Addition Rule

From the basic axioms, we can derive rules for calculating the probability of unions of events that are not mutually exclusive. The general addition rule for two events A and B is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

The term $P(A \cap B)$ is subtracted because it is counted in both $P(A)$ and $P(B)$, so we must remove the double-counting.

Remark. This leads to a useful inequality known as Boole's Inequality: $P(A \cup B) \leq P(A) + P(B)$.

2.2.1 Complement Rule

For any event A, its complement is $A^c = \Omega \setminus A$. Since A and A^c are mutually exclusive and their union is Ω, we have $P(A) + P(A^c) = P(\Omega) = 1$. This gives the complement rule:

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

2.2.2 Difference Rule

If event A implies event B (i.e., $A \subseteq B$), then the probability of the event "B and not A" is given by:

$$P(B \setminus A) = P(B) - P(A)$$

2.3 Examples of the Addition Rule

Example 2.3.1 (Both Heads and Tails in n Tosses)**.** A coin is tossed n times. All 2^n sequences of heads and tails are equally likely. What is the chance of getting at least one head and at least one tail?

Let A be the event of getting at least one head and at least one tail. The complement event, A^c, is that we *don't* get both faces. This means all the tosses result in the same face. The only two outcomes in A^c are "all heads" (HHHH...) and "all tails" (TTTT...). So, the number of outcomes in the complement event is $\#(A^c) = 2$. The probability of the complement is:

$$P(A^c) = \frac{\#(A^c)}{\#(\Omega)} = \frac{2}{2^n} = \frac{1}{2^{n-1}}$$

Using the complement rule, the probability of event A is:

$$P(A) = 1 - P(A^c) = 1 - \frac{1}{2^{n-1}}$$

Example 2.3.2 (Maximum of 12 Rolls of a Die)**.** A die is rolled 12 times. What is the probability that the maximum roll is 4?

Let M be the maximum value of the 12 rolls. We want to find $P(M = 4)$. This is equivalent to the event "the maximum is less than or equal to 4" AND "the maximum is not less than 4". Using the difference rule, we can write: $P(M = 4) = P(M \leq 4) - P(M \leq 3)$.

The event $M \leq 4$ is equivalent to all 12 rolls being less than or equal to 4. For each roll, there are 4 possible outcomes (1, 2, 3, 4). Since there are 12 rolls, the number of such sequences is 4^{12}. The total number of possible sequences is 6^{12}. So:

$$P(M \leq 4) = \frac{4^{12}}{6^{12}}$$

Similarly, the event $M \leq 3$ is equivalent to all 12 rolls being less than or equal to 3. There are 3^{12} such sequences.

$$P(M \leq 3) = \frac{3^{12}}{6^{12}}$$

Therefore, the probability that the maximum is exactly 4 is:

$$P(M = 4) = \frac{4^{12}}{6^{12}} - \frac{3^{12}}{6^{12}} = \frac{4^{12} - 3^{12}}{6^{12}}$$

2.4 The Multiplication Rule

2.4.1 Conditional Probability

Let A and B be two events. The **conditional probability of B given A**, denoted $P(B|A)$, is the probability that B occurs given that A has already occurred. It is defined by the division rule:

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

provided that $P(A) > 0$. Given that A happened, we restrict our focus to the outcomes in A. The chance that B also happens is the proportion of these outcomes that are also in B.

2.4.2 Multiplication Rule

By rearranging the definition of conditional probability, we get the multiplication rule, which is extremely useful for calculating the probability of the intersection of events:

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

This rule is used to calculate the chance of a sequence of events occurring.

Example 2.4.1 (Two Aces)**.** Two cards are dealt from a standard 52-card deck without replacement. What is the chance that both are aces? Let A_1 be the event that the first card is an ace, and A_2 be the event that the second card is an ace. We want to find $P(A_1 \cap A_2)$.

- The probability of the first card being an ace is $P(A_1) = \frac{4}{52}$.
- Given that the first card was an ace, there are now 51 cards left in the deck, and 3 of them are aces. So, $P(A_2|A_1) = \frac{3}{51}$.

Using the multiplication rule:

$$P(A_1 \cap A_2) = P(A_1)P(A_2|A_1) = \frac{4}{52} \cdot \frac{3}{51} \approx 0.0045$$

2.5 More Examples

Example 2.5.1 (One of Each Kind)**.** A box contains 6 dark chocolates and 4 milk chocolates. Two chocolates are picked at random without replacement. What is the chance of getting one of each kind?

The event ”one of each kind” can be partitioned into two mutually exclusive events:

1. First dark, then milk (DM)
2. First milk, then dark (MD)

We calculate the probability of each and add them.

$$P(DM) = P(\text{1st is dark}) \times P(\text{2nd is milk — 1st is dark}) = \frac{6}{10} \cdot \frac{4}{9} = \frac{24}{90}$$

$$P(MD) = P(\text{1st is milk}) \times P(\text{2nd is dark — 1st is milk}) = \frac{4}{10} \cdot \frac{6}{9} = \frac{24}{90}$$

The total probability is the sum of these two probabilities:

$$P(\text{one of each}) = P(DM) + P(MD) = \frac{24}{90} + \frac{24}{90} = \frac{48}{90} = \frac{8}{15}$$

2.6 Updating Probabilities: Bayes' Rule

Often, we start with an initial belief about an event, called a *prior probability.* Then, we observe some data, and we want to update our belief to a *posterior probability* based on this new information.

2.6.1 Derivation

Suppose we have a partition of the outcome space $A_1, A_2, \ldots, A_n$. Let B be another event. We want to find the "backwards in time" probability $P(A_i|B)$. From the definition of conditional probability, we know:

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

We can rewrite the numerator using the multiplication rule: $P(A_i \cap B) = P(A_i)P(B|A_i)$. The denominator $P(B)$ can be found by partitioning B based on the A_j's (the law of total probability): $P(B) = \sum_{j=1}^{n} P(A_j \cap B) = \sum_{j=1}^{n} P(A_j)P(B|A_j)$.

Combining these gives **Bayes' Rule**:

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{n} P(A_j)P(B|A_j)}$$

Example 2.6.1 (Rare Disease Testing)**.** A rare disease affects 0.4

- If a person has the disease, the test is positive 99
- If a person does not have the disease, the test is negative 99.5

A person is picked at random and tests positive. What is the probability they actually have the disease?

Let D be the event the person has the disease, and $+$ be the event the test is positive. We want to find $P(D|+)$. We are given the following probabilities:

- Prior probability of having the disease: $P(D) = 0.004$.

- Prior probability of not having the disease: $P(D^c) = 1 - 0.004 = 0.996$.
- Conditional probability of a positive test given disease (true positive rate): $P(+|D) = 0.99$.
- Conditional probability of a negative test given no disease: $P(-|D^c) = 0.995$.
- From this, the conditional probability of a positive test given no disease (false positive rate) is $P(+|D^c) = 1 - 0.995 = 0.005$.

Using Bayes' Rule, the denominator is the total probability of testing positive: $P(+) = P(D)P(+|D) + P(D^c)P(+|D^c) = (0.004)(0.99) + (0.996)(0.005) = 0.00396 + 0.00498 = 0.00894$.

The numerator is the probability of having the disease AND testing positive: $P(D \cap +) = P(D)P(+|D) = (0.004)(0.99) = 0.00396$.

The posterior probability is:

$$P(D|+) = \frac{P(D \cap +)}{P(+)} = \frac{0.00396}{0.00894} \approx 0.44295$$

So, even with a positive test, there is only a 44.3

Chapter 3

Random Variables

Many aspects of data science involve numerical quantities whose observed values are subject to chance. For instance, if we take a random sample of people, the number of individuals in a certain category is a random quantity. In probability theory, these numerical functions defined on an outcome space are called **random variables**. We typically denote them with uppercase letters like X and Y.

3.1 Functions on an Outcome Space

Definition 3.1.1 (Random Variable)**.** A **random variable** is a real-valued function defined on an outcome space Ω. That is, it is a mapping $X : \Omega \to \mathbb{R}$. For each outcome $\omega \in \Omega$, the random variable X assigns a numerical value $X(\omega)$.

Example 3.1.2 (Sum of Dice Rolls)**.** Consider an experiment of rolling a fair die two times. The outcome space is $\Omega = \{(i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}\}$. Let S be the random variable representing the sum of the two rolls. For an outcome $\omega = (i, j)$, the value of the random variable is $S(\omega) = i + j$. For example, if the outcome is $(3, 4)$, then $S((3, 4)) = 7$.

3.1.1 Events Determined by a Random Variable

For a random variable X and a set of real numbers A, the event $\{X \in A\}$ is the set of all outcomes ω in the outcome space such that the value $X(\omega)$ is in A. Formally:

$$\{X \in A\} = \{\omega \in \Omega : X(\omega) \in A\}$$

If we know the value of X, we can determine whether or not the event $\{X \in A\}$ has occurred. For simplicity, if A is a single value $\{k\}$, we write the event as $\{X = k\}$.

Example 3.1.3 (Difference of two rolls)**.** Consider two rolls of a die. Let X_1 be the result of the first roll and X_2 the result of the second. Define a new random variable $D = X_1 - X_2$. Let's find the outcomes corresponding to the event $\{D > 3\}$. The outcomes (i, j) for which $i - j > 3$ are: $(5, 1)$, $(6, 1)$, and $(6, 2)$. Thus, the event is $\{D > 3\} = \{(5, 1), (6, 1), (6, 2)\}$. Assuming all 36 outcomes are equally likely, the probability of this event is $P(D > 3) = \frac{3}{36} = \frac{1}{12}$.

3.2 Distributions

A random variable is fully characterized by its probability distribution.

Definition 3.2.1 (Probability Distribution)**.** The **probability distribution** (or simply **distribution**) of a random variable X is a specification of the set of all possible values of X along with the probabilities of these values.

If X is a discrete random variable with possible values $x_1, x_2, \ldots$, its distribution can be presented in a **probability distribution table**:

Value k	$P(X = k)$
x_1	p_1
x_2	p_2
$\vdots$	$\vdots$

The probabilities in a distribution must be non-negative and sum to 1:

$$\sum_k P(X = k) = 1$$

Example 3.2.2 (Sum of two dice)**.** Let S be the sum of two rolls of a fair die. The possible values for S are integers from 2 to 12. We can find the probability of each value by counting the number of outcomes that produce that sum. For instance, to get a sum of $S = 3$, the outcomes are $(1, 2)$ and $(2, 1)$. So, $P(S = 3) = 2/36$. The full distribution table for S is:

k	2	3	4	5	6	7	8	9	10	11	12
$P(S = k)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

The sum of these probabilities is $\frac{1+2+3+4+5+6+5+4+3+2+1}{36} = \frac{36}{36} = 1$.

3.2.1 Named Distributions

Certain distributions are encountered so frequently that they are given special names.

- **Bernoulli(p)**: This is the distribution of a random variable that takes value 1 with probability p and value 0 with probability $1 - p$. It often models a single "success/failure" trial.

- **Uniform on a finite set**: This distribution assigns equal probability to all possible values in a finite set. For example, the outcome of a single fair die roll has a uniform distribution on $\{1, 2, 3, 4, 5, 6\}$, where each value has a probability of 1/6.

3.3 Equality of Random Variables

There are two main ways in which random variables can be considered "equal".

3.3.1 Equality

Definition 3.3.1 (Equal)**.** Two random variables X and Y defined on the same outcome space Ω are said to be **equal**, written $X = Y$, if they are equal as functions. That is, for every outcome $\omega \in \Omega$, their values are the same:

$$X(\omega) = Y(\omega) \quad \text{for all } \omega \in \Omega$$

Example 3.3.2. Let an experiment consist of three coin tosses. Let N_H be the number of heads and N_T be the number of tails. The random variable $M = 3 - N_T$ is equal to N_H. For any outcome of three tosses, the number of heads is always equal to 3 minus the number of tails. So, $N_H = 3 - N_T$.

3.3.2 Equality in Distribution

A weaker form of equality compares only the probability distributions of the random variables.

Definition 3.3.3 (Equal in Distribution)**.** Two random variables X and Y are **equal in distribution**, written $X \overset{d}{=} Y$, if they have the same probability distribution. This means they have the same set of possible values, and the probabilities for these values are identical.

Example 3.3.4. In the three coin tosses experiment, the number of heads N_H and the number of tails N_T are *not* equal. For the outcome HHH, $N_H(\text{HHH}) = 3$ while $N_T(\text{HHH}) = 0$. However, their distributions are the same. Both can take values {0, 1, 2, 3}, and the probabilities are:

k	0	1	2	3
$P(N_H = k)$	1/8	3/8	3/8	1/8
$P(N_T = k)$	1/8	3/8	3/8	1/8

Since their probability distributions are identical, N_H and N_T are equal in distribution: $N_H \overset{d}{=} N_T$.

3.3.3 Relation Between the Two Equalities

Equality is a stronger condition than equality in distribution. If two random variables are equal, they must also be equal in distribution.

$$X = Y \implies X \overset{d}{=} Y$$

The converse is not true, as demonstrated by the N_H and N_T example.

3.4 Joint Distributions

To understand the relationship between two random variables, we need to look at them together.

Definition 3.4.1 (Joint Probability Distribution)**.** Let X and Y be two random variables defined on the same outcome space. Their **joint probability distribution** is a function that gives the probability of each pair of values (x, y). It is defined as:

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

for all possible values x of X and y of Y.

The joint distribution of two discrete random variables can be displayed in a **joint distribution table**. The values of X form the rows, the values of Y form the columns (or vice-versa), and the cells contain the probabilities $P(X = x, Y = y)$. The probabilities in the table must be non-negative and must sum to 1.

Example 3.4.2 (Two Draws Without Replacement)**.** Suppose we have a box with three tickets labeled 1, 2, and 3. We draw two tickets at random without replacement. Let X_1 be the number on the first ticket and X_2 be the number on the second. There are $3 \times 2 = 6$ equally likely outcomes: $(1, 2), (1, 3), (2, 1), (2, 3), (3, 1), (3, 2)$. Each has a probability of $1/6$. The joint distribution of X_1 and X_2 is given in the table below. Note that the diagonal entries are zero because we are drawing without replacement.

			X_2		
		1	2	3	Total
	1	0	1/6	1/6	2/6
X_1	2	1/6	0	1/6	2/6
	3	1/6	1/6	0	2/6
	Total	2/6	2/6	2/6	1

3.5 Marginal Distributions

The joint distribution contains all the probabilistic information about the two variables. From it, we can recover the distribution of each variable individually.

Definition 3.5.1 (Marginal Distribution)**.** The **marginal distribution** of a random variable X is its probability distribution, viewed in isolation from other variables. It can be calculated from a joint distribution by summing over the values of the other variable.

$$P(X = x) = \sum_{\text{all } y} P(X = x, Y = y)$$

This is often referred to as "summing out" or "marginalizing out" the variable Y.

In the joint distribution table, the marginal probabilities are found in the 'Total' row and column, which are also known as the margins of the table.

Example 3.5.2 (Marginals from the Two Draws Example)**.** Using the joint distribution table from the previous example: The marginal distribution of X_1 is found by summing the probabilities across the rows:

- $P(X_1 = 1) = 0 + 1/6 + 1/6 = 2/6 = 1/3$
- $P(X_1 = 2) = 1/6 + 0 + 1/6 = 2/6 = 1/3$
- $P(X_1 = 3) = 1/6 + 1/6 + 0 = 2/6 = 1/3$

So, X_1 has a uniform distribution on $\{1, 2, 3\}$.

The marginal distribution of X_2 is found by summing the probabilities down the columns:

- $P(X_2 = 1) = 0 + 1/6 + 1/6 = 2/6 = 1/3$
- $P(X_2 = 2) = 1/6 + 0 + 1/6 = 2/6 = 1/3$
- $P(X_2 = 3) = 1/6 + 1/6 + 0 = 2/6 = 1/3$

So, X_2 also has a uniform distribution on $\{1, 2, 3\}$. This is an interesting result, as one might think the second draw is different from the first.

3.6 Conditional Distributions

The core of understanding the relationship between variables lies in conditioning.

Definition 3.6.1 (Conditional Distribution)**.** Let X and Y be two random variables. The **conditional distribution of Y given X=x** is the distribution of Y under the condition that X has taken the value x. The conditional probabilities are given by:

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

This is defined only for values of x where $P(X = x) > 0$.

For a fixed value x, the function $P(Y = y|X = x)$ over all possible y is a valid probability distribution; its values are non-negative and sum to 1.

Example 3.6.2 (Conditional Distribution for the Two Draws)**.** Let's find the conditional distribution of X_2 given $X_1 = 3$. We use the formula with the values from our joint distribution table. We already know the marginal probability $P(X_1 = 3) = 1/3$.

- $P(X_2 = 1|X_1 = 3) = \frac{P(X_1=3,X_2=1)}{P(X_1=3)} = \frac{1/6}{1/3} = \frac{1}{2}$
- $P(X_2 = 2|X_1 = 3) = \frac{P(X_1=3,X_2=2)}{P(X_1=3)} = \frac{1/6}{1/3} = \frac{1}{2}$
- $P(X_2 = 3|X_1 = 3) = \frac{P(X_1=3,X_2=3)}{P(X_1=3)} = \frac{0}{1/3} = 0$

So, given that the first draw was a 3, the second draw is uniformly distributed on the remaining tickets, $\{1, 2\}$. This matches our intuition.

3.7 Dependence and Independence

The concept of independence is central to probability and statistics.

Definition 3.7.1 (Independent Random Variables)**.** Two random variables X and Y are said to be **independent** if for every pair of values (x, y), the events $\{X = x\}$ and $\{Y = y\}$ are independent. This means:

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

This condition is often called the **product rule for independent random variables**. It implies that the joint distribution table of two independent variables can be constructed by multiplying their marginal distributions.

An equivalent definition of independence can be stated using conditional probabilities: if X and Y are independent, then for any x with $P(X = x) > 0$,

$$P(Y = y|X = x) = P(Y = y)$$

This means that knowing the value of X does not change the distribution of Y. The conditional distribution of Y is the same as its marginal distribution.

In our running example of drawing tickets without replacement, X_1 and X_2 are **dependent**. We can see this because $P(X_1 = 1, X_2 = 1) = 0$, but $P(X_1 = 1)P(X_2 = 1) = (1/3)(1/3) = 1/9$. Since these are not equal, the variables are dependent.

Example 3.7.2 (Independent Random Variables: Two Dice Rolls)**.** Let a fair die be rolled twice. Let X_1 be the outcome of the first roll and X_2 be the outcome of the second. There are 36 equally likely outcomes. The marginal distribution for both X_1 and X_2 is uniform on $\{1, 2, 3, 4, 5, 6\}$, so $P(X_1 = i) = 1/6$ and $P(X_2 = j) = 1/6$ for any i, j in the set. The joint probability of any outcome (i, j) is $P(X_1 = i, X_2 = j) = 1/36$. We can check the product rule:

$$P(X_1 = i)P(X_2 = j) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

Since $P(X_1 = i, X_2 = j) = P(X_1 = i)P(X_2 = j)$ for all pairs (i, j), the random variables X_1 and X_2 are independent.

Chapter 4

Probabilities of Multiple Events

4.1 Bounding the Chance of a Union

Often, we are interested in the probability that at least one of a collection of events occurs, which corresponds to the probability of their union, $P(A_1 \cup A_2 \cup \cdots \cup A_n)$. Calculating this exactly can be complex, but we can start with a simple upper bound.

Theorem 4.1.1 (Boole's Inequality)**.** *For any finite or countably infinite collection of events $A_1, A_2, \ldots$, the following inequality holds:*

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

This inequality is also known as the union bound. It states that the probability of the union is no larger than the sum of the individual probabilities. The inequality is intuitive because if there is any overlap between the events, the sum on the right-hand side counts the probability of the intersections multiple times.

Proof. For two events A_1 and A_2, we know $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$. Since $P(A_1 \cap A_2) \geq 0$, we have $P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$. The proof for a general n events follows by induction. We assume the inequality holds for $n-1$ events.

$$P\left(\bigcup_{i=1}^{n} A_i\right) = P\left(\left(\bigcup_{i=1}^{n-1} A_i\right) \cup A_n\right) \leq P\left(\bigcup_{i=1}^{n-1} A_i\right) + P(A_n)$$

By the induction hypothesis, $P\left(\bigcup_{i=1}^{n-1} A_i\right) \leq \sum_{i=1}^{n-1} P(A_i)$. Substituting this in gives:

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n-1} P(A_i) + P(A_n) = \sum_{i=1}^{n} P(A_i)$$

This completes the proof. □

4.2 The Inclusion-Exclusion Formula

Boole's inequality gives an upper bound. To find the exact probability of a union, we need a formula that corrects for the over-counting. This is the Inclusion-Exclusion formula.

4.2.1 Formula for Two and Three Events

For two events, the formula is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

For three events, we add the individuals, subtract the pairs, and add back the triple intersection:

$$\begin{aligned} P(A \cup B \cup C) = &P(A) + P(B) + P(C) \\ &- [P(A \cap B) + P(A \cap C) + P(B \cap C)] \\ &+ P(A \cap B \cap C) \end{aligned}$$

4.2.2 General Formula

For a collection of n events $A_1, A_2, \ldots, A_n$, the general formula is:

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) = &\sum_{i=1}^{n} P(A_i) - \sum_{1\leq i<j\leq n} P(A_i \cap A_j) \\ &+ \sum_{1\leq i<j<k\leq n} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1} P(A_1 \cap \cdots \cap A_n) \end{aligned}$$

In this formula, we sum over all single events, then subtract the sum over all pairs, add the sum over all triples, and so on, alternating signs until we reach the intersection of all n events.

4.3 The Matching Problem

A classic application of the Inclusion-Exclusion formula is the matching problem, also known as the problem of derangements.

Problem: Suppose n letters, intended for n different recipients, are randomly placed into n pre-addressed envelopes. What is the probability that at least one letter is placed in the correct envelope?

Assume all $n!$ permutations of letters into envelopes are equally likely. Let A_i be the event that letter i goes into envelope i. We want to compute $P(A_1 \cup A_2 \cup \cdots \cup A_n)$. We use the Inclusion-Exclusion formula.

Step 1: Calculate the sum of single probabilities. The probability that a specific letter i goes into its correct envelope is $P(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n}$, since the other $n-1$ letters can be arranged in $(n-1)!$ ways. There are $\binom{n}{1}$ such terms, so the first sum is $\binom{n}{1}\frac{1}{n} = 1$.

Step 2: Calculate the sum of pairwise intersection probabilities. The probability that two specific letters, i and j, both go into their correct envelopes is $P(A_i \cap A_j) = \frac{(n-2)!}{n!}$. There are $\binom{n}{2}$ such pairs. The second sum is $\binom{n}{2}\frac{(n-2)!}{n!} = \frac{n(n-1)}{2}\frac{(n-2)!}{n!} = \frac{1}{2!}$.

Step 3: Calculate the sum of k-wise intersection probabilities. In general, for any k distinct indices $i_1, \ldots, i_k$, the probability of the intersection $P(A_{i_1} \cap \cdots \cap A_{i_k})$ is the chance that those k letters are in the correct envelopes. This is $\frac{(n-k)!}{n!}$. There are $\binom{n}{k}$ such intersections of size k. The k-th term in the Inclusion-Exclusion formula is:

$$\binom{n}{k} P(A_1 \cap \cdots \cap A_k) = \binom{n}{k}\frac{(n-k)!}{n!} = \frac{n!}{k!(n-k)!}\frac{(n-k)!}{n!} = \frac{1}{k!}$$

Step 4: Combine the terms. The probability of at least one match is:

$$P(\text{at least one match}) = \sum_{k=1}^{n}(-1)^{k-1}\frac{1}{k!} = 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n-1}\frac{1}{n!}$$

The probability of **no matches** is $P(\text{no match}) = 1 - P(\text{at least one match})$.

$$P(\text{no match}) = 1 - \left(1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n-1}\frac{1}{n!}\right) = \sum_{k=0}^{n}\frac{(-1)^k}{k!}$$

As $n \to \infty$, this sum converges to the Taylor series expansion of e^{-1}. So, for large n, the probability of no matches is approximately $1/e \approx 0.3679$.

4.4 Application to Sampling Without Replacement

Let's use the inclusion-exclusion principle to find the probability of getting at least one "good" element in a sample drawn without replacement.

Problem: A population of N items contains G good items and B bad items ($N = G + B$). We draw a simple random sample of size n without replacement. What is the probability that the sample contains at least one good item?

Let A_i be the event that the i-th good item (for $i = 1, \ldots, G$) is in the sample. We want to find $P(A_1 \cup A_2 \cup \cdots \cup A_G)$.

Step 1: Calculate intersection probabilities. Let's find the probability of the intersection of k of these events, say $P(A_1 \cap \cdots \cap A_k)$. This is the probability that k specific good items are all in our sample of size n. We can count the number of samples. The total number of samples of size n is $\binom{N}{n}$. To form a sample containing these k specific good items, we must choose them, and then choose the remaining $n-k$ members of the sample from the remaining $N-k$ items in the population. The number of ways to do this is $\binom{N-k}{n-k}$. So, the probability is:

$$P(A_1 \cap \cdots \cap A_k) = \frac{\binom{N-k}{n-k}}{\binom{N}{n}}$$

This probability is the same for any set of k specific good items.

Step 2: Apply the Inclusion-Exclusion Formula. The sum of all k-wise intersections is $\binom{G}{k}P(A_1 \cap \cdots \cap A_k) = \binom{G}{k}\frac{\binom{N-k}{n-k}}{\binom{N}{n}}$. The Inclusion-Exclusion formula gives:

$$P(\text{at least one good item}) = \sum_{k=1}^{G}(-1)^{k-1}\binom{G}{k}\frac{\binom{N-k}{n-k}}{\binom{N}{n}}$$

Alternative Method (Complement Rule) This problem is much more easily solved using the complement rule. The complement event is "no good items are in the sample," which means all n items must be drawn from the B bad items. The number of ways to do this is $\binom{B}{n}$. So, $P(\text{no good items}) = \frac{\binom{B}{n}}{\binom{N}{n}}$. And the desired probability is:

$$P(\text{at least one good item}) = 1 - \frac{\binom{B}{n}}{\binom{N}{n}}$$

Since $B = N - G$, this is $1 - \frac{\binom{N-G}{n}}{\binom{N}{n}}$. The fact that this simple expression is equal to the complicated sum from the Inclusion-Exclusion formula is a known combinatorial identity. This example highlights that while the Inclusion-Exclusion formula is always valid, a more direct method is often preferable if available.

Chapter 5

Discrete Variables

5.1 The Binomial Distribution

One of the simplest and most common types of random variables is a count of successes in a sequence of trials. This leads to the binomial distribution.

Definition 5.1.1 (Binomial Distribution)**.** A random variable X has the **binomial distribution** with parameters n and p, written $X \sim \text{Binomial}(n, p)$, if it represents the number of successes in n independent trials, where the probability of success in each trial is p.

The assumptions for a binomial model are:

- A fixed number of trials, n.
- The trials are independent of each other.
- Each trial has only two possible outcomes, ”success” and ”failure”.
- The probability of success, p, is the same for all trials.

The probability mass function of a binomial random variable is given by the **Binomial Formula**:

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

- $\binom{n}{k}$ is the binomial coefficient, representing the number of ways to choose k positions for the successes out of n trials.
- p^k is the probability of getting k successes.
- $(1 - p)^{n-k}$ is the probability of getting $n - k$ failures.

Example 5.1.2 (Flipping a Biased Coin)**.** A coin is biased with $P(\text{Heads}) = 0.7$. It is tossed 10 times. What is the probability of getting exactly 6 heads?

This is a binomial setting with $n = 10$ trials and success probability $p = 0.7$. Let X be the number of heads. We want to find $P(X = 6)$.

$$P(X = 6) = \binom{10}{6}(0.7)^6(0.3)^{10-6} = 210 \cdot (0.7)^6(0.3)^4$$

$$P(X = 6) = 210 \cdot (0.117649) \cdot (0.0081) \approx 0.2001$$

5.2 The Multinomial Distribution

The multinomial distribution generalizes the binomial distribution to the case where each trial can have more than two possible outcomes.

Definition 5.2.1 (Multinomial Distribution)**.** Suppose we have n independent trials. Each trial can result in one of m categories. For each trial, the probability of resulting in category i is p_i, where $\sum_{i=1}^{m} p_i = 1$. Let N_i be the random count of outcomes in category i after n trials. The vector of counts $(N_1, N_2, \ldots, N_m)$ has a multinomial distribution.

The joint probability mass function for the counts being $n_1, n_2, \ldots, n_m$ (where $\sum n_i = n$) is:

$$P(N_1 = n_1, \ldots, N_m = n_m) = \frac{n!}{n_1!n_2!\ldots n_m!} p_1^{n_1} p_2^{n_2} \ldots p_m^{n_m}$$

- The term $\frac{n!}{n_1!\ldots n_m!}$ is the multinomial coefficient, which counts the number of ways to arrange the n outcomes into the specified groups.
- The term $p_1^{n_1} \ldots p_m^{n_m}$ is the probability of any specific sequence with that number of outcomes in each category.

5.3 The Hypergeometric Distribution Revisited

The hypergeometric distribution describes the number of "good" elements in a simple random sample drawn without replacement. Let a population have N items, G of which are "good" and $B = N - G$ are "bad". If we draw a sample of size n, the number of good items in the sample, X, follows the hypergeometric distribution:

$$P(X = k) = \frac{\binom{G}{k}\binom{B}{n-k}}{\binom{N}{n}}$$

5.3.1 Binomial Approximation to the Hypergeometric

If the population size N is very large compared to the sample size n, then drawing without replacement is "almost" like drawing with replacement. In this case, the hypergeometric distribution can be well-approximated by the binomial distribution. The argument is that the probability of success on each draw does not change much. The probability of the first draw being good is $p = G/N$. The probability of the second being good, given the first was good, is $(G-1)/(N-1)$, which is very close to p if G and N are large. So, for large N and small n/N, we have:

$$\text{Hypergeometric}(N, G, n) \approx \text{Binomial}(n, p = G/N)$$

5.4 Odds and Odds Ratios

Probabilities can be expressed in terms of odds, which is common in some fields like gaming and epidemiology.

Definition 5.4.1 (Odds)**.** The **odds** of an event A are defined as the ratio of the probability of the event to the probability of its complement:

$$o(A) = \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)}$$

We can convert back from odds to probability using the formula: $P(A) = \frac{o(A)}{1+o(A)}$.

5.4.1 Bayes' Rule in Odds Form

A particularly elegant form of Bayes' Rule uses odds. Let A be an event and B be some data or evidence.

$$\text{Posterior Odds of } A = \frac{P(A|B)}{P(A^c|B)} = \frac{P(A)P(B|A)}{P(A^c)P(B|A^c)} = \frac{P(A)}{P(A^c)} \cdot \frac{P(B|A)}{P(B|A^c)}$$

This gives the rule:

$$\text{Posterior Odds} = \text{Prior Odds} \times \text{Likelihood Ratio}$$

The likelihood ratio $\frac{P(B|A)}{P(B|A^c)}$ measures how much the data B supports event A over its complement.

5.5 The Law of Small Numbers: Poisson Approximation

When the number of trials n is large and the probability of success p is small, the binomial distribution can be approximated by the Poisson distribution. This is often called the "law of small numbers".

Definition 5.5.1 (Poisson Distribution)**.** A random variable X has the **Poisson distribution** with parameter $\mu > 0$, written $X \sim \text{Poisson}(\mu)$, if its probability mass function is:

$$P(X = k) = e^{-\mu}\frac{\mu^k}{k!}, \quad \text{for } k = 0, 1, 2, \ldots$$

5.5.1 Derivation from the Binomial

Consider a binomial distribution where $n \to \infty$ and $p \to 0$ in such a way that the product $\mu = np$ remains constant.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1)\dots(n-k+1)}{k!} \left(\frac{\mu}{n}\right)^k \left(1 - \frac{\mu}{n}\right)^{n-k}$$

We rearrange the terms:

$$= \frac{1}{k!} \frac{n(n-1)\dots(n-k+1)}{n^k} \mu^k \left(1 - \frac{\mu}{n}\right)^n \left(1 - \frac{\mu}{n}\right)^{-k}$$

Now we take the limit as $n \to \infty$:

- The term $\frac{n(n-1)\dots(n-k+1)}{n^k} = \frac{n}{n} \frac{n-1}{n} \dots \frac{n-k+1}{n} \to 1 \cdot 1 \dots 1 = 1$.
- The term μ^k is constant with respect to n.
- The term $\left(1 - \frac{\mu}{n}\right)^n \to e^{-\mu}$.
- The term $\left(1 - \frac{\mu}{n}\right)^{-k} \to (1-0)^{-k} = 1$.

Putting it all together, we get the Poisson formula:

$$P(X = k) \to \frac{1}{k!} \cdot 1 \cdot \mu^k \cdot e^{-\mu} \cdot 1 = e^{-\mu} \frac{\mu^k}{k!}$$

This approximation is generally good when n is large (e.g., $n > 50$) and p is small (e.g., $p < 0.05$).

5.6 The Poisson Distribution

First, let's review the Poisson distribution and its key properties.

Definition 5.6.1 (Poisson Distribution)**.** A random variable X has the **Poisson distribution** with parameter $\mu > 0$, written $X \sim \text{Poisson}(\mu)$, if its probability mass function is given by:

$$P(X = k) = e^{-\mu} \frac{\mu^k}{k!}, \quad \text{for } k = 0, 1, 2, \dots$$

The parameter μ represents the expected number of events and is also approximately the mode of the distribution. The sum of the probabilities is 1, which follows from the Taylor series expansion of e^{μ}:

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

Theorem 5.6.2 (Sum of Independent Poissons)**.** *If $X \sim Poisson(\mu)$ and $Y \sim Poisson(\lambda)$ are independent random variables, then their sum $S = X + Y$ also follows a Poisson distribution with parameter $\mu + \lambda$.*

Proof. The possible values of S are non-negative integers. For any integer $s \geq 0$, we find $P(S = s)$ by conditioning on the value of X. Since X and Y are non-negative, X can take any integer value from 0 to s.

$$
\begin{aligned}
P(S = s) &= \sum_{k=0}^{s} P(X = k, Y = s - k) \\
&= \sum_{k=0}^{s} P(X = k) P(Y = s - k) \quad \text{(by independence)} \\
&= \sum_{k=0}^{s} \left(e^{-\mu} \frac{\mu^k}{k!} \right) \left(e^{-\lambda} \frac{\lambda^{s-k}}{(s-k)!} \right) \\
&= e^{-(\mu+\lambda)} \sum_{k=0}^{s} \frac{\mu^k \lambda^{s-k}}{k!(s-k)!} \\
&= \frac{e^{-(\mu+\lambda)}}{s!} \sum_{k=0}^{s} \frac{s!}{k!(s-k)!} \mu^k \lambda^{s-k} \\
&= \frac{e^{-(\mu+\lambda)}}{s!} (\mu + \lambda)^s \quad \text{(by the Binomial Theorem)}
\end{aligned}
$$

This is the probability mass function of a Poisson($\mu + \lambda$) distribution. □

5.7 Poissonizing the Binomial

The core idea of Poissonization is to take a fixed number of trials and make it random.

The Model:

1. Let N be a random variable with a Poisson(μ) distribution.
2. Let the conditional distribution of a random variable S given $N = n$ be Binomial(n, p).

This models a situation with a random number of trials, where each trial is an independent success with probability p.

Theorem 5.7.1. *In the model described above, the number of successes S and the number of failures $F = N - S$ have the following properties:*

1. *S has a Poisson(μp) distribution.*
2. *F has a Poisson($\mu(1 - p)$) distribution.*
3. *S and F are independent.*

Proof. **1. Distribution of S:** The possible values of S are $0, 1, 2, \ldots$. For an integer $s \geq 0$, we find $P(S = s)$ by conditioning on N. The sum must start from $n = s$ since we need at least s trials to have s successes.

$$\begin{aligned}
P(S = s) &= \sum_{n=s}^{\infty} P(N = n, S = s) = \sum_{n=s}^{\infty} P(N = n)P(S = s|N = n) \\
&= \sum_{n=s}^{\infty} \left(e^{-\mu}\frac{\mu^n}{n!}\right)\left(\binom{n}{s}p^s(1-p)^{n-s}\right) \\
&= \sum_{n=s}^{\infty} e^{-\mu}\frac{\mu^n}{n!}\frac{n!}{s!(n-s)!}p^s(1-p)^{n-s} \\
&= \frac{e^{-\mu}p^s}{s!}\sum_{n=s}^{\infty}\frac{\mu^n(1-p)^{n-s}}{(n-s)!} \\
&= \frac{e^{-\mu}(\mu p)^s}{s!}\sum_{n=s}^{\infty}\frac{(\mu(1-p))^{n-s}}{(n-s)!}
\end{aligned}$$

Let $j = n - s$. The sum becomes $\sum_{j=0}^{\infty}\frac{(\mu(1-p))^j}{j!} = e^{\mu(1-p)}$.

$$P(S = s) = \frac{e^{-\mu}(\mu p)^s}{s!}e^{\mu(1-p)} = e^{-\mu p}\frac{(\mu p)^s}{s!}$$

This is the PMF of a Poisson(μp) distribution.

2. Distribution of F: This follows by symmetry. If we call "failures" our new successes, their probability is $1 - p$. The number of failures F is thus Poisson with parameter $\mu(1 - p)$.

3. Independence of S and F: We compute the joint probability $P(S = s, F = f)$. This event is equivalent to having s successes and f failures, which means the total number of trials must have been $N = s + f$.

$$\begin{aligned}
P(S = s, F = f) &= P(N = s + f, S = s) \\
&= P(N = s + f)P(S = s|N = s + f) \\
&= \left(e^{-\mu}\frac{\mu^{s+f}}{(s+f)!}\right)\left(\binom{s+f}{s}p^s(1-p)^f\right) \\
&= e^{-\mu}\frac{\mu^{s+f}}{(s+f)!}\frac{(s+f)!}{s!f!}p^s(1-p)^f \\
&= e^{-\mu}\frac{(\mu p)^s(\mu(1-p))^f}{s!f!} \\
&= \left(e^{-\mu p}\frac{(\mu p)^s}{s!}\right)\left(e^{-\mu(1-p)}\frac{(\mu(1-p))^f}{f!}\right) \\
&= P(S = s)P(F = f)
\end{aligned}$$

Since the joint distribution is the product of the marginals, S and F are independent. This is a remarkable result: randomizing the number of trials with a Poisson distribution breaks the strong dependence between the number of successes and failures. □

5.8 Poissonizing the Multinomial

The results from the binomial case extend naturally to the multinomial case, where each trial can result in one of $m > 2$ categories.

The Model:

1. Let the total number of trials N be a Poisson(μ) random variable.
2. Conditional on $N = n$, the vector of counts in each category $(X_1, X_2, \dots, X_m)$ has a Multinomial($n, p_1, p_2, \dots, p_m$) distribution, where $\sum p_i = 1$.

Theorem 5.8.1. *In the multinomial model with a Poisson number of trials, the counts in each category, $X_1, X_2, \dots, X_m$, are mutually independent random variables, and each X_i follows a Poisson distribution with parameter μp_i.*

Proof. The proof is a direct extension of the binomial case. We find the joint PMF for a set of counts $n_1, n_2, \dots, n_m$. Let $n = \sum_{i=1}^m n_i$. This event requires that the total number of trials was $N = n$.

$$\begin{aligned}
P(X_1 = n_1, \dots, X_m = n_m) &= P(N = n) P(X_1 = n_1, \dots | N = n) \\
&= \left(e^{-\mu} \frac{\mu^n}{n!}\right) \left(\frac{n!}{n_1! n_2! \dots n_m!} p_1^{n_1} p_2^{n_2} \dots p_m^{n_m}\right) \\
&= e^{-\mu} \frac{(\mu p_1)^{n_1} (\mu p_2)^{n_2} \dots (\mu p_m)^{n_m}}{n_1! n_2! \dots n_m!} \\
&= \left(e^{-\mu p_1} \frac{(\mu p_1)^{n_1}}{n_1!}\right) \dots \left(e^{-\mu p_m} \frac{(\mu p_m)^{n_m}}{n_m!}\right) \\
&= \prod_{i=1}^m P(X_i = n_i) \quad \text{where } X_i \sim \text{Poisson}(\mu p_i)
\end{aligned}$$

The joint PMF factors into the product of individual Poisson PMFs, proving both the distributional form and the independence of the counts. □

Example 5.8.2 (Dice Rolling)**.** Suppose a fair die is rolled N times, where $N \sim \text{Poisson}(18)$. Let X_i be the number of times face i appears, for $i = 1, \dots, 6$. Here, $p_i = 1/6$ for all i. According to the theorem, the counts $X_1, \dots, X_6$ are independent, and each X_i is Poisson with parameter $\mu p_i = 18 \times (1/6) = 3$. What is the probability that each face appears at most twice? Because of independence, we can calculate this by finding the probability for one face and raising it to the power of 6.

$$P(X_1 \leq 2) = P(X_1 = 0) + P(X_1 = 1) + P(X_1 = 2) = e^{-3} \frac{3^0}{0!} + e^{-3} \frac{3^1}{1!} + e^{-3} \frac{3^2}{2!}$$

$$P(\text{All } X_i \leq 2) = [P(X_1 \leq 2)]^6 = \left(\sum_{k=0}^2 e^{-3} \frac{3^k}{k!}\right)^6$$

This calculation would be extremely difficult without Poissonization due to the complex dependencies in the multinomial distribution.

Chapter 6

Expectation

6.1 Definition of Expectation

Definition 6.1.1 (Expectation)**.** Let X be a random variable with a discrete set of possible values. The **expected value** of X, denoted $E(X)$, is defined as the sum of each possible value multiplied by its probability:

$$E(X) = \sum_x x \cdot P(X = x)$$

The sum is taken over all possible values x that the random variable X can assume. For the expectation to be well-defined, this sum must be absolutely convergent, i.e., $\sum_x |x| P(X = x) < \infty$. For any random variable with a finite number of possible values, the expectation is always well-defined.

Example 6.1.2 (Bernoulli Trial)**.** Let X be a random variable with a Bernoulli(p) distribution. X takes the value 1 with probability p and 0 with probability $1-p$. The expectation is:

$$E(X) = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = 1 \cdot p + 0 \cdot (1 - p) = p$$

Example 6.1.3 (Uniform on Integers)**.** Let X have a uniform distribution on the integers $\{1, 2, \ldots, n\}$. For each integer k in this range, $P(X = k) = 1/n$. The expectation is:

$$E(X) = \sum_{k=1}^{n} k \cdot P(X = k) = \sum_{k=1}^{n} k \cdot \frac{1}{n} = \frac{1}{n} \sum_{k=1}^{n} k$$

Using the formula for the sum of the first n integers, $\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$, we get:

$$E(X) = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}$$

This result makes intuitive sense; the center of the distribution is the midpoint of the range.

6.2 Expectation of a Function of a Random Variable

We often need to find the expectation of a function of a random variable, say $g(X)$. One way would be to find the distribution of the new random variable $Y = g(X)$ and then apply the definition. However, a more direct method is available.

Theorem 6.2.1 (Law of the Unconscious Statistician - LOTUS)**.** *Let X be a discrete random variable and g be a real-valued function. The expectation of the random variable $g(X)$ can be calculated directly without finding its distribution, using the following formula:*

$$E(g(X)) = \sum_{x} g(x)P(X = x)$$

The sum is over all possible values x of the original random variable X.

Example 6.2.2 ($E(X^2)$ for a single die roll)**.** Let X be the result of a single roll of a fair six-sided die. We know $P(X = k) = 1/6$ for $k = 1, \ldots, 6$. Let's find $E(X^2)$ using LOTUS with $g(x) = x^2$.

$$E(X^2) = \sum_{k=1}^{6} k^2 \cdot P(X = k) = \sum_{k=1}^{6} k^2 \cdot \frac{1}{6}$$

$$E(X^2) = \frac{1}{6}(1^2+2^2+3^2+4^2+5^2+6^2) = \frac{1}{6}(1+4+9+16+25+36) = \frac{91}{6} \approx 15.17$$

6.3 Linearity of Expectation

Expectation has a powerful linear property that is fundamental to many calculations in probability.

Theorem 6.3.1 (Additivity of Expectation)**.** *For any two random variables X and Y defined on the same outcome space, the expectation of their sum is the sum of their expectations:*

$$E(X + Y) = E(X) + E(Y)$$

This property holds regardless of whether the variables are independent or dependent.

Proof. Let $S = X + Y$. Using LOTUS for a function of two variables:

$$\begin{aligned} E(S) = E(X + Y) &= \sum_{\omega \in \Omega} (X(\omega) + Y(\omega))P(\omega) \\ &= \sum_{\omega \in \Omega} X(\omega)P(\omega) + \sum_{\omega \in \Omega} Y(\omega)P(\omega) \\ &= E(X) + E(Y) \end{aligned}$$

By induction, this extends to the sum of any finite number of random variables: $E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)$. □

Proposition 6.3.2 (Linearity of Expectation). *For any random variable X and any constants a, b:*

$$E(aX + b) = aE(X) + b$$

6.4 The Method of Indicators

The method of indicators is a particularly elegant application of the linearity of expectation. It simplifies the calculation of the expectation of complex random variables by breaking them down into a sum of simple indicator variables.

Definition 6.4.1 (Indicator Random Variable). An **indicator random variable** I_A for an event A is a random variable that takes the value 1 if event A occurs, and 0 otherwise.

$$I_A = \begin{cases} 1 & \text{if A occurs} \\ 0 & \text{if A does not occur} \end{cases}$$

The expectation of an indicator variable is simply the probability of the event it indicates.

$$E(I_A) = 1 \cdot P(A) + 0 \cdot P(A^c) = P(A)$$

Example 6.4.2 (Expectation of a Binomial Random Variable). Let $X \sim \text{Binomial}(n, p)$. We can calculate $E(X)$ using the definition, but it involves a cumbersome sum. Instead, we can represent X as a sum of indicators. Let I_j be the indicator of success on the j-th trial, for $j = 1, \ldots, n$. Since each trial is a success with probability p, we have $E(I_j) = p$ for all j. The total number of successes is $X = I_1 + I_2 + \cdots + I_n$. Using the additivity of expectation:

$$E(X) = E(I_1 + I_2 + \cdots + I_n) = E(I_1) + E(I_2) + \cdots + E(I_n)$$

$$E(X) = \underbrace{p + p + \cdots + p}_{n \text{ times}} = np$$

This derivation is remarkably simple and does not even require the trials to be independent, only that the probability of success on each trial is p.

Example 6.4.3 (Expectation of a Hypergeometric Random Variable). A population has N items, G of which are good. A simple random sample of size n is drawn without replacement. Let X be the number of good items in the sample. X has a hypergeometric distribution. Calculating $E(X)$ with the hypergeometric PMF is very difficult. Instead, we use indicators. Let I_j be the indicator that the j-th draw in the sample is a good item, for $j = 1, \ldots, n$. The total number of good items is $X = I_1 + I_2 + \cdots + I_n$. By additivity, $E(X) = \sum_{j=1}^{n} E(I_j)$. We need to find $E(I_j) = P(\text{draw } j \text{ is good})$. By symmetry in simple random sampling, any draw is equally likely to be any of the N items in the population. Therefore, the probability that any specific draw is good is the same as the probability for the first draw:

$$P(\text{draw } j \text{ is good}) = \frac{G}{N}$$

Thus, $E(I_j) = G/N$ for every $j = 1, \ldots, n$.

$$E(X) = \sum_{j=1}^{n} \frac{G}{N} = n\frac{G}{N}$$

This result is obtained with great ease compared to the combinatorial sums required by the definition. It highlights the power of linearity of expectation and the indicator method.

6.5 Expectation by Conditioning

We can find the expectation of a random variable T by conditioning on another random variable S.

Definition 6.5.1 (Conditional Expectation)**.** For two random variables S and T, the **conditional expectation of T given S=s**, denoted $E(T|S=s)$, is the expected value of the conditional distribution of T given $S = s$. The function that maps each value s to $E(T|S=s)$ defines a new random variable called the **conditional expectation of T given S**, denoted $E(T|S)$.

Theorem 6.5.2 (Law of Iterated Expectations)**.** *For any two random variables S and T, the expectation of T is the expectation of the conditional expectation of T given S.*

$$E(T) = E(E(T|S))$$

Proof. The proof proceeds by expanding the definitions:

$$\begin{aligned}
E(T) &= \sum_t tP(T=t) = \sum_t t \sum_s P(S=s, T=t) \\
&= \sum_t t \sum_s P(S=s)P(T=t|S=s) \\
&= \sum_s P(S=s) \left(\sum_t tP(T=t|S=s) \right) \\
&= \sum_s P(S=s)E(T|S=s) \\
&= E(E(T|S))
\end{aligned}$$

□

Example 6.5.3 (Random Sums)**.** Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ_X. Let N be a non-negative integer-valued random variable, independent of the X_i, with mean μ_N. Let $S = \sum_{i=1}^{N} X_i$ (with $S = 0$ if $N = 0$). Find $E(S)$.

We condition on N. Given $N = n$, S is the sum of n i.i.d. variables.

$$E(S|N = n) = E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) = n\mu_X$$

This defines the random variable $E(S|N) = N\mu_X$. Now we use iterated expectations:

$$E(S) = E(E(S|N)) = E(N\mu_X) = \mu_X E(N) = \mu_N \mu_X$$

The expected value of a random sum is the expected number of terms times the expected value of each term.

6.6 Expected Waiting Times

A common application of conditional expectation is to find the expected time until some event occurs. The strategy is to condition on the first step(s) of the process.

Example 6.6.1 (Waiting for a Success)**.** Let W_H be the number of tosses of a coin required to get the first head. The probability of heads is p. Find $E(W_H)$.

Let $x = E(W_H)$. We always make at least one toss. After the first toss:

- It is a head (probability p). The process stops. The total number of tosses is 1.
- It is a tail (probability $q = 1 - p$). The process must start over. The additional number of tosses needed is again a random variable with expectation x. So the total number of tosses is $1 + W_H^*$, where W_H^* is an independent copy of W_H.

By conditioning on the first toss:

$$x = E(W_H) = p \cdot (1) + q \cdot E(1 + W_H^*) = p + q(1 + x)$$

Since $p + q = 1$, this becomes $x = 1 + qx$. Solving for x:

$$x(1 - q) = 1 \implies xp = 1 \implies x = \frac{1}{p}$$

This confirms the known result for the geometric distribution.

Example 6.6.2 (Waiting for HH)**.** In tosses of a p-coin, let W_{HH} be the waiting time until two consecutive heads (HH) appear. Find $E(W_{HH})$.

Let $x = E(W_{HH})$. We condition on the first toss.

- First toss is T (prob q): We've wasted one toss and are back to the start. The expected total time is $1 + x$.

- First toss is H (prob p): Now we condition on the second toss.
  - Second toss is T (prob q): We have sequence HT. We've wasted two tosses and are back to the start. Expected total time is $2 + x$.
  - Second toss is H (prob p): We have sequence HH. The process stops. Total time is 2.

Putting this together in an equation for x:

$$x = q(1 + x) + p[q(2 + x) + p(2)]$$

$$x = q + qx + 2pq + pqx + 2p^2$$

$$x(1-q-pq) = q+2pq+2p^2 = q(1+2p)+2p^2 = (1-p)(1+2p)+2p^2 = 1+p-2p^2+2p^2 = 1+p$$

$$x(p - pq) = 1 + p \implies x(p(1 - q)) = 1 + p \implies xp^2 = 1 + p \implies x = \frac{1 + p}{p^2}$$

The expected waiting time can be written as $\frac{1}{p^2} + \frac{1}{p}$.

Chapter 7

Standard Deviation, Variance and Covariance

7.1 Variance and Standard Deviation

Definition 7.1.1 (Variance and Standard Deviation)**.** Let X be a random variable with expectation $E(X) = \mu$.

- The **variance** of X, denoted $\mathrm{Var}(X)$, is the expected squared deviation from the mean:
$$\mathrm{Var}(X) = E((X - \mu)^2)$$

- The **standard deviation** of X, denoted $SD(X)$ or σ_X, is the square root of the variance:
$$SD(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{E((X - \mu)^2)}$$

The standard deviation is in the same units as X, which makes it more interpretable than the variance. It represents a typical distance between the values of X and the mean μ.

7.1.1 Computational Formula for Variance

Calculating variance directly from the definition can be tedious. A more convenient formula is often used.

Theorem 7.1.2 (Computational Formula)**.** *The variance of a random variable X can be calculated as the expectation of the square of X minus the square of the expectation of X.*

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

Proof. Let $E(X) = \mu$.

$$\begin{aligned}
\text{Var}(X) &= E((X-\mu)^2) \\
&= E(X^2 - 2\mu X + \mu^2) \\
&= E(X^2) - E(2\mu X) + E(\mu^2) \quad \text{(by linearity of expectation)} \\
&= E(X^2) - 2\mu E(X) + \mu^2 \quad \text{(since } \mu \text{ is a constant)} \\
&= E(X^2) - 2\mu(\mu) + \mu^2 \\
&= E(X^2) - 2\mu^2 + \mu^2 \\
&= E(X^2) - \mu^2 = E(X^2) - (E(X))^2
\end{aligned}$$

This completes the proof. □

7.1.2 Properties of SD

1. **Shifting:** For any constant c, $SD(X + c) = SD(X)$. Adding a constant shifts the distribution but does not change its spread.
2. **Scaling:** For any constant a, $SD(aX) = |a|SD(X)$. Scaling the variable by a factor a scales the spread by $|a|$.

Example 7.1.3 (SD of a single die roll)**.** Let X be the result of rolling a fair six-sided die. We found earlier that $E(X) = 3.5$. To find the variance, we first need $E(X^2)$.

$$E(X^2) = \sum_{k=1}^{6} k^2 P(X = k) = \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = \frac{91}{6}$$

Now we use the computational formula for variance:

$$\text{Var}(X) = E(X^2) - (E(X))^2 = \frac{91}{6} - (3.5)^2 = \frac{91}{6} - (7/2)^2 = \frac{91}{6} - \frac{49}{4}$$

$$\text{Var}(X) = \frac{182 - 147}{12} = \frac{35}{12} \approx 2.917$$

The standard deviation is:

$$SD(X) = \sqrt{\frac{35}{12}} \approx 1.708$$

7.2 Prediction and Estimation

A key problem in data science is to predict the value of a random variable Y based on another random variable X. We can start by considering the simplest case: predicting Y with no information from other variables.

7.2.1 Mean Squared Error

Suppose we want to predict a random variable Y using a single constant value c. What is the best choice for c? A common way to measure the quality of a prediction is to use the **mean squared error (MSE)**.

Definition 7.2.1 (Mean Squared Error)**.** The mean squared error of predicting a random variable Y with a constant c is defined as:

$$MSE(c) = E((Y - c)^2)$$

The goal is to find the value of c that minimizes this MSE. The predictor c that achieves this minimum is called the **best constant predictor**.

Theorem 7.2.2. *The best constant predictor for a random variable Y is its expectation, $c = E(Y)$.*

Proof. Let $\mu_Y = E(Y)$. We can expand the MSE expression:

$$\begin{aligned} E((Y-c)^2) &= E(((Y-\mu_Y) + (\mu_Y - c))^2) \\ &= E((Y-\mu_Y)^2 + 2(Y-\mu_Y)(\mu_Y - c) + (\mu_Y - c)^2) \\ &= E((Y-\mu_Y)^2) + 2(\mu_Y - c)E(Y-\mu_Y) + (\mu_Y - c)^2 \end{aligned}$$

The middle term is zero because $E(Y - \mu_Y) = E(Y) - E(\mu_Y) = \mu_Y - \mu_Y = 0$. So we have:

$$MSE(c) = E((Y-\mu_Y)^2) + (\mu_Y - c)^2$$

The first term, $E((Y - \mu_Y)^2)$, is just the variance $\text{Var}(Y)$, which does not depend on our choice of c. The second term, $(\mu_Y - c)^2$, is a squared quantity and is always non-negative. To minimize the sum, we must make the second term as small as possible. This is achieved when $(\mu_Y - c)^2 = 0$, which implies $c = \mu_Y$. □

7.2.2 Root Mean Squared Error (RMSE)

When we use the best predictor $c = E(Y)$, the minimum possible MSE is:

$$\min_c MSE(c) = E((Y - E(Y))^2) = \text{Var}(Y)$$

The square root of this minimum error is called the **root mean squared error (RMSE)**.

$$\text{RMSE} = \sqrt{\min_c MSE(c)} = \sqrt{\text{Var}(Y)} = SD(Y)$$

This gives a powerful interpretation of the standard deviation:

The standard deviation of a random variable Y is the root mean squared error of predicting Y using the best constant predictor, E(Y).

The SD measures the inherent uncertainty or prediction error we have about a variable when we only know its distribution.

7.3 Covariance

Definition 7.3.1 (Covariance)**.** Let X and Y be two random variables with expectations $E(X) = \mu_X$ and $E(Y) = \mu_Y$. Their **covariance**, denoted $\mathrm{Cov}(X, Y)$, is the expected product of their deviations from their respective means:

$$\mathrm{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y))$$

A positive covariance indicates that the variables tend to move in the same direction (if one is above its mean, the other is likely to be as well). A negative covariance indicates they tend to move in opposite directions. A covariance of zero suggests no linear association.

Theorem 7.3.2 (Computational Formula for Covariance)**.** *A more practical formula for computing covariance is:*

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

Proof. Let $\mu_X = E(X)$ and $\mu_Y = E(Y)$.

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= E((X - \mu_X)(Y - \mu_Y)) \\
&= E(XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y) \\
&= E(XY) - E(\mu_X Y) - E(\mu_Y X) + E(\mu_X\mu_Y) \quad \text{(by linearity)} \\
&= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X\mu_Y \\
&= E(XY) - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y \\
&= E(XY) - \mu_X\mu_Y = E(XY) - E(X)E(Y)
\end{aligned}$$

This completes the proof. □

7.4 Properties of Covariance

7.4.1 Variance of a Sum

The primary motivation for introducing covariance is to find the variance of a sum of random variables.

Theorem 7.4.1 (Variance of a Sum)**.** *For any two random variables X and Y:*

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

Proof. Let $\mu_X = E(X)$, $\mu_Y = E(Y)$, and $\mu_{X+Y} = E(X + Y) = \mu_X + \mu_Y$.

$$\begin{aligned}
\mathrm{Var}(X + Y) &= E((X + Y - \mu_{X+Y})^2) \\
&= E(((X - \mu_X) + (Y - \mu_Y))^2) \\
&= E((X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)) \\
&= E((X - \mu_X)^2) + E((Y - \mu_Y)^2) + 2E((X - \mu_X)(Y - \mu_Y)) \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)
\end{aligned}$$

This result shows that the covariance term is the crucial correction factor. □

7.4.2 Other Properties

- **Symmetry:** $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
- **Covariance with a Constant:** For any constant c, $\mathrm{Cov}(X, c) = 0$.
- **Bilinearity:** For constants a, b, c, d: $\mathrm{Cov}(aX + b, cY + d) = ac\mathrm{Cov}(X, Y)$.
- **Covariance with Self:** $\mathrm{Cov}(X, X) = E(X^2) - (E(X))^2 = \mathrm{Var}(X)$.

7.5 Sums of Independent Variables

If two variables are independent, their covariance is zero.

Theorem 7.5.1. *If X and Y are independent random variables, then $Cov(X, Y) = 0$.*

Proof. If X and Y are independent, then $E(XY) = E(X)E(Y)$. Using the computational formula for covariance:

$$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$$

□

Remark. The converse is not true. A covariance of 0 does not imply independence.

Corollary 7.5.2 (Variance of a Sum of Independent Variables)**.** *If X and Y are independent, then:*

$$Var(X + Y) = Var(X) + Var(Y)$$

By induction, for a set of mutually independent random variables $X_1, \ldots, X_n$, $Var(\sum X_i) = \sum Var(X_i)$.

Example 7.5.3 (Variance of the Binomial Distribution)**.** Let $X \sim \mathrm{Binomial}(n, p)$. We can write X as a sum of n independent indicator variables, $X = I_1 + \cdots + I_n$, where each I_j is a Bernoulli(p) trial. We know $E(I_j) = p$ and $E(I_j^2) = p$. So, $\mathrm{Var}(I_j) = E(I_j^2) - (E(I_j))^2 = p - p^2 = p(1-p)$. Since the trials are independent, the indicators are independent. Therefore:

$$\mathrm{Var}(X) = \mathrm{Var}\left(\sum_{j=1}^{n} I_j\right) = \sum_{j=1}^{n} \mathrm{Var}(I_j) = \sum_{j=1}^{n} p(1 - p) = np(1 - p)$$

7.6 Symmetry and Indicators for Dependent Sums

When variables are dependent, we must compute the covariance terms. A powerful technique involves using indicators and exploiting symmetry.

Example 7.6.1 (Variance of the Hypergeometric Distribution)**.** A population has N items, G of which are good. A simple random sample of size n is drawn without replacement. Let X be the number of good items in the sample. We write X as a sum of indicators, $X = I_1 + \cdots + I_n$, where I_j is the indicator that draw j is good. These indicators are **dependent** because the outcome of one draw affects the others. The variance of the sum is:

$$\operatorname{Var}(X) = \operatorname{Var}\left(\sum_{j=1}^{n} I_j\right) = \sum_{j=1}^{n} \operatorname{Var}(I_j) + \sum_{i \neq j} \operatorname{Cov}(I_i, I_j)$$

By symmetry, the distribution of each I_j is the same, and the joint distribution of any pair (I_i, I_j) is the same for $i \neq j$. Let $p = G/N$.

1. **Variance of an indicator:** $P(I_j = 1) = G/N = p$. So, $\operatorname{Var}(I_j) = p(1-p) = \frac{G}{N}(1 - \frac{G}{N})$. There are n such terms.
2. **Covariance of two indicators:** For $i \neq j$, we need $\operatorname{Cov}(I_i, I_j) = E(I_i I_j) - E(I_i)E(I_j)$. $E(I_i I_j) = P(I_i = 1, I_j = 1) = P(I_1 = 1, I_2 = 1)$.

   $$P(I_1 = 1, I_2 = 1) = P(I_1 = 1)P(I_2 = 1 | I_1 = 1) = \frac{G}{N} \cdot \frac{G-1}{N-1}$$

   $$\operatorname{Cov}(I_i, I_j) = \frac{G}{N}\frac{G-1}{N-1} - \left(\frac{G}{N}\right)^2 = \frac{G}{N}\left(\frac{G-1}{N-1} - \frac{G}{N}\right) = -\frac{G(N-G)}{N^2(N-1)}$$

   The covariance is negative, which makes sense: if one draw is good, it slightly reduces the chance for another draw to be good. There are $n(n-1)$ such pairs.

Combining these results:

$$\begin{aligned}
\operatorname{Var}(X) &= n \cdot \frac{G}{N}\left(1 - \frac{G}{N}\right) + n(n-1) \cdot \left(-\frac{G(N-G)}{N^2(N-1)}\right) \\
&= n\frac{G(N-G)}{N^2} - n(n-1)\frac{G(N-G)}{N^2(N-1)} \\
&= n\frac{G(N-G)}{N^2}\left(1 - \frac{n-1}{N-1}\right) = n\frac{G}{N}\left(1 - \frac{G}{N}\right)\left(\frac{N-1-(n-1)}{N-1}\right) \\
&= n\frac{G}{N}\left(1 - \frac{G}{N}\right)\left(\frac{N-n}{N-1}\right)
\end{aligned}$$

7.7 The Finite Population Correction

The variance of the hypergeometric distribution we just derived is closely related to the variance of the corresponding binomial distribution. Let's compare sampling with replacement (Binomial) and without replacement (Hypergeometric).

- $\operatorname{Var}_{\text{with replacement}} = np(1-p) = n\frac{G}{N}(1 - \frac{G}{N})$

- $\text{Var}_{\text{without replacement}} = n\frac{G}{N}(1 - \frac{G}{N})\left(\frac{N-n}{N-1}\right)$

The factor $\frac{N-n}{N-1}$ is called the **finite population correction (fpc)**.

- The fpc is always less than 1, so sampling without replacement reduces the variance compared to sampling with replacement. This is because we cannot draw the same item twice, which removes some variability.
- If the population size N is very large compared to the sample size n, then the fpc is close to 1. In this case, there is little difference between the two sampling schemes, and the hypergeometric variance is well-approximated by the simpler binomial variance.

Chapter 8

The Central Limit Theorem

The **Central Limit Theorem (CLT)** is one of the most fundamental results in probability theory and statistics. It states that, under certain conditions, the probability distribution of the sum (or average) of a large number of independent random variables will be close to a normal distribution, regardless of the underlying distribution of the individual variables. This remarkable theorem explains the ubiquity of the bell-shaped normal curve in the natural and social sciences and provides the foundation for many statistical inference methods.

8.1 The Exact Distribution of a Sum

Before we approximate the distribution of a sum, let's look at a method for finding its exact distribution. This is feasible when the variables are non-negative integers.

Definition 8.1.1 (Probability Generating Function)**.** Let X be a random variable that takes non-negative integer values. Its **probability generating function (PGF)** is a function $G_X(s)$ defined as:

$$G_X(s) = E(s^X) = \sum_{k=0}^{\infty} s^k P(X = k)$$

The PGF is a polynomial (or power series) in the dummy variable s, where the coefficient of s^k is the probability $P(X = k)$.

Theorem 8.1.2 (PGF of a Sum)**.** *Let X and Y be independent, non-negative integer-valued random variables. The PGF of their sum $S = X + Y$ is the product of their individual PGFs:*

$$G_S(s) = G_X(s)G_Y(s)$$

Proof. By independence, $E(s^{X+Y}) = E(s^X s^Y) = E(s^X)E(s^Y)$. □

Remark. For a sum of n i.i.d. variables $S_n = X_1 + \cdots + X_n$, the PGF is $G_{S_n}(s) = (G_X(s))^n$. One can find the distribution of S_n by expanding this polynomial and collecting the coefficients. This can be done computationally using polynomial multiplication, for which libraries like NumPy can be useful.

8.2 The Central Limit Theorem

While PGFs provide an exact distribution, the calculation becomes unwieldy for large n. The CLT provides a powerful and elegant approximation.

Theorem 8.2.1 (Central Limit Theorem)**.** *Let $X_1, X_2, \dots$ be independent and identically distributed (i.i.d.) random variables with expectation $E(X_1) = \mu$ and standard deviation $SD(X_1) = \sigma$, where $0 < \sigma < \infty$. Let $S_n = X_1 + \cdots + X_n$ be the sum of the first n variables. For large n, the distribution of S_n is approximately normal with mean $E(S_n) = n\mu$ and variance $Var(S_n) = n\sigma^2$.*

To formalize this, we convert S_n to ***standard units:***

$$Z_n = \frac{S_n - E(S_n)}{SD(S_n)} = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

The CLT states that the cumulative distribution function (CDF) of Z_n converges to the standard normal CDF, $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$. For any number z:

$$P(Z_n \leq z) \to \Phi(z) \quad \text{as } n \to \infty$$

8.2.1 Using the Normal Approximation

The CLT allows us to approximate probabilities for S_n using the normal curve. For large n:

$$P(S_n \leq x) \approx \Phi\left(\frac{x - n\mu}{\sqrt{n}\sigma}\right)$$

Remark. Numerical libraries like SciPy in Python provide functions such as ‘scipy.stats.norm.cdf‘ to compute the values of $\Phi(z)$, making these calculations straightforward in practice.

8.3 The Sample Mean

The CLT is often applied to the sample mean of a random sample.

Definition 8.3.1 (Sample Mean)**.** Let $X_1, \dots, X_n$ be a random sample. The **sample mean** is defined as:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{S_n}{n}$$

8.3.1 Distribution of the Sample Mean

We can find the exact expectation and variance of the sample mean.

- **Expectation:** $E(\bar{X}_n) = E(\frac{1}{n}S_n) = \frac{1}{n}E(S_n) = \frac{1}{n}(n\mu) = \mu$. The sample mean is an unbiased estimator of the population mean.
- **Variance:** $\text{Var}(\bar{X}_n) = \text{Var}(\frac{1}{n}S_n) = (\frac{1}{n})^2\text{Var}(S_n) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$.
- **Standard Deviation:** $SD(\bar{X}_n) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$.

This last result is known as the **Square Root Law**. It implies that the variability of the sample mean decreases as the square root of the sample size. To double the accuracy of our estimate, we must collect four times as much data.

The CLT applies directly to the sample mean. For large n, the distribution of $\bar{X}_n$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$.

8.4 Confidence Intervals

One of the most important applications of the CLT is in constructing confidence intervals for an unknown population parameter, such as the mean μ.

Suppose we have a large random sample $X_1, \ldots, X_n$ from a population with unknown mean μ but known SD σ. The sample mean $\bar{X}_n$ is our estimate for μ. A confidence interval provides a range of plausible values for μ based on our sample data.

8.4.1 Derivation of a 95% Confidence Interval

By the CLT, the sample mean $\bar{X}_n$ is approximately normal with mean μ and SD $\sigma/\sqrt{n}$. In standard units, the variable

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

is approximately standard normal. For the standard normal distribution, approximately 95% of the probability lies between -2 and +2.

$$P\left(-2 \le \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le 2\right) \approx 0.95$$

We can rearrange the inequalities to isolate the unknown parameter μ:

$$-2\frac{\sigma}{\sqrt{n}} \le \bar{X}_n - \mu \le 2\frac{\sigma}{\sqrt{n}}$$

$$-\bar{X}_n - 2\frac{\sigma}{\sqrt{n}} \le -\mu \le -\bar{X}_n + 2\frac{\sigma}{\sqrt{n}}$$

$$\bar{X}_n - 2\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}_n + 2\frac{\sigma}{\sqrt{n}}$$

This gives us a random interval $[\bar{X}_n - 2\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 2\frac{\sigma}{\sqrt{n}}]$ that contains the true mean μ with approximately 95% probability.

Definition 8.4.1 (Confidence Interval)**.** An approximate **95% confidence interval for the population mean** μ is given by the interval:

$$\bar{X}_n \pm 2\frac{\sigma}{\sqrt{n}}$$

In practice, the population SD σ is often also unknown. For large samples, it can be replaced by the sample standard deviation, $SD(\text{sample})$, without much loss of accuracy.

Interpretation: A 95% confidence interval means that if we were to repeat the sampling process many times and construct an interval for each sample, about 95% of these intervals would contain the true, fixed population mean μ. It is a statement about the reliability of the interval-generating process.

Chapter 9

Continuous Distributions

So far, we have dealt with discrete random variables, where probabilities are found by summing over a countable set of values. We now turn our attention to **continuous random variables**, which can take any value in a continuous interval on the number line. For these variables, the concept of summing probabilities is replaced by integration. Instead of a probability mass function, we use a **probability density function (PDF)**, and the probability of an event is represented by the area under the curve of this function.

9.1 Density and CDF

Definition 9.1.1 (Probability Density Function (PDF))**.** A function $f(x)$ is a **probability density function** for a continuous random variable X if it satisfies two conditions:

1. **Non-negativity:** $f(x) \geq 0$ for all x.
2. **Total Area is 1:** $\int_{-\infty}^{\infty} f(x)dx = 1$.

The probability that X falls within an interval $[a, b]$ is the area under the density curve over that interval:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Definition 9.1.2 (Cumulative Distribution Function (CDF))**.** The **cumulative distribution function (CDF)** of a continuous random variable X, denoted $F(x)$, gives the total probability up to a value x:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

Conversely, the PDF can be recovered from the CDF by differentiation:

$$f(x) = \frac{d}{dx}F(x) = F'(x)$$

Example 9.1.3 (Uniform Distribution)**.** A random variable X has a **uniform distribution on the interval** (a, b), written $X \sim \text{Uniform}(a, b)$, if its PDF is constant over the interval and zero elsewhere. To make the total area 1, the height must be $1/(b - a)$.

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

The CDF is the integral of the PDF. For $x \in (a, b)$:

$$F(x) = \int_a^x \frac{1}{b-a} dt = \frac{x-a}{b-a}$$

9.2 The Meaning of Density

It is crucial to understand what the value of a density function $f(x)$ represents.

- For a continuous random variable, the probability of taking any single exact value is zero: $P(X = c) = \int_c^c f(x)dx = 0$.
- The density $f(x)$ is *not* a probability. It can be greater than 1. For example, a Uniform(0, 0.1) variable has a density of $f(x) = 10$ on its domain.
- The density represents "probability per unit length". For a very small interval dx around a point x, the probability that the variable falls in that interval is approximately the area of a thin rectangle with height $f(x)$ and width dx.
$$P(X \in [x, x + dx]) \approx f(x)dx$$

9.3 Expectation

The concept of expectation extends naturally from sums to integrals.

Definition 9.3.1 (Expectation)**.** The **expected value** of a continuous random variable X with PDF $f(x)$ is:

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

Theorem 9.3.2 (Law of the Unconscious Statistician - LOTUS)**.** *If g is a real-valued function, the expectation of the random variable $g(X)$ is:*

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

This allows us to calculate $E(g(X))$ without first finding the PDF of $g(X)$. Using this, the computational formula for variance remains the same: $\text{Var}(X) = E(X^2) – (E(X))^2$.

Example 9.3.3 (Expectation of a Uniform(a,b) variable)**.** Let $X \sim \text{Uniform}(a, b)$.

$$E(X) = \int_a^b x \cdot \frac{1}{b-a} dx = \frac{1}{b-a} \left[\frac{x^2}{2} \right]_a^b = \frac{1}{b-a} \frac{b^2 - a^2}{2} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}$$

The expectation is the midpoint of the interval, as expected.

9.4 The Exponential Distribution

The exponential distribution is often used to model waiting times until an event occurs in a random process.

Definition 9.4.1 (Exponential Distribution)**.** A random variable T has an **exponential distribution with rate parameter** $\lambda > 0$, written $T \sim \text{Exponential}(\lambda)$, if its PDF is:

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

9.4.1 CDF and Expectation

The CDF is found by integrating the PDF:

$$F(t) = P(T \leq t) = \int_0^t \lambda e^{-\lambda x} dx = [-e^{-\lambda x}]_0^t = 1 - e^{-\lambda t}, \quad \text{for } t \geq 0$$

The probability of ”surviving” past time t is $P(T > t) = 1 - F(t) = e^{-\lambda t}$.

The expectation is found using integration by parts ($u = t, dv = \lambda e^{-\lambda t} dt$):

$$E(T) = \int_0^\infty t(\lambda e^{-\lambda t}) dt = [-te^{-\lambda t}]_0^\infty - \int_0^\infty (-e^{-\lambda t}) dt = 0 - [-\frac{1}{\lambda} e^{-\lambda t}]_0^\infty = \frac{1}{\lambda}$$

If events arrive at a rate of λ per unit time, the average waiting time for an event is $1/\lambda$.

9.4.2 The Memoryless Property

The exponential distribution has a unique and crucial property among continuous distributions.

Theorem 9.4.2 (Memoryless Property)**.** *For any $s, t \geq 0$, if $T \sim$ Exponential(λ), then:*

$$P(T > t + s \mid T > s) = P(T > t)$$

This means that given the event has not occurred by time s, the remaining waiting time has the same distribution as the original waiting time. The process ”forgets” how long it has already waited.

Proof. By the definition of conditional probability:

$$\begin{aligned} P(T > t+s \mid T > s) &= \frac{P(\{T > t+s\} \cap \{T > s\})}{P(T > s)} \\ &= \frac{P(T > t+s)}{P(T > s)} \quad \text{(since if T¿t+s, then T¿s is implied)} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t - \lambda s + \lambda s} = e^{-\lambda t} \\ &= P(T > t) \end{aligned}$$

□

This property directly connects the exponential distribution to the Poisson process: the waiting times between successive events in a Poisson process with rate λ are independent and identically distributed Exponential(λ) random variables.

9.5 The Standard Normal Distribution

The foundation of the normal family is the standard normal distribution.

Definition 9.5.1 (Standard Normal)**.** A random variable Z has the **standard normal distribution**, written $Z \sim N(0,1)$, if its probability density function (PDF), denoted $\phi(z)$, is given by:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad \text{for } -\infty < z < \infty$$

The function $\phi(z)$ is a symmetric, bell-shaped curve centered at 0. It can be shown that $\int_{-\infty}^{\infty} \phi(z) dz = 1$.

Proposition 9.5.2 (Mean and Variance of a Standard Normal)**.** *For a standard normal random variable Z:*

1. $E(Z) = 0$
2. $Var(Z) = 1$

Proof. **1. Expectation:**

$$E(Z) = \int_{-\infty}^{\infty} z\phi(z) dz = \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

The integrand $g(z) = ze^{-z^2/2}$ is an odd function, meaning $g(-z) = -g(z)$. The integral of an odd function over a symmetric interval $(-\infty, \infty)$ is 0. Thus, $E(Z) = 0$.

2. Variance: Since $E(Z) = 0$, $\text{Var}(Z) = E(Z^2) - (E(Z))^2 = E(Z^2)$.

$$E(Z^2) = \int_{-\infty}^{\infty} z^2 \phi(z) dz = \int_{-\infty}^{\infty} z \cdot (z\phi(z)) dz$$

We use integration by parts with $u = z$ and $dv = z\phi(z)dz$. Note that $\frac{d}{dz}\phi(z) = -z\phi(z)$, so $dv = -d\phi(z)$.

$$\begin{aligned} E(Z^2) &= \int_{-\infty}^{\infty} -z\, d(\phi(z)) \\ &= [-z\phi(z)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} -\phi(z)dz \\ &= 0 + \int_{-\infty}^{\infty} \phi(z)dz = 1 \end{aligned}$$

The term $[-z\phi(z)]_{-\infty}^{\infty}$ evaluates to 0 because the exponential term $e^{-z^2/2}$ goes to zero much faster than z goes to infinity. Thus, $\text{Var}(Z) = 1$. □

Definition 9.5.3 (General Normal Distribution)**.** A random variable X has a normal distribution with mean μ and variance σ^2, written $X \sim N(\mu, \sigma^2)$, if it is a linear transformation of a standard normal variable Z:

$$X = \sigma Z + \mu$$

This implies $E(X) = \sigma E(Z) + \mu = \mu$ and $\text{Var}(X) = \sigma^2 \text{Var}(Z) = \sigma^2$.

9.6 The Gamma Family

The gamma family of distributions is a versatile two-parameter family defined on the positive real numbers.

Definition 9.6.1 (Gamma Function)**.** The **gamma function**, $\Gamma(r)$, is defined for $r > 0$ by the integral:

$$\Gamma(r) = \int_0^{\infty} t^{r-1} e^{-t} dt$$

It has the properties $\Gamma(r) = (r-1)\Gamma(r-1)$ and $\Gamma(n) = (n-1)!$ for integers $n \geq 1$.

Definition 9.6.2 (Gamma Distribution)**.** A random variable X has the **gamma distribution** with shape parameter $r > 0$ and rate parameter $\lambda > 0$, written $X \sim \text{Gamma}(r, \lambda)$, if its PDF is:

$$f(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, \quad \text{for } x > 0$$

Proposition 9.6.3 (Mean and Variance of a Gamma Distribution)**.** *If $X \sim$ Gamma(r, λ):*

1. $E(X) = r/\lambda$
2. *Var*$(X) = r/\lambda^2$

Proof. We compute $E(X^k)$ by recognizing the form of the gamma PDF in the integral.

$$E(X^k) = \int_0^\infty x^k \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} dx = \frac{\lambda^r}{\Gamma(r)} \int_0^\infty x^{r+k-1} e^{-\lambda x} dx$$

The integral $\int_0^\infty x^{r+k-1} e^{-\lambda x} dx$ is the kernel of a Gamma$(r+k, \lambda)$ density, which integrates to $\frac{\Gamma(r+k)}{\lambda^{r+k}}$.

$$E(X^k) = \frac{\lambda^r}{\Gamma(r)} \frac{\Gamma(r+k)}{\lambda^{r+k}} = \frac{\Gamma(r+k)}{\Gamma(r)\lambda^k}$$

For $k = 1$: $E(X) = \frac{\Gamma(r+1)}{\Gamma(r)\lambda} = \frac{r\Gamma(r)}{\Gamma(r)\lambda} = \frac{r}{\lambda}$. For $k = 2$: $E(X^2) = \frac{\Gamma(r+2)}{\Gamma(r)\lambda^2} = \frac{(r+1)r\Gamma(r)}{\Gamma(r)\lambda^2} = \frac{r(r+1)}{\lambda^2}$. The variance is $\text{Var}(X) = E(X^2) - (E(X))^2 = \frac{r(r+1)}{\lambda^2} - \left(\frac{r}{\lambda}\right)^2 = \frac{r^2+r-r^2}{\lambda^2} = \frac{r}{\lambda^2}$. □

Theorem 9.6.4 (Sum of Independent Gammas)**.** *If $X \sim Gamma(r, \lambda)$ and $Y \sim Gamma(s, \lambda)$ are independent, then their sum is also gamma distributed:*

$$X + Y \sim Gamma(r + s, \lambda)$$

The shape parameters add, provided the rate parameter is the same.

9.7 The Chi-Squared Distribution

The chi-squared distribution is a special case of the gamma distribution that is central to statistical hypothesis testing.

Definition 9.7.1 (Chi-Squared Distribution)**.** A random variable X has the **chi-squared distribution with n degrees of freedom**, written $X \sim \chi^2(n)$, if it has a Gamma distribution with shape parameter $r = n/2$ and rate parameter $\lambda = 1/2$.

$$X \sim \chi^2(n) \equiv \text{Gamma}(n/2, 1/2)$$

The mean is $E(X) = (n/2)/(1/2) = n$ and the variance is $\text{Var}(X) = (n/2)/(1/2)^2 = 2n$.

The primary importance of this distribution comes from its relationship to the standard normal.

Theorem 9.7.2. *If $Z_1, Z_2, \ldots, Z_n$ are independent standard normal random variables, then the sum of their squares has a chi-squared distribution with n degrees of freedom.*

$$\sum_{i=1}^n Z_i^2 \sim \chi^2(n)$$

Proof. The proof proceeds in two steps:

1. **Distribution of one squared normal, Z_1^2:** We showed in a previous lecture that if $Z_1 \sim N(0,1)$, its square $Y = Z_1^2$ has the PDF $f_Y(y) = \frac{1}{\sqrt{2\pi y}} e^{-y/2}$ for $y > 0$. This is exactly the PDF of a Gamma(1/2, 1/2) distribution, which is by definition a $\chi^2(1)$ distribution.
2. **Sum of squares:** The sum $S = \sum_{i=1}^n Z_i^2$ is a sum of n independent and identically distributed $\chi^2(1)$ random variables. Since $\chi^2(1)$ is just Gamma(1/2, 1/2), we are summing n independent Gamma variables with the same rate parameter $\lambda = 1/2$. By the additive property of the gamma family, the sum is distributed as:

$$S \sim \text{Gamma}\left(\frac{1}{2} + \cdots + \frac{1}{2}, \frac{1}{2}\right) = \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right) \equiv \chi^2(n)$$

This completes the proof. □

Chapter 10

Transformations of Random Variables

If we know the probability distribution of a random variable X, we are often interested in finding the distribution of a function of X, say $Y = g(X)$. For example, if X is a measurement in inches, we might want the distribution of $Y = 2.54X$, the same measurement in centimeters. This process is called finding the distribution of a **transformation** of a random variable. The goal is to derive the density of Y, $f_Y(y)$, from the density of X, $f_X(x)$. The general method involves first finding the cumulative distribution function (CDF) of Y and then differentiating it to get the probability density function (PDF).

10.1 Linear Transformations

Let's start with the simplest case: a linear transformation $Y = aX + b$.

Theorem 10.1.1. *Let X be a random variable with PDF $f_X(x)$, and let $Y = aX + b$ for constants $a \neq 0$ and b. The PDF of Y is given by:*

$$f_Y(y) = f_X\left(\frac{y-b}{a}\right)\frac{1}{|a|}$$

Proof. We find the CDF of Y first. Let's assume $a > 0$.

$$F_Y(y) = P(Y \leq y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right)$$

Now, we differentiate with respect to y using the chain rule to find the PDF of Y.

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X\left(\frac{y-b}{a}\right) = F_X'\left(\frac{y-b}{a}\right)\cdot\frac{1}{a} = f_X\left(\frac{y-b}{a}\right)\frac{1}{a}$$

If $a < 0$, the inequality flips:

$$F_Y(y) = P(aX + b \leq y) = P\left(X \geq \frac{y-b}{a}\right) = 1 - F_X\left(\frac{y-b}{a}\right)$$

Differentiating this gives:

$$f_Y(y) = -\frac{d}{dy}F_X\left(\frac{y-b}{a}\right) = -f_X\left(\frac{y-b}{a}\right)\cdot\frac{1}{a} = f_X\left(\frac{y-b}{a}\right)\frac{1}{|a|}$$

Both cases combine to give the general formula. □

10.2 Monotone Functions

The method used for linear transformations can be generalized to any strictly monotone (always increasing or always decreasing) function g.

Theorem 10.2.1 (Change of Variable Formula for Density)**.** *Let X be a random variable with PDF $f_X(x)$. Let g be a smooth, strictly monotone function, and let $Y = g(X)$. The PDF of Y is given by:*

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$$

where $x = g^{-1}(y)$ is the unique solution for x in terms of y.

Proof. Suppose g is strictly increasing. The inverse function $x = g^{-1}(y)$ is well-defined.

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

Differentiating with respect to y:

$$f_Y(y) = \frac{d}{dy}F_X(g^{-1}(y)) = f_X(g^{-1}(y))\cdot\frac{d}{dy}g^{-1}(y) = f_X(x)\frac{dx}{dy}$$

Since g is increasing, $dx/dy > 0$, so we can write this as $f_X(x)|dx/dy|$. If g is strictly decreasing, the inequality flips:

$$F_Y(y) = P(g(X) \leq y) = P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y))$$

Differentiating gives $f_Y(y) = -f_X(g^{-1}(y))\cdot\frac{d}{dy}g^{-1}(y) = f_X(x)(-\frac{dx}{dy})$. Since g is decreasing, $dx/dy < 0$, so $(-\frac{dx}{dy}) = |\frac{dx}{dy}|$. The formula holds in both cases. □

Example 10.2.2 (Density of X^3)**.** Let $X \sim \text{Uniform}(0,1)$, so $f_X(x) = 1$ for $x \in (0,1)$. Let $Y = X^3$. The function $g(x) = x^3$ is strictly increasing on $(0,1)$. The inverse is $x = y^{1/3}$. The derivative is $\frac{dx}{dy} = \frac{1}{3}y^{-2/3}$. The possible values of Y are in $(0,1)$. For $y \in (0,1)$:

$$f_Y(y) = f_X(y^{1/3})\left|\frac{1}{3}y^{-2/3}\right| = 1\cdot\frac{1}{3}y^{-2/3} = \frac{1}{3}y^{-2/3}$$

This is a Beta distribution.

10.3 Simulation via the Inverse CDF

A remarkable application of the change of variable formula is a method to generate random numbers from any desired distribution using only a standard uniform random number generator.

Theorem 10.3.1 (Probability Integral Transform)**.** *Let X be a continuous random variable with a strictly increasing CDF, F_X. Then the random variable $U = F_X(X)$ has a Uniform(0,1) distribution.*

Proof. Let $U = F_X(X)$. The possible values of U are in $(0,1)$. For any $u \in (0,1)$:

$$P(U \leq u) = P(F_X(X) \leq u) = P(X \leq F_X^{-1}(u))$$

The last step holds because F_X is strictly increasing, so its inverse F_X^{-1} exists. By definition, $P(X \leq x_0) = F_X(x_0)$. So,

$$P(U \leq u) = F_X(F_X^{-1}(u)) = u$$

This is the CDF of a Uniform(0,1) distribution. □

Theorem 10.3.2 (Inverse CDF Method for Simulation)**.** *Let $U \sim$ Uniform$(0,1)$, and let F be any CDF. Then the random variable $X = F^{-1}(U)$ has the CDF F.*

Proof.

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$$

The last step holds because $P(U \leq u) = u$ for a uniform variable. Thus, X has the desired CDF F. □

This theorem provides a practical method for simulation: to generate a random number from a distribution with CDF F, we generate a standard uniform number u and then compute $F^{-1}(u)$.

Example 10.3.3 (Simulating an Exponential Variable)**.** Let's generate a variable from an Exponential(λ) distribution. The CDF is $F(x) = 1 - e^{-\lambda x}$. We need to find the inverse $F^{-1}(u)$. Set $u = 1 - e^{-\lambda x}$. Then $e^{-\lambda x} = 1 - u$, so $-\lambda x = \log(1-u)$, which gives $x = -\frac{1}{\lambda}\log(1-u)$. If $U \sim \text{Uniform}(0,1)$, then $1 - U$ is also Uniform(0,1). So we can use the simpler formula $X = -\frac{1}{\lambda}\log(U)$ to generate an Exponential(λ) variate.

10.4 Two-to-One Functions

The change of variable formula does not directly apply if the function g is not one-to-one. In such cases, we must revert to the CDF method and sum the contributions from all branches of the inverse function.

Example 10.4.1 (Density of X^2)**.** Let X be a continuous random variable with PDF $f_X(x)$. Let $Y = X^2$. Find the PDF of Y. The function $g(x) = x^2$ is not monotone. The possible values of Y are non-negative. For any $y > 0$:

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y})$$

$$F_Y(y) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

Now we differentiate with respect to y using the chain rule:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}F_X(\sqrt{y}) - \frac{d}{dy}F_X(-\sqrt{y}) \\ &= f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} - f_X(-\sqrt{y}) \cdot \left(-\frac{1}{2\sqrt{y}}\right) \\ &= \frac{1}{2\sqrt{y}} \left(f_X(\sqrt{y}) + f_X(-\sqrt{y})\right), \quad \text{for } y > 0 \end{aligned}$$

This formula combines the contributions from the positive and negative values of X that map to the same value of Y.

Example 10.4.2 (Standard Normal Squared is Chi-Squared)**.** Let $Z \sim N(0, 1)$, so its PDF is $\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$. Let $Y = Z^2$. Using the formula derived above with $f_Z(z) = \phi(z)$: For $y > 0$,

$$\begin{aligned} f_Y(y) &= \frac{1}{2\sqrt{y}}(\phi(\sqrt{y}) + \phi(-\sqrt{y})) \\ &= \frac{1}{2\sqrt{y}} \left(\frac{1}{\sqrt{2\pi}}e^{-(\sqrt{y})^2/2} + \frac{1}{\sqrt{2\pi}}e^{-(-\sqrt{y})^2/2}\right) \\ &= \frac{1}{2\sqrt{y}} \left(2 \cdot \frac{1}{\sqrt{2\pi}}e^{-y/2}\right) \\ &= \frac{1}{\sqrt{2\pi y}}e^{-y/2} \end{aligned}$$

This is the PDF of the gamma distribution with shape parameter 1/2 and rate parameter 1/2. This specific distribution is also known as the **chi-squared distribution with 1 degree of freedom**.

Chapter 11

Joint Continuous Distributions

We now extend the concept of density from a single continuous random variable to two or more variables. This allows us to model the relationship between multiple continuous quantities. The **joint probability density function (joint PDF)** of two variables, say X and Y, is a surface in three-dimensional space. The probability that the pair (X, Y) falls into a certain region in the plane is the volume under this surface and over that region.

11.1 Joint Density Functions

Definition 11.1.1 (Joint PDF)**.** A function $f(x, y)$ is a **joint probability density function** for two continuous random variables X and Y if it satisfies two conditions:

1. **Non-negativity:** $f(x, y) \geq 0$ for all x, y.
2. **Total Volume is 1:** $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$.

11.1.1 Probabilities and Expectations

The probability that the pair (X, Y) falls into a region A in the plane is found by integrating the joint PDF over that region:

$$P((X, Y) \in A) = \iint_A f(x, y) dx dy$$

For a small region $dxdy$ around a point (x, y), the probability is approximately the volume of a small column:

$$P(X \in [x, x + dx], Y \in [y, y + dy]) \approx f(x, y) dx dy$$

The expectation of a function $g(X, Y)$ is found by integrating the function against the joint PDF over the entire plane.

Definition 11.1.2 (Expectation)**.** The expected value of a function $g(X, Y)$ is:

$$E(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy$$

Example 11.1.3 (Uniform Density on a Square)**.** Let the pair (X, Y) be chosen uniformly from the unit square $\{(x, y) : 0 < x < 1, 0 < y < 1\}$. To make the total volume 1, the joint PDF must be constant with height 1 inside the square and 0 elsewhere.

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

Let's find the probability $P(X > Y)$. This corresponds to the region where $0 < y < x < 1$.

$$P(X > Y) = \int_0^1 \int_0^x 1\,dydx = \int_0^1 [y]_0^x dx = \int_0^1 x\,dx = \left[\frac{x^2}{2}\right]_0^1 = \frac{1}{2}$$

This makes sense, as by symmetry, the chance that $X > Y$ should be the same as the chance that $Y > X$.

11.2 Independence

The concept of independence for continuous random variables is analogous to the discrete case.

Definition 11.2.1 (Independence)**.** Two continuous random variables X and Y with joint PDF $f(x, y)$ are **independent** if their joint PDF is the product of their individual (marginal) PDFs, $f_X(x)$ and $f_Y(y)$:

$$f(x,y) = f_X(x)f_Y(y) \quad \text{for all } x, y$$

A useful shortcut is that if the joint domain of the variables is a rectangle (or a product of intervals) and the function $f(x, y)$ factors into a part that only depends on x and a part that only depends on y, then the variables are independent.

Example 11.2.2 (Uniform on a Square, Revisited)**.** For (X, Y) uniform on the unit square, we have $f(x, y) = 1$ for $(x, y) \in (0, 1) \times (0, 1)$. The marginal density of X is $f_X(x) = \int_0^1 1\,dy = 1$ for $x \in (0, 1)$. Similarly, $f_Y(y) = 1$ for $y \in (0, 1)$. Thus, $f_X(x)f_Y(y) = 1 \cdot 1 = 1 = f(x, y)$ for (x, y) in the square. This confirms that X and Y are independent, and both are Uniform(0,1).

Example 11.2.3 (Uniform on a Triangle)**.** Let (X, Y) be chosen uniformly from the triangle $\{(x, y) : x \geq 0, y \geq 0, x + y \leq 1\}$. The area of this triangle is $1/2$. To make the total volume 1, the joint PDF must be $f(x, y) = 2$ inside the triangle and 0 elsewhere. The variables X and Y are **not independent**. The domain of possible values is not a rectangle; for instance, if $X = 0.8$, then Y is restricted to be less than 0.2. The value of one variable restricts the possible values of the other.

11.3 Marginal and Conditional Densities

11.3.1 Marginal Density

From the joint PDF, we can recover the PDF of a single variable by "integrating out" the other variable. This is analogous to summing out a variable in a discrete joint distribution table.

Definition 11.3.1 (Marginal PDF)**.** The **marginal probability density function** of X is given by:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

And similarly for Y:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

Example 11.3.2 (Marginals for Uniform on a Triangle)**.** Let's find the marginal PDF of X for the uniform distribution on the triangle $\{(x,y) : x \geq 0, y \geq 0, x+y \leq 1\}$, where $f(x,y) = 2$. The possible values for x are in the interval $(0,1)$. For a fixed x in this interval, y can range from 0 to $1-x$.

$$f_X(x) = \int_0^{1-x} 2\,dy = [2y]_0^{1-x} = 2(1-x), \quad \text{for } 0 < x < 1$$

By symmetry, the marginal PDF of Y is $f_Y(y) = 2(1-y)$ for $0 < y < 1$.

11.3.2 Conditional Density

Conditioning in continuous distributions follows the same division rule as in other contexts.

Definition 11.3.3 (Conditional PDF)**.** The **conditional probability density function** of Y given $X = x$ is defined as:

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

This is defined for all x where the marginal density $f_X(x) > 0$. For a fixed x, the function $f_{Y|X}(y|x)$ is itself a valid PDF in y, meaning it is non-negative and integrates to 1.

Example 11.3.4 (Conditionals for Uniform on a Triangle)**.** Let's find the conditional density of Y given $X = x$ for the uniform on a triangle example. We have $f(x,y) = 2$ and $f_X(x) = 2(1-x)$ for values in the appropriate ranges.

$$f_{Y|X}(y|x) = \frac{2}{2(1-x)} = \frac{1}{1-x}, \quad \text{for } 0 < y < 1-x$$

This result is very intuitive: given that $X = x$, the possible values of Y are restricted to the vertical line segment from $(x,0)$ to $(x,1-x)$. The conditional distribution of Y is uniform on this interval $(0,1-x)$.

11.3.3 Conditional Expectation

Once we have the conditional density, we can define the conditional expectation.

Definition 11.3.5 (Conditional Expectation)**.** The conditional expectation of Y given $X = x$ is the expectation calculated using the conditional density:

$$E(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy$$

Example 11.3.6 (Conditional Expectation for Uniform on a Triangle)**.** Given $X = x$, we know Y is uniform on $(0, 1 - x)$. The expectation of a uniform variable is its midpoint. So,

$$E(Y|X = x) = \frac{1 - x}{2}$$

This defines the conditional expectation of Y given X as a function of X: $E(Y|X) = (1-X)/2$. We can check the law of iterated expectations: $E(E(Y|X)) = E(\frac{1-X}{2}) = \frac{1}{2}(1 - E(X))$. From the marginal $f_X(x) = 2(1 - x)$, we can calculate $E(X) = \int_0^1 x \cdot 2(1 - x)dx = 1/3$. So $E(E(Y|X)) = \frac{1}{2}(1 - 1/3) = 1/3$. This matches $E(Y)$, as it should.

Chapter 12

Moments and Moment Generating Functions

12.1 Moment Generating Functions

A less direct but often simpler method uses moment generating functions (MGFs).

Definition 12.1.1 (Moment Generating Function)**.** The **moment generating function (MGF)** of a random variable X, denoted $M_X(t)$, is defined as:

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

for all real t for which the expectation exists.

The MGF is so named because its derivatives at $t = 0$ generate the moments of X.

$$M'_X(t) = \frac{d}{dt}E(e^{tX}) = E\left(\frac{d}{dt}e^{tX}\right) = E(Xe^{tX}) \implies M'_X(0) = E(X)$$

$$M''_X(t) = E(X^2e^{tX}) \implies M''_X(0) = E(X^2)$$

In general, $M_X^{(k)}(0) = E(X^k)$ is the k-th moment of X.

12.1.1 Properties of MGFs

1. **Uniqueness:** If two random variables have MGFs that are equal on an interval around $t = 0$, then they have the same distribution.
2. **MGF of a Sum:** If X and Y are independent, the MGF of their sum is the product of their MGFs.

   $$M_{X+Y}(t) = M_X(t)M_Y(t)$$

   Proof. $M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX}e^{tY}) = E(e^{tX})E(e^{tY})$ by independence. This is $M_X(t)M_Y(t)$. □

12.2 MGFs, the Normal Distribution, and the CLT

We can now use MGFs to prove fundamental results about the normal distribution.

Proposition 12.2.1 (MGF of a Normal Distribution)**.** *If $X \sim N(\mu, \sigma^2)$, its MGF is $M_X(t) = e^{t\mu + \frac{1}{2}\sigma^2 t^2}$.*

Proof. First, let $Z \sim N(0, 1)$.

$$M_Z(t) = E(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z^2 - 2tz)} dz$$

We complete the square in the exponent: $z^2 - 2tz = (z-t)^2 - t^2$.

$$M_Z(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}((z-t)^2 - t^2)} dz = e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2} dz$$

The integral is the total area under a normal density with mean t and variance 1, which is 1. Thus, $M_Z(t) = e^{t^2/2}$. Now for $X = \sigma Z + \mu$:

$$M_X(t) = E(e^{t(\sigma Z + \mu)}) = e^{t\mu} E(e^{(t\sigma)Z}) = e^{t\mu} M_Z(t\sigma) = e^{t\mu} e^{\frac{1}{2}(t\sigma)^2} = e^{t\mu + \frac{1}{2}\sigma^2 t^2}$$

□

Theorem 12.2.2 (Sum of Independent Normals)**.** *If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are independent, then their sum $S = X + Y$ is also normal with mean $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$.*

Proof. We use the MGF property for sums:

$$M_S(t) = M_X(t) M_Y(t) = \left(e^{t\mu_X + \frac{1}{2}\sigma_X^2 t^2}\right)\left(e^{t\mu_Y + \frac{1}{2}\sigma_Y^2 t^2}\right) = e^{t(\mu_X + \mu_Y) + \frac{1}{2}(\sigma_X^2 + \sigma_Y^2)t^2}$$

By the uniqueness property of MGFs, this is the MGF of a normal random variable with mean $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$. □

12.2.1 Proof of the Central Limit Theorem (Sketch)

MGFs provide a way to prove the Central Limit Theorem. Let $X_1, X_2, \ldots$ be i.i.d. with mean 0 and variance 1. Let $M_X(t)$ be their common MGF. Let $S_n = \sum_{i=1}^n X_i$, and let $Z_n = S_n/\sqrt{n}$ be the standardized sum. We want to show that the MGF of Z_n converges to $e^{t^2/2}$, the MGF of the standard normal.

$$M_{Z_n}(t) = E\left(e^{tS_n/\sqrt{n}}\right) = M_{S_n}\left(\frac{t}{\sqrt{n}}\right) = \left(M_X\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

Now we use the Taylor expansion of $M_X(s)$ around $s = 0$.

$$M_X(s) = M_X(0) + M_X'(0)s + \frac{M_X''(0)}{2!}s^2 + O(s^3)$$

Since $E(X) = 0$ and $E(X^2) = 1$, we have $M'_X(0) = 0$ and $M''_X(0) = 1$. Also $M_X(0) = 1$.

$$M_X(s) = 1 + \frac{1}{2}s^2 + O(s^3)$$

Let $s = t/\sqrt{n}$. For large n, s is small.

$$M_X\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{1}{2}\frac{t^2}{n} + O\left(\frac{t^3}{n^{3/2}}\right)$$

Now we take the n-th power:

$$M_{Z_n}(t) = \left(1 + \frac{t^2/2}{n} + \dots\right)^n$$

This expression is of the form $(1 + \frac{c}{n} + \text{smaller terms})^n$. As $n \to \infty$, this converges to e^c. In our case, $c = t^2/2$.

$$\lim_{n\to\infty} M_{Z_n}(t) = e^{t^2/2}$$

By the uniqueness property, the distribution of Z_n converges to the standard normal distribution.