Thinking like a Frequentist

Student

January 2026

This page is intentionally left blank.

Preface

"Are you watching closely?"

Alfred Borden, The Prestige (2006)

When I was learning Probability and Statistics (ProbStat for short), I did not really understand why things were distributed *normally*. Were they *normal* because they were, or were *we just assuming "Everything is normal for simplicity"*? Sometimes, I felt we were overusing this term and in many situations, we were separating our theory from the actual data due to the intital assumption of *normal* distribution.

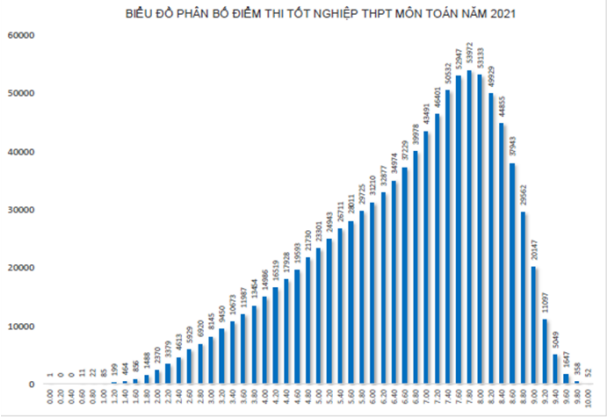


Figure 1: Distribution of Math scores in the 2021 National Entrance Exam

As you can see here, this graph **does not really resemble a bell-shaped curve**; and clearly, this data is not distributed normally. In fact, not much real-world statistical data fits a bell-shaped curve. But why are we still using it?
In my opinion (as a junior CE student), I think the main reason is due to **Central Limit Theorem** (CLT for short), every sample in the same population (regardless of how the data is distributed) has a **mean** that converges to a *normal distribution*, or **bell-shaped curve**. This topic will be explained in detail in **Chapter 8: Fundamentals of Statistics**.

This book approaches the fundamentals of probability and statistics in a very mathematically rigorous way to ensure maximum accuracy. I encourage readers to prove all of theorems, corollaries; and you should create your own examples for each theorem to gain a deeper understading of their origins.

—Student—

Acknowledgements

First of all, I want to thank you for reading this book. Although I am not a native English speaker, but I truly enjoy writing in English. I have tried my best to express my ideas in English, but minor grammatical or spelling errors are unavoidable. If you find any of them, I would be very happy to receive your feedback via my email address: tinvu1309@gmail.com

Secondly, I am extremly grateful and would like to express my thanks to the authors of the textbook "Probability and Statistics for Engineers and Scientists", 9th edition. This book truly saved my student life.

Finally, the idea of writing this book was inspired by Prof. Steve Brunton lectures on YouTube. You should check out his videos too!

—Student—

Contents

List of Figures

List of notations

Since much of my work is handwritten, so I have modified some commonly used notations for probability distribution funcions using curved script for convenience. You should notice that my conventions are not the international standards.

1. $\mathscr{I}$: Bernoulli distribution
2. $\mathscr{B}$: Binomial distribution
3. $\mathscr{B}^*$: Negative binomial distribution
4. $\mathscr{H}$: Hypergeometric distribution
5. $\mathscr{G}^*$: Geometric distribution
6. $\mathscr{P}$: Poisson distribution
7. $\mathscr{U}$: Uniform distribution
8. $\mathscr{N}$: Normal distribution
9. $\mathscr{G}$: Gamma distribution
10. $\mathscr{E}$: Exponential distribution
11. $\mathscr{C}$: Chi-squared distributon
12. $\mathscr{T}$: t-distribution (or Student distribution)

Chapter 1

Introduction to Probability and Statistics

1.1 Tossing a coin

Imagine you have a coin, like this one. It is an ordinary coin that can be found everywhere.

Figure 1.1: Tossing a coin

Everyone knows that the chance of heads (H) appearing after each toss is 50%, no one even doubts that. So, the **probability** of an **event** that heads appearing is 50%. But because you are a very curious person, you do not easily accept the truth like others, so you will find a way to test it. The fact "The probability of heads appearing is 50%" will be tested, and it can be called a **hypothesis**.

The simplest way to test your hypothesis is tossing a coin many times. You might toss a coin $N = 100$ times and see heads appear $n = 40$ times, so your actual probability is:

$$\hat{p} = \frac{n}{N} = 0.4$$

But your previous assumption (or hypothesis) states that the ideal probability is:

$$p = 0.5$$

Is there anything wrong here? Dissatisfied, you continue your experiment. This time, you toss a coin $N = 500$ times and see heads appear $n = 270$ times, so the new actual probability is:

$$\hat{p} = \frac{n}{N} = 0.54$$

This time the result is closer with initial hypothesis p, but your hand now must be very tired after tossing a coin 600 times. So do you think you have **enough** evidence to conclude that if $N \to +\infty$, then $\hat{p} \to p$? If so, your assumption is correct because *can not be refuted*; and if not, our common sense might be wrong because the evidence *strongly refutes it*.

Another strategy to toss a coin is fixing and dividing N. Instead of tossing $N = 600$ times and only getting the final p value, you can divide large number of coin tosses $N = 600$ to smaller tosses $N_1 = N_2 = \cdots = N_6 = 100$ (but do not too small, at least each $N_i > 30$), and get 6 values of $\hat{p}_i$. Suppose that you would get:

$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	$\hat{p}_5$	$\hat{p}_6$
0.43	0.46	0.56	0.53	0.49	0.51

You might see that the value $p = 0.5$ too idealistic to occur in our experiment, and $\hat{p}$ constantly changes. Therefore instead of determining the **exact value** of p, we **predict** that the p value lies inside a **closed interval** with a **certainty** of $(1-\alpha)100\%$. For example, we can confidently conclude that the value of p lies within the range (0.48,0.51) with 95% accuracy.

1.2 Weather forecasting

You do not need to know anything about geography to predict what the weather will be tomorrow. Everything you need is just knowledge about Poisson, exponential distributions and the **average number** of rainy days per month where you are living.
For example, in December there are **average** 5 rainy days. What are the probabilities of:

1. There will be 3 rainy days this month.
2. Tomorrow will be a rainy day, if today is the 10th and you have not seen any rain since the beginning of the month.

This problem will be covered in detail in **Chapter 5: Discrete Random Variables** and **Chapter 6: Continuous Random Variables**, but if you do have experience with random variables, you can try solving it!
The answer for the first question is:

$$P(X = 3) = \frac{e^{-5}5^3}{3!} = 14.03\%$$

The answer for the second question is:

$$P(X < 11|X > 10) = 1 - P(X > 11|X > 10) = 1 - e^{\frac{-5}{31}} = 14.89\%$$

Since 14.89% is quite low, so tomorrow you do not have to bring an umbrella.

1.3 Relationship between Probability and Statistics

The formal definitions of **probability** and **statistics** will be discussed later in **Chapter 2: Fundamentals of Probability** and **Chapter 8: Fundamentals of Statistics**, but now let's focus on the general model below:

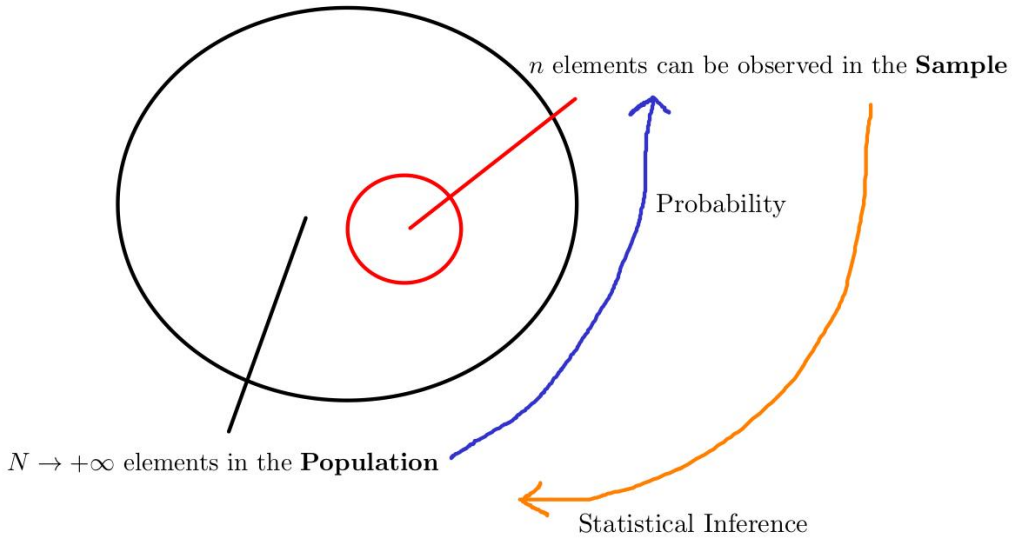


Figure 1.2: General model of simple ProbStat problems

You want to know some attributes of a very large **population** such as the probabilites, means,$\cdots$ of the quantities. But there is no way you can observe the entire population (in many cases, $N \to +\infty$) , so you have to take some **samples** (n elements) and use **probability rules** to process them. After that, you can try applying **statistical inference rules** to draw conclusions about the original population. That is how it works!

Referring to the previous example, you want to check if the probability of heads appearing is $p = 50\%$, so you toss a coin multiple times. But since you can only toss a coin for 600 times consecutively, so you conclude based on your final result $\hat{p} = 0.54$ (or closed interval with a certainty) that might be if $N \to +\infty$ then $\hat{p} \to p = 0.5$. You took the **samples** ($N = 600$ or $N_i = 100$), used **probability rule** to calculate the actual p value, and then applied **statistical inference rules** (hypothesis or closed interval) to draw conclusions respectively.

Chapter 2

Fundamentals of Probability

2.1 Sample Space and Events

2.1.1 Sample Space

If you toss a coin, you will see that there are 2 possible outcomes: heads and tails, denoted by capital letters H and T; or if you roll a die, you will see that there are 6 possible outcomes, from 1 to 6. Tossing a coin, or rolling a die are typical examples of **experiments**.

Definition 2.1.1. *An **experiment** is the process that generates a set of outcomes (or data).*

All of possible outcomes are collected into a single set.

$$S_1 = \{H, T\}$$

$$S_2 = \{1, 2, 3, 4, 5, 6\}$$

Definition 2.1.2. *The **sample space** is the set of all possible outcomes of an **experiment**.*

The sample space is usually represented by the symbol S or Ω. You can easily see that S is not always a countable or finite set. For instance, S_3 is the set of all random numbers you can choose within the range $(0, 1)$:

$$S_3 = \{x \mid 0 < x < 1\}$$

S_4 is the set of all number of coin tosses until first heads appear:

$$S_4 = \{1, 2, 3, \cdots\}$$

It is possible that you toss a coin forever, and heads never appear.

Definition 2.1.3. *A **sample point** is a single outcome of the sample space S.*

S_1, S_2 have 2 and 6 sample points respectively, while S_3, S_4 have infinite sample points.

2.1.2 Events

For any given experiment, we are often interested in the occurrence of certain **events** rather than a specific element (or **sample point**) in the sample space. For example, in the die roll experiment, you may want to know when the outcome is an even number. This will happen if the result is an element of the subset E_2 of the sample space S_2:

$$E_2 = \{2, 4, 6\}$$

Definition 2.1.4. *An **event** is the subset of a sample space.*

Events are always denoted by capital letters like $A, B, C, \cdots$. Similarly, an event is not always a countable or finite subset.
The **complement** of an event E_2 with respect to S_2 is the set of all *odd outcomes*, and can be represented as follows:

$$\overline{E}_2 = \{1, 3, 5\}$$

There are many ways to denote the **complement** of an event: $\overline{E}, E', E^c$, but in this book I choose overline notation for convenience and make it easy to relate with Boolean algebra.

Definition 2.1.5. *The **complement** of an event E with respect to S is the subset of all elements of S that are not in E, and can be denoted by the symbol $\overline{E}$.*

Applying set theory, we can perform many set operations like joint, disjoint, union,$\cdots$

Definition 2.1.6. *The **intersection** of two events A and B, denoted by the symbol $A \cap B$ is the event containing all elements that are common to A and B.*

For example, if A and B are the subset of S_3 and defined as:

$$\begin{aligned} A &= \{x \mid 0 < x < 0.7\} \\ B &= \{x \mid 0.2 < x < 0.9\} \\ \Rightarrow A \cap B &= \{x \mid 0.2 < x < 0.7\} \end{aligned}$$

Definition 2.1.7. *Two events A and B are **mutually exclusive**, or **disjoint**, if $A \cap B = \varnothing$, that is A and B have nothing in common.*

For example, E_2 and $\overline{E}_2$ are mutually exclusive.

Definition 2.1.8. *The **union** of the two events A and B, denoted by the symbol $A \cup B$ is the event containing all of the elements that belong to A or B or both.*

For example, $A \cup B = \{x \mid 0 < x < 0.9\}$, and $(A \cup B) \subset S_3$. Note that $E_2 \cup \overline{E}_2 = S_2$ is an useful result and can be generalized to corollary below:

Corollary 2.1.0.1. *If A is an event with respect to sample space S, then $A \cup \overline{A} = S$*

De Morgan's laws can also be applied to set theory.

Corollary 2.1.0.2. *If A and B are events with respect to sample space S, then*

$$\overline{A \cap B} = \overline{A} \cup \overline{B}$$

$$\overline{A \cup B} = \overline{A} \cap \overline{B}$$

These results above are really useful in many cases, but you do not have prove them since they are pretty simple. Proving them rigorously is not the focus of Probability and Statistics book.

2.2 Counting Sample Points

In this section, we develop some *counting techniques* to count the number of points in the sample space S and its event subset E without actually listing each elements. These techniques play an important role in solving some simple probability problems. Notice that these techniques can only be applied when your sample space is finite and countable.
Obviously, you can not count the number of elements of S_3 and S_4, since they are infinite and uncountable sets.

2.2.1 Rule of Product

You toss a pair of coins. How many sample points are there in the sample space? First you can try to list all of the possible outcomes:

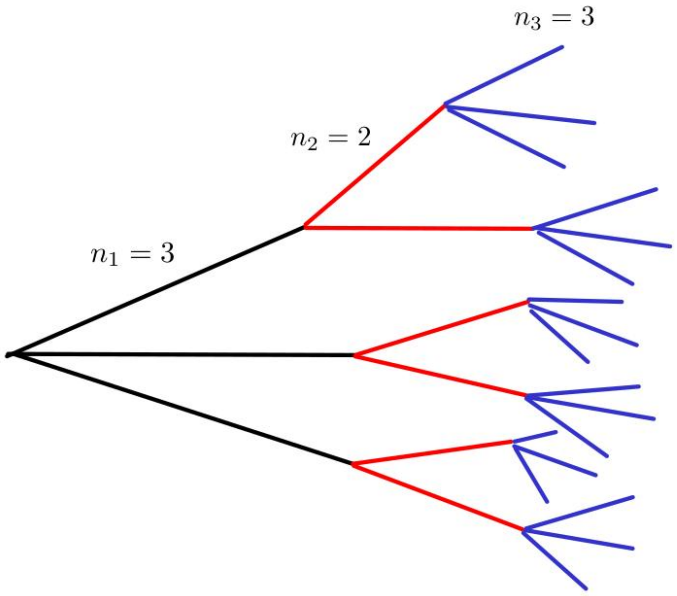
$$S = \{HH, HT, TH, TT\}$$

There are total 4 sample points in S. But listing all of the elements might be not so clever idea, so you use **rule of product**.
The first coin can land heads or tails, so can the second coin. We multiply the number of possible outcomes of the two coins:

$$n_1 n_2 = 2.2 = 4 \text{ (possible outcomes)}$$

Definition 2.2.1. *If an operation can be performed in n_1 ways, and if for each of these a second operation can be performed in n_2 ways, and for each of the first two a third operation can be performed in n_3 ways, and so fourth, the the sequence of k operations can be performed in:*

$$\prod_{i=1}^{k} n_i \text{ (ways)}$$



There are a total $n_1 n_2 n_3 = 18$ ways to perform this operation

Figure 2.1: Tree diagram for rule of product

2.2.2 Permutations

Definition 2.2.2. *A **permutation** is an arrangement of **all** or **part** of a set of objects.*

For instance, the number of permutations of n distinct objects can be counted by using rule of product:

$$n(n-1)(n-2)\cdots 3.2.1$$

We introduce a new notation for such a number.

Definition 2.2.3. *For any non-negative integer n, $n!$, called "n factorial", is defined as:*

$$n! = n(n-1)(n-2)\cdots 3.2.1$$

with special case $0! = 1$

Theorem 2.2.1. *The number of* ***permutations*** *of n distinct objects is $n!$*

In general, the number of permutations of n distinct objects taken r at a time can also be counted by using rule of product:

$$n(n-1)(n-2)\cdots(n-r+2)(n-r+1)$$

This product can be represented by the new symbol nPr.

Theorem 2.2.2. *The number of* ***permutations*** *of n distinct objects taken r at a time is:*

$$nPr = \frac{n!}{(n-r)!}$$

But how about our n objects are not distinct? Assume that n_1 are of one kind, n_2 are of a second kind,$\cdots$ and n_k of a kth kind.

Theorem 2.2.3. *The number of* ***permutations*** *of n things of which n_1 are of one kind, n_2 of a second kind, and so forth is:*

$$\frac{n!}{n_1!n_2!\cdots n_k!}$$

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

For example, from the digits 1 to 9, we can form 9! numbers made up of 9 distinct digits, or we can form $9P5$ five-digit numbers such that all the digits are different. If we allow repetation, five-digits numbers are now formed by 2 digits 1, 2 digits 2 and 1 digit 3, then the number of permutation that satisfy is:

$$\frac{5!}{2!2!1!} = 30$$

2.2.3 Combinations

Consider another problem, now you have a set of n elements. How many ways you can partition it into k cells with n_1 elements in the first cell, n_2 elements in the second cell, and so forth? Coincidentally, the equation in **Theorem 2.2.3** appears again.

Theorem 2.2.4. *The number of ways of partitioning a set of n objects into k cells with n_1 elements in the first cell, n_2 elements in the second, and so forth, is:*

$$\binom{n}{n_1, n_2, \cdots, n_k} = \frac{n!}{n_1!n_2!\cdots n_k!}$$

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

In many problems, we are interested in the number of ways of selecting k objects from n without regard to order. These selections are called **combinations**. It is not too hard to realize that a **combination** is just a partition with 2 cells, one cell containing the k objects and the other containing the $(n-k)$ objects.

$$\binom{n}{k, n-k} \text{ is often shortened to } \binom{n}{k}$$

Definition 2.2.4. *A **combination** is a selection of items from a set that has distinct elements, such that **the order of selection does not matter**.*

Theorem 2.2.5. *The number of combinations of n distinct objects taken k at a time is:*

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

For example, the number of subsets of 3 elements that can be obtained from an original set of 10 elements is:

$$\binom{10}{3} = 120$$

For the rest of this book, we mainly focus on **combinations**, and derive many probability distribution functions based on them. Like set theory, we do not go deeply into **counting techniques** since they are not the main concern of ProbStat.

2.3 Probability of an Event

Everyone knows the basic idea of probability. If I toss a **fair** coin once, and I want to know the probability of getting heads; how can I determine it? Strictly, I have to define S is the sample space of this experiment and A is the event "Getting heads after one toss".

$$S = \{H, T\}$$

$$A = \{H\}$$

So, the probability of event A occurring is:

$$P(A) = \frac{\text{number of sample points inside } A}{\text{number of sample points inside } S} = \frac{1}{2}$$

Very easy and intuitive. But how about tossing 4 **fair** coins once, and determining the probability of getting 2 heads? Now sample space S and its event set can be represented as:

$$S = \{HHHH, HHHT, HHTH, \cdots\}$$

$$A = \{TTHH, THTH, \cdots\}$$

Now we change our strategy using **counting techniques**:

$$P(A) = \frac{\text{number of sample points inside } A}{\text{number of sample points inside } S} = \frac{\binom{4}{2}}{2^4} = \frac{3}{8} = 0.375$$

In fact, the chance of getting heads is slightly greater than tails (you know, because coins are asymmetrical). Assume that the probability of heads appearing is 60%, and tails appearing is 40%. Since the role of **sample points** inside set S and subset A are not **equal** anymore, so we can not use **counting techniques** blindly.

$$P(A) = \binom{4}{2} 0.6^2 0.4^2 = 0.3456$$

Coin tossing is a typical example of **Bernoulli trial**, and the experiment tossing unfair coins is a **Bernoulli process**. The probability $P(A)$ can be calculated by using **Binomial distribution** formula. You do not have to worry about these terms, we will cover them in **Chapter 5: Discrete Random Variables** very carefully.

Consider another problem, what is the probability of getting 0.7 when randomly choosing a number (assume that the role of every numbers are equal) inside the interval $(0, 1)$?

$$S = \{x \mid 0 < x < 1\}$$

$$A = \{0.7\}$$

$$P(A) = \frac{\text{number of sample points inside } A}{\text{number of sample points inside } S} = \frac{1}{+\infty} = 0 \text{ (?????)}$$

Since $P(A) = 0$, so may we conclude 0.7 will never be chosen? Absolutely incorrect, even in common thinking. From previous examples, now we see the **limitation** of our "common definition" of probability in real life. Mathematically, our definition is just a very special case of the formal one.

Definition 2.3.1. *The **probability** of an event A is the sum of the weights (or probabilities) of all sample points in A. Therefore:*

$$0 \leq P(A) \leq 1, \quad P(\varnothing) = 0, \quad P(S) = 1$$

*If A and B are **mutually exclusive (or disjoint)**, then $P(A \cup B) = P(A) + P(B)$*

This definition can also be called Kolmogorov's axioms. An intuitive way to understand it is sketching a sample space with some events inside.

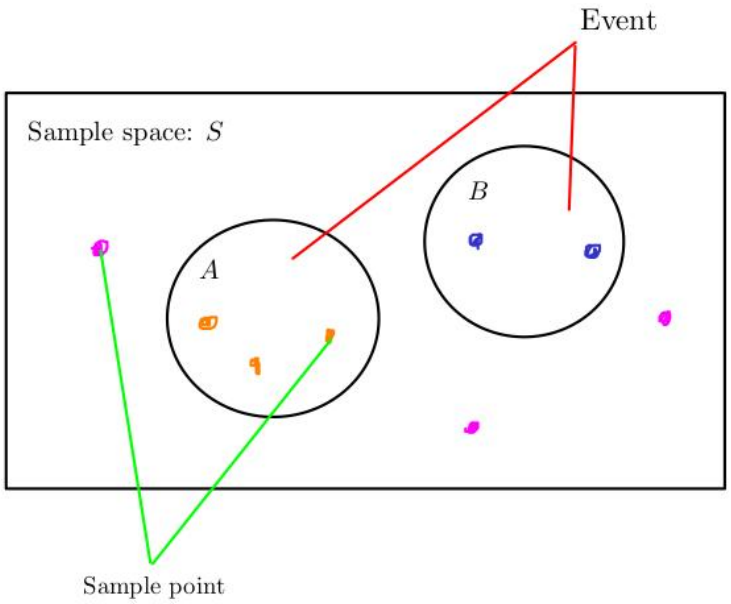


Figure 2.2: Visualize definition of probability

Literally, probability is just a number that describes how likely an event can occur. I often relate it with "weight" quantity. As you can see in the sample space, I can assign arbitrarily the "weights" (or probabilites) of sample points as follows:

- Each blue point is 0.1
- Each orange point is 0.2

- Each pink point is 0.066

Using **Definition 2.3.1**, now we can obtain these resuls:

$$P(A) = 3.0.2 = 0.6$$

$$P(B) = 2.0.1 = 0.2$$

$$P(A \cup B) = P(A) + P(B) = 0.6 + 0.2 = 0.8$$

$$P(\overline{A \cup B}) = 3.0.066 = 0.2$$

$$P(S) = 2.0.1 + 3.0.2 + 3.0.066 = 1$$

You should verify that our "special definition" of probabilty above satisfies the axioms of probability, and can be formalized by a theorem.

Theorem 2.3.1. *If an experiment can result in any one of N different **equally** likely outcomes, and if exactly n of these outcomes correspond to event A, then the probability of event A is:*

$$P(A) = \frac{n}{N}$$

Logically, you can view probability as a mapping from a set to a closed interval $[0, 1]$, then you can freely define the mapping P by yourself as long as it satisfies Kolmogorov's axioms.

$$P : \text{Set} \to [0, 1]$$
$$A \xrightarrow{P} P(A)$$

By applying set theory, we can derive several extremely useful theorems and corollaries:

Theorem 2.3.2. *If A and B are two events, then:*

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

Theorem 2.3.3. *If A is an event of sample space S, then:*

$$P(S) = P(A \cup \overline{A}) = P(A) + P(\overline{A}) = 1$$

Corollary 2.3.3.1. *If A is an event of sample space S, then:*

$$P(\overline{A}) = 1 - P(A)$$

Again, De Morgan's laws can also be applied:

Corollary 2.3.3.2. *If A and B are events of sample space S, then:*

$$P(\overline{A \cup B}) = P(\overline{A} \cap \overline{B})$$
$$P(\overline{A \cap B}) = P(\overline{A} \cup \overline{B})$$

2.4 Conditional Probability

2.4.1 Conditional Probability

Imagine you now have a perfectly fair coin. Obviously if you define events A and B are heads and tails appearing after one toss, respectively, you will conclude:

$$P(A) = P(B) = 0.5$$

But I might ask you "If event A **did** occur, might event B would have any chance to occur?" . Since there is no way heads and tails can simultaneously appear, so your answer must be "No!". Now we can form an equation to represent the probability of a "special event" (or formally, conditional event) that "If A occured, then B would occur.":

$$P(B|A) = 0$$

Now we shift our attention to a more complex problem. I give you a die, and you roll it. Events E_1 and E_2 can be defined as: "Landing a number that greater than 3" and "Landing a number that divisible by 2":

$$P(E_1) = P(E_2) = \frac{1}{3}$$

Sample points of 2 events can be listed:

$$E_1 = \{4, 5, 6\}$$

$$E_2 = \{2, 4, 6\}$$

If a number greater than 3 landed, what is the probability that it could be divisible by 2? You can count the number of sample points inside E_1 and draw a result:

$$P(E_2|E_1) = \frac{2}{3}$$

Conversely, if a number divisible by 2 landed, what is the probability that it could be greater than 3?

$$P(E_1|E_2) = \frac{2}{3}$$

As you can see here, if one event occurs before another event, probability will be completely **changed**. So we call the pairs of events A and B, E_1 and E_2 **dependent events** because they depend on each other. The "special" probabilities $P(B|A), P(E_2|E_1), P(E_1|E_2)$ are called **conditional probability**.

Definition 2.4.1. *The **conditional probability** of B, given A, denoted by $P(B|A)$, is defined:*

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

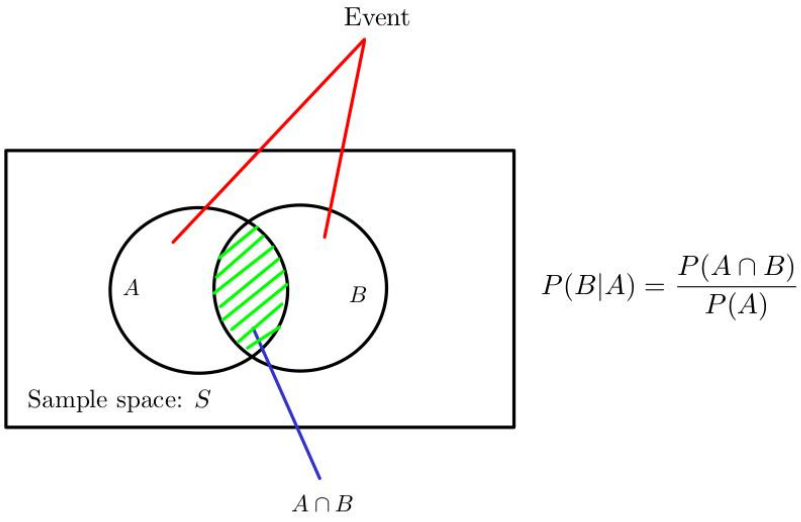


Figure 2.3: Visualize how conditional probability is calculated

Now back to our previous problems, these conditional probabilities can easily be determined using **Definition 2.4.1** above:

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{2/6}{1/2} = \frac{2}{3}$$

$$P(E_2|E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)} = \frac{2/6}{1/2} = \frac{2}{3}$$

2.4.2 Independent Events

Intuitively, we can see that if events A and B do not influence each other, then:

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

$$P(B|A) = P(B) \Leftrightarrow \frac{P(B \cap A)}{P(A)} = P(B) \Leftrightarrow P(A \cap B) = P(B)(A) \text{ (if } P(A) > 0)$$

Definition 2.4.2. *Two events A and B are independent if and only if:*

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

assuming $P(A) > 0$

Theorem 2.4.1. *Two events A and B are independent if and only if*

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

There are many examples of independent events, like if we define 2 events E_1: "Getting heads on the first toss." and E_2: "Getting heads on the second toss.". Intuitively you can see that E_1 and E_2 are unrelated, so they are **independent events**. You can also verify this fact:

$$P(E_1 \cap E_2) = \frac{1}{2}.\frac{1}{2} = \frac{1}{4} = P(E_1)P(E_2)$$

It is very important to note that determining the independence of events is completely **unrelated** to their mutual exclusion. Events are considered independent if and only if they satisfy **Theorem 2.4.1**.

2.5 Total Probability and Bayes' rule

2.5.1 Total Probability

In many situations, we do not know directly the information of $P(A)$ in **Definition 2.4.1**(or denomirator part); so in this section, we will develop a simple formula to handle this problem.

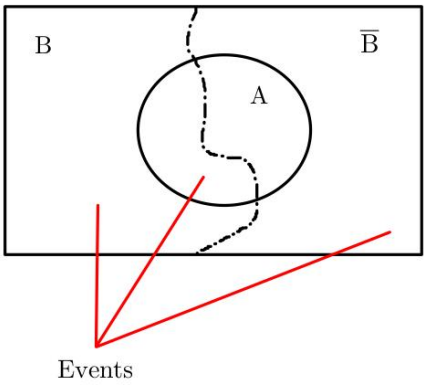


Figure 2.4: Total probability

$$A = A \cap S = A \cap (B \cup \overline{B}) = (A \cap B) \cup (A \cap \overline{B})$$

Because $A \cap B$ and $A \cap \overline{B}$ are independent events, so:

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

Theorem 2.5.1. *If A and B are two events of sample space S, then:*

$$P(A) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B})$$

Total probability is an useful formula, especially when you are conducting surveys in practice. For example, in my university, there are 2 types of students: those who "studied the whole semester" and those who "studied only one night before the exam". After the final exam, I asked everyone in my class about their scores and totaled the results. Event A: "Got $A+$" and event $\overline{A}$: "Did not get $A+$"; event B: "Studied the whole semester" and event $\overline{B}$: "Studied only one night".

$$P(A|B) = 0.8,\ P(A|\overline{B}) = 0.3,\ P(B) = 0.4$$

Using total probability formula, I obtained the probability of A:

$$P(A) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B}) = 0.8.0.4 + 0.3.(1 - 0.4) = 0.5$$

Wow, that was an impressive ratio. Half of a class received perfect score. Was this course too easy?

2.5.2 Bayes' rule

If I studied hard, I would get $A+$. But how about me, who was not keen on studying boring courses like Computer Architecture but *still survived after final test and even got "A+"*? I could questioned myself "Was I too lucky?". To answer myself, I had to calculate $P(\overline{B}|A)$, if the result is not so high, perhaps I was lucky. Bayes' rule was what I needed.

Theorem 2.5.2. *If A and B are two events of sample space S, then:*

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

Bayes' rule seems very simple. Indeed an average high school student has no difficulty finding it, but its idea is brilliant. Before Bayes, we typically only reasoned about problems in terms of cause first, then effect. But Bayes' rule opens up a completely new way of thinking for us: knowing the effect beforehand, then understanding how cause influences it.

Back to my previous problem, I applied Bayes' rule:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\overline{B})P(\overline{B})} = \frac{0.8.0.4}{0.5} = 0.64 \Rightarrow P(\overline{B}|A) = 1 - P(B|A) = 0.36$$

Since 0.36 was not a high number, indeed I was incredibly lucky and the course was not as easy as I thought.

Chapter 3

Random Variables and Probability Distributions

3.1 Definition of Random Variables

Back to our coin tossing game, now you toss a fair coin three times. This experiment has sample space S:

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

Since writing all of the sample points is quite lengthy and time-consuming, sometimes unnecessarily, so how about we assign each point with a **numerical value**? Because the assignment of values is entirely based on our own conventions, there are no constraints whatsoever. But for this reason, we should choose the **smartest** and **most convenient** way to assign values to suit our concern. For example, if I define an event A: "Getting 2 heads", then the cleverest way is assigning each sample point with its number of heads.

$$\begin{aligned} HHH &\to 3 \\ HHT, HTH, THH &\to 2 \\ HTT, TTH, THT &\to 1 \\ TTT &\to 0 \end{aligned}$$

Or event B: "Getting both tails and heads":

$$\begin{aligned} HHT, HTH, HTT, THH, THT, TTH, TTT &\to 1 \quad \text{(Yes)} \\ HHH, TTT &\to 0 \quad \text{(No)} \end{aligned}$$

These values may be viewed as values assumed by the **random variable** X, X can be "number of heads appearing" or "getting both tails and heads state", but *not both*.

Definition 3.1.1. *A **random variable** is a function that associates a real number with each element in the sample space.*

We shall use a capital letter X, to denote the random variable and its corresponding smaller letter x, for **one of its values**. A typical mistake is forgetting to define the meaning of random variable X, so *please do not forget it in your exam.* Let's continue by looking at some examples of random variables and sample spaces.
Roll the die and observe the landing value. X is the random variable defined as the value we observed:

$$S_1 = \{1, 2, 3, 4, 5, 6\}$$

Roll the die repeatedly until the sixth appears 6 times. Y is the random varibale defined as the number of rolling:

$$S_2 = \{1, 2, 3, \cdots\}$$

Use **exponential distribution** to predict if tomorrow will be a rainy day. Z is the random variable defined as the probability of the event "Tomorrow will be a rainy day":

$$S_3 = \{z \mid 0 < z < 1\}$$

The random variable X can take one of the values $x_1 = 1, x_2 = 2, \cdots, x_6 = 6$, and similarly with Y and Z can take one of their own values in their sample spaces S_2, S_3 respectively. Now we interested in classifying 2 types of sample spaces and their random variables.

Definition 3.1.2. *If a sample space contains a finite number of possibilities or an unending sequence with as many elements as there are whole numbers, it is called a* ***discrete sample space****.*

Definition 3.1.3. *If a sample space contains an infinite number of possibilities equal to the number of points on a line segment, it is called a* ***continuous sample space****.*

So S_1 and S_2 are **discrete sample spaces** and X, Y are called **discrete random variables**; S_3 is **continuous sample space** and Z is called **continuous random variable**.

3.2 Discrete Probability Distributions

Consider coin tossing experiment, now if we assign random variable X as the number of heads appearing, I can obtain some useful results:

$$\begin{aligned}
P(X = 3) &= P(\{HHH\}) = \frac{1}{8} = 0.125 \\
P(X = 2) &= P(\{HHT, HTH, THH\}) = \frac{3}{8} = 0.375 \\
P(X = 1) &= P(\{TTH, THT, HTT\}) = \frac{3}{8} = 0.375 \\
P(X = 0) &= P(\{TTT\}) = \frac{1}{8} = 0.125
\end{aligned}$$

Or in table form:

x	0	1	2	3
$P(X = x)$	0.125	0.375	0.375	0.125

Frequently, it is much convenient to represent all of the probabilities of a random variable X by a **formula**.

Definition 3.2.1. *The function $f(x)$ is a* ***probability density function*** *(pdf) of the discrete random variable X if, for each possible outcome X:*

$$\begin{cases}
f(x) \geq 0 \\
\sum_x f(x) = 1 \\
P(X = x) = f(x)
\end{cases}$$

You can derive that the pdf of coin tossing experiment above is:

$$f(x) = P(X = x) = \frac{\binom{3}{x}}{2^3} \quad (x = 0, 1, 2, 3)$$

In many cases, we want to know the probability of $(X \leq x)$ (you will see it clearly in the next section). For instance, I want to know the probability of heads appearing a maximum of 2 times.

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0.125 + 0.375 + 0.375 = 0.875$$

In general, we define a new function $F(x)$ to handle these cases as follows:

Definition 3.2.2. *The **cumulative distribution function** (cdf) $F(x)$ of a discrete random variable X with probability distribution $f(x)$ is:*

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t) \quad (-\infty < x < +\infty)$$

For example, in the above experiment:

$$F(x) = \begin{cases} 0 & (x < 0) \\ 0.125 & (0 \leq x < 1) \\ 0.5 & (1 \leq x < 2) \\ 0.875 & (2 \leq x < 3) \\ 1 & (x \geq 3) \end{cases}$$

3.3 Continuous Probability Distributions

Choosing randomly a number within the range $(0, 1)$. X is the random variable, defined as the chosen one. In the previous chapter, we have discussed that the probability of getting **a single number** in the range $(0, 1)$ is 0.

$$P(X = 0.7) = 0$$

So 0.7 will never be chosen since $P(X = 0.7) = 0$? Now think carefully about the reasons why the probability of an event might be zero. Recall the **Theorem 2.3.1**:

$$P(A) = \frac{n}{N}$$

There are 2 main reasons that could explain why $P(A)$ can be zero; the first one is $n = 0$ **and** N **is a finite number**, and the second one is n **is a finite number and** N **is an infinite number**. This might be the big misconception, since people always claim that the only reason for $P(A) = 0$ is the first one, and forgot the second. But now you can clearly see that if $n > 0$, event will always have a chance of happening. So now we conclude certainly: "$P(A) = 0$ does not mean event A will never happen."
If sample space S is *continuous sample space*, which contains *an infinite number of possibilities equal to the number of points on a line segment*, we do not care about the probability of **a single sample point** occurring (because it is always equal 0). We shift our attention to the probability of **the interval that our concern sample point may be fallen inside** occurring.

Figure 3.1: What is the probability that a number will fall within this range?

Intuitively you can conclude that:

$$P(0.6 < X < 0.78) = 0.78 - 0.6 = 0.18$$

Similarly with the discrete random variables, we can also define:

Definition 3.3.1. *The function $f(x)$ is a **probability distribution (or density) function** (pdf) for the continuous random variable X, defined over the set of real numbes, if:*

$$\begin{cases} f(x) \geq 0, \text{ for all } x \in R \\ \int_{-\infty}^{+\infty} f(x)dx = 1 \\ P(a < X < b) = \int_a^b f(x)dx \end{cases}$$

Verifying yourself that the pdf of number choosing experiment is:

$$f(x) = \begin{cases} 1 & (0 < x < 1) \\ 0 & (\text{elsewhere}) \end{cases}$$

This pdf is the simplest case of **uniform distribution function**. As I mentioned before, in the continuous sample space case, concerning on the probability of $(X = x)$ does not not make any sense since it is equal 0 and we can not use any information about it. Instead, we turn our focus to the probability of $(X < x)$ and define the **cumulative distribution function** $F(x)$ as follows:

Definition 3.3.2. *The **cumulative distribution function** (cdf) $F(x)$ of a continuous random variable X with pdf $f(x)$ is:*

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

The cdf of number choosing experiment is:

$$F(x) = \begin{cases} 0 & (x < 0) \\ x & (0 \leq x < 1) \\ 1 & (x \geq 1) \end{cases}$$

3.4 Joint Probability Distributions

3.4.1 Case of Discrete Random Variables

In the previous sections, we have considered **single random variable** and its **probability distribution function** case, and restricted ourshelves to **one-dimension sample space**. But now we are interested in observing the **simultaneous outcomes** of **several random**

variables. For example, you toss 2 fair coins simultaneously and observe their appearing faces. You define 2 random variables, X for the first coin face, and Y for the second one. The heads is assigned value 1, and the tails is assigned value 0. These are some results obtained from this experiment:

$$P(X=0,Y=0)=\frac{1}{2}.\frac{1}{2}=\frac{1}{4}=0.25$$
$$P(X=0,Y=1)=\frac{1}{2}.\frac{1}{2}=\frac{1}{4}=0.25$$
$$P(X=1,Y=0)=\frac{1}{2}.\frac{1}{2}=\frac{1}{4}=0.25$$
$$P(X=1,Y=1)=\frac{1}{2}.\frac{1}{2}=\frac{1}{4}=0.25$$

These results can be written in table form:

$P(X=x,Y=y)$	$Y=0$	$Y=1$
$X=0$	0.25	0.25
$X=1$	0.25	0.25

Another classic example of **discrete joint probability distribution** is the problem of picking balls from a basket. Now you have a basket with many colorful balls inside; there are 3 red, 4 green and 5 blue balls. You choose randomly 3 balls from the basket, and you define 2 random variables X and Y; X is the number of red balls and Y is the number of green balls. Using counting techniques and combinations, you can write the **joint pdf**:

$$f(x,y)=P(X=x,Y=y)=\frac{\binom{3}{x}\binom{4}{y}\binom{5}{3-x-y}}{\binom{12}{3}}$$

Or in table form:

$P(X=x,Y=y)$	$Y=0$	$Y=1$	$Y=2$	$Y=3$
$X=0$	$\frac{1}{22}$	$\frac{2}{11}$	$\frac{3}{22}$	$\frac{1}{55}$
$X=1$	$\frac{3}{22}$	$\frac{3}{11}$	$\frac{9}{110}$	-
$X=2$	$\frac{3}{44}$	$\frac{3}{55}$	-	-
$X=3$	$\frac{1}{220}$	-	-	-

Definition 3.4.1. *The function $f(x,y)$ is a **joint pdf** of the **discrete random variables** X and Y if:*

$$\begin{cases} f(x,y)\geq 0 \\ \sum_x \sum_y f(x,y)=1 \\ P(X=x,Y=y)=f(x,y) \end{cases}$$

Corollary 3.4.0.1. *For any region A in the xy plane:*

$$P[(X,Y)\in A]=\sum\sum_{A} f(x,y)$$

After considering the case where both variables X and Y are both varying, what if we **fix** one variable and vary the other? You can see this idea appearing very naturally in the process of finding the values of the above table using a calculator. We introduce the new functions called **marginal distributions** of X and Y alone.

Definition 3.4.2. *The **marginal distributions** of X alone and of Y alone for the **discrete case** are:*

$$g(x) = \sum_y f(x,y); \quad h(y) = \sum_x f(x,y)$$

Since the general form of $g(x)$ and $h(y)$ are not easy to be generalized, and the range of discrete random variables X and Y is very narrow, so we should reuse the table above to obtain values of these functions.

$P(X=x, Y=y)$	$Y=0$	$Y=1$	$Y=2$	$Y=3$	$g(x)$
$X=0$	$\frac{1}{22}$	$\frac{2}{11}$	$\frac{3}{22}$	$\frac{1}{55}$	$g(0)=\frac{21}{55}$
$X=1$	$\frac{3}{22}$	$\frac{3}{11}$	$\frac{9}{110}$	-	$g(1)=\frac{27}{55}$
$X=2$	$\frac{3}{44}$	$\frac{3}{55}$	-	-	$g(2)=\frac{27}{220}$
$X=3$	$\frac{1}{220}$	-	-	-	$g(3)=\frac{1}{220}$
$h(y)$	$h(0)=\frac{14}{55}$	$h(1)=\frac{28}{55}$	$h(2)=\frac{12}{55}$	$h(3)=\frac{1}{55}$	1

Corollary 3.4.0.2. *If $g(x)$ and $h(y)$ are marginal distributions of X alone and Y alone for the **discrete case**, then:*

$$\sum_x g(x) = \sum_y h(y) = 1$$

Corollary 3.4.0.2 is directly derived from the **Definition 3.4.1**, and you can verify them by using the probability distribution table.
After defining the **marginal distribution**, now we can see clearly the connection between them and regular pdfs are:

$$P(X=x) = g(x); \quad P(Y=y) = h(y)$$

Now if I choose 3 balls from the basket, and I know two of them are green, what is the probability that the remaining ball is red?

$$P(X=1|Y=2) = \frac{P(X=1,Y=2)}{P(Y=2)} = \frac{\frac{9}{110}}{\frac{12}{55}} = \frac{3}{8} = 0.375$$

In general, it is not hard to deduce these formulas:

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

Definition 3.4.3. *Let X and Y be two **discrete random variables**. The **conditional distribution** of the variable Y given that $(X=x)$ is:*

$$f(y|x) = \frac{f(x,y)}{g(x)}$$

Similarly for the reverse case:

$$f(x|y) = \frac{f(x,y)}{h(y)}$$

From the previous section **Chapter 2.4: Conditional Probability**, you already knew how to identify two independent events, by checking this condition:

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

Now we use it again, with random variables X and Y:

$$P(X=x, Y=y) = P(X=x)P(Y=y) \Leftrightarrow f(x,y) = g(x)h(y)$$

Definition 3.4.4. *Let X and Y be two* ***discrete random variables****, with joint pdf $f(x,y)$ and marginal distributions $g(x)$, $h(y)$, respectively. They are said to be* ***statistically independent*** *if and only if:*

$$f(x,y) = g(x)h(y)$$

If we check the case $(X = 1, Y = 1)$:

$$\left(f(1,1) = \frac{3}{11}\right) \neq \left(g(1)h(1) = \frac{27}{55}.\frac{28}{55}\right)$$

So X and Y are **not** statistically independent, they are interdependent. But how interdependent are they? Are they highly or minimally interdependent? This question will be answered at the end of the next chapter.

3.4.2 Case of Continuous Random Variables

Similarly, the joint pdf of **continuous random variable** can also be defined as follows:

Definition 3.4.5. *The function $f(x,y)$ is a* ***joint pdf*** *of the* ***continuous random variables*** *X and Y if:*

$$\begin{cases} f(x,y) \geq 0 \\ \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy = 1 \\ P(a < X < b, c < Y < d) = \int_c^d \int_a^b f(x,y)dxdy \end{cases}$$

Corollary 3.4.0.3. *For any region A in the xy plane:*

$$P[(X,Y) \in A] = \iint_A f(x,y)dxdy$$

Almost all the formulas in the above section are redefined exactly the same, with a slight difference in the notation for the integral (you can read more about Riemann sum to see the relationship between integrals and infinite discrete sums).

Definition 3.4.6. *The* ***marginal distributions*** *of X alone and of Y alone for the* ***continuous case*** *are:*

$$g(x) = \int_{-\infty}^{+\infty} f(x,y)dy; \quad h(y) = \int_{-\infty}^{+\infty} f(x,y)dx$$

Corollary 3.4.0.4. *If $g(x)$ and $h(y)$ are marginal distributions of X alone and Y alone for the* ***continuous case****, then:*

$$\int_{-\infty}^{+\infty} g(x)dx = \int_{-\infty}^{+\infty} h(y)dy = 1$$

Definition 3.4.7. *Let X and Y be two* ***continuous random variables****. The* ***conditional distribution*** *of the variable Y given that $(X = x)$ is:*

$$f(y|x) = \frac{f(x,y)}{g(x)}$$

Similarly for the reverse case:

$$f(x|y) = \frac{f(x,y)}{h(y)}$$

But notice that if X and Y are continuous random variables, then $P(X = x|Y = y)$ is always equal zero! The correct way to obtain conditional probability value is shown below:

Corollary 3.4.0.5. *If X and Y be two* ***continuous random variables****, then:*

$$P(a < X < b|Y = y) = \int_a^b f(x|y)dx$$

Definition 3.4.8. *Let X and Y be two* ***continuous random variables****, with joint pdf $f(x, y)$ and marginal distributions $g(x)$, $h(y)$, respectively. They are said to be* ***statistically independent*** *if and only if:*

$$f(x, y) = g(x)h(y)$$

Chapter 4

Mathematical Expectation

Due to the relatively high degree of similarity between the formulas in the case of discrete and continuous random variables, this chapter approaches them **in parallel**, rather than seperating them like previous chapter.

4.1 Mean of a Random Variable

After conducting experiment, now it is time to process your obtained results. Toss a fair coin for 3 times, and define the discrete random variable X as the number of heads appearing. The pdf of this experiment is:

$$f(x) = P(X = x) = \frac{\binom{3}{x}}{2^3} \quad (x = 0, 1, 2, 3)$$

Or in table form:

x	0	1	2	3
$P(X = x)$	0.125	0.375	0.375	0.125

What is the **mean** or **expected value** of X? Intuitively, you will use the formula:

$$\mu_X = E(X) = \sum_x xf(x) = 0.0.125 + 1.0.375 + 2.0.375 + 3.0.125 = 1.5$$

By illustrating the above results with a graph, the position of μ is shown.

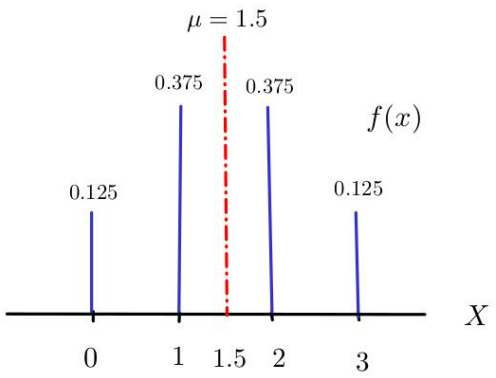


Figure 4.1: Pdf of fair coin tossing experiment with its mean

If our coin is unfair, assume that the probability of getting heads is $p = 0.3$, then the probability of getting tails is $q = 1 - p = 0.7$. Now the pdf is:

$$f(x) = P(X = x) = \binom{3}{x} p^x q^{3-x} = \binom{3}{x} 0.3^x 0.7^{3-x} \quad (x = 0, 1, 2, 3)$$

Or in table form:

x	0	1	2	3
$P(X = x)$	0.343	0.441	0.189	0.027

In this experiment, we **expect** to get the **average value** (or mean) of X:

$$\mu_X = E(X) = \sum_x x f(x) = 0.0.343 + 1.0.441 + 2.0.189 + 3.0.027 = 0.9$$

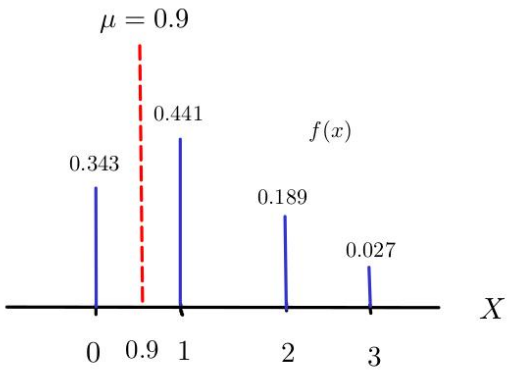


Figure 4.2: Pdf of unfair coin tossing experiment with its mean

Definition 4.1.1. *Let X be a random variable with pdf $f(x)$. The **expected value** or **mean** of X is:*

$$\mu_X = E(X) = \sum_x x f(x) \quad (\textit{if } X \textit{ is discrete})$$

$$\mu_X = E(X) = \int_{-\infty}^{+\infty} x f(x) dx \quad (\textit{if } X \textit{ is continuous})$$

Now imagine you are playing coin tossing game (do not treat it like an experiment). As I mention the rule above, you have 3 turns to toss a coin. If you see 3 heads appear, you are extremely lucky today; but if you do not see any, do not be sad because life is long. A very natural thought occurred to me that we should quantify a player's "luck level" with a quantitative value. Random variable Y takes this role and can be defined as:

$$Y = \frac{X}{3}$$

Now we are intersted in average "luck level" of coin tossing game, so reuse the previous table that we have created:

x	0	1	2	3
$y = \frac{x}{3}$	0	0.333	0.666	1
$P(X = x)$	0.343	0.441	0.189	0.027

The **expected value** or **mean** of "luck level" is:

$$\mu_Y = E(Y) = \sum_y yf(x) = 0.0.343 + 0.333.0.441 + 0.666.0.189 + 1.0.027 = 0.3$$

Since 0.3 is not so high value, so the **mean** of "luck level" is pretty small and players should not look forward to their chances in this game. Furthermore, you should notice the subtle connection between two random variables X and Y as:

$$Y = \frac{X}{3} \Leftrightarrow E(Y) = \frac{E(X)}{3}$$

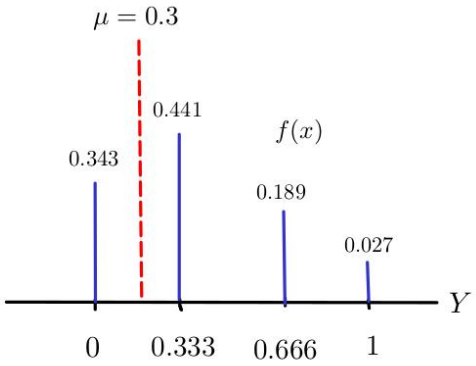


Figure 4.3: The mean value of "luck level"

If Y is defined as **function of random variable** X or $(Y = g(X))$, then the **expected value** of it can be calculated using the theorem below:

Theorem 4.1.1. *Let X be a random variable with pdf $f(x)$, and the **expected value** of the random variable $g(X)$ is:*

$$\mu_{g(X)} = E(g(X)) = \sum_x g(x)f(x) \quad \textit{(if } X \textit{ is discrete)}$$

$$\mu_{g(X)} = E(g(X)) = \int_{-\infty}^{+\infty} g(x)f(x)dx \quad \textit{(if } X \textit{ is continuous)}$$

Now I want to make our game funnier because just tossing a coin is too boring, so I came up with an idea. How about toss 3 unfair coins and the bottle cap **at the same time**? The random variable X is defined as the number of heads appearing with the probability of getting heads is $p = 0.3$, and getting tails is $q = 1 - p = 0.7$, so the pdf is:

$$g(x) = P(X = x) = \binom{3}{x} p^x q^{3-x} = \binom{3}{x} 0.3^x 0.7^{3-x} \quad (x = 0, 1, 2, 3)$$

A 'PEPSI' bottle cap has just been founded, so I define a new random variable Y as its face after tossing. The 'PEPSI' face is assigned the value 1, and the other is assigned the value 0. Assume that:

$$P(Y = 1) = p' = 0.8$$
$$P(Y = 0) = q' = 1 - p' = 0.2$$

So the pdf of bottle cap tossing experiment is:

$$h(y) = P(Y = y) = (p')^y (q')^{1-y} = 0.8^y 0.2^{1-y} \quad (y = 0, 1)$$

Intuitively, you can certainly conclude that X and Y are **statistically independent** since coins and bottle cap tossing process do not affect each other. So the **joint pdf** of them is:

$$f(x, y) = g(x)h(y) = \binom{3}{x} 0.3^x 0.7^{3-x} 0.8^y 0.2^{1-y} \quad (x = 0, 1, 2, 3;\ y = 0, 1)$$

Or in table form:

$P(X = x, Y = y)$	$X = 0$	$X = 1$	$X = 2$	$X = 3$	$h(y)$
$Y = 0$	0.0686	0.0882	0.0378	0.0054	$h(0) = 0.2$
$Y = 1$	0.2744	0.3528	0.1512	0.0216	$h(1) = 0.8$
$g(x)$	$g(0) = 0.343$	$g(1) = 0.441$	$g(2) = 0.189$	$g(3) = 0.027$	1

Now "luck level" can be defined as the value of $(X + Y)$. Thinking according to the same logic, the **mean** value of $(X + Y)$ is:

$$\begin{aligned}
\mu_{(X+Y)} = E(X + Y) &= \sum_x \sum_y (x + y) f(x, y) = \sum_x \left(\sum_y (x + y) f(x, y) \right) \\
&= \sum_x (x f(x, 0) + x f(x, 1) + 0 f(x, 0) + 1 f(x, 1)) \\
&= \sum_x (x f(x, 0) + x f(x, 1) + f(x, 1)) \\
&= 0 f(0, 0) + 1 f(1, 0) + 2 f(2, 0) + 3 f(3, 0) + 0 f(0, 1) + 1 f(1, 1) \\
&+ 2 f(2, 1) + 3 f(3, 1) + f(0, 1) + f(1, 1) + f(2, 1) + f(3, 1) \\
&= 0.0882 + 2.0.0378 + 3.0.0054 + 0.3528 + 2.0.1512 \\
&+ 3.0.00216 + 0.2744 + 0.3528 + 0.1512 + 0.0216 \\
&= 1.7
\end{aligned}$$

So now, the players can confidently look forward for their opportunity in this game!

Definition 4.1.2. *Let X and Y be random variables with **joint pdf** $f(x, y)$. The **mean** or **expected value** of the random variable function $g(X, Y)$ is:*

$$\mu_{g(X,Y)} = E(g(X, Y)) = \sum_x \sum_y g(x, y) f(x, y) \quad \textit{(if } X \textit{ and } Y \textit{ are discrete)}$$

$$\mu_{g(X,Y)} = E(g(X, Y)) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) f(x, y) dx dy \quad \textit{(if } X \textit{ and } Y \textit{ are continuous)}$$

If $g(X, Y) = X$, then we obtain some useful results and can be represented as corollaries:

Corollary 4.1.1.1. *Let X and Y be discrete random variables with **joint pdf** $f(x, y)$, the **mean** of X is:*

$$\mu_X = E(X) = \sum_x \sum_y x f(x, y) = \sum_x x \left(\sum_y f(x, y) \right) = \sum_x x g(x)$$

Similarly, the mean of Y is:

$$\mu_Y = E(Y) = \sum_y y h(y)$$

Corollary 4.1.1.2. *Let X and Y be continuous random variables with* ***joint pdf*** *$f(x,y)$, the* ***mean*** *of X is:*

$$\mu_X = E(X) = \int_{-\infty}^{+\infty} xg(x)dx$$

Similarly, the mean of Y is:

$$\mu_Y = E(Y) = \int_{-\infty}^{+\infty} yh(y)dy$$

4.2 Variance and Covariance of Random Variables

4.2.1 Variance of Random Variables

The concept of **mean** or **expected value** is very important in ProbStat. In any probability distribution graph, we use the **mean value** as **main reference point** (although the mean is not always in the center of the graph). We are very interested in how the data is distributed **around the mean value**. Is it distributed far or close to it? How does it spread?

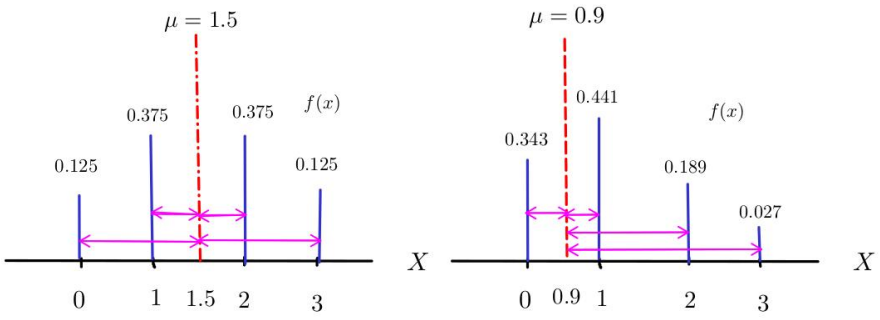


Figure 4.4: How is the data distributed around the **mean**?

The sums of total squares of the distances from the mean are:

$$\sum = (0-1.5)^2 + (1-1.5)^2 + (2-1.5)^2 + (3-1.5)^2 = 5 \text{ (fair coin case)}$$
$$\sum = (0-0.9)^2 + (1-0.9)^2 + (2-0.9)^2 + (3-0.9)^2 = 6.44 \text{ (unfair coin case)}$$

Because $6.44 > 5$, and you can also see on the graph, the data from the unfair coin toss experiment is more **widely distributed** compare to the fair coin case. For simplicity, we often take **the average of sums** and call it **variance**.

Definition 4.2.1. *Let X be a random variable with pdf $f(x)$ and mean μ. The* ***variance*** *of X is:*

$$\sigma_X^2 = E((X-\mu_X)^2) = \sum_x (x-\mu_X)^2 f(x) \quad \textit{(if } X \textit{ is discrete)}$$

$$\sigma_X^2 = E((X-\mu_X)^2) = \int_{-\infty}^{+\infty} (x-\mu_X)^2 f(x)dx \quad \textit{(if } X \textit{ is continuous)}$$

The positive square root of the variance, σ_X, is called the ***standard deviation*** *of X.*

Theorem 4.2.1. *The* ***variance*** *of a random variable X is:*

$$\sigma_X^2 = E(X^2) - \mu_X^2$$

Proof. We only prove for the discrete case because proving for the continuous case using exactly the same logic:

$$\sigma_X^2 = E((X-\mu_X)^2) = \sum_x (x-\mu_X)^2 f(x) = \sum_x x^2 f(x) - 2\mu_X \sum_x x f(x) + \mu_X^2 \sum_x f(x)$$
$$= E(X^2) - 2\mu_X^2 + \mu_X^2 = E(X^2) - \mu_X^2$$

□

At this point you may wonder "Why do not we care about the sum of total **distances** itself?" Well, there are 3 main reasons; the first is absolutely avoiding **negative values** from miscalculation, the second is working with σ_X^2 is much easier than σ_X and you will understand the final reason after reading **Chapter 11: Simple Linear Regression**. Using a completely similar line of thinking as before, we can also develop the following formulas:

Theorem 4.2.2. *Let X be a random variable with pdf $f(X)$. The **variance** of the random variable $g(X)$ is:*

$$\sigma_{g(X)}^2 = E((g(X)-\mu_{g(X)})^2) = \sum_x (g(x)-\mu_{g(X)})^2 f(x) \quad \textit{(if } X \textit{ is discrete)}$$

$$\sigma_{g(X)}^2 = E((g(X)-\mu_{g(X)})^2) = \int_{-\infty}^{+\infty} (g(x)-\mu_{g(X)})^2 f(x)dx \quad \textit{(if } X \textit{ is continuous)}$$

Corollary 4.2.2.1. *The **variance** of random variable $g(X)$ is:*

$$\sigma_{g(X)}^2 = E(g(X)^2) - \mu_{g(X)}^2$$

Proof. We only prove for the discrete case:

$$\sigma_{g(X)}^2 = E((g(X)-\mu_{g(X)})^2) = \sum_x (g(x)-\mu_{g(X)})^2 f(x)$$
$$= \sum_x g(x)^2 f(x) - 2\mu_{g(X)} \sum_x g(x) f(x) + \mu_{g(X)}^2 \sum_x f(x)$$
$$= E(g(X)^2) - 2\mu_{g(X)}^2 + \mu_{g(X)}^2 = E(g(X)^2) - \mu_{g(X)}^2$$

□

4.2.2 Covariance of Random Variables

In this subsection, we will answer the question: "How are 2 random variables X and Y interdependent?" after checking the **statistically independent criterion**:

$$f(x,y) = g(x)h(y)$$

We define a new concept called **covariance**, which is **a measure of the nature of the association between the two**.

Definition 4.2.2. *Let X and Y be random variables with joint pdf $f(x,y)$. The covariance of X and Y is:*

$$\sigma_{XY} = E((X-\mu_X)(Y-\mu_Y)) = \sum_x \sum_y (x-\mu_X)(y-\mu_Y) f(x,y) \quad \textit{(if } X \textit{ and } Y \textit{ are discrete)}$$

$$\sigma_{XY} = E((X-\mu_X)(Y-\mu_Y)) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x-\mu_X)(y-\mu_Y) f(x,y)dxdy \quad \textit{(if } X \textit{ and } Y \textit{ are continuous)}$$

Applying directly **Definition 4.2.2** is not convenient in many situations, so we should rewrite its forumula:

Corollary 4.2.2.2. *The **covariance** of random variable X and Y is:*

$$\sigma_{XY} = E(XY) - \mu_X \mu_Y$$

Proof. We only prove for the discrete case:

$$\begin{aligned}
\sigma_{XY} &= \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y) \\
&= \sum_x \sum_y xy f(x, y) - \mu_Y \sum_x \sum_y x f(x, y) - \mu_X \sum_x \sum_y y f(x, y) + \mu_X \mu_Y \sum_x \sum_y f(x, y) \\
&= E(XY) - \mu_Y \sum_x x g(x) - \mu_X \sum_y y h(y) + \mu_X \mu_Y \\
&= E(XY) - \mu_Y \mu_X - \mu_X \mu_Y + \mu_X \mu_Y \\
&= E(XY) - \mu_X \mu_Y
\end{aligned}$$

□

Although the **covariance** between two random variables does provide information regarding the nature of the relationship, the magnitude of σ_{XY} does not indicate anything regarding the strength of the relationship, since σ_{XY} is not scale-free quantity. There is a scale-free version of the covariance called the **correlation coefficient** that is widely used in ProbStat.

Definition 4.2.3. *Let X and Y be random variables with **covariance** σ_{XY} and **standard deviations** σ_X and σ_Y, respectively. The **correlation coefficient** of X and Y is:*

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Theorem 4.2.3. *The **absolute value** of **correlation coefficient** ρ_{XY} is always smaller or equal* 1.

$$|\rho_{XY}| \leq 1$$

Proof. Proving this theorem above is quiet difficult and requires knowledge of **Linear Algebra**. I recommend skipping this proof and moving on to the next section to learn some **useful theorems** first, and then back here later.
In the **random variables vector space** sharing the same joint pdf, we now define the **inner product** is:

$$\langle X, Y \rangle = E(XY)$$

With X, Y and W are three vectors of this vector space and for any constant $c \in \mathbb{R}$, now we check if our **definition** satisfies 5 axioms of inner product.

$$\begin{gathered}
\langle X, Y \rangle = \langle Y, X \rangle \leftrightarrow E(XY) = E(YX) \\
c \langle X, Y \rangle = \langle cX, Y \rangle \leftrightarrow cE(XY) = E(cXY) \\
\langle X, Y + W \rangle = \langle X, Y \rangle + \langle X, W \rangle \leftrightarrow E(X(Y + W)) = E(XY) + E(XW) \\
\langle X, X \rangle \geq 0 \leftrightarrow E(X^2) \geq 0 \\
\langle X, X \rangle = 0 \Leftrightarrow X = 0 \leftrightarrow E(X^2) = 0 \Leftrightarrow X = 0
\end{gathered}$$

The original inequality is equivalent to:

$$\begin{gathered}
|\rho_{XY}| \leq 1 \Leftrightarrow \sigma_{XY}^2 \leq (\sigma_X \sigma_Y)^2 \\
\Leftrightarrow E^2((X - \mu_X)(Y - \mu_Y)) \leq E((X - \mu_X)^2) E((Y - \mu_Y)^2)
\end{gathered}$$

Define 2 new random variables A and B (or vectors):

$$A = X - \mu_X$$

$$B = Y - \mu_Y$$

So we have to prove:

$$E^2(AB) \leq E(A^2)E(B^2) \Leftrightarrow |E(AB)| \leq \sqrt{E(A^2)E(B^2)}$$

Rewrite in vector form, this is exactly Cauchy-Schwarz inequality:

$$|\langle A, B\rangle| \leq ||A||.||B||$$

Now **Theorem 4.2.3** has been completely proven. The equality of the inequality occurs if and only if $B = kA$, where k is any real constant.

$$B = kA \Leftrightarrow Y - \mu_Y = k(X - \mu_X) \Leftrightarrow Y = kX + (\mu_Y - k\mu_X)$$

Since k is an arbitrary value in $\mathbb{R}$, so we can conclude that if and only if $|\rho_{XY}| = 1$, then $Y = aX + b$ $(a, b \in \mathbb{R})$ or in other words, they have a linear relationship. □

Corollary 4.2.3.1. *If $\rho_{XY} = 0$, then X and Y are **statistically independent**.*

Proof. $\rho_{XY} = 0 \Leftrightarrow \sigma_{XY} = E(XY) - \mu_X\mu_Y = 0 \Leftrightarrow E(XY) = \mu_X\mu_Y$. Or in equivalent:

$$\begin{aligned} & E(XY) = \mu_X\mu_Y \\ \Leftrightarrow & \sum_x \sum_y xyf(x,y) = \sum_x xg(x) \sum_y yh(y) \\ \Leftrightarrow & \sum_x \sum_y xyf(x,y) = \sum_x \sum_y xyg(x)h(y) \end{aligned}$$

Finally, $f(x, y) = g(x)h(y)$, and we conclude X and Y are statistically independent. □

From the equality condition of the Cauchy-Schwarz inequality, we can deduce:

Corollary 4.2.3.2. *If $\rho_{XY} = 1$, then:*

$$Y = aX + b \quad (a > 0)$$

And if $\rho_{XY} = -1$, then:

$$Y = aX + b \quad (a < 0)$$

4.3 Means and Variances of Linear Combinations of Random Variables

In this section, we will derive some exetremely useful theorems and corollaries. You should note that all results obtained from this section will be reused multiple times throughout the rest of this book.

Theorem 4.3.1. *If a, b are 2 constants, then: $E(aX + b) = aE(X) + b$*

Proof. By definition of the **mean**:

$$E(aX+b) = \sum_x (ax+b)f(x) = a\sum_x f(x) + b\sum_x f(x) = aE(X) + b$$

□

Corollary 4.3.1.1. *Setting $a = 0$, we see that $E(b) = b$*

Corollary 4.3.1.2. *Setting $b = 0$, we see that $E(aX) = 0$*

Theorem 4.3.2. *If $g(X)$ and $h(X)$ are two functions of random variable X, then:*

$$E(g(X) + h(X)) = E(g(X)) + E(h(X))$$

Proof. Also by the definition of the **expected value**:

$$E(g(X)+h(X)) = \sum_x (g(x)+h(x))f(x) = \sum_x g(x)f(x) + \sum_x h(x)f(x) = E(g(X)) + E(h(X))$$

□

Theorem 4.3.3. *If $g(X,Y)$ and $h(X,Y)$ are two functions of random variable (X,Y), then:*

$$E(g(X,Y) + h(X,Y)) = E(g(X,Y)) + E(h(X,Y))$$

Proof. Applying directy the definition of the **expected value**, we have:

$$\begin{aligned} E(g(X,Y)+h(X,Y)) &= \sum_x \sum_y (g(x,y)+h(x,y))f(x,y) \\ &= \sum_x \sum_y g(x,y)f(x,y) + \sum_x \sum_y h(x,y)f(x,y) \\ &= E(g(X,Y)) + E(h(X,Y)) \end{aligned}$$

□

Corollary 4.3.3.1. *Setting $g(X,Y) = g(X)$ and $h(X,Y) = h(Y)$, then:*

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

Corollary 4.3.3.2. *Setting $g(X,Y) = X$ and $h(X,Y) = Y$, then:*

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

Theorem 4.3.4. *If X and Y are statistically independent, then:*

$$E(XY) = \mu_X \mu_Y$$

Proof. We already have $f(x,y) = g(x)h(y)$.

$$E(XY) = \sum_x \sum_y xyf(x,y) = \sum_x \sum_y xyg(x)h(y) = \left(\sum_x xg(x)\right)\left(\sum_y yh(y)\right) = \mu_X\mu_Y$$

□

You should note that X and Y are independent random variables $\Leftrightarrow E(XY) = \mu_X\mu_Y$

Theorem 4.3.5. *If X and Y are two **independent random variables** with the joint pdf $f(x, y)$ with three constants a, b, c, then:*

$$\sigma^2_{aX+bY+c} = a^2\sigma^2_X + b^2\sigma^2_Y$$

Proof. Using definition, we have:

$$\begin{aligned}
\sigma^2_{aX+bY+c} &= E((aX + bY + c - \mu_{aX+bY+c})^2) \\
&= E((aX + bY + c - a\mu_X - b\mu_Y - c)^2) \\
&= E((a(X - \mu_X) + b(Y - \mu_Y))^2) \\
&= E(a^2(X - \mu_X)^2) + 2abE(X - \mu_X)(Y - \mu_Y) + E(b^2(Y - \mu_Y)^2) \\
&= a^2\sigma^2_X + 2ab\sigma_{XY} + b^2\sigma^2_Y \\
&= a^2\sigma^2_X + b^2\sigma^2_Y \quad (\sigma_{XY} = 0)
\end{aligned}$$

□

Corollary 4.3.5.1. *Setting $b = 0$, we see that: $\sigma^2_{aX+c} = a^2\sigma^2_X$*

Corollary 4.3.5.2. *Setting both $b = c = 0$, we see that: $\sigma^2_{aX} = a^2\sigma^2_X$*

Corollary 4.3.5.3. *Setting $c = 0$, we see that:*

$$\sigma^2_{aX+bY} = a^2\sigma^2_X + b^2\sigma^2_Y$$

*This corollary is only true when X and Y are **statistically independent**.*

4.4 Markov's and Chebyshev's Inequalities

4.4.1 Markov's Inequality

Theorem 4.4.1. *If X is a **non-negative** random variable with pdf $f(x)$, then:*

$$P(X \geq a) \leq \frac{E(X)}{a}$$

Proof. From the intitial condition $X \geq 0$:

$$E(X) = \sum_x xf(x) \geq \sum_{x \geq a} xf(x) \geq \sum_{x \geq a} af(x) = aP(X \geq a) \Rightarrow P(X \geq a) \leq \frac{E(x)}{a}$$

□

4.4.2 Chebyshev's Inequality

Theorem 4.4.2. *If X is any random variables with pdf $f(x)$, then:*

$$P(\mu_X - k\sigma_X < X < \mu_X + k\sigma_X) \geq 1 - \frac{1}{k^2}$$

Proof. Applying Markov's inequality with $(X - \mu_X)^2$ as non-negative random variable, now we obtain:

$$\begin{aligned}
P((X - \mu_X)^2 \geq a^2) &\leq \frac{E((X - \mu_X)^2)}{a^2} = \frac{\sigma^2_X}{a^2} \\
\Rightarrow P((X - \mu_X)^2 < a^2) &\geq 1 - \frac{\sigma^2_X}{a^2}
\end{aligned}$$

Using the positive scaling factor k: $a = k\sigma_X$ from the **standard deviation**, now we obtain:

$$P(\mu_X - k\sigma_X < X < \mu_X + k\sigma_X) \geq 1 - \frac{1}{k^2}$$

□

Chebyshev's Inequality can also be written as:

$$P(|X - \mu_X| \geq k\sigma_X) \leq \frac{1}{k^2}$$

This result is very important, especially in **case of continuous random variable**, the range of X usually extends to infinity; it shows us that the first and last ends of any pdf are **bounded**. In general, the shape of a valid pdf graph is always **flattened** at both ends according to the inequality above.

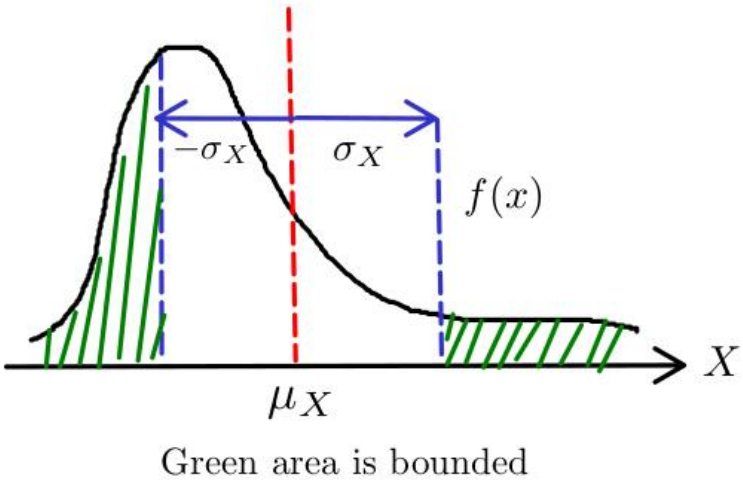


Figure 4.5: Example of a valid pdf graph

Chapter 5

Some Discrete Probability Distributions

Form this chapter to the end of this book, if X is a random variable with pdf $f(x)$, we can denote it as:

$$X \sim f(x)$$

5.1 Bernoulli, Binomial and Poisson Distributions

5.1.1 Bernoulli Distribution

Because tossing a coin many times might be so boring, so in this chapter we will begin with more vivid examples. Let's plan to plant some mung bean plants (or maybe just imagine it)! Now you have to buy a packet of mung bean seeds from the agriculture store.

Figure 5.1: Mung bean seeds

All seed packets clearly state the germination rate of the seeds. Assume that the germination rate is $p = 0.8$. Now if we define the random variable X as the germination state of **a single seed**, then the pdf of X can be represented as follows, where $(X = 1)$ describes the germinating state of the seed, and $(X = 0)$ describes the opposite state:

x	0	1
$P(X=x)$	0.2	0.8

Or in equation form:

$$f(x) = P(X = x) = 0.8^x 0.2^{1-x} \quad (x = 0, 1)$$

The experiment to test the germination state of **a single mung bean seed** is a typical exmaple of **Bernoulli trial**. Formally, a **Bernoulli trial** is a random experiment with exactly **two possible outcomes**: "success" and "failure". The success rate is always denoted by the letter p, and failure rate is q. Because they complement each other, so: $q = 1 - p$.

Definition 5.1.1. *The **Bernoulli trial** is a random experiment with exactly **two** possible outcomes: "success" and "failure".*

Theorem 5.1.1. *The pdf of a **Bernoulli trial** is:*

$$\mathscr{I}(x; 1, p) = p^x q^{1-x} \quad (x = 0, 1)$$

Corollary 5.1.1.1. *If $X \sim \mathscr{I}(x; 1, p)$, then:*

$$\mu_X = p$$
$$\sigma_X^2 = pq$$

Proof. For the mean:

$$\mu_X = \sum_x x f(x) = p^1 q^0 = p$$

For the variance:

$$\sigma_X^2 = E(X^2) - \mu_X^2 = \sum_x x^2 f(x) - p^2 = p - p^2 = pq$$

□

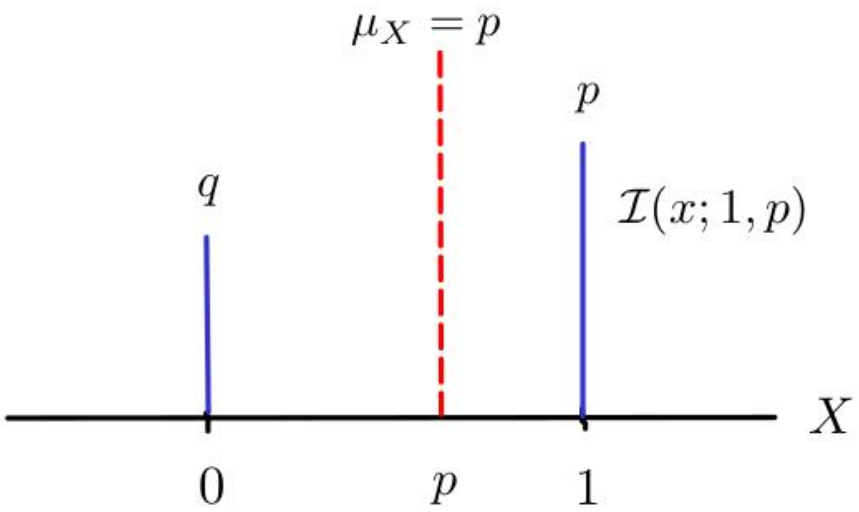


Figure 5.2: The pdf of Bernoulli distribution

5.1.2 Binomial Distribution

We know that the germination rate is $p = 0.8$. Now you pick randomly 10 mung bean seeds from the pocket and put them on the wet tissue.

Figure 5.3: Test the germination ability of 10 seeds

Perhaps after waiting 3 days, you will find out that only 6 seeds have germinated, while the rest will start to smell. You might ask "But why did only 6 seeds germinated? I thought it would be 8?" Now let me explain.
Firstly I change our previous definition of the random variable X. X is now defined as the number of seeds that successfully germinated. The probability of 6 seeds germinating is:

$$P(X = 6) = \binom{10}{6} 0.8^6 0.2^4 = 0.088$$

And the probability of 8 seeds germinating is:

$$P(X = 8) = \binom{10}{8} 0.8^8 0.2^2 = 0.301$$

Although $0.301 > 0.088$, it does not mean the event "6 seeds germinated" will never happen (because $0.088 > 0$); but $1 > 0.301$, so the event "8 seeds germinated" will not always happen. This is the paradox of ProbStat, we evaluate everything through the question **"How likely it will occur?"**, and our certainty is never absolute.
The pdf of our experiment is:

$$f(x) = P(X = x) = \binom{10}{x} 0.8^x 0.2^{10-x} \quad (x \in \{0, 1, 2 \cdots, 10\})$$

Or in table form:

x	0	1	2	3	4
$P(X = x)$	10^{-7}	4.10^{-6}	$7.3.10^{-5}$	$7.8.10^{-4}$	$5.5.10^{-3}$

x	5	6	7	8	9	10
$P(X = x)$	0.026	0.088	0.201	0.301	0.268	0.107

From the data table above, you can easily see that even if $P(X = 8)$ reaches its maximum value, the probability of occurrence is only about 30%. Testing the germination ability of 10 seeds experiment is a classic example of the **Bernoulli process**, which consists of **repeated Bernoulli trials**.

Definition 5.1.2. *The **Bernoulli process** is the process that consists **repeated Bernoulli trials**.*

Theorem 5.1.2. *A single Bernoulli trial can result is a sucess with probability p and a failure with probability $q = 1 - p$. Then the pdf of **binomial random variable** X, the **number of successes** in n independent trials is:*

$$\mathscr{B}(x; n, p) = \binom{n}{x} p^x q^{n-x} \quad (x = 0, 1, 2, \cdots, n)$$

Corollary 5.1.2.1. *If $X \sim \mathscr{B}(x; n, p)$, then:*

$$\mu_X = np$$
$$\sigma_X^2 = npq$$

Proof. For the mean:

$$\begin{aligned}
\mu_X &= \sum_x x f(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x} = \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x q^{n-x} \\
&= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x} = np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} q^{n-x} \\
&= np(p+q)^{n-1} = np
\end{aligned}$$

Finding the variance directly is not easy, so we have to perform small trick here:

$$\begin{aligned}
E(X(X-1)) &= \sum_x x(x-1) f(x) = \sum_{x=1}^{n} x(x-1) \binom{n}{x} p^x q^{n-x} \\
&= p^2 n(n-1) \sum_{x=2}^{n} \frac{(n-2)!}{(n-x)!(x-2)!} p^{x-2} q^{n-x} \\
&= p^2 n(n-1) \sum_{x=2}^{n} \binom{n-2}{x-2} p^{x-2} q^{n-x} \\
&= p^2 n(n-1)(p+q)^{n-2} = p^2 n(n-1)
\end{aligned}$$

Now we apply the result above to the defintion of variance:

$$\sigma_X^2 = E(X^2) - \mu_X^2 = E(X(X-1)) + E(X) - \mu_X^2 = p^2 n(n-1) + np - n^2 p^2 = npq$$

□

Proof. There is another subtle way to prove **Corollary 5.1.2.1**, if $X \sim \mathscr{B}(x; n, p)$ then:

$$X = I_1 + I_2 + \cdots + I_n$$

where each $I_i \sim \mathscr{I}(x; 1, p)$. Since they are **independent random variables**, we can apply some useful results:

$$\mu_X = E(I_1) + E(I_2) + \cdots + E(I_n) = np$$
$$\sigma_X^2 = \sigma_{I_1}^2 + \sigma_{I_2}^2 + \cdots + \sigma_{I_n}^2 = npq$$

□

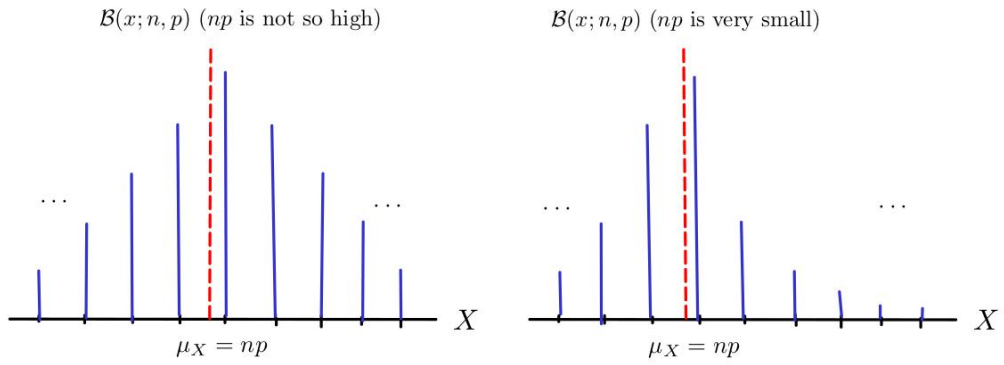


Figure 5.4: The pdf of Binomial distribution

You should note that the position of the **mean value** is very close to the value of the random variable X at which $P(X = x)$ reaches its **highest value**. Or in other words, μ_X is very close to the **peak** of the pdf graph.

5.1.3 Poisson Distribution

Using Bell-shaped Curve to approx the Binomial Distribution

Let's begin with a specific example: $X \sim \mathscr{B}(x; 100, 0.6)$. Now we can see that the mean value $\mu_X = np = 60$ is relatively close to the center $X = 50$.

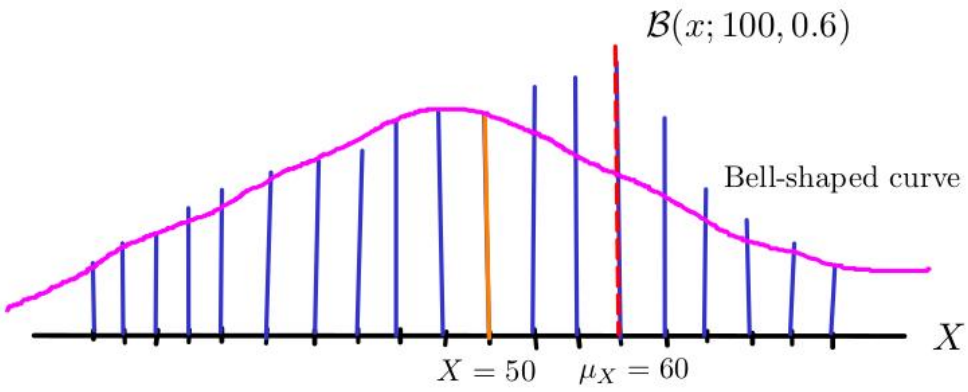


Figure 5.5: $\mathscr{B}(x; 100, 0.6)$ graph

Because $np = 60$ value is medium, neither too high nor low compared to the center value $X = 50$; and plotting $n = 100$ points is high enough for us to connect all of them to obtain a relatively smooth curve. This smooth curve **can be approximated** by the **bell-shaped curve**, so we will perform operations on it instead of the original curve.
In this subsection, we will focus on mathematical concepts rather than delving into calculation methods (you can try it yourself if you want). We will return to specific calculation methods in the next chapter.

Theorem 5.1.3. *If $X \sim \mathscr{B}(x; n, p)$ with np value is medium compared to the center value, then:*

$$\textit{Original Curve} \approx \textbf{\textit{Bell-shaped Curve:}}\ \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where: $\mu = np$ *and* $\sigma^2 = npq$

Now let's compare 2 methods: caculating directly and using approximation.

$$P(X \leq 70) = 1 - P(X \geq 71) = 1 - \sum_{x=71}^{100} \binom{100}{x} 0.6^x 0.4^{100-x} = 0.985$$

$$P(X \leq 70) = P\left(Z < \frac{70 - 100.0.6 + 0.5}{\sqrt{100.0.6.0.4}}\right) = P(Z < 2.143) = 0.983$$

Or we can also try:

$$\begin{aligned} P(50 \leq X \leq 65) &= P(X \leq 65) - P(X \leq 49) \\ &= P\left(Z < \frac{65 - 100.0.6 + 0.5}{\sqrt{100.0.6.0.4}}\right) - P\left(Z < \frac{49 - 100.0.6 + 0.5}{\sqrt{100.0.6.0.4}}\right) \\ &= P(Z < 1.122) - P(Z < -2.143) \\ &= P(Z < 1.122) - 1 + P(Z < 2.143) \\ &= 0.868 - 1 + 0.983 = 0.851 \end{aligned}$$

$$P(50 \leq X \leq 65) = \sum_{x=50}^{65} \binom{100}{x} 0.6^x 0.4^{100-x} = 0.852$$

The approximation method works very well and yields resutls very close to calculating directly. But how about the $X \sim \mathscr{B}(x; 100, 0.05)$ case? Can we still use the **bell-shaped curve**? Let's check it!

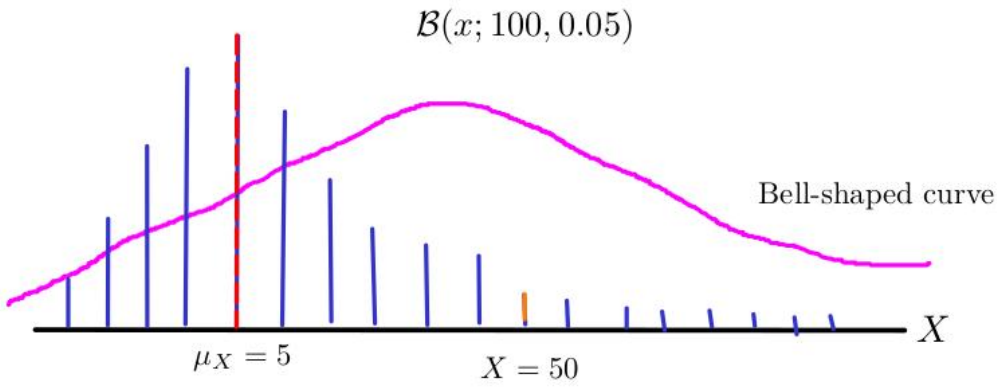


Figure 5.6: $\mathscr{B}(x; 100, 0.05)$ graph

For example:

$$P(X \leq 7) = \sum_{x=0}^{7} \binom{100}{x} 0.05^x 0.95^{100-x} = 0.872$$

$$P(X \leq 7) = P\left(Z < \frac{7 - 100.0.05 + 0.5}{\sqrt{100.0.05.0.95}}\right) = 0.9292$$

The approximate result differs from the actual result nearly 6%. Therefore, for cases where $\mu_X = np$ is located very far from the center, the bell-shaped curve approximation method is no longer suitable.

Poisson Distribution

How do we handle extreme value cases of np? We must find better way to approximate original curve. Because $n >> p$, so we can perform some equivalent transformations with assumptions $n \to +\infty$ and $p \to 0$. We begin with $X \sim \mathscr{B}(x; n, p)$:

$$\begin{aligned} P(X = x) = \mathscr{B}(x; n, p) &= \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!x!} p^x q^{n-x} \\ &= \frac{n(n-1)(n-2)\cdots(n-x+1)}{n^x} \frac{(np)^x}{x!} (1-p)^{n-x} \end{aligned}$$

Because $n \to +\infty$ and $p \to 0$:

$$\lim_{n \to +\infty} \frac{n(n-1)(n-2)\cdots(n-x+1)}{n^x} = 1$$

Our approximation curve will be more accurate if we only care about **the first tail** values, it means $n >> x$:

$$\lim_{n \to +\infty} (1-p)^{n-x} = (1-p)^n$$

By the definition of the constant e:

$$\lim_{p \to 0} (1-p)^{\frac{-1}{p}} = e \Rightarrow \lim_{p \to 0} (1-p) = e^{-p} \Rightarrow (1-p) \approx e^{-p}$$

Now we obtain:

$$(1-p)^n \approx e^{-np}$$

Finally we have:

$$P(X = x) = \frac{(np)^x}{x!} e^{-np}$$

Or in shortened form with $\mu = np$:

$$P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$$

Test the new function with the case $X \sim \mathscr{B}(x; 100, 0.05)$:

$$P(X \leq 7) = \sum_{x=0}^{7} \frac{(100.0.05)^x e^{-100.0.05}}{x!} = 0.866$$

Obviously, the approximated result is very close to the actual result 0.872.
How about the case $X \sim \mathscr{B}(x; 100, 0.9)$? Since p is clearly close to 1, not zero, so we must **inverse the role** of p and q. We define the new random variable Y as the **number of failures** to satisfy the condition of valid approximation.

$$P(X \geq 90) = \sum_{90}^{100} \binom{100}{x} 0.9^x 0.1^{100-x} = 0.583$$

$$P(X \geq 90) = P(Y \leq 10) = \sum_{y=0}^{10} \frac{(100.0.1)^x e^{-100.0.1}}{y!} = 0.583$$

So relatively speaking, we conclude the following, if $\mu = np$ position is very **far** from the **left side** of the center, then the Binomial distribution can be approximated by the formula:

$$P(X = x) = \mathscr{B}(x; n, p) \approx \frac{\mu^x e^{-\mu}}{x!}$$

Because of two main constraints $n >> p$ and p is very close to 0, the probability of a success event is relatively **low**. You can think about p as the **average number** of outcomes per unit time, distance or volume; and n as a given **time interval** or **specified region**. In practice, to avoid being confused with Binomial distribution and emphasize the fact that $p \approx 0$, we usually change our notations to:

$$p \leftrightarrow \lambda$$

$$n \leftrightarrow t$$

Definition 5.1.3. *The **Poisson process** is the **Bernoulli process** where $n >> p$ and $p \approx 0$.*

Definition 5.1.4. *A **Poisson event** is an event with a **low** probability of occurring.*

Theorem 5.1.4. *The pdf of the **Poisson random variable** X, representing the number of outcomes occurring in a **given time interval** or **specified region** denoted by t, is:*

$$\mathscr{P}(x;\mu) = \frac{\mu^x e^{-\mu}}{x!} \quad (x = 0, 1, 2, \cdots)$$

*where $\mu = \lambda t$, λ is the **average** number of outcomes per unit time, distance area or volume.*

Theorem 5.1.5. *If $X \sim \mathscr{B}(x; n, p)$, when $n \to +\infty$, $p \to 0$ and $np \to \mu$ remains constant:*

$$\mathscr{B}(x; n, p) \to \mathscr{P}(x; \mu)$$

Corollary 5.1.5.1. *If $X \sim \mathscr{P}(x;\mu)$, then:*

$$\mu_X = \mu$$

$$\sigma_X^2 = \mu$$

Proof. For the mean:

$$\mu_X = \sum_x x f(x) = \sum_{x=0}^{n} x \frac{\mu^x e^{-\mu}}{x!} = \sum_{x=1}^{n} \frac{\mu^x e^{-\mu}}{(x-1)!} = \mu e^{-\mu} \sum_{x=1}^{n} \frac{\mu^{x-1}}{(x-1)!} = \mu e^{-\mu} e^{\mu} = \mu$$

For the variance, perform the same trick:

$$\begin{aligned} E(X(X-1)) &= \sum_x x(x-1) f(x) = \sum_{x=1}^{n} x(x-1) \frac{\mu^x e^{-\mu}}{x!} \\ &= \mu^2 e^{-\mu} \sum_{x=2}^{n} \frac{\mu^{x-2}}{(x-2)!} = \mu^2 e^{-\mu} e^{\mu} \\ &= \mu^2 \end{aligned}$$

Apply directly our previous result:

$$\sigma_X^2 = E(X^2) - \mu_X^2 = E(X(X-1)) + E(X) - \mu_X^2 = \mu^2 + \mu - \mu^2 = \mu$$

□

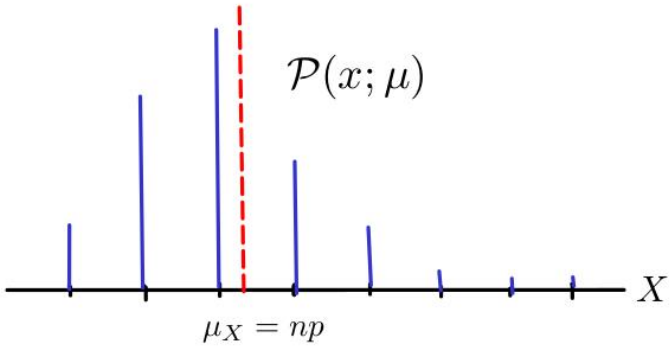


Figure 5.7: The pdf of Poisson distribution

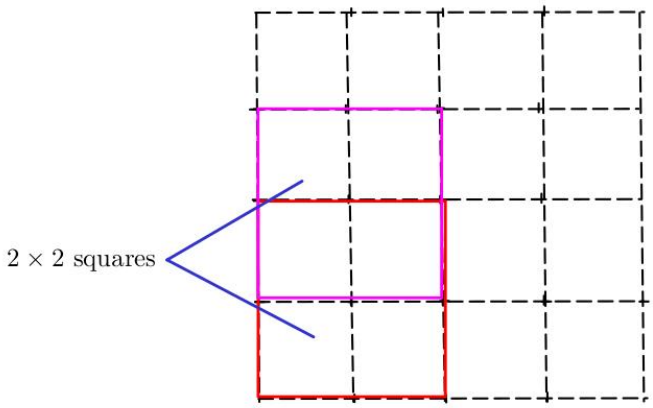


Figure 5.8: A 4×4 grid

Back to our mung bean seeds, now you will see how Poisson distribution can be applied in many situations in real life. If you take some seeds from the packet and throw them on a 4×4 grid and you observe that there are 10 seeds in the 2×2 square closet to you, what are the probabilities of:

1. There are a total of 40 seeds on that 4×4 grid.
2. Suppose you knew the exact number of seeds that you had thrown is 40. Evaluate the chance of observing 10 seeds in the 2×2 square.

The random variable X is defined as the number of seeds on a 4×4 grid. Using the formula above, we have:

$$\mu = \lambda t = \frac{10}{4}.16 = 40$$

$$\Rightarrow P(X = 40) = \frac{e^{-\mu}\mu^x}{x!} = \frac{e^{-40}40^{40}}{40!} = 0.0629$$

The random variable Y is defined as the number of seeds in the 2×2 square. Similarly, we have:

$$\mu = \lambda t = \frac{40}{16}.4 = 10$$

$$\Rightarrow P(Y = 10) = \frac{e^{-10}10^{10}}{10!} = 0.125$$

So the chance of observing 10 seeds in the 2×2 square if you throw 40 seeds on a 4×4 grid is pretty low, just 12.5%.

5.2 Negative Binomial and Geometric Distributions

5.2.1 Negative Binomial Distribution

What is the probability of tossing a coin x times **until** getting k heads, if we know that the probability of heads appearing is p and X is a random variable, defined as the number of tosses? This question is very easy so I will write down the answer:

$$P(X = x) = \binom{x-1}{k-1} p^k q^{x-k}$$

The function above is a **Negative Binomial Distribution** formula. Formally, we state that:

Theorem 5.2.1. *If repeated **Bernoulli trials** can result in a success with probability p and a failure with probability $q = 1 - p$, then the pdf of random variable X, defined as the **number of the trial** on which the **kth success occurs** is:*

$$\mathscr{B}^*(x; k, p) = \binom{x-1}{k-1} p^k q^{x-k} \quad (x = k, k+1, k+2, \cdots)$$

5.2.2 Geometric Distribution

If we only care about the probability of tossing a coin x times until getting **first** heads, it means $k = 1$ now and we can rewrite our formula as:

$$P(X = x) = pq^{x-1}$$

The $k = 1$ case of **Negative Binomial Distribution** can also be called as **Geometric Distribution**.

Theorem 5.2.2. *If repeated **Bernoulli trials** can result in a success with probability p and a failure with probability $q = 1 - p$, then the pdf of random variable X, defined as the **number of trials** on which the **first success occurs** is:*

$$\mathscr{G}^*(x; p) = pq^{x-1} \quad (x = 1, 2, 3, \cdots)$$

Corollary 5.2.2.1. *If $X \sim \mathscr{G}^*(x; p)$, then:*

$$\mu_X = \frac{1}{p}$$

$$\sigma_X^2 = \frac{q}{p^2}$$

Proof. For the mean:

$$\mu_X = \sum_x x f(x) = p \sum_{x=1}^{+\infty} x q^{x-1} = p \sum_{x=1}^{+\infty} x(1-p)^{x-1}$$

Now we use the Geometric series and take the derivative once with respect to p:

$$\sum_{x=0}^{+\infty} (1-p)^x = \frac{1}{1-(1-p)} = \frac{1}{p}$$

$$\Leftrightarrow \left(\sum_{x=0}^{+\infty} (1-p)^x \right)' = \left(\frac{1}{p} \right)'$$

$$\Leftrightarrow -\sum_{x=0}^{+\infty} x(1-p)^{x-1} = \frac{-1}{p^2} \Rightarrow \sum_{x=1}^{+\infty} x(1-p)^{x-1} = \frac{1}{p^2}$$

After substituting our result to the previous equation, we obtain:

$$\mu_X = p\sum_{x=1}^{+\infty} x(1-p)^{x-1} = p.\frac{1}{p^2} = \frac{1}{p}$$

Take the second derivative with respect to p from the previous equation:

$$\left(\sum_{x=1}^{+\infty} x(1-p)^{x-1}\right)' = \left(\frac{1}{p^2}\right)' \Leftrightarrow \sum_{x=2}^{+\infty} x(x-1)(1-p)^{x-2} = \frac{2}{p^3}$$

We want to determine:

$$E(X(X-1)) = \sum_x x(x-1)f(x) = p\sum_{x=1}^{+\infty} x(x-1)(1-p)^{x-1} = \frac{2p(1-p)}{p^3} = \frac{2q}{p^2}$$

For the variance:

$$\sigma_X^2 = E(X^2) - \mu_X^2 = E(X(X-1)) + E(X) - \frac{1}{p^2} = \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2} = \frac{q}{p^2}$$

□