STA 131A: Introduction to Probability

Xiaodong Li

Statistics Department

University of California, Davis

September, 2017

Chapter 1

Overview

What is probability, what is statistics, and what is their relationship?

Statistics:

- Fitting a statistical model to the data;
- Finding patterns in the data;
- Making decisions based on the data;
- Prediction based on the data;
- Finding the relationship between variable based on data;
- Quantify the uncertainty based on the data;
- Visualization of data;
- Remove noise from data;
- Finding the principal components in the data;
- Design experiments;
- Design surveys;
- Etc.

In general, statistics is a science regarding data, so sometimes we call statistics "data science" (Jeff Wu).

Probability:

- Understanding concepts of "uncertainty";
- Discovery and evaluation of "certainty" among uncertainty.

Therefore, I usually call probability "uncertainty science".

STA 131A is an introductory course for probability.

- It is not a course of statistics, but very fundamental and useful for statistics;
- It is not a course regarding data or data analysis.

Relation to other probability courses provided by the statistics department at Davis

- STA 130A: Basic probability concepts/results and estimation theory;
- STA 200A: More serious in the mathematics of probability (graduate course);
- STA 235A: Probability theory based on measure theory (Ph.D course);

Logistics: See the tentative syllabus.

1.1 Concepts

- What is the probability that it will rain next Monday?
- What is the expected temperature next Monday?
- How to quantify the uncertainty of the temperature next Monday?
- How to quantify the correlation between temperature next Monday and the temperature next Tuesday?

We will introduce or explain some concepts in clear and mathematical ways, such that we can have systematical approaches to find the certainties in uncertainty. Typical quantities include probability, expectation, variance, covariance, correlation, etc.

In order to explain these concepts clearly, we need to introduce some basic concepts in probability:

- Events and sample space: how to define the events "It will rain tomorrow" and "A will win B in the game tomorrow" mathematically?
- Random variable: how to define "the temperature next Monday" mathematically?
- Distribution, joint distribution.
- The mathematical definitions of probability, expectation, variance, covariance, correlation.

In real life, we also hear some other probabilistic concepts, such as independent events and conditional independent events.

- Suppose team A and team B play three games in July, August and September, and it turns out A win all the games. One said: "They are independent events, but the result is not a coincidence." How to interpret it?
- Suppose team A and team B will play two games tomorrow and the day after tomorrow. There is an important player K in team A, and there is a rumor that he may get injured. Then can we say that the results of these two games will be dependent? What is your intuition?
- One said:"If we know K gets injured, then the two events are independent; if we know K does not get injured, then the two events are independent. But now we don't know whether K gets injured or not, so these two events are dependent." Could this statement be reasonable?

More complicatedly, we have concepts like independent random variables and conditionally independent random variables:

- Consider some stock prices in Day 1, 2, 3. Are the stock price in Day 1 and that in Day 3 dependent?
- Is it possible that conditional on the stock price of Day 2, the stock prices of Day 1 and Day 3 are independent?

Other complicated concepts include conditional distribution, conditional probability, conditional expectation, conditional variance.

- What is the probability that it will rain next Monday given it rains this Sunday?
- What is the probability that it will rain next Monday given the temperature this Sunday?
- What is the expected temperature next Monday given it rains this Sunday?
- What is the expected temperature next Monday given the temperature this Sunday?
- How to quantify the uncertainty of the temperature next Monday given it rains this Sunday?
- How to quantify the uncertainty of the temperature next Monday given the temperature this Sunday?

1.2 Computing and calculation

After understanding the relevant concepts, the next question becomes how to compute/calculate the relevant quantities: probability/distribution/expectation/variance/covariance/correlation, conditional probability/distribution/expectation/variance/covariance/correlation.

We need the following tools to compute:

- Counting;
- Bayes's Formula;
- Independence;
- Mass function of discrete random variables;
- Pdf of continuous random variables, joint density;
- Definitions of expectation, variance;
- Examples of discrete and continuous random variables;
- Properties of expectation, variance, covariance;
- Properties of conditional expectation, variance, covariance;
- Moment generating functions;
- Law of large numbers and central limit theorem.

1.3 Mathematics

- Counting and combinatorics;
- Set theory, Axioms of probability, events, sample space, probability;
- Random variables, distribution, density, expectation, variance/covariance, independence, conditioning;
- Familiarity with calculus, particularly integrals, double integrals, integration by parts, etc;
- Familiarity with the important examples of random variables;
- Correct understanding and application of key properties and formula;
- Rigorous mathematical proofs are required, including proof by induction/contradiction.

Chapter 2

Counting methods

Evaluating probabilities by counting. We can an event A. If there are totally N possible outcomes, and all outcomes share the same likelihood, then

$$p(A) = \frac{\text{number of ways } A \text{ can occur}}{\text{total number of outcomes}}.$$

The number of ways A can occur is the same as the number of possible outcomes in the event A.

2.1 Basics

One-to-one correspondence If A is a set consisting of finite elements there is a one-to-one correspondence between the elements in A and elements in B. Then we know that B is finite and A and B have the same number of elements.

Example: How many games in total in a single-elimination tournament with 16 teams?
Answer: Each game in the tournament corresponds to an eliminated team, and the correspondence is one-to-one. Therefore, there are 15 games.

The Multiplication Principle: if experiment A has m outcomes and B has n outcomes, then the sequence of experiments (A, B) has mn outcomes.

Extended multiplication principle: There are p experiments. If n_j outcomes for the j-th experiments, then there are $n_1 \times n_2 \times \ldots \times n_p$ outcomes for the sequence of p experiments.

Example 1: How many different 8-bit words are there? $2 \times 2 \times \ldots \times 2 = 256$.

Example 2 [Sampling with ordering and replacement]: Given a set $C = \{c_1, c_2, \ldots, c_n\}$, if we choose r elements with replacement and put them in order, how many ways do we have?
Answer: n^r.

Permutations:
Motivating question 1: *How many possible outcomes of the champion and runner-up among* 8 *people?*

Motivating question 2: *How to count the number of possible outcomes for* ***sampling with ordering but without replacement****?*

A permutation is an ordered arrangement of objects. It is well-known that the number of orderings

for n elements is

$$n! = n(n-1)(n-2)\dots 1.$$

Now let's extend this simple ordering to a more general setup. Given a set $C = \{c_1, c_2, \dots, c_n\}$, if we choose r elements and put them in order, there are $n(n-1)\dots(n-r+1)$ orderings for sampling without replacement.

Example: Suppose that a room contains n people. What's the probability that at least two of them have a common birthday? Denote by A the event that at least two people have a common birthday. We want to evaluate $\mathbb{P}(A)$. However, it is not easy to count the element in A. We should try whether it is easier to count the elements in A^c. Now we just assume there are 365 days per year. Each outcome is the sequence of birthdays of the n people: (Birthday of person 1, Birthday of person 2, ... Birthday of person n). There are 365^n elements in the sample space, while $365 \times \dots \times (365-n+1)$ elements in A^c, so

$$\mathbb{P}(A) = 1 - P(A^c) = 1 - \frac{365 \times 364 \times \dots \times (365-n+1)}{365^n}.$$

When $n = 56$, $P(A) = 98.8\%$.

Combinations:

How to count the number of possible outcomes for sampling without replacement and disregarding order?

Choosing r objects from $\{C_1, \dots, C_n\}$ without replacement and disregarding order, how many ways? Answer:

$$\frac{n(n_1)\dots(n-r+1)}{r!} = \frac{n!}{(n-r)!r!} = \binom{n}{r} = \binom{n}{n-r}.$$

The answer is called binomial coefficient.

Example 1: How many ways to formulate a committee with 3 people from a group of 10 people?
Answer: $\binom{10}{3} = \frac{10\times 9\times 8}{3\times 2\times 1} = 120$. It is surprising to me that the uncertainty for this problem is so huge.

Example 2: Suppose there are 6 items in a lot that contains 3 defective items. Choose a random sample of size 3, what is the probability that the sample contains exactly 1 defective?

Answer 1: Let's first answer this problem by enumeration. We denote these 6 items as A, B, C, D, E, F, among which A, B, C are defectives while D, E, F are not defectives. If we choose random sample of size 3 (without replacement and disregarding the order), the possible combinations include

(A, B, C), (A, B, D), (A, B, E), (A, B, F), (A, C, D), (A, C, E), (A, C, F), (A, D, E), (A, D, F), (A, E, F)
(B, C, D), (B, C, E), (B, C, F), (B, D, E), (B, D, F), (B, E, F), (C, D, E), (C, D, F), (C, E, F), (D, E, F).

We can clearly see that among all the combinations, there is exactly one defective in each of the following:

(A, D, E), (A, D, F), (A, E, F), (B, D, E), (B, D, F), (B, E, F), (C, D, E), (C, D, F), (C, E, F).

Therefore, the probability is 9/20.

Answer 2: Based on the previous example, we can see that each outcome in the event consists one sample from the detectives and two from the non-defectives. So the number of outcomes in the event is $\binom{3}{1}\binom{3}{2} = 9$, while the number of all possible outcomes is $\binom{6}{3} = 20$. Then the probability of the event is $9/20 = 0.45$.

Example 3: Suppose there are n items in a lot that contains k defective items. Choose a random

sample of size r, what is the probability that the sample contains exactly m defectives?

Answer: Call the event in question A. How many elements are there in A? First, choose m defective items from k items: $\binom{k}{m}$; Second, choose $r-m$ non-defective items from k items: $\binom{n-k}{r-m}$. By multiplication law: there are $\binom{k}{m}\binom{n-k}{r-m}$ elements in A. How many elements in the sample space? Answer: $\binom{n}{r}$. Therefore,

$$\mathbb{P}(A) = \frac{\binom{k}{m}\binom{n-k}{r-m}}{\binom{n}{r}}.$$

Example 4: Suppose there are 10 items in a lot that contains 3 defective items. Choose a random sample of size 2, what is the probability that the sample contains at least one defective item?

Answer 1: The number of outcomes with exactly 1 defective item in the random sample is

$$\binom{3}{1}\binom{10-3}{1} = 21.$$

The number of outcomes with exactly 2 defective items in the random sample is

$$\binom{3}{2}\binom{7}{0} = 3.$$

The number of all outcomes is

$$\binom{10}{2} = \frac{10}{2\times 1} = 45.$$

Therefore, the probability that the sample contains at least one defective item is

$$(21+3)/45 = 8/15.$$

Answer 2: The probability that the sample contains no defective item is

$$\frac{\binom{3}{0}\binom{7}{2}}{\binom{10}{2}} = \frac{21}{45} = \frac{7}{15}.$$

Therefore, the probability that the sample contains at least one defective item is $1-\frac{7}{15} = \frac{8}{15}$.

Answer 3 (The rigorous definition of conditioning will be given later) To obtain this random sample with size 2, we choose the first one randomly from the 10 items, and then choose one randomly from the rest 9 items. The probability that the first one is defective is $3/10$. The probability that the first one is not defective but the second one is defective is $7/10\times 3/9 = \frac{7}{30}$. Then the probability that at least one item is defective is

$$\frac{3}{10} + \frac{7}{30} = \frac{8}{15}.$$

2.1.1 Theory of binomial coefficients

Motivating questions

How to evaluate the binomial coefficients $\binom{n}{r}$?
What properties do the binomial coefficients enjoy?

Some observations of binomial coefficients:

1	$1;$
$\binom{1}{0}, \binom{1}{1}$	$1,\ 1;$
$\binom{2}{0}, \binom{2}{1}, \binom{2}{2}$	$1,\ 2,\ 1;$
$\binom{3}{0}, \binom{3}{1}, \binom{3}{2}, \binom{3}{3}$	$1,\ 3,\ 3,\ 1;$
$\binom{4}{0}, \binom{4}{1}, \binom{4}{2}, \binom{4}{3}, \binom{4}{4}$	$1,\ 4,\ 6,\ 4,\ 1;$

Two observations:

- Pascal's triangle;
- Coefficients for the binomial $(x+y)^n$.

Question: Are these observations always true? Can we prove them?

Properties of binomial coefficients:

$$\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r}, \quad 1 \le r \le n.$$

Proof: Algebraic proof; Combinatorial proof.

The binomial theorem

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

Proof by Induction. When $n = 1$,

$$x + y = \binom{1}{0} x^0 y^1 + \binom{1}{1} x^1 y^0 = y + x.$$

Assume the equation is true of n-1. Now

$$\begin{aligned}
(x+y)^n &= (x+y)(x+y)^{n-1} \\
&= (x+y)\sum_{k=0}^{n-1}\binom{n-1}{k}x^k y^{n-1-k} \\
&= \sum_{k=0}^{n-1}\binom{n-1}{k}x^{k+1}y^{n-1-k} + \sum_{k=0}^{n-1}\binom{n-1}{k}x^k y^{n-k} \\
&= \sum_{k=1}^{n}\binom{n-1}{k-1}x^k y^{n-k} + \sum_{k=0}^{n-1}\binom{n-1}{k}x^k y^{n-k} \\
&= x^n + \sum_{k=1}^{n-1}\binom{n-1}{k-1}x^k y^{n-k} + \sum_{k=1}^{n-1}\binom{n-1}{k}x^k y^{n-k} + y^n \\
&= x^n + \sum_{k=1}^{n-1}\left(\binom{n-1}{k-1} + \binom{n-1}{k}\right)x^k y^{n-k} + y^n \\
&= x^n + \sum_{k=1}^{n-1}\binom{n}{k}x^k y^{n-k} + y^n \\
&= \sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}.
\end{aligned}$$

2.2 The number of integer solutions of equations

How many positive integer solutions to the equation

$$x_1 + x_2 + x_3 = 10?$$

Furthermore, how many non-negative integer solutions to the equation

$$x_1 + x_2 + x_3 = 10?$$

In general, how many positive integer solutions and how many non-negative integer solutions to the equation

$$x_1 + x_2 + \ldots + x_r = n?$$

Answer 1 : There are $\binom{n-1}{r-1}$ positive integer solutions of $(x_1, \ldots, x_r)$ satisfying

$$x_1 + x_2 + \ldots + x_r = n.$$

Idea: Choose $r-1$ of the spaces $\llcorner$:

$$\star \llcorner \star \llcorner \star \llcorner \ldots \llcorner \star \llcorner \star$$

Answer 2 : There are $\binom{n+r-1}{r-1}$ non-negative integer solutions of $(x_1, \ldots, x_r)$ satisfying

$$x_1 + x_2 + \ldots + x_r = n.$$

Idea: Each non-negative integer solution $(x_1, \ldots, x_r)$ to $x_1 + x_2 + \ldots + x_r = n$ corresponds to a positive integer solution $(y_1, \ldots, y_r)$ to $y_1 + y_2 + \ldots + y_r = n + r$ by $y_i = x_i + 1$, $i = 1, \ldots, r$. One can see that the correspondence is one-to-one.

Example: Consider a set of n antennas of which m are defective and $n-m$ are functional and assume that all of the defectives and all of the functionals are considered indistinguishable. How many linear orderings are there in which no two defectives are consecutive?

Answer:

$$x_1 \square x_2 \square x_3 \square \ldots \square x_m \square x_{m+1},$$

where $x_1, \ldots, x_m$ represent the number of functional items in each space. Then we have the following equation:

$$x_1 + \ldots + x_{m+1} = n - m, \quad x_1 \geq 0, \quad x_{m+1} \geq 0, \quad x_i > 0, i = 2, \ldots, m.$$

If we write $y_1 = x_1 + 1$, $y_{m+1} = x_{m+1} + 1$, and $y_i = x_i$ for $i = 2, \ldots, m$, the problem becomes to find the number off positive integer solutions of the equation

$$y_1 + y_2 + \ldots + y_{m+1} = n - m + 2.$$

There there are

$$\binom{n-m+1}{m}$$

outcomes.

2.3 Multinomial coefficients

Distribute n objects into Class 1 to Class r, such that there are n_i objects in the i-th group (so $n = n_1 + \ldots + n_r$). How many ways of assignments?

Answer:

$$\binom{n}{n_1 n_2 \ldots, n_r} = \frac{n!}{n_1! n_2! \ldots n_r!}$$

Example In the first round of a knockout tournament involving $n = 2^m$ players, the n players are divided into $\frac{n}{2}$ pairs, with each of these pairs then playing a game. The losers of the games are eliminated while the winners go on to the next round, where the process is repeated until only a single player remains. Suppose we have a knockout tournament of 8 players.

(a) How many possible outcomes are there for the initial round?

(b) How many outcomes of the tournament are possible, where an outcome gives complete information for all rounds?

Solution

(a) We first pair the 8 players together. This job consists of two steps

- We first pair the 8 players into 4 "labeled" pairs, e.g., the first, second, third, fourth pair. The answer is $\binom{8}{2,2,2,2} = \frac{8!}{2^4}$.
- The number of possible pairings without ordering is $\frac{8!}{2^4 4!}$.

Next, each pair generates 2 possible outcomes, so the total number of possible outcomes is $\frac{8!}{2^4 4!} 2^4 = \frac{8!}{4!}$

(b) There are $\frac{8!}{4!}$ different outcomes in the first round, $\frac{4!}{2!}$ outcomes in the second round, and 2 outcomes in the final round. So totally

$$\frac{8!}{4!}\frac{4!}{2!}2 = 8!$$

totally different outcomes. The result can be interpreted in the following way: There is a one-to-one correspondence between each outcome of the tournament and a ranking as follows:

- The final winner is ranks the first;
- The runner-up ranks the second;
- The one coming in the second round losing to the winner is ranks the 3rd;
- The one coming in the second round losing to the runner-up is ranks the 4th;
- The one stopped in the first round losing to the first ranks the 5th;
- The one stopped in the first round losing to the second ranks the 6th;
- The one stopped in the first round losing to the 3th ranks the 7th;
- The one stopped in the first round losing to the 4th ranks the 8th.

Why is this correspondence one-to-one? Give the outcome based on the rank E, A, G, B, C, F, D, H.

Chapter 3

Axioms of Probability

The probability of an event is the measure of the chance that the event will occur as a result of an experiment.

3.1 Sample spaces and events

3.1.1 Sample spaces

The set of all possible outcomes corresponding experiments.

Example: Suppose there are 5 items a, b, c, d and e in a lot. Choose a random sample of size 3, what is the sample space? The possible combinations of size 3 include

{{a, b, c}, {a, b, d}, {a, b, e}, {a, c, d}, {a, c, e}, {a, d, e}, {b, c, d}, {b, c, e}, {b, d, e}, {c, d, e}}.

Example: Number of people in a queue: $\Omega = \{0, 1, 2, 3, \ldots\}$. If an upper limit, N is known: $\Omega = \{1, \ldots, N\}$.

Example: The length of time between successive earthquakes: $\Omega = \{t | t \geq 0\}$.

3.1.2 Events

Subsets of the sample space:

Example 1: Suppose there are 5 items a, b, c, d, and e in a lot. Choose a random sample of size 3. Figure out the following events:
(a) A= The sample contains a;
(b) B= The sample contains b;
(c) C= The sample contains a or b;
(d) D= The sample contains both a and b;
(e) E= The sample does not contain a.
(f) F= contains both c and d;

Answer:
(a) A = {{a, b, c}, {a, b, d}, {a, b, e}, {a, c, d}, {a, c, e}, {a, d, e}};
(b) B = {{a, b, c}, {a, b, d}, {a, b, e}, {b, c, d}, {b, c, e}, {b, d, e}};
(c) C = {{a, b, c}, {a, b, d}, {a, b, e}, {a, c, d}, {a, c, e}, {a, d, e}, {b, c, d}, {b, c, e}, {b, d, e}};
(d) D = {{a, b, c}, {a, b, d}, {a, b, e}};

(e) E = {{b, c, d}, {b, c, e}, {b, d, e}, {c, d, e}}.
(f) F = {{a, c, d}, {b, c, d}, {c, d, e)} };

Example: The length of time between successive earthquakes is
(a) A: no less than 10 days;
(b) B: no greater than 300 days;
(c) C: no less than 10 days and no greater than 300 days.

Answer (a) $A = \{t|t \geq 10\}$;
(b) $B = \{t|t \leq 300\}$;
(c) $C = \{t|10 \leq t \leq 300\}$.

3.2 Union, Intersection, and Complement

3.2.1 Basic Definitions

Union of events: "A occurs or B occurs or both occur": $A \bigcup B$.

In Example 1, $C = A \bigcup B$.

Intersection of events: "Both A and B occur": $A \bigcap B$.

In Example 1, $D = A \bigcap B$.

Complement of an event: "A does not occur": A^c.

In Example 1, $E = A^c$.

Impossible event: $\phi = \Omega^c$.

Full probability event: The sample space Ω.

Disjoint events: $A \bigcap B = \phi$.

In Example 1, $D \bigcap F = \phi$.

One event implies another event: $A \subset B$ if and only if any occurrence of event A implies any occurrence of the event B.

In Example 1, $A \subset A$, $A \subset C$, $D \subset A$, $A \subset \Omega$, $\phi \subset A$.

3.2.2 Basic Properties

1. $A \bigcap \phi = \phi$, $A \bigcup \phi = A$;
2. If $A \subset B$, then $A \cap B = A$ and $A \cup B = B$;
3. $A \bigcap B \subset A$, $A \subset A \cup B$;
4. $A \bigcup A^c = \Omega$; $A \bigcap A^c = \phi$;

5. $A \subset \Omega$, $\phi \subset A$;

6. $A \subset A$;

7. If $A \subset B$ and $B \subset A$, then $A = B$.

3.2.3 Laws of set theory

1. **Commutative laws**: $A \bigcup B = B \bigcup A$, $A \bigcap B = B \bigcap A$;

2. **Associative laws**: $(A \bigcup B) \bigcup C = A \bigcup (B \bigcup C)$, $(A \bigcap B) \bigcap C = A \bigcap (B \bigcap C)$;

By the commutative laws and associative laws, we can define

$$\bigcap_{i=1}^{n} A_i = A_1 \bigcap A_2 \bigcap \ldots \bigcap A_n,$$

$$\bigcap_{i=1}^{\infty} A_i = A_1 \bigcap A_2 \bigcap \ldots \bigcap A_n \bigcap \ldots,$$

$$\bigcup_{i=1}^{n} A_i = A_1 \bigcup A_2 \bigcup \ldots \bigcup A_n,$$

$$\bigcup_{i=1}^{\infty} A_i = A_1 \bigcup A_2 \bigcup \ldots \bigcup A_n \bigcup \ldots,$$

Example: Consider the number of people in the queue. We have $\Omega = \{0, 1, 2, 3, \ldots, n, \ldots, \}$. The event the number of people is greater than or equal to k is $A_k = \{k, k+1, k+2, \ldots\}$. Then

$$\bigcup_{i=1}^{\infty} A_i = A_0 = \Omega,$$

and

$$\bigcap_{i=1}^{\infty} A_i = \phi.$$

3. **Distributive laws**: $(A \bigcup B) \bigcap C = (A \bigcap C) \bigcup (B \bigcap C)$, $(A \bigcap B) \bigcup C = (A \bigcup C) \bigcap (B \bigcup C)$; Venn Diagram

In Example 1, by careful comparison, we have

$$(A \bigcup B) \bigcap F = C \bigcap F = \{\{a, c, d\}, \{b, c, d\}\}.$$

On the other hand, we know $A \bigcup F = \{a, c, d\}$ and $B \bigcup F = \{b, c, d\}$. Then by distributive law, we have

$$(A \bigcup B) \bigcap F = (A \bigcap F) \bigcup (B \bigcap F) = \{\{a, c, d\}, \{b, c, d\}\}.$$

A useful corollary: For any events A and B

$$B = B \bigcap \Omega = B \bigcap (A \bigcup A^c) = (B \bigcap A) \bigcup (B \bigcap A^c).$$

If $A \subset B$, we have $B \bigcap A = A$, so $B = A \bigcup (B \bigcap A^c)$.

In Example 1, $D \subset A$, so $A = D \bigcup (A \bigcap D^c)$. Here $A \bigcap D^c$ means A occurs while D does not occur, i.e., the sample contains a but does not contain b.

3.3 Axioms of Probability

Question: Why do we need to introduce axioms, instead of just admitting a bunch of facts?

My answer: To understand the connections between facts, and to derive new facts by correct logic, it is important to know the root/fundamentality. This is like first principle driven project: We need to always relate our ongoing job/thinking to the motivation.

Some common senses:

- For any event, the probability should be at least 0;
- If the event A implies the event B, then the probability of the occurrence of A should be less than or equal to that of B;
- If we know the probability of A and know the probability of B, moreover, let's assume we know the probability of the joint occurrence of both A and B. Do we have enough information to the probability either A occurs or B occurs?
- ...

Definition 3.3.1 *A probability measure $P : \{\text{subsets of } \Omega\} \longrightarrow \mathbb{R}$ satisfies the following axioms:*

1. *$P(\Omega) = 1$;*
2. *For any $A \subset \Omega$, $P(A) \geq 0$;*
3. *If $A_1 \bigcap A_2 = \phi$, then $P(A_1 \bigcup A_2) = P(A_1) + P(A_2)$.*

Property 3 is equivalent to if $A_1, \ldots, A_n, \ldots$ are mutually disjoint, then

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

Basic (and intuitive) properties:

1. $p(A^c) = 1 - p(A)$;
   Proof: Since $A^c \cap A = \phi$ and $A^c \cup A = \Omega$, by Axioms 3 and 1, we have
   $$p(A) + p(A^c) = p(\Omega) = 1.$$
2. $p(\phi) = 0$;
   Proof: Since $\phi = \Omega^c$, by the first property and Axiom 1, we have
   $$p(\phi) = 1 - p(\Omega) = 1 - 1 = 0.$$
3. For any $A, B \subset \Omega$, $p(B) = p(B \bigcap A) + p(B \bigcap A^c)$.
   Proof: We have introduced the partition of $B = (B \bigcap A) \bigcup (B \bigcap A^c)$. Moreover, by the commutative and associative laws of intersection, we know
   $$(B \bigcap A) \bigcap (B \bigcap A^c) = (A \bigcap A^c) \bigcap (B \bigcap B) = \phi \bigcap B = \phi.$$
   Then by Axiom 3,
   $$p(B) = p(B \bigcap A) + p(B \bigcap A^c).$$
4. $A \subset B \Rightarrow p(A) \leq p(B)$;
   Proof: By Property 3 and Axiom 2, we know $p(B \bigcap A^c) \geq 0$ and $p(B) \geq p(B \bigcap A)$. Since $A \subset B$, we have $B \bigcap A = A$. Then $p(B) \geq p(A)$.

Inclusion-exclusion identity

1. **Addition law**: $p(A \bigcup B) = p(A) + p(B) - p(A \bigcap B)$.

   Proof: By property 3, we have

   $$p(A \bigcup B) = p((A \bigcup B) \bigcap B) + ((A \bigcup B) \bigcap B^c) = p(B) + p((A \bigcap B^c) \bigcup (B \bigcap B^c)) = p(B) + p(A \bigcap B^c).$$

   On the other hand, we know

   $$p(A) = p(A \bigcap B) + p(A \bigcap B^c).$$

   Take the difference of the two equations, we have

   $$p(A \bigcup B) - p(A) = p(B) - p(A \bigcap B),$$

   which implies $p(A \bigcup B) = p(A) + p(B) - p(A \bigcap B)$.

2. **Generalized addition law**:

   $$p(A \bigcup B \bigcup C) = p(A) + p(B) + p(C) - p(A \bigcap B) - p(B \bigcap C) - p(A \bigcap C) + p(A \bigcap B \bigcap C).$$

   Proof: Since $A \bigcup B \bigcup C = (A \bigcup B) \bigcup C$, by the addition law, we have

   $$p(A \bigcup B \bigcup C) = p(A \bigcup B) + p(C) - p((A \bigcup B) \bigcap C).$$

   By the addition law again, we have

   $$p(A \bigcup B) = p(A) + p(B) - p(A \bigcap B).$$

   Furthermore, by the distributive law and the addition law, we have

   $$\begin{aligned} p((A \bigcup B) \bigcap C) &= p\left((A \bigcap C) \bigcup (B \bigcap C)\right) = p(A \bigcap C) + p(B \bigcap C) - p((A \bigcap C) \bigcap (B \bigcap C)) \\ &= p(A \bigcap C) + p(B \bigcap C) - p(A \bigcap B \bigcap C). \end{aligned}$$

   We can obtain the final result by plug-in.

3. In general:

   $$\begin{aligned} p(E_1 \bigcup E_2 \bigcup \ldots \bigcup E_n) = \sum_{i=1}^{n} p(E_i) - \sum_{i_1 < i_2} p(E_{i1} \bigcap E_{i2}) + \ldots + (-1)^{r+1} \sum_{i_1 < i_2 < \ldots < i_r} p(E_{i_1} \bigcap E_{i_2} \bigcap \ldots \bigcap E_{i_r}) \\ + \ldots + (-1)^{n+1} p(E_1 \bigcap E_2 \bigcap \ldots \bigcap E_n). \end{aligned}$$

It is amazing to see we can derive useful and nonintuitive results by simple and intuitive axioms!

3.4 Examples

Example 1: A total of 36 members of a club play tennis, 28 play squash, and 18 play badminton. Furthermore, 22 of the members play both tennis and squash, 12 play both tennis and badminton, 9 play both squash and badminton, and 4 play all three sports. How many members of this club play at least one of the three sports?

Solution: (Define the experiment and the associated sample space.) Let N denote the number of the members of the club. Introduce the probability by assuming that a member of the club is randomly selected. For any subset C of the members of the club, we let $p(C)$ denote the probability that the selected member is contained in C, then

$$p(C) = \frac{|C|}{N}.$$

Now, with T being the set of members that plays tennis, S being the set that plays squash, and B being the set that plays badminton, we have

$$\begin{aligned} p(T \bigcup S \bigcup B) &= p(T) + p(S) + p(B) - p(T \bigcap S) - p(T \bigcap B) - p(S \bigcap B) + p(T \bigcap S \bigcap B) \\ &= \frac{1}{N}(36 + 28 + 18 - 22 - 12 - 9 + 4) = \frac{43}{N}. \end{aligned}$$

Then there are 43 members play at least one of the sports.

Example 2: Suppose that each of N men at a party throws his hat into the center of the room. The hats are first mixed up, and then each man randomly selects a hat. What is the probability that none of the men selects his own hat?

Solution: Denote by E_i the event that the ith man selects his own hat, where $i = 1, \ldots, N$. Then we have the inclusion-exclusion identity:

$$\begin{aligned} p(E_1 \bigcup E_2 \bigcup \ldots \bigcup E_n) = &\sum_{i=1}^{n} p(E_i) - \sum_{i_1 < i_2} p(E_{i1} \bigcap E_{i2}) + \ldots + (-1)^{r+1} \sum_{i_1 < i_2 < \ldots < i_r} p(E_{i_1} \bigcap E_{i_2} \bigcap \ldots \bigcap E_{i_r}) \\ &+ \ldots + (-1)^{n+1} p(E_1 \bigcap E_2 \bigcap \ldots \bigcap E_n). \end{aligned}$$

It is easy to see that

$$p(E_{i_1} \bigcap E_{i_2} \bigcap \ldots \bigcap E_{i_r}) = \frac{(N-r)!}{N!},$$

so

$$\sum_{i_1 < i_2 < \ldots < i_r} p(E_{i_1} \bigcap E_{i_2} \bigcap \ldots \bigcap E_{i_r}) = \binom{N}{r} \frac{(N-r)!}{N!} = \frac{N!(N-r)!}{(N-r)!r!N!} = \frac{1}{r!}.$$

Then the probability that at least one man selects his own hat is $\sum_{i=0}^{n} \frac{(-1)^i}{i!}$. Notice the taylor expansion $e^x = \sum_{i=0}^{N} x^i/i!$, we know when N large,

$$p(E_1 \bigcup E_2 \bigcup \ldots \bigcup E_n) \approx e^{-1} = .3679.$$

Then with probability approximately .6321 no man selects his own hat.

Chapter 4

Conditional probability and independence

4.1 Conditional probability

Let A and B be two events with $P(B) \neq 0$. The conditional probability of A given B is defined to be

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

or

$$P(A \bigcap B) = P(A|B)P(B).$$

Example: Suppose there are 5 items a, b, c, d and e in a lot. Choose a random sample of size 3, what is the sample space? The possible combinations of size 3 include

{{a, b, c}, {a, b, d}, {a, b, e}, {a, c, d}, {a, c, e}, {a, d, e}, {b, c, d}, {b, c, e}, {b, d, e}, {c, d, e}}.

(a) A= The sample contains a;
(b) B= The sample contains b;
(c) C= The sample contains a or b;
(d) D= The sample contains both a and b;
(e) E= The sample does not contain a.
(f) F= contains both c and d;

(a) A = {{a, b, c}, {a, b, d}, {a, b, e}, {a, c, d}, {a, c, e}, {a, d, e}};
(b) B = {{a, b, c}, {a, b, d}, {a, b, e}, {b, c, d}, {b, c, e}, {b, d, e}};
(c) C = {{a, b, c}, {a, b, d}, {a, b, e}, {a, c, d}, {a, c, e}, {a, d, e}, {b, c, d}, {b, c, e}, {b, d, e}};
(d) D = {{a, b, c}, {a, b, d}, {a, b, e}};
(e) E = {{b, c, d}, {b, c, e}, {b, d, e}, {c, d, e}}.
(f) F = {{a, c, d}, {b, c, d}, {c, d, e)} };
Then
1. $P(A|B) = P(A \cap B)/P(B) = 1/2$;
2. $P(A|A \cup B) = P(A)|P(A \cup B) = 2/3$;
3. $P(A \cap B|A \cup B) = 1/3$;
4. $P(A|E) = 0$;
5. $P(A|F) = 1/3$.

4.2 Independence

A and B are independent if knowing that one had occurred gave us no information about whether the other had occurred: if $P(B) \neq 0$ and $P(A) \neq 0$, then

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

Therefore, we usually define

$$A \text{ and } B \text{ are independent if and only if } P(A)P(B) = P(A \cap B).$$

Notice that here A and B can have zero probability.

Example: Flip a coin for twice. A: heads up in the first time; B: tails up in the second time; C at least heads up for once; D at least tails up for once. Are A and B independent? Are C and D independent?

Solution: $P(A) = 1/2$, $P(B) = 1/2$, $P(A \cap B) = 1/4$, so $P(A \cap B) = P(A)P(B)$.
$P(C) = 3/4$, $P(D) = 3/4$, $P(C \cap D) = 1/2$, so $P(C \cap D) \neq P(C)P(D)$.

Property

If A and B are independent, then A and B^c are independent, A^c and B^c are independent.
Proof

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)(1 - P(B)) = P(A)P(B^c). \\ P(A^c \cap B^c) &= P((A \cup B)^c) = 1 - P(A \cup B) = 1 - (P(A) + P(B) - P(A \cap B)) \\ &= 1 - P(A) - P(B) + P(A)P(B) = (1 - P(A))(1 - P(B)) = P(A^c)P(B^c). \end{aligned}$$

Mutual independence and pairwise independence Suppose we have events $A_1, A_2, \ldots, A_n$ pairwise independent, i.e., $P(A_i \bigcup A_j) = P(A_i)P(A_j)$ for all $i \neq j$. In contrast, we say $A_1, \ldots, A_n$ are mutually independent if for any sub collection $A_{i_1}, \ldots, A_{i_m}$ where $1 \leq i_1 < i_2 < \ldots < i_m \leq n$, we have

$$\mathbb{P}(A_{i_1} \bigcap \ldots \bigcap A_{i_m}) = P(A_{i_1}) \ldots P(A_{i_m}).$$

From this definition, we can conclude that if a collection of events are mutually independent, then they are pairwise independent; In contrast, if a collection of events are pairwise independent, they are not necessarily mutually independent.

Example: Throw two fair dice. Consider the events:

A:={the sum of the points is 7}
B:={the first die rolled a 3}
C:={the second die rolled a 4}

Solution: $P(A) = 1/6$, $P(B) = 1/6$, $P(C) = 1/6$, $P(A \cap B) = 1/36$, $P(A \cap C) = 1/36$, $P(B \cap C) = 1/36$, but $P(A \cap B \cap C) = 1/36$.

4.3 Bayes rule

Law of total probability Let $B_1, \ldots, B_n$ be such that $\bigcup_{i=1}^n B_i = \Omega$ and $B_i \bigcap B_j \neq \phi$ for $i \neq j$. Assume $\mathbb{P}(B_i) > 0$ for all i. Then for any event A,

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

Bayes Rule Let A and $B_1, \ldots, B_n$ be events where the B_i are disjoint, $\bigcup_{i=1}^n B_i = \Omega$, and $P(B_i) > 0$ for all i. Then

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^n P(A|B_i)P(B_i)}.$$

Example 1: A true-false question is to be posed to a husband-and-wife team on a quiz show. Both the husband and the wife will independently give the correct answer with probability p. Which of the following is a better strategy for the couple?
(a) Choose one of them and let that person answer the question.
(b) Have them both consider the question, and then either give common answer if they agree or, if they disagree, flip a coin to determine which answer to give.

Moreover, if they choose (b), what is the probability both the husband and wife are correct given they finally provided the correct answer?

Solution:
Strategy (a): $P(\text{the couple give the correct answer}) = p$
Strategy (b): Denote by H the husband is correct. Denote by W the wife is correct. Denote by C the couple give the correct answer. Then

$$\begin{aligned} P(C) &= P(C|H \cap W)P(H \cap W) + P(C|H \cap W^c)P(H \cap W^c) && (4.3.1)\\ &\quad + P(C|H^c \cap W)P(H^c \cap W) + P(C|H^c \cap W^c)P(H^c \cap W^c) && (4.3.2)\\ &= p^2 + \frac{1}{2}(p(1-p) + (1-p)p) = p && (4.3.3) \end{aligned}$$

If the choose strategy (b),

$$P(H \cap W|C) = \frac{P(C|H \cap W)P(H \cap W)}{P(C)} = \frac{p^2}{p} = p.$$

Example 2: A deck of cards is shuffled and then divided into two halves of 26 cards each. A card is drawn from one of the halves; it turns out to be an ace. The ace is then placed in the second half-deck. The half is then shuffled, and a card is drawn from it.

1. Compute the probability that this drawn card is an ace;
2. Compute the probability that this drawn card is the same card that is drawn before given it is an ace.

Solution: Suppose the interchanged card is a, and the selected card is x. Then

$$P(\text{x is ace}) = P(\text{x is ace}|x = a)P(x = a) + P(\text{x is ace}|x \neq a)P(x \neq a) = 1 \times \frac{1}{27} + \frac{3}{51} \times \frac{26}{27} = 0.0937$$

Moreover,

$$P(x = a|\text{x is ace}) = \frac{P(\text{x is ace}|x = a)P(x = a)}{P(\text{x is ace})} = 0.40.$$

Example 3: Let $S = \{1, 2, \ldots, n\}$ and suppose that A and B are, independently, equally likely to be any of the 2^n subsets (including the null set and S itself) of S.
(a) Show that

$$P(A \subset B) = \left(\frac{3}{4}\right)^n.$$

(b) Show that

$$P(A \cap B = \phi) = \left(\frac{3}{4}\right)^n.$$

Solution: (a)

$$\begin{aligned} P(A \subset B) &= \sum_{i=0}^{n} P(A \subset B | N(B) = i) P(N(B) = i) \\ &= \sum_{i=0}^{n} \frac{2^i}{2^n} \times \frac{\binom{n}{i}}{2^n} \\ &= \frac{1}{4^n} \sum_{i=0}^{n} 2^i \times 1^{n-i} \binom{n}{i} \\ &= \frac{(2+1)^n}{4^n} \\ &= \left(\frac{3}{4}\right)^n \end{aligned}$$

(b) Similarly to (a), we have $P(A \cap B = \phi) = P(A \subset B^c) = \left(\frac{3}{4}\right)^n$.

Example 4: The color of a person's eyes is determined by a single pair of genes. If they are both blue-eyed genes, then the person will have blue eyes; if they are both brown-eyed genes, then the person will have brown eyes; and if one of them is blue-eyed gene and the other a brown-eyed gene, then the person will have brown eyes. A newborn child independently receives one eye gene from each of its parents, and the gene it receives from a parent is equally likely to be either of the two eye genes of that parent. Suppose that Smith and both of his parents have brown eyes, but Smith's sister has blue eyes.

- (a) What is the probability that Smith possesses a blue-eyed gene?
- (b) Suppose that Smith's wife has blue eyes. What is the probability that their first child will have blue eyes?
- (c) If their first child has brown eyes, what is the probability that their next child will also have brown eyes?

Solution: Smith: brown; Mother: brown; Father: brown; Sister: blue; Wife: blue.

We know Sister (blue and blue), Wife (blue and blue), Mother (blue and brown), Father (blue and brown).

Denote

$$A = \text{Smith (blue and blue)},\ B = \text{Smith (blue and brown)},\ C = \text{Smith (brown and brown)}.$$

Then the unconditional probabilities are

$$P(A) = \frac{1}{4},\ P(B) = \frac{1}{2},\ P(C) = \frac{1}{4}.$$

(a)

$$P(B | B \cup C) = \frac{P(B \cap (B \cup C))}{P(B \cup C)} = \frac{P(B)}{P(B \cup C)} = \frac{2}{3}.$$

(b) Given the available information, we first update our belief

$$P(B) = \frac{2}{3},\ P(C) = \frac{1}{3}.$$

Denote

$$A_1 = \text{The first child (blue and blue)},\ B_1 = \text{the first child (blue and brown)}.$$

We have

$$P(A_1|B) = \frac{1}{2},\ P(B_1|B) = \frac{1}{2},\ P(A_1|C) = 0,\ P(B_1|C) = 1.$$

Then

$$P(A_1) = P(A_1|B)P(B) + P(A_1|C)P(C) = \frac{1}{2} \times \frac{2}{3} = \frac{1}{3}$$

(c) Given the first child has brown eyes, i.e., B_1, we need to first update our belief of B and C:

$$P(B|B_1) = \frac{P(B_1|B)P(B)}{P(B_1)} = \frac{\frac{1}{2} \times \frac{2}{3}}{1 - \frac{1}{3}} = \frac{1}{2}$$

and thus $P(C|B_1) = 1/2$. Therefore, given the known information, we update our belief as

$$P(B) = 1/2, \quad P(C) = 1/2.$$

Denote

$$A_2 = \text{The second child (blue and blue)},\ B_2 = \text{the second child (blue and brown)}.$$

we still have

$$P(A_2|B) = \frac{1}{2},\ P(B_2|B) = \frac{1}{2},\ P(A_2|C) = 0,\ P(B_2|C) = 1.$$

Then

$$P(B_2) = P(B_2|B)P(B) + P(B_2|C)P(C) = \frac{3}{4}.$$

Property of Mutual Independence

If E_1, E_2, ..., E_n are mutually independent, state true or false for each of the following:

- For any $1 \le i_1 < i_2 < \ldots < i_r \le n$, we know E_{i_1}, E_{i_2}, ..., E_{i_r} are mutually independent. This is a straightforward corollary by definition.

- $E_1 \cap E_2$, E_3, ..., E_n are mutually independent.
  Proof It suffices to show for any sub-collection $E_{i_1}, \ldots, E_{i_r}$ for $2 \le i_1 < \ldots < i_r \le n$,

  $$P((E_1 \cap E_2) \cap E_{i_1} \cap \ldots \cap E_{i_r}) = P(E_1 \cap E_2)P(E_{i_1}) \ldots P(E_{i_r}).$$

  In fact

  $$\begin{aligned} P((E_1 \cap E_2) \cap E_{i_1} \cap E_{i_2} \cap \ldots \cap E_{i_r}) &= P(E_1 \cap E_2 \cap E_{i_1} \cap E_{i_2} \cap \ldots \cap E_{i_r}) \\ &= P(E_1)P(E_2)P(E_{i_1})P(E_{i_2}) \ldots P(E_{i_r}) \\ &= P(E_1 \cap E_2)P(E_{i_1})P(E_{i_2}) \ldots P(E_{i_r}). \end{aligned}$$

  ■

- Similarly, for any $1 \le i_{11} < i_{12} < \ldots < i_{1m_1} \le n$, $1 \le i_{21} < i_{22} < \ldots < i_{2m_2} \le n$, ..., $1 \le i_{k1} < i_{k2} < \ldots < i_{km_k} \le n$ and these $m_1 + m_2 + \ldots + m_k$ indices are all distinct, we have that $E_{i_{11}} \cap E_{i_{12}} \cap \ldots \cap E_{i_{1m_1}}$, $E_{i_{21}} \cap E_{i_{22}} \cap \ldots \cap E_{i_{2m_2}}$, ..., $E_{i_{k1}} \cap E_{i_{k2}} \cap \ldots \cap E_{i_{km_k}}$ are independent.

- E_1^c, E_2, E_3, ..., E_n are mutually independent.
  Proof It suffices to show for any sub-collection $E_{i_1}, \ldots, E_{i_r}$ for $2 \leq i_1 < \ldots < i_r \leq n$,

$$P(E_1^c \cap E_{i_1} \cap \ldots \cap E_{i_r}) = P(E_1^c)P(E_{i_1}) \ldots P(E_{i_r}).$$

  In fact

$$P(E_{i_1} \cap \ldots \cap E_{i_r}) = P(E_1 \cap (E_{i_1} \cap \ldots E_{i_r})) + P(E_1^c \cap (E_{i_1} \cap \ldots E_{i_r})).$$

  Then

$$\begin{aligned} P(E_1^c \cap E_{i_1} \cap \ldots E_{i_r}) &= P(E_{i_1} \cap \ldots \cap E_{i_r}) - P(E_1 \cap E_{i_1} \cap \ldots E_{i_r}) \\ &= P(E_{i_1} \ldots P(E_{i_r}) - P(E_1)P(E_{i_1} \ldots P(E_{i_r}) \\ &= (1 - P(E_1))P(E_{i_1} \ldots P(E_{i_r}) \\ &= P(E_1^c)P(E_{i_1}) \ldots P(E_{i_r}) \end{aligned}$$

  ■

- Similarly, for any $1 \leq i_1 < i_2 < \ldots < i_m \leq n$ and $1 \leq j_1 < j_2 < \ldots < j_r \leq n$ and these $m + r$ indices are all distinct, we have that $E_{i_1}^c, \ldots E_{i_m}^c, E_{j_1}, \ldots, E_{j_r}$ are mutually independent.
- By the results above, can we claim that $E_1 \cap E_2$, E_3, ..., E_n are mutually independent?

Chapter 5

Random Variables

5.1 Topics covered by Miles

- Definition of random variable
- Definition of pmf
- Plot/properties of pmf
- Derivation of pmf for Bernoulli, Binomial, and Poisson (including showing that Poisson pmf is the limit of a Binomial pmf)

5.2 Review the concepts of random variables and probability mass functions

Random variables: Real-valued functions defined on Ω.

A random variable that can taken on at most a countable number of possible values is said to be discrete.

probability mass function:

$$p(a) = P\{X = a\}.$$

Suppose X must assume one of the values $x_1, x_2, \ldots$, then we have

$$p(x_i) \geq 0, i = 1, 2, \ldots, \quad p(x) = 0, \text{ otherwise.}$$

and

$$\sum_{i=1}^{\infty} p(x_i) = 1.$$

Example 1: Suppose there are 5 items a, b, c, d and e in a lot. Choose a random sample of size 3, what is the sample space? The possible combinations of size 3 include

{{a, b, c}, {a, b, d}, {a, b, e}, {a, c, d}, {a, c, e}, {a, d, e}, {b, c, d}, {b, c, e}, {b, d, e}, {c, d, e}}.

We can define number of variables: X: the cardinality of the intersection of the random sample and the

set $\{a, b, c\}$. Then

$$\begin{aligned}
X(\{a,b,c\}) &= 3 \\
X(\{a,b,d\}) &= 2 \\
X(\{a,b,e\}) &= 2 \\
X(\{a,c,d\}) &= 2 \\
X(\{a,c,e\}) &= 2 \\
X(\{a,d,e\}) &= 1 \\
X(\{b,c,d\}) &= 2 \\
X(\{b,c,e\}) &= 2 \\
X(\{b,d,e\}) &= 1 \\
X(\{c,d,e\}) &= 1
\end{aligned}$$

If we further assume all outcomes have equal probabilities, then we also have

$$p(1) = P(X = 1) = 0.3, \quad p(2) = P(X = 2) = 0.6, \quad p(3) = P(X = 3) = 0.1.$$

Example 2

Five men and 5 women are ranked according to their scores on an examination. Assume that no two scores are alike and all 10! possible rankings are equally likely. Let X denote the highest ranking achieved by a woman. (For instance, $X = 1$ if the top ranked person in female.) Find $\mathbb{P}(X = i), i = 1, \ldots, 10$.

Solution:
$P(X=1) = \frac{5\times 9!}{10!} = \frac{1}{2}$
$P(X=2) = \frac{5\times 5\times 8!}{10!} = \frac{5}{18}$
$P(X=3) = \frac{5\times 4\times 5\times 7!}{10!} = \frac{5}{36}$
$P(X=4) = \frac{5\times 4\times 3\times 5\times 6!}{10!} = \frac{5}{84}$
$P(X=5) = \frac{5\times 4\times 3\times 2\times 5\times 5!}{10!} = \frac{5}{252}$
$P(X=6) = \frac{5\times 4\times 3\times 2\times 1\times 5!}{10!} = \frac{5}{1260}$

Example 3

Suppose that a die is rolled twice. First, determine all possible values that the following random variables can take on. Second, calculate the probabilities associated with the random variables assuming that the die is fair.
(a) X: the maximum value to appear in the two rolls.
(b) Y: the value of the first roll minus the value of the second roll.

Solution:
(a)
$P(X=1) = \frac{1}{6}\times\frac{1}{6} = \frac{1}{36}$
$P(X=2) = \frac{1}{6}\times\frac{1}{6} + \frac{1}{6}\times\frac{2-1}{6} + \frac{1}{6}\times\frac{2-1}{6} = \frac{1}{12}$
$P(X=3) = \frac{1}{6}\times\frac{1}{6} + \frac{1}{6}\times\frac{3-1}{6} + \frac{1}{6}\times\frac{3-1}{6} = \frac{5}{36}$
$P(X=4) = \frac{1}{6}\times\frac{1}{6} + \frac{1}{6}\times\frac{4-1}{6} + \frac{1}{6}\times\frac{4-1}{6} = \frac{7}{36}$
$P(X=5) = \frac{1}{6}\times\frac{1}{6} + \frac{1}{6}\times\frac{5-1}{6} + \frac{1}{6}\times\frac{5-1}{6} = \frac{9}{36}$
$P(X=6) = \frac{1}{6}\times\frac{1}{6} + \frac{1}{6}\times\frac{6-1}{6} + \frac{1}{6}\times\frac{6-1}{6} = \frac{11}{36}$

(b)
$P(Y=5)=P(Y=-5)=\frac{1}{6}\times\frac{1}{6}=\frac{1}{36}$
$P(Y=4)=P(Y=-4)=\frac{2}{6}\times\frac{1}{6}=\frac{2}{36}$
$P(Y=3)=P(Y=-3)=\frac{3}{6}\times\frac{1}{6}=\frac{3}{36}$
$P(Y=2)=P(Y=-2)=\frac{4}{6}\times\frac{1}{6}=\frac{4}{36}$
$P(Y=1)=P(Y=-1)=\frac{5}{6}\times\frac{1}{6}=\frac{5}{36}$
$P(Y=0)=\frac{6}{6}\times\frac{1}{6}=\frac{6}{36}$

5.3 Expectation of a discrete random variable

In general, the expectation of random variable can be informally written as

$$E(X)=\sum_{\omega\in\Omega}X(\omega)p(\omega).$$

For a discrete random variable, its expectation is defined as

$$E(X)=\sum_{x:p(x)>0}xp(x).$$

Example 1 above: By definition,

$$E(X)=0.3\times1+0.6\times2+0.1\times3=1.8.$$

In fact, by the general definition, we have

$$E(X)=\sum_{\omega\in\Omega}X(\omega)p(\omega)=\frac{1}{10}(3+2+2+2+2+1+2+2+1+1)=1.8.$$

The second method seems to be more naive, but it is conceptually important in statistics, such as estimation and Monte Carlo methods.

Example 3 above:
Method 1: By definition,

$$E(Y)=(-5)\times\frac{1}{36}+(-4)\times\frac{2}{36}+(-3)\times\frac{3}{36}+(-2)\times\frac{4}{36}+(-1)\times\frac{5}{36}+0\times\frac{6}{36}+1\times\frac{5}{36}+2\times\frac{4}{36}+3\times\frac{3}{36}+4\times\frac{2}{36}+5\times\frac{1}{36}=0$$

Method 2: Notice that the graph of pmf is symmetric about the y-axis, so we can claim that the expectation is 0.
Method 3: We represent the sample space as $\Omega=\{(a,b):a,b=1,2,\ldots,6\}$. Then $Y((a,b))=a-b$. We know that if $a=b$, then $Y((a,b))=0$. If $a<b$, then $Y((a,b))p((a,b))+Y((b,a))p((b,a))=0$. Then by the general definition, $E(Y)=0$.

Example

A and B play the following game: A writes down either number 1 or number 2, and B must guess which one. If the number that A has written down is i and B has guessed correctly, B receives i units from A. If B makes a wrong guess, B pays $\frac{3}{4}$ unit to A. If B randomizes his decision by guessing 1 with probability p and 2 with probability $1-p$, determine his expected gain if (a) A has written down number 1 and (b) A has written down number 2.

- What values of p maximizes the minimum possible value of B's expected gain, and what is this maximum value? (Note that B's expected gain depends not only on p, but also on what A does.) Consider now player A. Suppose that she also randomize her decision, writing down number 1 with probability q. What is A's expected loss if (c) B chooses number 1 and (d) B chooses number 2?
- What value of q minimizes A's maximum expected loss? Show that the minimum if A's maximum expected loss is equal to the maximum of B's minimum expected gain.

This result, known as the minimax theorem was first established in generality by the mathematician Jon von Neumann and is the fundamental result in mathematical discipline know as the theory of games. The common value is called the value of the game to player B.

Solution:
(a) $E(\text{B's gain}|\text{A writes down 1}) = p - (1-p)\frac{3}{4} = \frac{7}{4}p - \frac{3}{4}$
(b) $E(\text{B's gain}|\text{A writes down 2}) = p \times (-\frac{3}{4}) + (1-p) \times 2 = 2 - \frac{11}{4}p$

To compute $maximin(\frac{7}{4}p - \frac{3}{4}, 2 - \frac{11}{4}p)$, we let $\frac{7}{4}p - \frac{3}{4} = 2 - \frac{11}{4}p$. Then $p = \frac{11}{18}$, and $maximin(\frac{7}{4}p - \frac{3}{4}, 2 - \frac{11}{4}p) = \frac{23}{72}$.

(c) $E(\text{A's loss}|\text{B chooses 1}) = q - (1-q)\frac{3}{4} = \frac{7}{4}q - \frac{3}{4}$
(d) $E(\text{A's loss}|\text{B chooses 2}) = q \times (-\frac{3}{4}) + (1-q) \times 2 = 2 - \frac{11}{4}q$

To compute $minimax(\frac{7}{4}q - \frac{3}{4}, 2 - \frac{11}{4}q)$, we let $\frac{7}{4}q - \frac{3}{4} = 2 - \frac{11}{4}q$. Then $q = \frac{11}{18}$, and $minimax(\frac{7}{4}q - \frac{3}{4}, 2 - \frac{11}{4}q) = \frac{23}{72}$. We see that the minimum of A's maximum expected loss is equal to the maximum of B's minimum expected gain.

5.4 Expectation of Binomial and Poisson random variables

Let's recall the pmf of Binomial distribution: If $X \sim \text{Binomial}(n, p)$, then

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

Then

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=0}^{n} kp(k) = k\binom{n}{k} p^k (1-p)^{(n-k)} = k\frac{n!}{k!(n-k)!} p^k (1-p)^{(n-k)} \\
&= (np)\frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} p^{k-1} (1-p)^{((n-1)-(k-1))} \\
&= np(p + (1-p))^{n-1} = np.
\end{aligned}$$

Examples: Coins flipping, Dies rolling.

Let's recall the pmf of Poisson distribution: If $X \sim \text{Poisson}(\lambda)$,

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots.$$

Then

$$\begin{aligned}\mathbb{E}(X) &= \sum_{k=0}^{\infty} kp(k) = \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} = \sum_{k=1}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} \\ &= \sum_{k=1}^{\infty} \lambda\frac{\lambda^{(k-1)}}{(k-1)!}e^{-\lambda} = \sum_{k=1}^{\infty} \lambda\frac{\lambda^{(k-1)}}{(k-1)!}e^{-\lambda} \\ &= \sum_{k=1}^{\infty} \lambda\frac{\lambda^{(k-1)}}{(k-1)!}e^{-\lambda} = (\lambda e^{-\lambda})\sum_{k=0}^{\infty}\frac{\lambda^{(k)}}{(k)!} \\ &= (\lambda e^{-\lambda})e^{\lambda} = \lambda.\end{aligned}$$

5.5 Expectation of a function of random variables, Variance

Suppose X is a discrete random variable with pmf $p_X(x)$, then for any real-valued function f, $Y = g(X)$ is also a discrete random variable, with some pmf $p_Y(y)$. Then by definition, we have

$$\mathbb{E}(Y) = \sum_{y:p_Y(y)>0} yp_Y(y).$$

However, this formula is not very practically convenient, since we need to evaluate the pmf of Y. The motivating question becomes

Can be evaluate $\mathbb{E}(Y)$ without evaluating $p_Y(y)$?

To this end, we can come back to the "intrinsic" definition of $E(Y)$ from the sample space:

$$E(Y) = \sum_{\omega\in\Omega} Y(\omega)p(\omega) = \sum_{\omega\in\Omega} f(X(\omega))p(\omega).$$

Combining like terms, i.e., different ω with the same $X(\omega)$, we get

$$E(Y) = \sum_{x:p_X(x)>0} g(x)p_X(x).$$

This implies that if X only takes on values x_1, x_2, ..., then

$$E(g(X)) = \sum_i g(x_i)p_X(x_i).$$

Example

Flip a fair coin independently for 5 times. Let Y be the difference between the number of heads and the number of tails. Find $\mathbb{E}[Y]$.

Solution: Denote by X the number of heads, we know $X \sim$ Binomial$(5, \frac{1}{2})$. Then the number of tails is $5 - X$. This implies that $Y = X - (5 - X) = 2X - 5$. By the formula of the expectation of a function of a discrete random variable, we have

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{k=0}^{5}(2k-5)p_X(k) \\ &= 2\sum_{k=0}^{5} kp_X(k) - 5\sum_{k=0}^{5} p_X(k) \\ &= 2\,\mathbb{E}(X) - 5 = 2(5\times\frac{1}{2}) - 5 = 0.\end{aligned}$$

Corollary

This example implies that if $Y = aX + b$, then $\mathbb{E}(Y) = aE(X) + b$. In fact, we have

$$\begin{aligned}\mathbb{E}[Y] &= \sum_i (ax_i + b)p_X(x_i) \\ &= a\sum_i x_i p_X(x_i) + b\sum_i p_X(x_i) \\ &= a\,\mathbb{E}(X) + b.\end{aligned}$$

Example

A communication channel transmits the digits 0 and 1. However due to static, the digit transmitted is incorrectly received with probability .2. Suppose that we want to transmit an important message consisting of one binary digit. To reduce the chance of error, we transmit 00000 instead of 0 and 11111 instead of 1. If the receiver of the message uses "majority" decoding, what is the probability that the message will be wrong when decoded?

Solution: Define the number of incorrectly received digits as X. Denote the event "the message will be wrong when decoded" as A. Then we have

$$X \sim \text{Binomial}(5, 0.1), \quad A = \{X \geq 3\}.$$

Define a new random variable $Y = f(X)$, where $f(x) = \begin{cases} 0, & \text{if } X \leq 2 \\ 1, & \text{if } X \geq 3 \end{cases}$. Then

$$\mathbb{P}(A) = \mathbb{E}(Y) = \sum_{x=1}^{5} f(x)p_X(x) = \binom{5}{3}(0.2)^3(0.8)^2 + \binom{5}{4}(0.2)^4(0.8)^1 + \binom{5}{5}(0.2)^5 = 0.0579.$$

5.6 Variance

For a random variable X, assuming $\mathbb{E}(X) = \mu$, then its variance is defined as

$$\text{Var}(X) = \mathbb{E}(X - \mu)^2.$$

Variance of a Binomial random variable

If $X \sim \text{Binomial}(n, p)$, then

$$p_X(k) = \binom{n}{k} p^k (1-p)^{(n-k)}, \quad k = 0, 1, \ldots, n.$$

Moreover, $\mu = \mathbb{E}(X) = np$. Then

$$\begin{aligned}
\mathbb{E}(X-\mu)^2 &= \sum_{k=0}^{n}(k-np)^2 p_X(k) \\
&= \sum_{k=0}^{n}(k-np)^2 p_X(k) \\
&= \sum_{k=0}^{n}(k(k-1)-(2np-1)k+n^2p^2)p_X(k) \\
&= \sum_{k=0}^{n} k(k-1)p_X(k) - (2np-1)\sum_{k=0}^{n} kp_X(k) + n^2p^2\sum_{k=0}^{n} p_X(k) \\
&= \sum_{k=0}^{n} k(k-1)p_X(k) - (2np-1)\,\mathbb{E}(X) + n^2p^2 \\
&= \sum_{k=0}^{n} k(k-1)\frac{n!}{k!(n-k)!}p^k(1-p)^{(n-k)} - n^2p^2 + np \\
&= \sum_{k=2}^{n} \frac{n!}{(k-2)!(n-k)!}p^k(1-p)^{(n-k)} - n^2p^2 + np \\
&= \sum_{k=0}^{n-2} \frac{n!}{(k)!((n-2)-k)!}p^{k+2}(1-p)^{(n-2-k)} - n^2p^2 + np \\
&= n(n-1)p^2\sum_{k=0}^{n-2} \frac{(n-2)!}{(k)!((n-2)-k)!}p^k(1-p)^{(n-2-k)} - n^2p^2 + np \\
&= n(n-1)p^2(p+1-p)^{n-2} - n^2p^2 + np \\
&= np(1-p).
\end{aligned}$$

Variance of a Poisson random variable

If $X \sim \text{Poisson}(\lambda)$,

$$p(k) = \frac{\lambda^k}{k!}e^{-\lambda}, \quad k = 0, 1, 2, \ldots.$$

Moreover, $\mu = \mathbb{E}(X) = \lambda$. Then

$$\begin{aligned}
\mathbb{E}(X-\mu)^2 &= \sum_{k=0}^{\infty}(k-\lambda)^2 p_X(k) \\
&= \sum_{k=0}^{\infty}(k^2 - 2\lambda k + \lambda^2) p_X(k) \\
&= \sum_{k=0}^{\infty}(k(k-1) + (1-2\lambda)k + \lambda^2 p_X(k) \\
&= \sum_{k=0}^{\infty} k(k-1)p_X(k) + (1-2\lambda)\sum_{k=0}^{\infty} k p_X(k) + \lambda^2 \sum_{k=0}^{\infty} p_X(k) \\
&= \sum_{k=0}^{\infty} k(k-1)p_X(k) + (1-2\lambda)\lambda + \lambda^2 \\
&= \sum_{k=2}^{\infty} k(k-1)\frac{\lambda^k}{k!}e^{-\lambda} + (\lambda - \lambda^2) \\
&= \sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!}e^{-\lambda} + (\lambda - \lambda^2) \\
&= \sum_{k=0}^{\infty} \frac{\lambda^{k+2}}{k!}e^{-\lambda} + (\lambda - \lambda^2) \\
&= \lambda^2 \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}e^{-\lambda} + (\lambda - \lambda^2) \\
&= \lambda^2 + (\lambda - \lambda^2) \\
&= \lambda.
\end{aligned}$$

Property

In general, for a discrete variable X taking on values x_1, x_2, ..., we have

$$\mathrm{Var}(X) = \mathbb{E}\,X^2 - \mathbb{E}(X)^2.$$

In fact,

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}(X-\mu)^2 \\
&= \sum_i (x_i - \mu)^2 p_X(x_i) \\
&= \sum_i (x_i^2 - 2\mu x_i + \mu^2) p_X(x_i) \\
&= \sum_i x_i^2 p_X(x_i) - 2\mu \sum_i x_i p_X(x_i) + \mu^2 \sum_i p_X(x_i) \\
&= \mathbb{E}(X^2) - 2\mu^2 + \mu^2 \\
&= \mathbb{E}\,X^2 - \mathbb{E}(X)^2.
\end{aligned}$$

5.7 Expectation of sum of random variables

Given discrete random variables X_1, X_2, ... X_k defined on the sample space. Let $Y = X_1 + \ldots + X_k$. How to find $\mathbb{E}[Y]$ without evaluating its pmf?

Informally,

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_{\omega_i \in \Omega} Y(\omega_i) p(\omega_i) \\
&= \sum_{\omega_i \in \Omega} \left(X_1(\omega_i) + \ldots + X_k(\omega_i)\right) p(\omega_i) \\
&= \sum_{\omega_i \in \Omega} \left(X_1(\omega_i) p(\omega_i) + \ldots + X_k(\omega_i) p(\omega_i)\right) \\
&= \sum_{\omega_i \in \Omega} X_1(\omega_i) p(\omega_i) + \ldots + \sum_{\omega_i \in \Omega} X_k(\omega_i) p(\omega_i) \\
&= \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_k].
\end{aligned}$$

Example

Suppose there are 6 items $\{a, b, c, d, e, f\}$ in a lot. Choose a random sample of size 3. Let X be the cardinality of the intersection of the random sample and $\{a, b, c\}$. Find $\mathbb{E}(X)$.

Method 1 Let

$$Z_a := \begin{cases} 1 & \text{if a is sampled} \\ 0 & \text{if a is not sampled.} \end{cases}$$

Similarly, we can define Z_b and Z_c. Then $X = Z_a + Z_b + Z_c$. Notice that here we have

$$\mathbb{E}[Z_a] = \mathbb{P}(\text{a is sampled}) = \frac{\binom{5}{2}}{\binom{6}{3}} = \frac{1}{2}.$$

Then

$$\mathbb{E}\, X = \mathbb{E}(Z_a + Z_b + Z_c) = \frac{3}{2}.$$

Method 2 Let Y be the cardinality of the intersection between the random sample and $\{d, e, f\}$. By symmetry, we know X and Y have the same pmf, which implies that $\mathbb{E}(X) = \mathbb{E}(Y)$. Moreover, we know $X + Y = 3$. This implies

$$3 = \mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 2\,\mathbb{E}(X) \Longrightarrow \mathbb{E}(X) = \frac{3}{2}.$$

Chapter 6

Continuous random variables

6.1 Probability density function

For a continuous random variable X, define its probability density function as

$$f_X(x) = \lim_{\delta \to 0} \frac{\mathbb{P}(x - \delta/2 < X < x + \delta/2)}{\delta}$$

- Local distributional property;
- The "unit" is not probability, but "probability divided by length".

From this definition, if $X \sim \text{Unif}([0,1])$, we have

$$f_X(x) = \begin{cases} 1, & \text{if } x \in (0,1) \\ \frac{1}{2} & \text{if } x \in \{0,1\} \\ 0, & \text{if } x < 0 \text{ or } x > 1 \end{cases}.$$

For simplicity, we can just write

$$f_X(x) = \begin{cases} 1, & \text{if } x \in (0,1) \\ 0, & \text{if } x \leq 0 \text{ or } x \geq 1 \end{cases}.$$

or

$$f_X(x) = \begin{cases} 1, & \text{if } x \in [0,1] \\ 0, & \text{if } x < 0 \text{ or } x > 1 \end{cases}.$$

6.2 How to calculate probabilities by pdf

Let's see how the local property can be integrated into global property, e.g., $\mathbb{P}(a < X < b)$. For any random variable X, consider its discretization

$$X_n := \frac{2k-1}{2n} \quad \text{if and only if} \quad \frac{k-1}{n} \leq X < \frac{k}{n},$$

where $k = \ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$. Then as long as n goes to infinity, we have (nonrigorously)

$$\mathbb{P}(a < X_n < b) \to P(a < X < b).$$

On the other hand, we know

$$\begin{aligned}\mathbb{P}(a < X_n < b) &= \sum_{k:a<\frac{2k-1}{2n}<b} p_{X_n}\left(\frac{2k-1}{2n}\right) \\ &= \sum_{k:a<\frac{2k-1}{2n}<b} \mathbb{P}\left(\frac{k-1}{n} \le X < \frac{k}{n}\right) \\ &\approx \sum_{k:a<\frac{2k-1}{2n}<b} f_X\left(\frac{2k-1}{2n}\right)\frac{1}{n} \\ &\to \int_a^b f_X(x)dx.\end{aligned}$$

Therefore, we have

$$p(a < X < b) = \int_a^b f_X(x)dx.$$

6.3 How to find expectations by pdf

For any random variable X, consider its discretization

$$X_n := \frac{2k-1}{2n} \quad \text{if and only if} \quad \frac{k-1}{n} \le X < \frac{k}{n},$$

where $k = \ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$. Then as long as n goes to infinity, we have (nonrigorously)

$$\mathbb{E}(X_n) \to \mathbb{E}(X).$$

On the other hand, we know

$$\begin{aligned}\mathbb{E}(X_n) &= \sum_{k=-\infty}^{\infty} \frac{2k-1}{2n} p_{X_n}\left(\frac{2k-1}{2n}\right) \\ &= \sum_{k=-\infty}^{\infty} \frac{2k-1}{2n}\,\mathbb{P}\left(\frac{k-1}{n} \le X < \frac{k}{n}\right) \\ &\approx \sum_{k=-\infty}^{\infty} \frac{2k-1}{2n} f_X\left(\frac{2k-1}{2n}\right)\frac{1}{n} \\ &\to \int_{-\infty}^{\infty} x f_X(x)dx.\end{aligned}$$

Therefore, we have

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx.$$

Example

If $X \sim \text{Unif}(0,1)$, then

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx = \int_0^1 x dx = \frac{1}{2}.$$

6.4 How to find pdf

Think of the question

$$\textit{If } X \sim \textit{Unif}([0,1]), \textit{ find the pdf of } Y = 2X,\ Z = X^2.$$

Definition 6.4.1 *The cumulative distributive function (cdf) is defined as*

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du = P(X \le x).$$

The fundamental theorem of calculus gives

$$F_X'(x) = f_X(x).$$

By this tool, we have

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(2X \le y) = \mathbb{P}\left(X \le \frac{y}{2}\right) = F_X\left(\frac{y}{2}\right)$$

Then

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X\left(\frac{y}{2}\right) = F'\left(\frac{y}{2}\right)\frac{1}{2} = \frac{1}{2}f_X\left(\frac{y}{2}\right).$$

In the case $X \sim \text{Unif}([0,1])$,

$$f_Y(y) = \begin{cases} \frac{1}{2}, & 0 < y < 1 \\ 0 & \text{o/w.} \end{cases}$$

As for $Z = X^2$, let's focus on the case $Z = X^2$. We know $Z \in (0,1)$ with probability one, so

$$f_Z(z) = 0, \quad \text{if } z \le 0 \text{ or } z \ge 1.$$

For any $0 < z < 1$,

$$F_Z(z) = \mathbb{P}(Z \le z) = \mathbb{P}(X \le \sqrt{z}) = \sqrt{z}.$$

Then $f_Z(z) = F_Z'(z) = \frac{1}{2\sqrt{z}}$. In sum

$$f_Z(z) = \begin{cases} \frac{1}{2\sqrt{z}} \text{ if } 0 < z < 1 \\ 0, \quad \text{if } z \le 0 \text{ or } z \ge 1. \end{cases}$$

6.5 How to evaluate $\mathbb{E}(g(X))$

Given a continuous random variable X with pdf $f_X(x)$, how to find $\mathbb{E}(g(X))$ without evaluating the pdf of $Y = g(X)$?

For any random variable X, consider its discretization

$$X_n := \frac{2k-1}{2n} \quad \text{if and only if} \quad \frac{k-1}{n} \le X < \frac{k}{n},$$

where $k = \ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$. Then as long as n goes to infinity, we have (nonrigorously)

$$\mathbb{E}(g(X_n)) \to \mathbb{E}(g(X)).$$

On the other hand, we know

$$\begin{aligned}\mathbb{E}(g(X_n)) &= \sum_{k=-\infty}^{\infty} g\left(\frac{2k-1}{2n}\right) p_{X_n}\left(\frac{2k-1}{2n}\right) \\ &= \sum_{k=-\infty}^{\infty} g\left(\frac{2k-1}{2n}\right) \mathbb{P}\left(\frac{k-1}{n} \leq X < \frac{k}{n}\right) \\ &\approx \sum_{k=-\infty}^{\infty} g\left(\frac{2k-1}{2n}\right) f_X\left(\frac{2k-1}{2n}\right) \frac{1}{n} \\ &\to \int_{-\infty}^{\infty} g(x) f_X(x) dx.\end{aligned}$$

Therefore, we have

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Example

If $X \sim \text{Unif}(0,1)$, then

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_0^1 x^2 dx = \frac{1}{3}.$$

6.6 Variances

For any continuous random variable X, as with discrete cases, define

$$\text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right]$$

Let $\mu_X := \mathbb{E}(X)$. We have

$$\text{Var}(X) = \mathbb{E}\left[(X - \mu_X)^2\right] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx.$$

Furthermore,

$$\begin{aligned}\int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx &= \int_{-\infty}^{\infty} (x^2 - 2x\mu_X + \mu_X^2) f_X(x) dx \\ &= \int_{-\infty}^{\infty} x^2 f_X(x) dx - 2\mu_X \int_{-\infty}^{\infty} x f_X(x) dx + \mu_X^2 \int_{-\infty}^{\infty} f_X(x) dx \\ &= \mathbb{E}[X^2] - 2\mu_X \mathbb{E}[X] + \mu_X^2 \\ &= \mathbb{E}[X^2] - (E[X])^2.\end{aligned}$$

Then we have another formula

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

Example

If $X \sim \text{Unif}(0,1)$, then

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

6.7 Basic Properties of Expectations and Variances

Now we are ready to introduce the following property

Theorem 6.7.1 *Let X be a discrete or continuous random variable. Let $Y = aX + b$ where a and b are deterministic. Then*

$$\mathbb{E}[Y] = a\,\mathbb{E}[X] + b, \quad Var(Y) = a^2\,Var(X).$$

Proof

X is discrete

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_{x:p_X(x)>0} (ax+b)p_X(x) \\
&= a \sum_{x:p_X(x)>0} x p_X(x) + b \sum_{x:p_X(x)>0} p_X(x) \\
&= a\,\mathbb{E}[X] + b.
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}[Y] &= \sum_{x:p_X(x)>0} [(ax+b) - (a\,\mathbb{E}[X]+b)]^2 p_X(x) \\
&= \sum_{x:p_X(x)>0} a^2 (x - \mathbb{E}[X])^2 p_X(x) \\
&= a^2 \sum_{x:p_X(x)>0} (x - \mathbb{E}[X])^2 p_X(x) \\
&= a^2\,\mathbb{E}(X - \mathbb{E}[X])^2 \\
&= a^2 \mathrm{Var}[X].
\end{aligned}$$

X is continuous

$$\begin{aligned}
\mathbb{E}[Y] &= \int_{-\infty}^{\infty} (ax+b) f_X(x)dx \\
&= a \int_{-\infty}^{\infty} x f_X(x)dx + b \int_{-\infty}^{\infty} f_X(x)dx \\
&= a\,\mathbb{E}[X] + b.
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}[Y] &= \int_{-\infty}^{\infty} [(ax+b) - (a\,\mathbb{E}[X]+b)]^2 f_X(x)dx \\
&= \int_{-\infty}^{\infty} a^2 (x - \mathbb{E}[X])^2 f_X(x)dx \\
&= a^2 \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 f_X(x)dx \\
&= a^2\,\mathbb{E}(X - \mathbb{E}[X])^2 \\
&= a^2 \mathrm{Var}[X].
\end{aligned}$$

■

Notice that we have explained that as long as $X_1, \ldots, X_n$ are discrete and defined on the same sample space, then

$$\mathbb{E}[X_1 + \ldots + X_n] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n].$$

This is also true when they are continuous random variables (or even partially discrete and partially continuous). Since it is difficulty to explain the sample space for continuous random variables, we will explain this in the chapter of "joint distribution".

6.8 Exponential random variables

We say $X \sim \text{Exp}(\lambda)$, if its density is

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

One can easily obtain

$$\int_{-\infty}^{\infty} f_X(x)dx = \int_0^{\infty} \lambda e^{-\lambda x} dx = 1,$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx = \int_0^{\infty} \lambda x e^{-\lambda x} dx = \frac{1}{\lambda},$$

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)dx = \int_0^{\infty} \lambda x^2 e^{-\lambda x} dx = \frac{2}{\lambda^2},$$

and

$$\text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{1}{\lambda^2}.$$

cumulative distribution function

If $x < 0$

$$F_X(x) = \int_{-\infty}^{x} 0du = 0.$$

If $x > 0$

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du = \int_0^{x} \lambda e^{-\lambda u} du = 1 - e^{-\lambda x}.$$

Memoryless property

If the life time of some battery is exponentially distributed and we know with probability 0.75 it will last for at least 1 year. Given the knowledge it has been in effect for 2 years, what is the probability that it can be used for another year?

Answer:

$$\begin{aligned} \mathbb{P}(X \geq 3 | X \geq 2) &= \frac{\mathbb{P}(\{X \geq 3\} \cap \{X \geq 2\})}{\mathbb{P}(X \geq 2)} = \frac{\mathbb{P}(X \geq 3)}{\mathbb{P}(X \geq 2)} = \frac{1 - F_X(3)}{1 - F_X(2)} \\ &= \frac{e^{-3\lambda}}{e^{-2\lambda}} = e^{-\lambda} = 1 - F_X(1) = P(X \geq 1) = 0.75. \end{aligned}$$

6.9 Normality

We say $X \sim \mathcal{N}(\mu, \sigma^2)$, if its pdf is

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

If $X \sim \mathcal{N}(0,1)$ with pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} := \phi(x),$$

we say X is a standard normal random variable, and its cdf is denoted as

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du := \Phi(x).$$

Proposition 6.9.1 *If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Y := \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.*

Proof Since

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}\left(\frac{X-\mu}{\sigma} \le y\right) = \mathbb{P}\left(X \le \sigma y + \mu\right) = F_X(\sigma y + \mu),$$

we have

$$f_Y(y) = F_Y'(y) = \frac{d}{dy} F_X(\sigma y + \mu) = F_X'(\sigma y + \mu)\sigma = \sigma f_X(\sigma y + \mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}.$$

■

Proposition 6.9.2 *If $X \sim \mathcal{N}(0,1)$, then $Y := \sigma X + \mu \sim \mathcal{N}(\mu, \sigma^2)$.*

Proof Since

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}\left(\sigma X + \mu \le y\right) = \mathbb{P}\left(X \le \frac{y-\mu}{\sigma}\right) = F_X\left(\frac{y-\mu}{\sigma}\right),$$

we have

$$f_Y(y) = F_Y'(y) = \frac{d}{dy} F_X\left(\frac{y-\mu}{\sigma}\right) = F_X'\left(\frac{y-\mu}{\sigma}\right)\frac{1}{\sigma} = \phi\left(\frac{y-\mu}{\sigma}\right)\frac{1}{\sigma} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

■

Proposition 6.9.3 *If X is a normal random variable, then $Y = aX + b$ is a normal random variable.*

Expectations and variances of normal random variables

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} \frac{x}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\
&= \int_{-\infty}^{\infty} \frac{\sigma y + \mu}{\sigma\sqrt{2\pi}} e^{-\frac{y^2}{2}} d(\sigma y + \mu) \quad (x := \sigma y + \mu) \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma y + \mu) e^{-\frac{y^2}{2}} dy \\
&= \frac{1}{\sqrt{2\pi}} \left(\sigma \int_{-\infty}^{\infty} y e^{-\frac{y^2}{2}} dy + \mu \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy\right) \\
&= \frac{1}{\sqrt{2\pi}} \left(\sigma \left(-e^{-\frac{y^2}{2}}\right)\Big|_{-\infty}^{\infty} + \mu \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy\right) = \mu.
\end{aligned}$$

$$
\begin{aligned}
\mathrm{Var}[X] &= \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\
&= \int_{-\infty}^{\infty} \frac{\sigma^2 y^2}{\sigma\sqrt{2\pi}} e^{-\frac{y^2}{2}} d(\sigma y + \mu) \quad (x := \sigma y + \mu) \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-\frac{y^2}{2}} dy \\
&= -\frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y de^{-\frac{y^2}{2}} \\
&= -\frac{\sigma^2}{\sqrt{2\pi}} y\, e^{-\frac{y^2}{2}} \Big|_{-\infty}^{\infty} + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy = \sigma^2.
\end{aligned}
$$