

Mathematical Statistics

Mingheng Su

April 2025

Contents

1	Probability	3
1.1	Introduction	3
1.2	Conditional Probability	5
1.3	Random Variables	9
2	Univariate Distributions	12
2.1	Probability Functions	12
2.2	Expectation and Variance	16
2.3	Moment Generating Functions	26
2.4	Distribution Transformation	28
3	Multivariate Distributions	31
3.1	Probability Functions	31
3.2	Independence	33
3.3	Expectation and Covariance	35
3.4	Moment Generating Functions	39
3.5	Max and Min Distributions	40
3.6	Conditional Distributions	42
4	Known Distributions	50
4.1	Normal Distribution	50
4.2	Gamma Distribution	53
4.3	Beta Distribution	55
4.4	Chi Squared Distribution	55
4.5	Student t Distribution	55
5	Known Multivariate Distributions	56
5.1	Multinomial Distribution	56
5.1.1	Definitions	56
5.1.2	MGF	57
5.1.3	Marginals	58
5.1.4	Conditional Multinomial Distribution	59
5.2	Multivariate Uniform Distribution	61
5.3	Multivariate Normal Distribution	62
5.3.1	Definitions	62
5.3.2	Stochastic Representations	62
5.3.3	Independence of Marginals	65
5.3.4	MGF	65
5.3.5	Covariance	67

5.3.6	Conditional MVN	68
5.4	Complex Distributions	69
6	Limiting (Asymptotic) Distributions	71
6.1	Convergence in Probability	71
6.2	Convergence in Distribution	71
6.3	Central Limit Theorem (CLT)	72
7	Maximum Likelihood Estimation	76
7.1	Introduction	76
7.2	Maximum Likelihood Estimate	77
A	Calculus	81
A.1	Fundamentals	81
A.2	Change of Variable	83
A.3	Taylor Series	86
A.4	Optimization	89
B	Linear Algebra	91
C	Indices	93
D	References	96

1 Probability

1.1 Introduction

Suppose that we conduct some experiments with randomness, such as tossing a coin, a die, buying lotteries, or spinning the lucky turntable. We will analyze the experiments in the following way.

Definition 1.1.1 (Sample Spaces)

The sample space Ω of an experiment is the collection of all possible outcomes.

The outcomes outside sample space will never happen (i.e. with probability 0). For example, if we toss a fair die, the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$, and each number has a probability of $\frac{1}{6}$ to occur.

Definition 1.1.2 (Events)

An event A is a subset of outcomes in sample space Ω .

In the previous example, we can define events $O = \{1, 3, 5\}$ to be the outcomes of tossing odd numbers and $E = \{2, 4, 6\}$ to be the outcomes of tossing even numbers. Note that $O, E \subset \Omega$. As well, we can define $A = \{1, 2, 3\}$ to be the set of outcomes less than or equal to 3, and $B = \{3, 4, 5\}$ to be those greater than 2. If an event contains outcomes that are not in the sample space, we consider that it will not happen.

Definition 1.1.3 (Disjoint Events)

Two events $A, B \subseteq \Omega$ are disjoint if and only if $A \cap B = \emptyset$.

Clearly, $O = \{1, 3, 5\}$ and $E = \{2, 4, 6\}$ are disjoint events since they do not have elements in common. However, $A = \{1, 2, 3\}$ and $B = \{3, 4, 5\}$ are not disjoint events since $A \cap B = \{3\} \neq \emptyset$.

Definition 1.1.4 (Probability Functions)

The probability function $P(\cdot)$ is a numerical measurement of the likelihood of some event. It maps an event to a value between 0 and 1, and a larger probability means the it is more likely to happen.

The likelihood of event O , $P(O)$, is equal to $\frac{3}{6} = 0.5$ in the context, because it takes $\{1, 3, 5\}$ which are 3 outcomes in the sample space $\{1, 2, 3, 4, 5, 6\}$. One can understand $P(O)$ as the probability of tossing 1 or 3 or 5, and in a broader way, for discrete sample space Ω , $P(A)$ is the proportion of outcomes that event A takes from Ω , namely

$$P(A) = \frac{|A|}{|\Omega|}$$

We go through some examples of calculating probabilities. When facing combinatorics problems such as seating, ordering, or selections, always first find out the sample space, and count the number of qualified outcomes to compute the probability.

Example 1.1.1 (Seats Selection)

Adam and Ben are invited to watch a movie. They can select any two seats in the first row. Due to some reason, they cannot sit adjacent to each other. Suppose that the first row has 14 seats, what is the probability that there are at least two seats between them?

Solution

We consider two steps: the distribution of the two friends and the order of them. First consider the distribution of them. Some distributions are like

— — ★ — — ★ ★ — ★ — — — ★ — — — — — ★ ...

Note that we did not specify which star is who. This is equivalent to choose 2 seats from the 14 seats, where there are a total of $\binom{14}{2} = 91$ choices. Note that some of them are not eligible, since they cannot sit adjacent to each other. The adjacent choices are like

$\star\star-----\star\star----\star\star---\dots-----\star\star$

These totals $14 - (2 - 1) = 13$ choices, which should be excluded from the 91 choices. Thus the sample space has size $|\Omega| = 91 - 13 = 78$. Our target is to find the number of choices that they sit at least two seats away from other (denoted event A), namely subtracting the number of choices that they sit one seat away from each other (denoted \bar{A}). These are like

$\star-\star-----\star-\star---\dots-----\star-\star$

There are $|\bar{A}| = 14 - (3 - 1) = 12$ choices in this case. Thus the target is $|A| = 78 - 12 = 66$ choices. We also need to count the order. For any distribution of seats, there are $2!$ orders for Adam and Ben. So the sizes of ordered sample space and events could be evaluated via $|S| = 2|\Omega|, |B| = 2|A|$. The probability is then computed as

$$P(B) = \frac{|B|}{|S|} = \frac{2(66)}{2(78)} \approx 0.85$$

We can see that the order does not affect the probability, which can be canceled out in the fraction.

Proposition 1.1.5 (Properties of Probability Function)

For sample space Ω ,

1. $P(\emptyset) = 0$ and $P(\Omega) = 1$.
2. $0 \leq P(A) \leq 1$ for any event $A \subseteq \Omega$.
3. $P(A) \leq P(B)$ for $A \subseteq B \subseteq \Omega$.
4. $P(A) = 1 - P(\bar{A})$, where \bar{A} is the complement of event A .

Note that the result will only be one of the elements in Ω , that is, the probability that tossed number is in Ω , $P(\Omega)$, is 1. Also notice that $P(O) + P(E) = \frac{1}{2} + \frac{1}{2} = 1 = P(\Omega)$, because E is the complement set of O .

Proposition 1.1.6 (Probability of Union of Events)

Let A, B be two events of Ω . Then

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

This also leads to the following result.

Corollary 1.1.7 (Probability of Union of Disjoint Events)

Let A, B be two disjoint events of Ω . Then

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

Proof.

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

□

We can extend this statement to a general case. Let events B_1, \dots, B_n be n **disjoint** decompositions of event A such that

$$\bigcup_{i=1}^n B_i = A \quad \text{and} \quad B_i \cap B_j = \emptyset, \forall i \neq j \in \{1, \dots, n\}$$

Then

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

Let $B_1 = \{1\}, B_2 = \{2, 3, 4\}, B_3 = \{5, 6\}$ be 3 disjoint decompositions of sample space $\Omega = \{1, \dots, 6\}$, where $P(B_1) = \frac{1}{6}, P(B_2) = \frac{1}{2}, P(B_3) = \frac{1}{3}$. We can clearly see that $P(\Omega) = P(B_1) + P(B_2) + P(B_3) = \frac{1}{6} + \frac{1}{2} + \frac{1}{3} = 1$.

Proposition 1.1.8 (Probability of Independent Events)

For independent events A, B ,

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

This is often leaded by two independent experiments, and the sample space is a crossed set $\Omega = \{(A, B)\}$, where A, B refer to the experiments respectively. For example, toss a coin and a die independently, the sample space will be $\Omega = \{(Head, 1), \dots, (Head, 2), \dots, (Tail, 6)\}$. Independence means their behavior does not affect each other, as in example, there is no relationship between obtaining a head and 6.

What if the sample space is continuous? Suppose that we flip the pointer of a clock, and the pointer will stop at some random position, pointing to a number within $[0, 12]$. How would you guess the number it points to when it stops?

There is no way you can guess the resulting number if we count it like the discrete sample space, since there are infinite numbers on the clock. However, it makes sense to oversee the interval that the pointer ultimately lies in, for instance, $[0, 3]$, the right above quarter of the clock, which has probability $\frac{1}{4}$. This is the idea of discretizing the continuous domain. We can find that the probability of lying within $[a, b]$ where $0 \leq a \leq b \leq 12$ is

$$P([a, b]) = \frac{b - a}{12 - 0} \tag{Clock}$$

As well, for disjoint intervals $[a, b], [c, d]$,

$$P([a, b] \cup [c, d]) = P([a, b]) + P([c, d]) = \frac{(b - a) + (d - c)}{12}$$

Or for the joint intervals, such as $[3, 5], [4, 7]$ the probability will be

$$P([3, 5] \cup [4, 6]) = P([3, 7]) = \frac{7 - 3}{12} = \frac{1}{3}$$

A more rigorous definition of continuous distribution will be shown in later chapters.

1.2 Conditional Probability

The conditional sample space is the set of possible outcomes given that the conditional event has happened. Again with the die example, but this time someone else sees the result of the die and tells you whether it is even or odd, and then you make a guess based on the given information. What would you guess after your friend tells you that the result is odd? You can reasonably guess 1, 3, 5, each with $\frac{1}{3}$ chance to be correct. Why are they no longer $\frac{1}{6}$? Because the sample space has shrink to $\Omega' = \{1, 3, 5\}$ and you need not guess other numbers, yielding a larger probability for the correct guess.

Definition 1.2.1 (Conditional Probability)

The probability of event A given that B has happened is

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

Note that $0 \leq P(A|B) \leq 1$. It denotes the probability of A **after** B has occurred, thus the outcome of the conditional experiment will be an element of $A \cap B$. Also note that

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

It explains the relationship between the conditional probability and joint probability. Let's see how this works. Recall that $P(\text{die outcome is } 5) = \frac{1}{6}$ and $P(\text{die outcome is } 5|\text{outcome is odd}) = \frac{1}{3}$, then

$$\begin{aligned} P(A|B)P(B) &= P(\text{die outcome is } 5|\text{outcome is odd})P(\text{die outcome is odd}) = \frac{1}{3} \cdot \frac{1}{2} \\ &= P(\text{die outcome is } 5, \text{ outcome is odd}) = \frac{1}{6} \\ &= P(\text{die outcome is } 5) \end{aligned}$$

where we have obtained $\frac{1}{6}$, the true probability of obtaining 5. In this example, we can see that the probability of obtaining 5 is $P(\text{die outcome is } 5|\text{outcome is odd})$ multiplied by the likelihood of obtaining an odd number, which is $\frac{1}{2}$. This example better explains the conditional sample space,

Example 1.2.1 (Boy or Girl?)

Suppose a family has two children, given that one of them is a boy, what is the probability that the other child is a girl?

Solution

The wording is tricky. Intuitively, the probability should be 0.5, since the sexes of two children are independent of each other. However, it didn't specify which child is a boy, who might be the first child, or second. The conditional sample space is

$$\{(B, G), (B, B), (G, B)\}$$

Clearly, the probability is $\frac{2}{3}$.

For independent events A, B ,

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

such that the occurrence of B does not influence A . In the last example, if the question states that, for example, the first child is a boy, then the second child's sex is independent of the first, which is 0.5 for both sexes, denoted by

$$P(\text{second child is a girl}|\text{first child is a boy}) = P(\text{second child is a girl}) = 0.5$$

Theorem 1.2.2 (Law of Total Probability)

Let events B_1, \dots, B_n be n **complete** and **disjoint** decompositions of sample space Ω such that

$$\bigcup_{i=1}^n B_i = \Omega \quad \text{and} \quad B_i \cap B_j = \emptyset, \forall i \neq j \in \{1, \dots, n\} \tag{*}$$

Let A be an event of Ω . Then

$$\begin{aligned}
 P(A) &= P(A \cap \Omega) \\
 &= \sum_{i=1}^n P(A \cap B_i) \\
 &= \sum_{i=1}^n P(A|B_i)P(B_i)
 \end{aligned}
 \tag{**}$$

$$\tag{***}$$

The (**) states that marginal probability of $P(A)$ is $\sum_{i=1}^n P(A \cap B_i)$, B_i defined as in (*), namely summing over all related B conditions. When talking about $P(A)$, we do not care about B and any outcome of B counts. For example, we independently toss a coin and roll a die. Let A denote the event of tossing a head of the coin, $A = \{\text{Head}\}$, and $B_1 = \{1, 3, 5\}$, $B_2 = \{2, 4, 6\}$ denote the odd and even outcomes of the die. By independence $P(A \cap B_1) = 0.25$, $P(A \cap B_2) = 0.25$. For $P(A)$, we don't worry about what outcome of die is, so just sum them up to get $P(A) = 0.5$, which matches the theoretical probability of tossing a head.

Example 1.2.2 (Law of Total Probability 1)

Suppose that we have three bags, A, B, C . A has 3 Green and 2 Red balls, B has 2 Green and 4 Red balls, C has 1 Green and 3 Red balls. If we randomly choose a bag and draw 2 balls from the bag (blindly), what is the probability drawing 1 Green and 1 Red balls?

Solution

We first consider the probability of drawing balls from a bag. Suppose the bag has x Green and y Red balls, and the probability of drawing a Green and b Red balls, where $a \leq x, b \leq y$, is

$$P(\text{drawing } a \text{ Green and } b \text{ Red}) = \frac{\binom{x}{a} \binom{y}{b}}{\binom{x+y}{a+b}}$$

Note that this is an instance of **Hypergeometric Distribution**. $\binom{x}{a}$ is the number of ways to choose a balls from x the Green balls, $\binom{y}{b}$ is the number of ways to choose b balls from the y Red balls, $\binom{x}{a} \binom{y}{b}$ is the total number of ways to choose 1 Green and 1 Red balls, and $\binom{x+y}{a+b}$ is the total number of ways to choose any two balls from the overall pool which are $x+y$ balls. In this case,

$$\begin{aligned}
 P(\text{draw 1 Green 1 Red} \mid A \text{ is chosen}) &= \frac{\binom{3}{1} \binom{2}{1}}{\binom{5}{2}} = \frac{3}{5} \\
 P(\text{draw 1 Green 1 Red} \mid B \text{ is chosen}) &= \frac{\binom{2}{1} \binom{4}{1}}{\binom{6}{2}} = \frac{8}{15} \\
 P(\text{draw 1 Green 1 Red} \mid C \text{ is chosen}) &= \frac{\binom{1}{1} \binom{3}{1}}{\binom{4}{2}} = \frac{1}{2}
 \end{aligned}$$

Therefore by **Law of Total Probability**,

$$\begin{aligned}
 P(\text{draw 1 Green 1 Red}) &= P(\text{draw 1 Green 1 Red} \mid A \text{ is chosen})P(A \text{ is chosen}) + \\
 &\quad P(\text{draw 1 Green 1 Red} \mid B \text{ is chosen})P(B \text{ is chosen}) + \\
 &\quad P(\text{draw 1 Green 1 Red} \mid C \text{ is chosen})P(C \text{ is chosen}) \\
 &= \frac{3}{5} \frac{1}{3} + \frac{8}{15} \frac{1}{3} + \frac{1}{2} \frac{1}{3} \\
 &\approx 0.544
 \end{aligned}$$

Example 1.2.3 (Law of Total Probability 2)

Suppose that bag A has 3 white and 4 red balls and bag B has 1 white and 2 red balls. Now we randomly draw two balls from A and put them into B , and then randomly draw two balls from B . What is the probability of drawing 2 red balls?

Solution

There are two steps: draw two balls from A and put into B , and then draw two balls from B . We need to consider how the first event would affect the second one. Let x denote the number of red balls drawn from A ^a, which has the following cases: $x = 0, 1, 2$. Let y denote the number of red balls drawn from B , which also has cases $y = 0, 1, 2$. Condition on x , note that $P(\text{draw } x \text{ red from } A) = \frac{\binom{4}{x} \binom{3}{2-x}}{\binom{7}{2}}$, $x = 0, 1, 2$, and

$$P(x = 0) = \frac{\binom{4}{0} \binom{3}{2}}{\binom{7}{2}} = \frac{1}{7}$$

$$P(x = 1) = \frac{\binom{4}{1} \binom{3}{1}}{\binom{7}{2}} = \frac{4}{7}$$

$$P(x = 2) = \frac{\binom{4}{2} \binom{3}{0}}{\binom{7}{2}} = \frac{2}{7}$$

We see that they sum to 0, since these are the complete conditions. Now consider that x is given, i.e. fixed^b. Note that y conditioning on x has probability $P(\text{draw } y \text{ red from } B \mid \text{given } x \text{ red from } A) = \frac{\binom{2+x}{y} \binom{1+(2-x)}{2-y}}{\binom{3+x}{2}}$, $y = 0, 1, 2$, then

$$P(y = 2|x) = \frac{\binom{2+x}{2} \binom{3-x}{0}}{\binom{3+x}{2}} = \frac{\binom{2+x}{2}}{\binom{3+x}{2}} \quad (\text{Target Probability})$$

By Law of Total Probability,

$$\begin{aligned} P(y = 2) &= \sum_{x=0}^2 P(y = 2|x)P(x) \\ &= \frac{\binom{2+0}{2}}{\binom{3+0}{2}} \frac{1}{7} + \frac{\binom{2+1}{2}}{\binom{3+1}{2}} \frac{4}{7} + \frac{\binom{2+2}{2}}{\binom{3+2}{2}} \frac{2}{7} \quad \text{by substituting in } x \\ &= \frac{1}{3} \frac{1}{7} + \frac{3}{6} \frac{4}{7} + \frac{6}{10} \frac{2}{7} \\ &\approx 0.5 \end{aligned}$$

^aagain this is also a hypergeometric distribution

^bneed not worry about the x right now, just use it as a number

Warning

Do not confuse conditional probability $P(A|B)$ with joint probability $P(A, B)$, where the later one refers to the probability that A, B happen together, while conditional probability assumes that A happens after B . Sometimes it is easier to derive the conditional probability and then re-generate the joint probability.

Theorem 1.2.3 (Bayes' Theorem)

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

A generalized version is

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^n P(B|A_i)P(A_i)}, i = 1, \dots, n \quad (\text{Bayes})$$

for complete and disjoint decompositions A_1, \dots, A_n of A .

Example 1.2.4 (Infection Probability)

Suppose the probability of getting some infection is 0.02. The infection can be detected by a test, such that if a person has this infection, the probability of being tested positive (true positive) is 0.9, while if the person does not have the infection, the probability of being tested positive is 0.025 (false positive). If given that someone is tested positive, what is the probability of having the disease?

Solution

Let A denote the event that the person has the disease, and B the event of being tested positive. The target probability is

$$P(A|B)$$

The known probabilities are

$$P(A) = 0.02, P(B|A) = 0.9, P(B|\bar{A}) = 0.025$$

We first need to know $P(B)$. By **Law of Total Probability**,

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) = 0.9(0.02) + 0.025(1 - 0.02) = 0.0425$$

By **Bayes' Theorem**,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.9(0.02)}{0.0425} \approx 0.42$$

It seems that the test is not fully reliable.

1.3 Random Variables

We typically use a variable to denote the *outcome* of an experiment, known as the **Random Variable**. It is not a fixed number, but will generate outputs based on corresponding probabilities. The outcomes of a discrete experiment could be numbers, colors, words, etc., while for convenience, we map all outcomes to be real numbers.

Definition 1.3.1 (σ -algebra)

Let $\mathcal{P}(\Omega)$ denote the power set of sample space Ω . The σ -algebra \mathcal{B} satisfies

1. $\emptyset \in \mathcal{B}$ and $\Omega \in \mathcal{B}$.
2. If $A \in \mathcal{B}$ then $\bar{A} \in \mathcal{B}$.
3. If $A_1, \dots, A_n \in \mathcal{B}$ then $\cup_{i=1}^n A_i \in \mathcal{B}$.

Definition 1.3.2 (Probability Space)

A probability measure with respect to σ -field \mathcal{B} , denoted by $P : \mathcal{B} \rightarrow [0, 1]$, satisfy

1. $P(A) \geq 0, \forall A \in \mathcal{B}$.

2. $P(\Omega) = 1$.

3. For disjoint events $A_1, A_2, \dots \in \mathcal{B}$,

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

The **Probability Space** is defined as $\mathbb{P} = (\Omega, \mathcal{B}, P)$.

We can say that a random variable X with probability function P belongs to the probability space induced by P , namely $X \in \mathbb{P}$, not \mathbb{R} . The formal definition of random variable is

Definition 1.3.3 (Random Variables)

A random variable $X : \Omega \rightarrow \mathbb{R}$ has

$$P(X \leq x) = P(\{w \in \mathcal{B} : X(w) \leq x\})$$

defined $\forall x \in \mathbb{R}$.

In fact the random variable just maps the outcomes of a distribution to numerical values. We go through some examples.

We are tossing 10 identical coins independently, where each coin has a 0.3 probability of tossing a head, a.k.a success probability. Let X denote the number of heads. What is the probability of having 4 heads, namely $P(X = 4)$? Consider the following eligible example

1, 1, 1, 1, 0, 0, 0, 0, 0, 0

where 1 denotes the head and 0 the tail. The previous four trials are heads, while the rest are tails. Since they are independent, the probability is thus

$$P(\text{Head})P(\text{Head})P(\text{Head})P(\text{Head}) \underbrace{P(\text{Tail}) \cdots P(\text{Tail})}_{6 \text{ } P(\text{Tail})\text{'s}} = P(\text{Head})^4 P(\text{Tail})^6 = 0.3^4 0.7^6$$

where 0.3^4 is the probability of obtaining 4 heads, and 0.7^6 is the probability of obtaining 6 tails. Note that it can also be written as

$$0.3^4(1 - 0.3)^{10-4}$$

However, we didn't specify which 4 coins are heads. That is, this could also work

1, 0, 1, 1, 1, 0, 0, 0, 0, 0

or this

1, 0, 1, 0, 1, 0, 1, 0, 0, 0

even this

1, 1, 0, 0, 0, 0, 0, 0, 1, 1

Although all of them have the same probability $0.3^4(1 - 0.3)^{10-4}$, it is impossible to enumerate each of them and count how many are there. Interestingly, this is equivalent to choosing 4 positions from the total 10 positions, namely

$$\binom{10}{4} = 210$$

That said, there are 210 cases that have 4 heads and 6 tails. The total probability is thus

$$P(X = 4) = \binom{10}{4} 0.3^4(1 - 0.3)^{10-4} \approx 0.2$$

Definition 1.3.4 (Binomial Distribution)

A RV X follows binomial distribution with n trials and success probability p , denoted by $X \sim Bin(n, p)$, if X has probability function

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, \dots, n$$

Note that the cases $X = x$ where $x = 0, 1, \dots, n$ are *disjointly* and *completely* decomposed events, thus their probabilities should sum to 1. Note that by *Binomial Theorem*,

$$\sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} = (p + (1-p))^n = 1$$

which verifies our hypothesis.

Suppose you are waiting for a bus at a bus station, and the average waiting time for a bus is 5 minutes. What is the probability of having 3 buses coming?

Definition 1.3.5 (Poisson Distribution)

X follows Poisson distribution with mean λ , denoted by $X \sim Poi(\lambda)$, if

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, x = 0, 1, \dots$$

2 Univariate Distributions

2.1 Probability Functions

Discrete Distributions For a random variable X with discrete sample space Ω , its **Probability Mass Function (PMF)** is defined as

$$P_X(x) = P(X = x) = \begin{cases} p_x & \text{if } x \in \Omega \\ 0 & \text{otherwise} \end{cases} \quad (\text{PMF})$$

Now assume it has numerical sample space^{2.1}. Its **Cumulative Distribution Function (CDF)** is

$$F_X(x) = P(X \leq x) = \sum_{-\infty}^x P(X = x) \quad (\text{CDF})$$

And similarly its **Tailed Probability Function (TPF)**^{2.2} is defined as

$$\bar{F}_X(x) = P(X > x) = 1 - P(X \leq x) \quad (\text{TPF})$$

Example 2.1.1 (Discrete Distribution Example 1)

Suppose we toss a fair coin 5 times independently. Let X denote the number of heads tossed, which follows a $\text{Bin}(5, 0.5)$ distribution. The sample space is

$$\Omega = \{0, 1, 2, 3, 4, 5\}$$

which is discrete, and the general probability mass function is

$$P(X = x) = \binom{5}{x} 0.5^x 0.5^{5-x}, x = 0, \dots, 5$$

Note that $\binom{5}{x}$ is the number of ways to toss x heads out of the total 5 tosses, 0.5^x is the probability of tossing x heads, and 0.5^{5-x} is the probability that the rest $5 - x$ tosses are tails. Therefore,

$$P(X = x) = \begin{cases} 0.03125 & \text{if } x = 0 \\ 0.15625 & \text{if } x = 1 \\ 0.3125 & \text{if } x = 2 \\ 0.3125 & \text{if } x = 3 \\ 0.15625 & \text{if } x = 4 \\ 0.03125 & \text{if } x = 5 \\ 0 & \text{otherwise} \end{cases}$$

Note that

$$\sum_{x=0}^5 P(X = x) = 1$$

where since X must be a result in $\{0, 1, \dots, 5\}$, the probability of X being in the set is 1. Also note that all probabilities must be between 0 and 1.

^{2.1}map to numerical data if not

^{2.2}sometimes called *Survival Function*

For cumulative distribution function, we just "stack" the probabilities until step x , such that

$$P(X \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ 0.03125 & \text{if } 0 \leq x < 1 \\ 0.1875 & \text{if } 1 \leq x < 2 \\ 0.5 & \text{if } 2 \leq x < 3 \\ 0.8125 & \text{if } 3 \leq x < 4 \\ 0.96875 & \text{if } 4 \leq x < 5 \\ 1 & \text{if } x \geq 5 \end{cases}$$

Note that

$$P(X = x) = P(X \leq x) - \lim_{h \rightarrow x^-} P(X \leq h)$$

such that the function $P(X \leq x)$ "updates" only at points x where $P(X = x) > 0$.

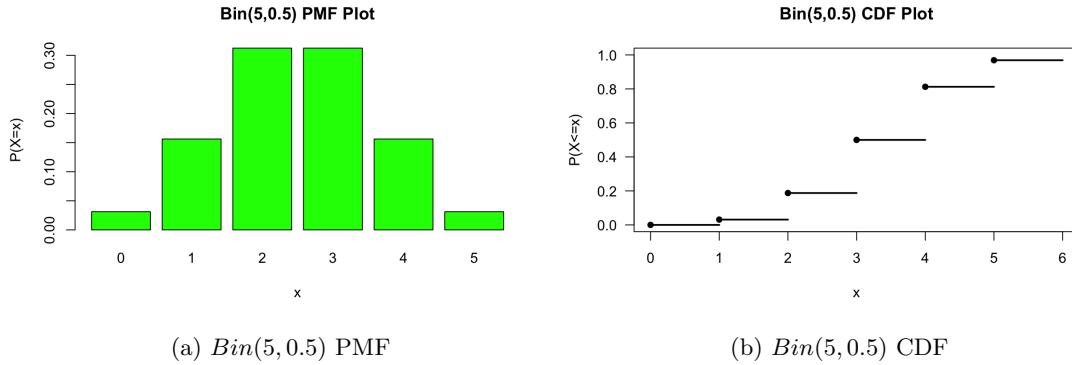


Figure 2.1: $\text{Bin}(5, 0.5)$ Probability Functions

Some properties follow,

Proposition 2.1.1 (Properties of Probability Mass Function)

Suppose discrete random variable X has probability mass function $P(X = x)$. Then

1. $\sum_{x=-\infty}^{\infty} P(X = x)^a = 1$
 2. $0 \leq P(X = x) \leq 1, \forall x \in \mathbb{R}$
-
- ^aThis is the same as $P(X < \infty)$.

Example 2.1.2

The Poisson distribution with mean λ , denoted by $X \sim \text{Poi}(\lambda)$, has PMF

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, x = 0, 1, \dots$$

Show that

$$\sum_{x=-\infty}^{\infty} P(X = x) = 1$$

Solution

See [Taylor Series](#) for more information.

$$\sum_{x=-\infty}^{\infty} P(X = x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda} \lambda^x}{x!} = e^{-\lambda} \underbrace{\sum_{x=0}^{\infty} \frac{\lambda^x}{x!}}_{e^{\lambda}} = 1$$

Continuous Distributions Recall that the **Cumulative Distribution Function** for some random variable X is defined as $F_X(x) = P(X < x)$. Continuous distributions have continuous CDF as their characteristic function, such as in the ([Clock](#)) example, its CDF is in fact

$$P(X < x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{x-0}{12-0} = \frac{x}{12} & \text{if } 0 \leq x \leq 12 \\ 1 & \text{if } x > 12 \end{cases} \quad (\text{Clock CDF})$$

where X refers to be the number that the pointer points to. As mentioned before, for some continuous random variable X defined on some continuous sample space, it makes no sense to have PMF $P(X = x)$ which is 0 for all x over domain. However, also as we have said that the probability is defined for X within some interval such as $P(a < X < b)$ ^{2.3}, if we consider to let X fall in some sufficiently small interval adjacent to x , then X is approximately at x , and is infinitely close to x if the interval is infinitely small. Let dx be an infinitely small increment in x , we can define its "abstract $P(X = x)$ " via

$$f_X(x) = \frac{P(x < X < x + dx)}{dx} \quad (\text{PDF})$$

Note that it is not $P(X = x)$, but is referred to be the **Probability Density Function (PDF)** of X . Note that $f_X(a), a \in \mathbb{R}$ is the instantaneous rate of change of $F_X(x)$ at $x = a$, and in other words, $f_X(x)$ is the derivative of $F_X(x)$, such that

$$f_X(x) = \lim_{dx \rightarrow 0} \frac{F_X(x + dx) - F_X(x)}{dx}$$

$f_X(a)dx$ is the probability mass added to the CDF as $x = a$ increases to $a + dx$, or equivalently it is the probability of X being in the interval $[a, a + dx]$. From calculus we know that

$$f_X(x) = \frac{d}{dx} F_X(x) \quad F_X(x) = \int_{-\infty}^x f_X(t) dt \quad (\text{PDF} \rightleftharpoons \text{CDF})$$

Refer to [Fundamental Theorem of Calculus 1](#) for more information. For example, we have that the PDF of ([Clock](#)) is

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx} \frac{x}{12} = \begin{cases} \frac{1}{12} & \text{if } 0 \leq x \leq 12 \\ 0 & \text{otherwise} \end{cases}$$

Observe that it follows a continuous uniform distribution, that is, $X \sim \text{Uni}(0, 12)$.

Proposition 2.1.2 (Properties of Probability Density Function)

Suppose continuous random variable X has probability density function $f_X(x)$. Then

1. $f_X(x)$ is continuous or piece-wisely continuous.

^{2.3}It works since its CDF is well defined.

$$2. \int_{-\infty}^{\infty} f_X(x) dx^a = 1$$

$$3. f_X(x) \geq 0, \forall x \in \mathbb{Z}$$

^aThis is the same as $\lim_{x \rightarrow \infty} F_X(x)$.

Again, note that as $f_X(x)$ is not $P(X = x)$ but the derivative of $F_X(x)$, it can exceed 1, while in general it needs to be non-negative. Most of times we only look at the space with positive probability, so define the *support region* as

Definition 2.1.3 (Support)

For RV X , its support is defined as

$$\text{supp}(X) = \{x \in \mathbb{R} : f(x) > 0\}$$

The following is the general property of CDF, not for specific discrete or continuous CDFs.

Proposition 2.1.4 (Properties of Cumulative Distribution Function)

For some RV X with CDF $F(x)$, the following properties holds true:

1. $F(x)$ is non-decreasing^a, i.e. $F(x_1) \leq F(x_2)$ for $x_1 < x_2$.
2. For $x_1 \leq x_2$, $P(x_1 < X \leq x_2) = P(X \leq x_2) - P(X \leq x_1) = F(x_2) - F(x_1)$.
3. $\lim_{x \rightarrow -\infty} F(x) = 0$ $\lim_{x \rightarrow \infty} F(x) = 1$.
4. $F(x)$ is right-continuous^b, i.e. $\lim_{x \rightarrow a^+} F(x) = F(a)$.
5. $P(X = a) = F(a) - \lim_{x \rightarrow a^-} F(x)$ ^c.

^aThis requires probability function $f(x) \geq 0$.

^bFor dis. RV, F is right-continuous, while for cts RV, by the continuity requirement of differentiation $\lim_{x \rightarrow a} F(x) = F(a)$, it is continuous in both directions, i.e. $\lim_{x \rightarrow a^-} F(x) = \lim_{x \rightarrow a^+} F(x) = F(a)$.

^cThis is well defined for dis. RV, while for cts. RV, $P(X = a) = F(a) - \lim_{x \rightarrow a^-} F(x) = F(a) - F(a) = 0$.

With the CDF for some distribution X , $F_X(x)$, if we are given some probability p , we can use the CDF to inversely find the position x such that $P(X \leq x) = p$. A formal definition is

Definition 2.1.5 (Quantile Function)

For some distribution X with CDF $F_X(x)$, the quantile function $F_X^- : [0, 1] \rightarrow \mathbb{R}$ is defined as

$$F_X^-(p) = \inf \{x \in \mathbb{R} : F_X(x) = p\}$$

where p is some probability.

That said, the output of quantile function given some probability p is the smallest position x such that $F_X(x) = p$. Note that continuous and monotonic CDF $F(x)$'s quantile function is their inverse such that

$$F^-(p) = F^{-1}(p), p \in [0, 1] \quad (\text{cts. Quantile})$$

since

$$F(F^{-1}(x)) = x$$

Here the prerequisite is that F must be monotone, or differentiable everywhere in support, otherwise the inverse F^{-1} does not exist.

Although the big brackets are beautiful, we need a more concise way to represent the probability functions, like in one line. We introduce a powerful tool, the **indicator variable**.

$$\mathbf{1}_A = \begin{cases} 1 & \text{if event } A \text{ is true} \\ 0 & \text{otherwise} \end{cases} \quad (\text{Indicator Variable})$$

Note that $\mathbf{1}_A \sim \text{Bern}(P(A))$. We typically multiply the probability function by it to specify the range. Probability functions have to be defined everywhere on \mathbb{R} , which means we also need to indicate the region with probability 0. Consider two distributions, $X \sim U(0, 3)$ and $Y \sim U(4, 7)$, where their pdfs look the same without domain

$$f(u) = \frac{1}{3}$$

However, they are two completely different distributions. Specifying the domain would help distinguish between them, where

$$\begin{aligned} f_X(x) &= \frac{1}{3}\mathbf{1}_{\{x \in [0, 3]\}} \\ f_Y(y) &= \frac{1}{3}\mathbf{1}_{\{y \in [4, 7]\}} \end{aligned}$$

Then it is safer to identify them, and also able to compute their CDF, expectation, MGF, etc. For example, the CDF could be written as

$$F_X(x) = \int_{-\infty}^{\infty} \frac{1}{3}\mathbf{1}_{\{x \in [0, 3]\}} dx + \mathbf{1}_{\{x > 3\}}$$

That said $F(x)$ is 1 if x passes 3 (beyond support), and 0 if $x < 0$ (below support).

2.2 Expectation and Variance

The expectation (or mean) of a distribution is the outcome on average. For example, if you conduct an experiment multiple times, the average result is likely to be close to the expectation value, a.k.a "convergence to the mean". Note that a single or very few samples are not guaranteed to converge to the theoretical expectation. The mathematical definition is

Definition 2.2.1 (Expectation)

For a distribution X ,

$$\mathbb{E}(X) = \begin{cases} \sum_{x=-\infty}^{\infty} xP(x=x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} xf(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

Remark

The expectation DNE if $\mathbb{E}(|X|) = \infty$. From now on, we always assume $\mathbb{E}(|X|) < \infty$ (i.e. converges absolutely, see [Convergence](#)) if not specified.

However, note that the outcomes not in support of X have probability 0 such that $P(X = x) = 0$ or $f(x) = 0$ for $x \notin \text{supp}(X)$, so we can shrink the summation range to

$$\mathbb{E}(X) = \begin{cases} \sum_{x \in \text{supp}(X)} xP(X=x) & \text{if } X \text{ is discrete} \\ \int_{\text{supp}(X)} xf(x) dx & \text{if } X \text{ is continuous} \end{cases} \quad (\text{Expectation})$$

We have that $\inf(\text{supp}(X)) \leq \mathbb{E}(X) \leq \sup(\text{supp}(X))$, such that the expectation is always within the support. Another interesting fact is that

Theorem 2.2.2 (Expectation with TPF)

Suppose X is non-negative. Then

$$\mathbb{E}(X) = \begin{cases} \sum_{x=1}^{\infty} P(X \geq x) & \text{if } X \text{ is discrete} \\ \int_0^{\infty} \bar{F}(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

That is, the expectation of X is the area above its CDF and under 1.

Proof. For discrete case,

$$\begin{aligned} \sum_{x=1}^{\infty} P(X \geq x) &= \sum_{x=1}^{\infty} \sum_{y=x}^{\infty} P(X = y) = \sum_{y=1}^{\infty} \sum_{x=1}^y P(X = y) \\ &= \sum_{y=1}^{\infty} P(X = y) \sum_{x=1}^y 1 = \sum_{y=1}^{\infty} P(X = y)y \\ &= \mathbb{E}(X) \end{aligned}$$

For continuous case,

$$\begin{aligned} \int_0^{\infty} P(X > x) dx &= \int_0^{\infty} \int_x^{\infty} f(t) dt dx \quad \text{note that } 0 \leq x \leq t < \infty \\ &= \int_0^{\infty} \int_0^t f(t) dx dt \quad \text{by Fubini's Theorem} \\ &= \int_0^{\infty} f(t) \int_0^t 1 dx dt = \int_0^{\infty} f(t)t dt \\ &= \mathbb{E}(X) \end{aligned}$$

□

Let's start with some easy examples.

Example 2.2.1 (Discrete Expectation)

Suppose you are rolling a fair six-sided die. Let X denote the outcome of the die, with probabilities

$$P(X = x) = \begin{cases} \frac{1}{6} & \text{if } x = 1, 2, 3, 4, 5, 6 \\ 0 & \text{otherwise} \end{cases}$$

The expectation is

$$\mathbb{E}(X) = \sum_{x=1}^6 xP(X = x) = \frac{1}{6}1 + \frac{1}{6}2 + \frac{1}{6}3 + \frac{1}{6}4 + \frac{1}{6}5 + \frac{1}{6}6 = \frac{21}{6} = 3.5$$

On average, the result of rolling the die is likely to be around 3.

Now, what if we modify the probabilities?

$$P(X = x) = \begin{cases} \frac{1}{2} & \text{if } x = 1 \\ \frac{1}{4} & \text{if } x = 2 \\ \frac{1}{8} & \text{if } x = 4 \\ \frac{1}{8} & \text{if } x = 6 \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\mathbb{E}(X) = \sum_{x \in \text{supp}(X)} xP(X = x) = \frac{1}{2}1 + \frac{1}{4}2 + \frac{1}{8}4 + \frac{1}{8}6 = \frac{9}{4} = 2.25$$

Example 2.2.2 (Poisson Distribution Mean)

The Poisson Distribution $X \sim Poi(\lambda)$ has pmf

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \mathbf{1}_{\{x=0,1,\dots\}}$$

Find $\mathbb{E}(X)$.

Solution

$$\begin{aligned} \mathbb{E}(X) &= \sum_{x=-\infty}^{\infty} xP(X = x) \\ &= \sum_{x=0}^{\infty} \frac{x e^{-\lambda} \lambda^x}{x!} \\ &= 0 + \sum_{x=1}^{\infty} \frac{e^{-\lambda} \lambda^{x-1} \lambda}{(x-1)!} \end{aligned} \tag{*}$$

note that $x - 1$ above ranges from 0 to ∞

$$\begin{aligned} &= \lambda \sum_{y=0}^{\infty} \underbrace{\frac{e^{-\lambda} \lambda^y}{y!}}_{Y \sim Poi(\lambda) \text{ PMF}} \quad \text{taking } y = x - 1 = 0, 1, \dots \\ &= \lambda \end{aligned} \tag{**}$$

Note that in (**), $\sum_{y=0}^{\infty} \frac{e^{-\lambda} \lambda^y}{y!}$ is the summation of $Poi(\lambda)$'s PMF over support, which is 1. Also note that in (*), this is a common way to modify the summand and turn it into some known PMF, Expectation, or variance, where here we extract one λ and the summand becomes $P(X = x - 1)$, hence we can proceed with the change of variable to eliminate it.

Example 2.2.3 (Binomial Distribution Mean)

Let $X \sim Bin(n, p)$. Find $\mathbb{E}(X)$.

Solution

Let $q = 1 - p$. The PMF is

$$P_X(x) = \binom{n}{x} p^x q^{n-x}, x = 0, 1, \dots, n$$

So

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{x=0}^n x \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^n x \frac{n! p^x q^{n-x}}{x!(n-x)!} = 0 + \sum_{x=1}^n \frac{n! p^x q^{n-x}}{(x-1)!(n-x)!} \\
&= \frac{p}{q} \sum_{x=1}^n \frac{(n-x+1)n! p^{x-1} q^{n-x+1}}{(x-1)!(n-x+1)!} = \frac{p}{q} \sum_{y=0}^n \frac{(n-y)n! p^y q^{n-y}}{y!(n-y)!} \quad \text{taking } y = x-1 \\
&= \frac{p}{q} \sum_{y=0}^n (n-y) \binom{n}{y} p^y q^{n-y} = \frac{p}{q} \left\{ n \sum_{y=0}^n \underbrace{\binom{n}{y} p^y q^{n-y}}_{Y \sim \text{Bin}(n, p) \text{ PMF}} - \sum_{y=0}^n y \underbrace{\binom{n}{y} p^y q^{n-y}}_{\text{Bin}(n, p) \text{ mean}} \right\} \\
&= \frac{p}{q} (n - \mathbb{E}(X))
\end{aligned}$$

Solve for the equation,

$$\mathbb{E}(X) = \frac{\frac{np}{q}}{1 + \frac{p}{q}} = \frac{\frac{np}{q}}{\frac{1}{q}} = np$$

Example 2.2.4 (Exponential Distribution Mean)

The exponential distribution $X \sim Exp(\lambda)$ has PDF

$$f(x) = \lambda e^{-\lambda x} \mathbf{1}_{\{x>0\}}$$

and CDF

$$F_X(x) = 1 - e^{-\lambda x} \mathbf{1}_{\{x \geq 0\}}$$

Find $\mathbb{E}(X)$.

Solution

$$\begin{aligned}
\mathbb{E}(X) &= \int_{-\infty}^{\infty} x f(x) dx = \int_0^{\infty} x \lambda e^{-\lambda x} dx = \lambda \int_0^{\infty} x e^{-\lambda x} dx \\
&= \lambda \left(\frac{x e^{-\lambda x}}{-\lambda} \Big|_0^\infty - \int_0^\infty \frac{e^{-\lambda x}}{-\lambda} dx \right) \quad \text{Integration By Parts} \\
&= \lambda \left(0 + \frac{1}{\lambda^2} \int_0^\infty \underbrace{\lambda e^{-\lambda x}}_{X \sim Exp(\lambda) \text{ PDF}} dx \right) \\
&= \frac{1}{\lambda}
\end{aligned}$$

Note that

$$\lim_{x \rightarrow \infty} x e^{-\lambda x} = 0$$

by L'Hôpital's rule. Now we try [Expectation with TPF](#),

$$\mathbb{E}(X) = \int_0^\infty e^{-\lambda x} dx = -\frac{e^{-\lambda x}}{\lambda} \Big|_0^\infty = \left(\frac{1}{\lambda} - 0 \right) = \frac{1}{\lambda}$$

which also aligns with the above result.

The expectation function maps a random variable to a constant, which is its expected outcome. That is, $\mathbb{E}(X) \in \mathbb{R}$. A classic example will be the [Indicator Variable](#), which follows a Bernoulli distribution. Its mean is

$$\mathbb{E}(\mathbf{1}_A) = 0 \cdot (1 - P(A)) + 1 \cdot P(A) = P(A)$$

Due to its idempotent property, the n th moment is also

$$\mathbb{E}(\mathbf{1}_A^n) = 0^n \cdot (1 - P(A)) + 1^n \cdot P(A) = P(A), \forall n \in \mathbb{N} \quad (\star)$$

such that the moments are uniform over \mathbb{N} . A constant (degenerate) distribution $X = c, c \in \mathbb{R}$ has no randomness, so the expectation is always itself. For instance, you roll a die whose faces are all 1, then the number facing up is always 1, and you will not expect other numbers. To be rigorous, the constant distribution $X = c$ is defined as

$$P(X = x) = \mathbf{1}_{\{x=c\}} = \begin{cases} 1 & \text{if } x = c \\ 0 & \text{otherwise} \end{cases}$$

So the expectation is

$$\mathbb{E}(X) = \sum_{x \in \text{supp}} x P(X = x) = c P(X = c) = c \quad \text{by (Expectation)}$$

It is sufficient to show that for a constant $c \in \mathbb{R}$, its expectation is

$$\mathbb{E}(c) = c \quad (\text{Expectation of Constants})$$

Note that the expectation function can also be applied to some transformation of distribution.

Definition 2.2.3 (Expectation Function)

For a function $h(X)$ on some distribution X , its expectation is

$$\mathbb{E}(h(X)) = \begin{cases} \sum_{x=-\infty}^{\infty} h(x) P(X = x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} h(x) f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

Proposition 2.2.4 (Linearity of Expectation)

The expectation function $\mathbb{E}(\cdot)$ is linear. Mathematically, for a distribution X and some linear (affine) function $h(x) = kx + b$ with $k, b \in \mathbb{R}$, we have

$$\mathbb{E}(h(X)) = h(\mathbb{E}(X)) = k\mathbb{E}(X) + b$$

Proof.

$$\begin{aligned}
\mathbb{E}(h(X)) &= \mathbb{E}(kX + b) = \int_{-\infty}^{\infty} (kX + b)f(x) dx \\
&= k \int_{-\infty}^{\infty} xf(x) dx + b \int_{-\infty}^{\infty} f(x) dx \\
&= k\mathbb{E}(X) + b
\end{aligned}$$

□

This proof also leads to

Corollary 2.2.5 (Linearity of Expectation)

For n functions h_1, \dots, h_n and constants $c_1, \dots, c_n \in \mathbb{R}$,

$$\mathbb{E}\left(\sum_{i=1}^n h_i(X) + c_i\right) = \sum_{i=1}^n \mathbb{E}(h_i(x)) + \sum_{i=1}^n c_i$$

As mean is the average outcome of X , it does not provide whole information on the distribution. For example, $X \sim \text{Uni}(5, 15)$ and $Y \sim \text{Uni}(0, 20)$ have the same mean 10, but the active region $\text{supp}(Y) = [0, 20]$ is larger than $\text{supp}(X) = [5, 15]$, as in 2.2.

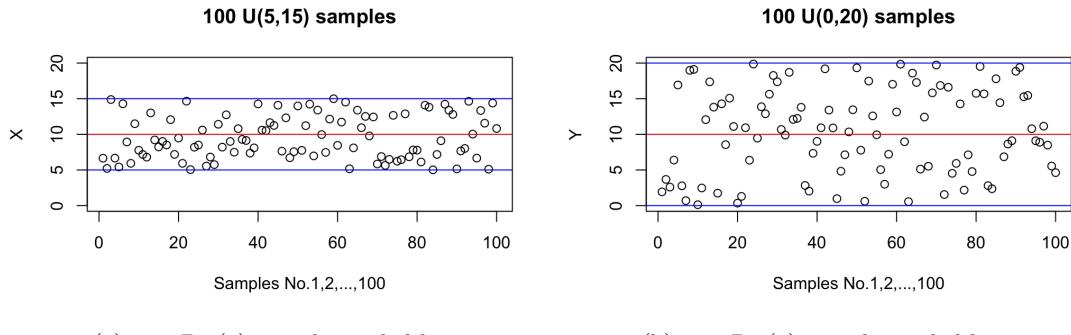


Figure 2.2: Same Mean Different Spread

This discrepancy information around mean can be measured by *variance*.

Definition 2.2.6 (Variance)

The variance of distribution X is

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}((X - \mu)^2) \\
&= \mathbb{E}(X^2) - \mathbb{E}(X)^2
\end{aligned}$$

where $\mu = \mathbb{E}(X)$.

Note that

$$\begin{aligned}
 \mathbb{E}((X - \mu)^2) &= \mathbb{E}(X^2 - 2\mu\mathbb{E}(X) + \mu^2) \\
 &= \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) + \mathbb{E}(\mu^2) \\
 &= \mathbb{E}(X^2) - 2\mu^2 + \mu^2 \\
 &= \mathbb{E}(X^2) - \mathbb{E}(X)^2
 \end{aligned}$$

It measures the average spread of the outcomes around mean, which is the average squared distance between X and mean.

Example 2.2.5 (Poisson Distribution Variance)

Find the variance of $X \sim Poi(\lambda)$.

Solution

We need

$$\mathbb{E}(X^2) - \mathbb{E}(X)^2$$

From [Poisson Distribution Mean](#), we know that $\mathbb{E}(X) = \lambda$. Now proceed with

$$\begin{aligned}
 \mathbb{E}(X^2) &= \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda} \lambda^x}{x!} = 0 + \sum_{x=1}^{\infty} x \frac{e^{-\lambda} \lambda^x}{(x-1)!} = \lambda \sum_{x=1}^{\infty} x \frac{e^{-\lambda} \lambda^{x-1}}{(x-1)!} \\
 &= \lambda \sum_{y=0}^{\infty} (y+1) \frac{e^{-\lambda} \lambda^y}{y!} \quad \text{taking } y = x-1 \\
 &= \lambda \left\{ \underbrace{\sum_{y=0}^{\infty} y \frac{e^{-\lambda} \lambda^y}{y!}}_{\mathbb{E}(Y), Y \sim Poi(\lambda)} + \underbrace{\sum_{y=0}^{\infty} \frac{e^{-\lambda} \lambda^y}{y!}}_{\text{PMF of } Y} \right\} \\
 &= \lambda(\lambda + 1) = \lambda^2 + \lambda
 \end{aligned}$$

Thus,

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda$$

This derivation is nice and neat.

Example 2.2.6 (Binomial Distribution Variance)

Find the variance of $X \sim Bin(n, p)$.

Solution

Recall from [Binomial Distribution Mean](#), $\mathbb{E}(X) = np$.

$$\begin{aligned}
\mathbb{E}(X^2) &= \sum_{x=0}^{\infty} x^2 \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{\infty} x^2 \frac{n!}{x!(n-x)!} p^x q^{n-x} \\
&= \sum_{x=1}^{\infty} x \frac{n!}{(x-1)!(n-x)!} p^x q^{n-x} = \frac{p}{q} \sum_{x=1}^{\infty} x \frac{(n-x+1)n!}{(x-1)!(n-x+1)!} p^{x-1} q^{n-x+1} \\
&= \frac{p}{q} \sum_{y=0}^{\infty} (y+1) \frac{(n-y)n!}{y!(n-y)!} p^y q^{n-y} \quad \text{taking } y = x-1 \\
&= \frac{p}{q} \sum_{y=0}^{\infty} (ny + n - y^2 - y) \binom{n}{y} p^y q^{n-y} = \frac{p}{q} \sum_{y=0}^{\infty} ((n-1)y + n - y^2) \binom{n}{y} p^y q^{n-y} \\
&= \frac{p}{q} \left\{ (n-1) \underbrace{\sum_{y=0}^{\infty} y \binom{n}{y} p^y q^{n-y}}_{\text{Bin}(n, p) \text{ Mean}} + n \underbrace{\sum_{y=0}^{\infty} \binom{n}{y} p^y q^{n-y}}_{\text{Bin}(n, p) \text{ PMF}} - \underbrace{\sum_{y=0}^{\infty} y^2 \binom{n}{y} p^y q^{n-y}}_{\text{same as } E(X^2)} \right\} \\
&= \frac{p}{q} ((n-1)np + n - \mathbb{E}(X^2))
\end{aligned}$$

Solve,

$$\mathbb{E}(X^2) = \frac{\frac{p}{q}((n-1)np + n)}{1 + \frac{p}{q}} = p(n^2p - np + n) = n^2p^2 - np^2 + np$$

Thus

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = n^2p^2 - np^2 + np - n^2p^2 = npq$$

Example 2.2.7 (Exponential Distribution Variance)

Let $X \sim \text{Exp}(\frac{1}{\lambda})$. Find $\text{Var}(X)$.

Solution

Recall

$$f_X(x) = \lambda e^{-\lambda x} \mathbf{1}_{\{x \geq 0\}}$$

So

$$\begin{aligned}
\mathbb{E}(X^2) &= \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx = \lambda \int_0^{\infty} x^2 e^{-\lambda x} dx \\
&= \lambda \left(\frac{x^2 e^{-\lambda x}}{-\lambda} \Big|_0^\infty + \int_0^{\infty} \frac{2x e^{-\lambda x}}{\lambda} dx \right) \\
&= \lambda \left(0 + \frac{2}{\lambda^2} \underbrace{\int_0^{\infty} x \lambda e^{-\lambda x} dx}_{\mathbb{E}(X) = \frac{1}{\lambda}} \right) \\
&= \frac{2}{\lambda^2}
\end{aligned}$$

Thus

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

If variance is zero, we can reasonably assume that the RV has no variability, and is thus constant.

Proposition 2.2.7 (Zero Variance)

For some RV X , $\text{Var}(X) = 0$ if and only if X is a constant $c \in \mathbb{R}$.

Proof. If X is constant, then

$$\text{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = 0 \quad \text{since } \mathbb{E}(X) = X = c$$

Conversely, say $\text{Var}(X) = 0$. Thus

$$\begin{aligned}\text{Var}(X) &= 0 \\ \mathbb{E}(X^2) - \mathbb{E}(X)^2 &= 0 \\ \mathbb{E}(X^2) &= \mathbb{E}(X)^2\end{aligned}$$

By [Expectation of Separable Independent Functions](#), we see that X is independent to itself. It can only be the case that X is a constant. \square

Its square root is a rough estimate of how far X is away from mean, on average, and is roughly unbiased.

Definition 2.2.8 (Standard Deviation)

The standard deviation of distribution X is

$$\sigma = \sqrt{\text{Var}(X)}$$

This graph 2.3 helps visualize the spread of $\text{Poi}(2)$ samples, whose mean, variance and standard deviation are 2.05, 1.93 and 1.39 respectively. For convenience we use the theoretical mean and variance, 2. Note that the majority of data points fall in the [1, 3] and evenly spread around the mean 2. The average distance from a data point to the mean (spread) is also close to the theoretical standard deviation, $\sqrt{2}$.

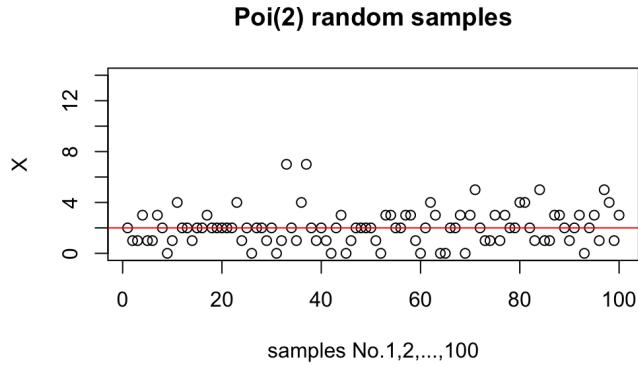


Figure 2.3: 100 $\text{Poi}(2)$ samples

Now consider a linear transformation on some distribution X , $Y = aX + b$ where $a, b \in \mathbb{R}$, a.k.a. scale X by a , and shift by b . What will be Y 's variance? As variance is a measurement of the spread of X around mean, intuitively, the

shift of mean (i.e. b) would not affect the variance, but the scale factor a can increase or decrease the variance, except for $a = 1$. For instance, $|a| > 1$ would extend the spread range, whereas $|a| < 1$ would shorten the range. See 2.4 for examples of scaled by 2 and 0.5 respectively, with corresponding variances 7.71 and 0.48, such that the original variance is scaled by 4 and 0.25. We can see that if X is scaled by a , then its variance is scaled by a^2 . We prove it formally.

Proposition 2.2.9 (Variance of Linearly Transformed Distribution)

For some RV X , and $a, b \in \mathbb{R}$,

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

Proof. Let $Y = aX + b$, where $\mu_Y = \mathbb{E}(aX + b) = a\mu_X + b$ by linearity of expectation.

$$\begin{aligned} \text{Var}(Y) &= \mathbb{E}((Y - \mu_Y)^2) \\ &= \mathbb{E}(((aX + b) - (a\mu_X + b))^2) \\ &= \mathbb{E}((aX - a\mu_X)^2) \\ &= \mathbb{E}(a^2(X - \mu_X)^2) \\ &= a^2 \mathbb{E}((X - \mu_X)^2) \quad \text{by linearity of expectation} \\ &= a^2 \text{Var}(X) \end{aligned}$$

□

Note that the shift b here is canceled out. Therefore the shift will not affect variance.

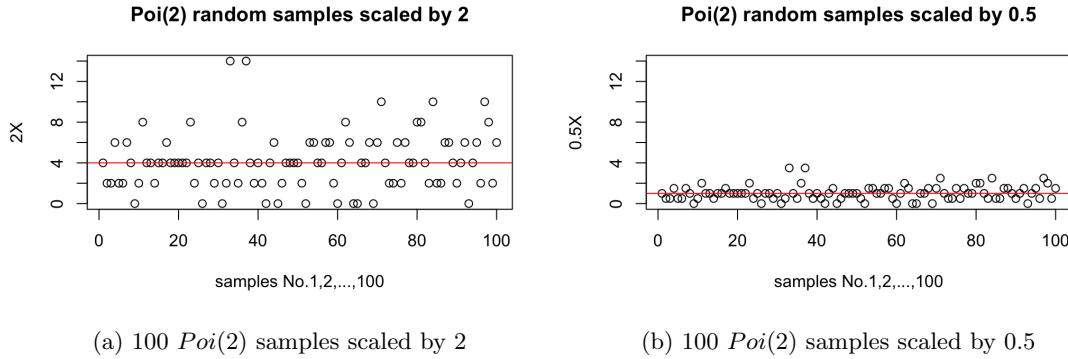


Figure 2.4: scaled 100 *Poi*(2) samples

2.3 Moment Generating Functions

The n th moment of distribution X is

$$\mathbb{E}(X^n) = \begin{cases} \sum_{x=-\infty}^{\infty} x^n P(X=x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^n f(x) dx & \text{if } X \text{ is continuous} \end{cases} \quad (\text{nth moment})$$

Moments can provide information on the distribution, for instance,

1st moment: mean, $\mathbb{E}(X)$

2nd moment: variance, $\mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

3rd moment: skewness, $\mathbb{E}\left(\left(\frac{X - \mathbb{E}(X)}{\sigma}\right)^3\right)$

4th moment: kurtosis, $\mathbb{E}\left(\left(\frac{X - \mathbb{E}(X)}{\sigma}\right)^4\right)$

Higher moments give better information on tailedness of the distribution. For example, the n th moment of standard uniform distribution is

$$\mathbb{E}(U^n) = \int_0^1 u^n du = \frac{u}{n+1} \Big|_0^1 = \frac{1}{n+1}$$

The variance is thus

$$\text{Var}(U^n) = \mathbb{E}(U^{2n}) - \mathbb{E}(U^n)^2 = \frac{1}{2n+1} - \frac{1}{(n+1)^2}$$

The *Moment Generating Function (MGF)* of distribution X is defined as

Definition 2.3.1 (Moment Generating Function (MGF))

For RV X , its MGF is defined as

$$\mathcal{M}_X(t) = \mathbb{E}(e^{tX}) = \begin{cases} \sum_{x=-\infty}^{\infty} e^{tx} P(X=x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{tx} f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

where $t \in (-h, h)$ for some $h > 0$.

Recall the Taylor Series Expansion of e^{tx} ,

$$e^{tx} = \sum_{n=0}^{\infty} \frac{t^n x^n}{n!} \quad (\text{Taylor Expansion of } e^{tx})$$

Apply expectation function and by linearity

$$\mathbb{E}(e^{tX}) = \mathbb{E}\left(\sum_{n=0}^{\infty} \frac{t^n X^n}{n!}\right) = \sum_{n=0}^{\infty} \frac{t^n \mathbb{E}(X^n)}{n!} \quad (\text{MGF Expansion})$$

Note that

$$\left. \frac{d}{dt} \right|_{t=0} \sum_{n=0}^{\infty} \frac{t^n \mathbb{E}(X^n)}{n!} = \mathbb{E}(X^n)$$

which gives the n th moment of X , by property of Taylor Series. Therefore, differentiating the MGF gives the moments,

Theorem 2.3.2 (Generating Moments)

For distribution X with MGF $\mathcal{M}_X(t)$, its n th moment is

$$\mathbb{E}(X^n) = \left. \frac{d}{dt} \right|_{t=0} \mathcal{M}_X(t)$$

We look at some examples.

Example 2.3.1 (Binomial Distribution MGF)

Find the MGF of $X \sim \text{Bin}(n, p)$.

Solution

$$\begin{aligned}\mathcal{M}_X(t) &= \mathbb{E}(e^{tX}) \\ &= \sum_{x=0}^n e^{tx} \binom{n}{x} p^x q^{n-x} \\ &= \sum_{x=0}^n \binom{n}{x} (pe^t)^x q^{n-x} \\ &= (pe^t + q)^n \quad \text{by Binomial Theorem}\end{aligned}$$

Note that the n th moment is

$$\mathbb{E}(X^n) = \left. \frac{d^n}{dt^n} \right|_{t=0} \mathcal{M}_X(t)$$

For example,

$$\begin{aligned}\mathbb{E}(X) &= \left. \frac{d}{dt} \right|_{t=0} (pe^t + q)^n \\ &= n(pe^t + q)^{n-1} pe^t \Big|_{t=0} \\ &= np \\ \mathbb{E}(X^2) &= \left. \frac{d^2}{dt^2} \right|_{t=0} (pe^t + q)^n \\ &= n(n-1)(pe^t + q)^{n-2} p^2 e^{2t} + n(pe^t + q)^{n-1} pe^t \Big|_{t=0} \\ &= n^2 p^2 - np^2 + np \\ &\dots\end{aligned}$$

By the way, note that

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = n^2 p^2 - np^2 + np - n^2 p^2 = npq$$

This matches the result in [Binomial Distribution Variance](#), but is **MUCH EASIER**.

Example 2.3.2 (Poisson Distribution MGF)

Find the MGF of $X \sim \text{Poi}(\lambda)$.

Solution

$$\begin{aligned}
\mathcal{M}_X(t) &= \mathbb{E}(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda} \lambda^x}{x!} \\
&= \sum_{x=0}^{\infty} \frac{e^{-\lambda} (\lambda e^t)^x}{x!} = \frac{e^{-\lambda}}{e^{-\lambda e^t}} \sum_{x=0}^{\infty} \underbrace{\frac{e^{-\lambda e^t} (\lambda e^t)^x}{x!}}_{Poi(\lambda e^t) \text{ PMF}} \\
&= e^{\lambda(e^t - 1)}
\end{aligned}$$

MGF uniquely characterizes the distribution, or equivalently, any distribution has its unique MGF.

Theorem 2.3.3 (Uniqueness Theorem of MGF)

Let X and Y be RVs with MGFs $\mathcal{M}_X(t), \mathcal{M}_Y(t)$ respectively. If

$$\mathcal{M}_X(t) = \mathcal{M}_Y(t)$$

then X and Y have the same distribution such that $X \sim Y$.

Proof. Suppose X, Y are discrete. Then $\mathcal{M}_X(\log(t)) = \mathcal{M}_Y(\log(t))$ such that

$$\mathbb{E}_X(e^{\log(t)X}) = \sum_{x=-\infty}^{\infty} t^x P(X=x) = \mathbb{E}_Y(e^{\log(t)Y}) = \sum_{y=-\infty}^{\infty} t^y P(Y=y)$$

Two generating series are equal if and only if all coefficients are equal, which means

$$P(X=x) = P(Y=y) \text{ for all } x=y \in (-\infty, \infty)$$

Side: For continuous RVs, consider

$$\int_{-\infty}^{\infty} t^x f_X(x) dx = \int_{-\infty}^{\infty} t^y f_Y(y) dy$$

□

Other than computing the moments, MGF could also be used to determine the distribution.

Lemma 2.3.4 (Characterizing Distributions)

Any distribution X has unique

1. PMF/PDF: $P(X=x)$ or $f_X(x)$
2. CDF: $P(X < x)$
3. MGF: $\mathcal{M}_X(t) = \mathbb{E}(e^{tX})$

Remark

Note that if two distributions have the same PF/CDF/MGF, then it would be sufficient to deduce that they are the same distribution.

2.4 Distribution Transformation

Typically, a transformation of a random variable is still a random variable, except for transformation to constant. A transformation is a function applied on the random variable. Let $h(\cdot)$ be the transformation function, and we determine the distribution of transformed random variable with the following criteria.

Algorithm 2.4.1 (Determine Distribution)

Let $Y = h(X)$ be a transformation of RV X with some h . To find the distribution of Y ,

1. Find out the support of Y .
2. For discrete distribution: Derive $P(Y = y)$ for each outcome y in support.
For continuous distribution: Derive $P(Y < y)$ for y in support.
3. Optional: Use MGF if easier.

We always convert unknown into known, dependent into independent. Consider the following examples.

Example 2.4.1 (Discrete Distribution Transformation 1)

Let $X \sim \text{Bern}(p)$ for $0 < p < 1$. Show that $1 - X \sim \text{Bern}(1 - p)$.

Solution

Note that

$$P(X = x) = p\mathbf{1}_{\{x=1\}} + (1 - p)\mathbf{1}_{\{x=0\}}$$

Let $Y = 1 - X$.

$$P(Y = y) = P(1 - X = y) = P(X = 1 - y) = p\mathbf{1}_{\{y=0\}} + (1 - p)\mathbf{1}_{\{y=1\}}$$

which is the PMF of $\text{Bern}(1 - p)$.

Example 2.4.2 (Continuous Distribution Transformation 1)

Let $X \sim \text{Exp}(\frac{1}{\lambda})$. Find the distribution of $Y = X^2$.

Solution

Recall

$$f_X(x) = \lambda e^{-\lambda x} \mathbf{1}_{\{x \geq 0\}} \quad F_X(x) = (1 - e^{-\lambda x}) \mathbf{1}_{\{x \geq 0\}}$$

We first find the support of Y . Since support of X is $x \geq 0$, we know that Y also has support $y \geq 0$. Now

$$\begin{aligned} F_Y(y) &= P(Y < y) \\ &= P(X^2 < y) \text{ convert unknown } Y \text{ into known } X \\ &= P(|X| < \sqrt{y}) \text{ since } \sqrt{\cdot} \text{ is monotone} \\ &= P(X < \sqrt{y}) \text{ since } X \geq 0 \\ &= (1 - e^{-\lambda \sqrt{y}}) \mathbf{1}_{\{y \geq 0\}} \end{aligned}$$

This is sufficient to show the distribution of Y .

Example 2.4.3 (Continuous Distribution Transformation 2)

Let $X \sim U(0, 1)$. Show that $1 - X \sim X \sim U(0, 1)$.

Solution

Note that

$$P(X < x) = x\mathbf{1}_{\{x \in [0, 1]\}}$$

Let $Y = 1 - X$. Then

$$P(Y < y) = P(1 - X < y) = P(X > 1 - y) = 1 - P(X < 1 - y) = 1 - (1 - y) = y\mathbf{1}_{\{y \in [0, 1]\}}$$

which is the CDF of $U(0, 1)$.

The following is a classic and significant theorem using distribution transformation,

Theorem 2.4.2 (Inverse Probability Integral Transformation)

Let continuous X have invertible CDF $F(x)$, denoted by $X \sim F$. Then

$$F(X) \sim U(0, 1)$$

Conversely, let $U \sim U(0, 1)$. Then

$$F^{-1}(U) \sim X$$

Proof. First,

$$\begin{aligned} P(F(X) \leq u) &= P(F^{-1}(F(X)) \leq F^{-1}(u)) \quad \text{since } F^{-1} \text{ is monotone} \\ &= P(X \leq F^{-1}(u)) = F(F^{-1}(u)) \\ &= u \end{aligned}$$

Note that this is the CDF of $U(0, 1)$. Thus $F(X) \sim U(0, 1)$.

Conversely,

$$\begin{aligned} P(F^{-1}(U) \leq x) &= P(F(F^{-1}(U)) \leq F(x)) \quad \text{since } F \text{ is monotone} \\ &= P(U \leq F(x)) \\ &= F(x) \end{aligned}$$

Thus $F^{-1}(U) \sim X$.

Remark

For discrete RVs, replace F^{-1} with their quantile F^- .

□

This can be used to sample distributions.

Example 2.4.4 (Sampling Exponential Distribution)

You are given access to sample independent $U_1, \dots, U_n \stackrel{i.i.d.}{\sim} U(0, 1)$. Use them to sample n independent $Exp(1/2)$ samples.

Solution

The quantile function of $Exp(1/2)$ is

$$F^{-1}(u) = \frac{\log(1-u)}{-2}$$

Using Inverse Probability Integral Transformation, the pseudo-code for sampling $Exp(1/2)$ is

1. Sample $U_1, \dots, U_n \stackrel{i.i.d.}{\sim} U(0, 1)$.
2. Set $X_i = F^{-1}(U_i) = \frac{\log(1-u_i)}{-2}, i = 1, \dots, n$.
3. Return X_1, \dots, X_n .

3 Multivariate Distributions

3.1 Probability Functions

Now suppose we have multiple random variables. Similar to the univariate case, their probability function is

Definition 3.1.1 (Joint Probability Function)

For some RVs X, Y , their joint probability function is

$$f(x, y) = \begin{cases} P(X = x, Y = y) := P_{X,Y}(x, y) & \text{if } X, Y \text{ dis.} \\ \frac{P(x \leq X \leq x+dx, y \leq Y \leq y+dy)}{dxdy} := f_{X,Y}(x, y) & \text{if } X, Y \text{ cts.} \end{cases}$$

Proposition 3.1.2 (Properties of Joint Probability Function)

If X, Y dis.:

1. $0 \leq P(X = x, Y = y) \leq 1, \forall x, y \in \mathbb{Z}$

2. $\sum_{x=-\infty}^{\infty} \sum_{y=-\infty}^{\infty} P(X = x, Y = y) = 1.$

If X, Y cts.:

1. $f(x, y) \geq 0, \forall x, y \in \mathbb{R}$.

2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1.$

Now look at **Marginal Probability Functions**. For examples of marginal probabilities, suppose a weather condition that follows the following probabilities Note that they are valid probabilities since they sum to 1. Each numerical entry

Temperature \ Weather	Sunny	Cloudy	Rainy
High	0.23	0.08	0.01
Moderate	0.13	0.19	0.08
Low	0.07	0.09	0.12

Table 3.1.1: Weather Probability

(Temperature, Weather) indicates the joint probability of having both the specific temperature and weather. We see that these probabilities are *NOT INDEPENDENT*, where given sunny weather, the temperature is likely to be high, and low if it is rainy. However, the extreme condition that sunny weather and low temperature appear together is still possible, though with a small probability 0.07. The question is, what is the probability of being sunny? As we are not asking about temperature, any temperature should count. That said,

$$P(\text{Sunny}) = P(\text{High, Sunny}) + P(\text{Moderate, Sunny}) + P(\text{Low, Sunny}) = 0.23 + 0.13 + 0.07 = 0.43$$

As we can see, it is the summation of the *Sunny* column. Recall from **Law of Total Probability**, $P(A) = \sum_{i=1}^n P(A \cup B_i)$ for all disjoint and complement subsets B_i of sample space Ω .

Definition 3.1.3 (Marginal Probability Function)

For some RV X, Y with joint probability function $f(x, y)$, the marginal distribution of X has probability function

$$f_X(x) = \begin{cases} \sum_{y=-\infty}^{\infty} P(X = x, Y = y) & \text{if dis.} \\ \int_{-\infty}^{\infty} f(x, y) dy & \text{if cts.} \end{cases}$$

Similarly, the marginal distribution of Y ,

$$f_Y(y) = \begin{cases} \sum_{x=-\infty}^{\infty} P(X = x, Y = y) & \text{if dis.} \\ \int_{-\infty}^{\infty} f(x, y) dx & \text{if cts.} \end{cases}$$

For example, in [Weather Probability](#), the marginal probability function of *Weather* is

$$P(Weather) = \begin{cases} 0.43 & \text{for Sunny} \\ 0.36 & \text{for Cloudy} \\ 0.21 & \text{for Rainy} \end{cases}$$

and for *Temperature* is

$$P(Temperature) = \begin{cases} 0.32 & \text{for High} \\ 0.4 & \text{for Moderate} \\ 0.28 & \text{for Low} \end{cases}$$

In summary,

Temperature \ Weather	Sunny	Cloudy	Rainy	Temperature
High	0.23	0.08	0.01	0.32
Moderate	0.13	0.19	0.08	0.4
Low	0.07	0.09	0.12	0.28
Weather	0.43	0.36	0.21	1

Table 3.1.2: Weather Probability

The marginal probability functions are valid probability functions, where you can verify that

$$\sum_{x=-\infty}^{\infty} P_X(x) = \sum_{y=-\infty}^{\infty} P_Y(y) = \sum_{x=-\infty}^{\infty} \sum_{y=-\infty}^{\infty} P(X = x, Y = y) = 1$$

and also for continuous functions

$$\int_{-\infty}^{\infty} f_X(x) dx = \int_{-\infty}^{\infty} f_Y(y) dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$$

Like other distributions, the marginal distributions have means

$$\mathbb{E}(X) = \begin{cases} \sum_{x \in \text{supp}(X)} x P_X(x) \\ \int_{\text{supp}(X)} x f_X(x) dx \end{cases} \quad \mathbb{E}(Y) = \begin{cases} \sum_{y \in \text{supp}(Y)} y P_Y(y) \\ \int_{\text{supp}(Y)} y f_Y(y) dy \end{cases}$$

Same way to compute variances or higher moments.

The joint CDF is defined as

Definition 3.1.4 (Joint Cumulative Distribution Function)

For RVs X, Y , their joint CDF is

$$F_{X,Y}(x,y) = \begin{cases} \sum_{t=-\infty}^x \sum_{k=-\infty}^y P(X=t, Y=k) & \text{if } X, Y \text{ dis.} \\ \int_{-\infty}^x \int_{-\infty}^y f(t,k) dt dk & \text{if } X, Y \text{ cts.} \end{cases}$$

Proposition 3.1.5 (Properties of Joint CDF)

If X, Y cts.:

1. $F(x,y)$ is non-decreasing in both arguments. i.e.

$$F(x_1, y) \leq F(x_2, y) \text{ for } x_1 < x_2 \quad F(x, y_1) \leq F(x, y_2) \text{ for } y_1 < y_2$$

2. $F(x,y)$ has limiting probability

$$\lim_{x \rightarrow -\infty} F(x,y) = \lim_{y \rightarrow -\infty} F(x,y) = 0 \quad \lim_{(x,y) \rightarrow (\infty,\infty)} F(x,y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) dx dy = 1^a$$

3. The marginal CDFs are

$$F_X(x) = \lim_{y \rightarrow \infty} F(x,y) = \int_{-\infty}^{\infty} f_X(t) dt \quad F_Y(y) = \lim_{x \rightarrow \infty} F(x,y) = \int_{-\infty}^y f_Y(k) dk$$

^aThe joint cumulative probability is 0 when either x or y approaches $-\infty$, but is 1 only when both of them go to ∞ .

3.2 Independence

Two RVs X, Y are independent if the behavior of one of them does not affect the other. For example, if $Y = 2X$ then they are not independent, since the larger X is, the larger is Y , and vice-versa.

First, for discrete sample space, we review the definition of cross set of two sets A, B , where

$$A \times B = \{(a,b) : a \in A, b \in B\}$$

which is the total number of combinations. Intuitively, if X has sample space A and Y has sample space B , and they are independent, then the sample space of joint X, Y distribution, which is $A \times B$, should have "full dimension" such that it should have

$$|A \times B| = |A| |B|$$

outcomes. This can be demonstrated by the following example. Suppose we independently roll a die and flip a coin (both fair), then the sample space should be

$$\begin{bmatrix} (1,H) & (2,H) & (3,H) & (4,H) & (5,H) & (6,H) \\ (1,T) & (2,T) & (3,T) & (4,T) & (5,T) & (6,T) \end{bmatrix}$$

which has $2 \times 6 = 12$ outcomes, and looks like a rectangle. However, if we modify the rule, and the die will not roll even numbers if the coin is a head due to some reason, then they will be dependent and the sample space will shrink to

$$\begin{bmatrix} (1,H) & (3,H) & (5,H) \\ (1,T) & (2,T) & (3,T) & (4,T) & (5,T) & (6,T) \end{bmatrix}$$

such that the cases $(2, H), (4, H), (6, H)$ will not happen now. That said, the outcome of the die is constrained by the outcome of the coin, since if the coin is a head, then it cannot freely pick outcomes in its own sample space, i.e. no longer independent.

Now move on to continuous distributions, and consider $X \in [0, a], Y \in [0, b]$. Similar to the discrete case, the joint support of X, Y is

$$\{(x, y) : 0 \leq x \leq a, 0 \leq y \leq b\}$$

if they are independent, which is exactly a $a \times b$ rectangle. If we add a constraint, for instance, $Y \leq X$, then the support will be a triangle since Y cannot pick values that are greater than X . This will be expanded later.

From above, we can show that

Definition 3.2.1 (Independence of RVs)

Let RVs X, Y have joint probability function $f(x, y)$ and marginal probability functions^a $f_X(x), f_Y(y)$. They are independent, denoted by $X \perp Y$, if and only if

$$f(x, y) = f_X(x)f_Y(y)$$

^aThe generic probability function $f(\cdot)$ stands for PMF for discrete distributions and PDF for continuous distributions.

Again, see [Weather Probability](#), and as we have stated that *Weather* and *Temperature* are not independent, we can now prove it formally. For example, we know that $P(\text{Sunny}) = 0.43$ and $P(\text{High}) = 0.32$, but $P(\text{Sunny}, \text{High}) = 0.23 \neq P(\text{Sunny})P(\text{High})$, which are not independent.

Example 3.2.1 (Independent RVs)

RVs X, Y have joint PDF

$$f(x, y) = \lambda e^{-\lambda(x+y)} \mathbf{1}_{\{x \geq 0, y \geq 0\}}, \lambda > 0$$

Are they independent?

Solution

The joint support is a rectangle, and the constraint $x \geq 0, y \geq 0$ seems independent. To formally prove this, we first find the marginal PDFs

$$\begin{aligned} f_X(x) &= \int_0^\infty \lambda e^{-\lambda(x+y)} dy = -e^{-\lambda(x+y)}|_0^\infty = 0 - (-e^{-\lambda x}) \\ &= e^{-\lambda x} \mathbf{1}_{\{x \geq 0\}} \end{aligned}$$

$$\begin{aligned} f_Y(y) &= \int_0^\infty \lambda e^{-\lambda(x+y)} dx = -e^{-\lambda(x+y)}|_0^\infty = 0 - (-e^{-\lambda y}) \\ &= e^{-\lambda y} \mathbf{1}_{\{y \geq 0\}} \end{aligned}$$

where we verify that

$$f(x, y) = f_X(x)f_Y(y)$$

Therefore, they are independent.

3.3 Expectation and Covariance

Definition 3.3.1 (Expectation of Multivariate Distribution)

The expectation of joint distribution of RVs X, Y is

$$\mathbb{E}(XY) = \begin{cases} \sum_{x=-\infty}^{\infty} \sum_{y=-\infty}^{\infty} xyP(X=x, Y=y) & \text{if } X, Y \text{ dis.} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y) dx dy & \text{if } X, Y \text{ cts.} \end{cases}$$

It also follows that

Definition 3.3.2 (Expectation function of Multivariate Distribution)

For some function $h : \text{supp}(X) \times \text{supp}(Y) \rightarrow \mathbb{R}$, the expectation of joint distribution of RVs $h(X, Y)$ is

$$\mathbb{E}(h(X, Y)) = \begin{cases} \sum_{x=-\infty}^{\infty} \sum_{y=-\infty}^{\infty} h(x, y)P(X=x, Y=y) & \text{if } X, Y \text{ dis.} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y)f(x, y) dx dy & \text{if } X, Y \text{ cts.} \end{cases}$$

Proposition 3.3.3 (Expectation of Separable Independent Functions)

For **independent** RVs $X \perp Y$ and some functions $h(\cdot), g(\cdot)$,

$$\mathbb{E}(h(X)g(Y)) = \mathbb{E}(h(X))\mathbb{E}(g(Y))$$

Proof.

$$\begin{aligned} \mathbb{E}(h(X)g(Y)) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x)g(y)f(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x)g(y)f_X(x)f_Y(y) dx dy \quad \text{by independence of } X, Y \text{ 3.2.1} \\ &= \int_{-\infty}^{\infty} h(x)f_X(x) dx \int_{-\infty}^{\infty} g(y)f_Y(y) dy \\ &= \mathbb{E}(h(X))\mathbb{E}(g(Y)) \end{aligned}$$

□

The covariance of X, Y measures the strength and direction of the *linear relationship* between them.

Definition 3.3.4 (Covariance)

The covariance between RVs X, Y is defined as

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}((X - \mu_X)(Y - \mu_Y)) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \end{aligned}$$

Note that

$$\begin{aligned}
 \text{Cov}(X, Y) &= \mathbb{E}((X - \mu_X)(Y - \mu_Y)) \\
 &= \mathbb{E}(XY - \mu_X Y - X\mu_Y + \mu_X\mu_Y) \\
 &= \mathbb{E}(XY) - 2\mu_X\mu_Y + \mu_X\mu_Y \quad \text{by linearity of expectation} \\
 &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)
 \end{aligned}$$

From the definition, $\text{Var}(X) = \text{Cov}(X, X)$. A positive covariance reveals a positive linear relationship, and although it might not be exact linear relationship, the correlated RVs regress to each other. Figure [RVs with covariance 3.43](#) shows that X, Y have a positive linear relationship, with covariance 3.43.

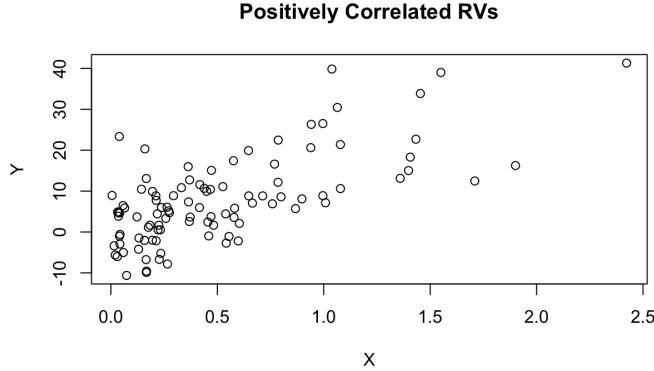


Figure 3.1: RVs with covariance 3.43

Proposition 3.3.5 (Covariance of Scaled RVs)

For some RVs X, Y and constants $a, b \in R$,

$$\text{Cov}(aX, bY) = ab \text{Cov}(X, Y)$$

Proof. Recall $\mathbb{E}(aX) = a\mu_X, \mathbb{E}(bX) = b\mu_Y$.

$$\begin{aligned}
 \text{Cov}(aX, bY) &= \mathbb{E}((aX - a\mu_X)(bY - b\mu_Y)) \\
 &= \mathbb{E}(ab(X - \mu_X)(Y - \mu_Y)) \\
 &= ab\mathbb{E}((X - \mu_X)(Y - \mu_Y)) \quad \text{by linearity of expectation} \\
 &= ab \text{Cov}(X, Y)
 \end{aligned}$$

□

Proposition 3.3.6 (Independence Implies No Covariance)

For independent $X \perp Y$, their covariance $\text{Cov}(X, Y) = 0$.

Proof. They have no relationship, thus no covariance (is that enough?). □

This is a more rigorous proof.

Proof.

$$\begin{aligned}
 \text{Cov}(X, Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\
 &= \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) \quad \text{by Expectation of Separable Independent Functions} \\
 &= 0
 \end{aligned}$$

□

Warning

This statement is *sufficient* but not *necessary*. Zero covariance can only deduce that there is no linear relationship between the two RVs, while independence is not guaranteed.

Proposition 3.3.7 (Variance of Linear Combination of RVs)

For some RVs X, Y and constants $a, b, c \in R$,

$$\text{Var}(aX + bY + c) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y)$$

Proof. Let $Z = aX + bY + c$, and $\mu_Z = a\mu_X + b\mu_Y + c$.

$$\begin{aligned}
 \text{Var}(Z) &= \mathbb{E}((Z - \mu_Z)^2) \\
 &= \mathbb{E}(((aX + bY + c) - (a\mu_X + b\mu_Y + c))^2) \\
 &= \mathbb{E}((aX + bY - a\mu_X - b\mu_Y)^2) \\
 &= \mathbb{E}((a(X - \mu_X) + b(Y - \mu_Y))^2) \\
 &= \mathbb{E}(a^2(X - \mu_X)^2 + 2ab(X - \mu_X)(Y - \mu_Y) + b^2(Y - \mu_Y)^2) \\
 &= a^2\mathbb{E}((X - \mu_X)^2) + 2ab\mathbb{E}((X - \mu_X)(Y - \mu_Y)) + b^2\mathbb{E}((Y - \mu_Y)^2) \quad \text{by linearity of Expectation} \\
 &= a^2 \text{Var}(X) + 2ab \text{Cov}(X, Y) + b^2 \text{Var}(Y)
 \end{aligned}$$

□

Remark

By Independence Implies No Covariance, if $X \perp Y$,

$$\text{Var}(aX + bY + c) = a^2 \text{Var}(X) + b^2 \text{Var}(Y)$$

Although covariance can provide information on whether two RVs are correlated and the direction (positive or negative) of regression, the magnitude of covariance depends on the standard deviation of X, Y , which are σ_X, σ_Y , respectively. That said, if we are given the covariance of X, Y without information on σ_X, σ_Y , we cannot deduce the strength of the linear relationship. In fact,

Lemma 3.3.8 (Range of Covariance)

For RVs X, Y ,

$$-\sigma_X\sigma_Y \leq \text{Cov}(X, Y) \leq \sigma_X\sigma_Y$$

Proof. We first claim that for two events A, B ,

$$P(A, B) \leq P(A)P(B)$$

such that the joint probability of X, Y is at most of the product of their marginal probability (equals when they independent). The proof is simple, since $A \cap B$ is a subset of A and a subset of B , $P(A \cap B) \leq P(A)$ and

$P(A \cap B) \leq P(B)$, which completes the proof. This implies that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |h(x)g(y)| f(x,y) dx dy \leq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |h(x)g(y)| f_X(x)f_Y(y) dx dy^a$$

By [Cauchy-Schwarz Inequality](#),

$$\left(\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |h(x)g(y)| f_X(x)f_Y(y) dx dy \right)^2 \leq \left(\int_{-\infty}^{\infty} h(x)^2 f_X(x) dx \right) \left(\int_{-\infty}^{\infty} g(y)^2 f_Y(y) dy \right)^b$$

Therefore

$$\begin{aligned} \mathbb{E}((X - \mu_X)(Y - \mu_Y))^2 &\leq \mathbb{E}((X - \mu_X)^2)\mathbb{E}((Y - \mu_Y)^2) \\ |\mathbb{E}((X - \mu_X)(Y - \mu_Y))| &\leq \sqrt{\mathbb{E}((X - \mu_X)^2)}\sqrt{\mathbb{E}((Y - \mu_Y)^2)} \\ |\text{Cov}(X, Y)| &\leq \sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)} \end{aligned}$$

Hence, we can prove that

$$|\text{Cov}(X, Y)| \leq \sigma_X \sigma_Y$$

as desired. \square

^aBy Bounded Function Theory in integration.

^bThis suggests that $\mathbb{E}(h(X)g(Y))^2 \leq \mathbb{E}(h(X)^2)\mathbb{E}(g(Y)^2)$.

The linear relationship is stronger if the magnitude $|\text{Cov}(X, Y)|$ is closer to $\sigma_X \sigma_Y$. If the magnitude is exactly $\sigma_X \sigma_Y$, then there exists an exact linear relationship between X, Y such that $Y = aX + b$ for some $a, b \in \mathbb{R}$. For convenience we can scale the covariance by $\sigma_X \sigma_Y$, then we obtain a measurement between 0 and 1, like a probability.

Definition 3.3.9 (Correlation Coefficient)

The correlation coefficient between RVs X, Y is defined as

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

Here we assumed that their standard deviation $\sigma_X = \sqrt{\text{Var}(X)}, \sigma_Y = \sqrt{\text{Var}(Y)}$ are non-zero. Let $r = \text{Corr}(X, Y)$, then

1. The sign of r measures the direction of correlation between X, Y .
2. $|r|$ measures the strength of correlation between X, Y , where

$$|r| \approx \begin{cases} 1 & \text{iff } X, Y \text{ have an approximate linear relationship} \\ 0 & \text{iff } X, Y \text{ have almost no linear relationship} \end{cases}$$

Bonus: What is the correlation between X and itself? From the definition we know that $\text{Var}(X) = \text{Cov}(X, X)$, so the correlation will be

$$\text{Corr}(X, X) = \frac{\text{Var}(X)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(X)}} = 1$$

A pretty strong and positive correlation! Although it is trivial, to be honest, there exists a linear relationship between X and itself, namely $X = 1 \cdot X$. This does not violate the completeness and correctness of the definition.

3.4 Moment Generating Functions

Definition 3.4.1 (Multivariate MGF)

The multivariate MGF of X, Y with respect to variables t, k is defined as

$$\mathcal{M}_{X,Y}(t, k) = \mathbb{E}(e^{tX+kY})$$

where $t, k \in (-h, h)$ for some $h > 0$.

Proposition 3.4.2 (Properties of Multivariate MGF)

Let $\mathcal{M}_{X,Y}(t, k)$ denote the MGF of X, Y .

1. The α, β -th moment of X, Y is

$$\mathbb{E}(X^\alpha Y^\beta) = \left. \frac{\partial^{\alpha+\beta}}{\partial t^\alpha \partial k^\beta} \right|_{t=k=0} \mathcal{M}_{X,Y}(t, k)$$

2. The marginal MGFs of X, Y are

$$\mathcal{M}_X(t) = \mathcal{M}_{X,Y}(t, 0) \quad \mathcal{M}_Y(k) = \mathcal{M}_{X,Y}(0, k)$$

The MGF is also useful for determining the distribution of transformation on a distribution or some distributions, since it uniquely characterizes a distribution.

Lemma 3.4.3 (MGF of Sum Distribution)

For **independently** distributed random variables X_1, \dots, X_n with corresponding MGF $\mathcal{M}_{X_1}(t), \dots, \mathcal{M}_{X_n}(t)$, the sum distribution $Y = \sum_{i=1}^n X_i$ has MGF

$$\mathcal{M}_Y(t) = \prod_{i=1}^n \mathcal{M}_{X_i}(t)$$

Proof.

$$\begin{aligned} \mathcal{M}_Y(t) &= \mathbb{E}(e^{tY}) = \mathbb{E}(e^{t(\sum_{i=1}^n X_i)}) = \mathbb{E}\left(\prod_{i=1}^n e^{tX_i}\right) \\ &= \prod_{i=1}^n \mathbb{E}(e^{tX_i}) \quad \text{by Expectation of Separable Independent Functions} \\ &= \prod_{i=1}^n \mathcal{M}_{X_i}(t) \end{aligned}$$

□

Remark

Note that the MGFs of the summand distributions are all based on the **same** t as for Y . This proof states that the MGF of the sum of independent distributions is the product their own MGF.

Example 3.4.1 (MGF of Sum Bernoulli RVs)

For $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} Bern(p)$, determine the distribution of

$$Y = \sum_{i=1}^n X_i$$

Solution

Note that the MGF of $X \sim Bern(p)$ is $\mathcal{M}_X(t)X = pe^t + q, q = 1 - p$. Thus by [MGF of Sum Distribution](#),

$$\mathcal{M}_Y(t) = \prod_{i=1}^n \mathcal{M}_X(t)X_i = \prod_{i=1}^n pe^t + q = (pe^t + q)^n$$

We notice that this is the MGF of $Bin(n, p)$. Therefore, we can conclude that the distribution of the sum of n independent Bernoulli $Bern(p)$ distributions is a Binomial $Bin(n, p)$ distribution.

Example 3.4.2 (MGF of Sum Poisson RVs)

For independent $X_i \sim Poi(\lambda_i), i = 1, \dots, n$, determine the distribution of

$$Y = \sum_{i=1}^n X_i$$

Solution

Recall from [Poisson Distribution MGF](#),

$$\mathcal{M}_{X_i}(t) = e^{\lambda_i(e^t - 1)}$$

By [MGF of Sum Distribution](#),

$$\mathcal{M}_Y(t) = \prod_{i=1}^n e^{\lambda_i(e^t - 1)} = e^{(e^t - 1) \sum_{i=1}^n \lambda_i}$$

which we see is the MGF of $Poi(\sum_{i=1}^n \lambda_i)$. We deduce that the sum of independent Poisson distributions is still a Poisson distribution whose mean is the sum of the marginals' means.

3.5 Max and Min Distributions

Lemma 3.5.1 (Product and Max Distribution)

Let $X_1 \sim F_1, \dots, X_n \sim F_n$ be continuous RVs with common their own function $F_1(x), \dots, F_n(x)$ respectively. Then their max

$$Y = \max \{X_1, \dots, X_n\}$$

has probability function

$$F_Y(y) = \prod_{i=1}^n F_i(y)$$

Proof.

$$\begin{aligned}
F_Y(y) &= P(Y < y) \\
&= P(\max \{X_1, \dots, X_n\} < y) \\
&= P(X_1 < y, \dots, X_n < y) \\
&= \prod_{i=1}^n F_i(y) \quad \text{by independence}
\end{aligned}$$

□

Remark

A special case is that: Let $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} F_X$ be continuous RVs with common probability function $F_X(x)$. Then their max

$$Y = \max \{X_1, \dots, X_n\}$$

has probability function

$$F_Y(y) = F(y)^n$$

Proof.

$$\begin{aligned}
F_Y(y) &= P(Y < y) \\
&= P(\max \{X_1, \dots, X_n\} < y) \\
&= P(X_1 < y, \dots, X_n < y) \\
&= P(X_1 < y) \cdots P(X_n < y) \quad \text{by independence} \\
&= F_X(y)^n
\end{aligned}$$

Also note that

$$f_Y(y) = \frac{d}{dx} y F_X(y)^n = n F_X(y)^{n-1} f_X(y)$$

They are valid since $f_Y(y) \geq 0$ and

$$\lim_{y \rightarrow \infty} F_Y(y) = 1 \quad \lim_{y \rightarrow -\infty} F_Y(y) = 0$$

□

Lemma 3.5.2 (Min Distribution)

Let $X_1 \sim F_1, \dots, X_n \sim F_n$ be continuous RVs with their own probability function $F_1(x), \dots, F_n(x)$. Then their min

$$Y = \min \{X_1, \dots, X_n\}$$

has probability function

$$F_Y(y) = 1 - \prod_{i=1}^n (1 - F_i(y))$$

Proof. We start with the TPF $\bar{F}_Y(y) = P(Y > y)$,

$$\begin{aligned} P(Y > y) &= P(\min \{X_1, \dots, X_n\} > y) \\ &= P(X_1 > y, \dots, X_n > y) \\ &= P(X_1 > y) \cdots P(X_n > y) \quad \text{by independence} \\ &= \prod_{i=1}^n (1 - F_i(y)) \\ P(Y < y) &= 1 - \prod_{i=1}^n (1 - F_i(y)) \end{aligned}$$

□

Remark

Let $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} F_X$ be continuous RVs with common probability function $F_X(x)$. Then their min

$$Y = \min \{X_1, \dots, X_n\}$$

has probability function

$$F_Y(y) = 1 - (1 - F_X(y))^n$$

Proof. We start with the TPF $\bar{F}_Y(y) = P(Y > y)$,

$$\begin{aligned} P(Y > y) &= P(\min \{X_1, \dots, X_n\} > y) \\ &= P(X_1 > y, \dots, X_n > y) \\ &= P(X_1 > y) \cdots P(X_n > y) \quad \text{by independence} \\ &= (1 - F_X(y))^n \\ P(Y < y) &= 1 - (1 - F_X(y))^n \end{aligned}$$

And

$$f_Y(y) = n(1 - F_X(y))^{n-1} f_X(y)$$

Note that they are both valid since $f_Y(y) \geq 0$ and

$$\lim_{y \rightarrow \infty} F_Y(y) = \lim_{y \rightarrow \infty} 1 - (1 - F_X(y))^n = 1 \quad \lim_{y \rightarrow -\infty} F_Y(y) = 0$$

□

3.6 Conditional Distributions

The conditional distribution of Y given X is denoted as $Y|X$. Here, X is the given condition, and Y is the conditional distribution itself, where the $|X$ is just a tail. For example, any function applied to this conditional distribution should be applied to the front Y , denoted by

$$h(Y)|X \quad \text{for some function } h(\cdot)$$

When talking about "given conditions", we assume that they have happened, such that they are already fixed events. For example, suppose X randomly generates an integer between 1 and 10, and Y randomly generates an integer between 2 and 8. Define

$$Z = \mathbf{1}_{\{Y+X>10\}}$$

such that Z is 1 if $X + Y > 10$, and 0 otherwise. Then Z is a joint distribution on X, Y , where the randomness of Z depends on both X and Y , i.e. normally you cannot determine Z unless you know the values of both X and Y . Now

suppose X has happened and generated a number 3, then the conditional distribution of Z is

$$Z|X=3 \equiv (Y+X)|X=3$$

Here, X is no longer random, and we can treat it as a constant, thus the distribution becomes

$$(Y+3)|X=3 \implies Z|X=3 = \mathbf{1}_{\{Y>7\}}$$

where now Z is 1 only when $Y > 7$ (only depending on Y), or equivalently, $Y = 8$. We will find this useful when finding the probability function of Z later. However, if X happens to be 9, that is,

$$(Y+X)|X=9 \implies (Y+9)|X=9$$

then

$$Z|X=10 = 1$$

such that Z is now a constant^{3.1} regardless the value of Y , since any value of Y would result in $Y+X > 10$.

Definition 3.6.1 (Conditional Distribution)

The conditional distribution of Y given $X = x$ has probability function

$$f(y|x) = \begin{cases} P_{Y|X}(y|x) = P(Y = y|X = x) = \frac{P(X=x, Y=y)}{P(X=x)} & \text{if } X, Y \text{ are dis.} \\ f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} & \text{if } X, Y \text{ are cts.} \end{cases}$$

and CDF

$$F_{Y|X}(y|x) = P(Y \leq y|X = x) = \begin{cases} \sum_{t=-\infty}^y P(Y = t|X = x) & \text{if dis.} \\ \int_{-\infty}^y f_{Y|X}(t|x) dt & \text{if cts.} \end{cases}$$

Definition 3.6.2 (Conditional Expectation)

The expectation of $X|Y = y$ is

$$\mathbb{E}(X|Y = y) = \begin{cases} \sum_{x=-\infty}^{\infty} xP(X = x|Y = y) & \text{if dis} \\ \int_{-\infty}^{\infty} xf_{X|Y}(x|y) dx & \text{if cts} \end{cases}$$

Do not expand the conditioned variable $Y = y$.

Theorem 3.6.3 (Law of Total Expectation)

For (possibly related) distributions X, Y , the expectation of X is

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

^{3.1}also called degenerate distribution

Proof. We only prove for continuous distributions.

$$\begin{aligned}
\mathbb{E}(X) &= \int_{-\infty}^{\infty} xf_X(x) dx \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f(x,y) dy dx \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f(x|y)f_Y(y) dy dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf(x|y) dx f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \mathbb{E}(X|Y=y) f_Y(y) dy \\
&= \mathbb{E}(\mathbb{E}(X|Y))
\end{aligned} \tag{*}$$

□

Remark

It also follows that, for some function $g(\cdot)$,

$$\mathbb{E}(g(X)) = \mathbb{E}(\mathbb{E}(g(X)|Y))$$

Notice that in (*), $h(y) = \mathbb{E}(X|Y = y)$ is a function on $Y = y$, that is, in $\mathbb{E}(X|Y)$, any *random variable other than Y* ^{3.2} has been "expected" (i.e. mapped to constant), and Y is the only source of randomness. Now look at the variance,

Theorem 3.6.4 (Law of Total Variance)

For (possibly related) distributions X, Y , the variance is

$$\text{Var}(X) = \mathbb{E}(\text{Var}(X|Y)) + \text{Var}(\mathbb{E}(X|Y))$$

Proof. A straightforward proof will be

$$\begin{aligned}
\mathbb{E}(\text{Var}(X|Y)) + \text{Var}(\mathbb{E}(X|Y)) &= \mathbb{E}(\mathbb{E}(X^2|Y) - \mathbb{E}(X|Y)^2) + \text{Var}(\mathbb{E}(X|Y)) \\
&= \mathbb{E}(\mathbb{E}(X^2|Y)) - \mathbb{E}(\mathbb{E}(X|Y)^2) + \text{Var}(\mathbb{E}(X|Y)) \quad \text{by Linearity of Expectation} \\
&= \mathbb{E}(X^2) - \mathbb{E}(\mathbb{E}(X|Y)^2) + \text{Var}(\mathbb{E}(X|Y)) \quad \text{by Law of Total Expectation} \\
&= \mathbb{E}(X^2) - \mathbb{E}(\mathbb{E}(X|Y)^2) + (\mathbb{E}(\mathbb{E}(X|Y)^2) - \mathbb{E}(\mathbb{E}(X|Y))^2) \quad \text{recall } \mathbb{E}(X|Y) \text{ is a RV} \\
&= \mathbb{E}(X^2) - \mathbb{E}(\mathbb{E}(X|Y))^2 \\
&= \mathbb{E}(X^2) - \mathbb{E}(X)^2 \quad \text{by Law of Total Expectation} \\
&= \text{Var}(X)
\end{aligned}$$

□

Remark

The total variance $\text{Var}(X)$ consists of two components: The expected value of variance of $X|Y$, $\text{Var}(X|Y)$, and variance of the expectation of $X|Y$, $\mathbb{E}(X|Y)$.

^{3.2}Not necessarily X ! We can reasonably assume that X is a transformation of Y if they are dependent, and it might also involve some other random variables.

Example 3.6.1 (Law of Total Expectation)

A manufacturer produces commemorative coins, while due to the bad quality of the outdated product line, most coins have different probabilities of tossing heads and tails. To detect if a randomly picked coin is fair, the examiners will toss the coin multiple times and record the proportion of heads to see how far it is away from $\frac{1}{2}$, and decide whether to dispose it based on the error. Let $X \sim U(0, 0.6)$ denote the randomly picked coin's probability of tossing a head, and with given X , let Y denote the number of heads tossed. Using law of total expectation and variance, find $\mathbb{E}(Y)$ and $\text{Var}(Y)$ with $n = 10$. Compare your result with the expected number of heads of tossing a fair coin, which follows $\text{Bin}(n, 0.5)$.

Solution

We have

$$Y|X \sim \text{Bin}(n, X) \quad \text{or} \quad Y|X = x \sim \text{Bin}(n, x)$$

By law of total expectation,

$$\begin{aligned}\mathbb{E}(Y) &= \mathbb{E}(\mathbb{E}(Y|X)) \\ &= \mathbb{E}(nX) \quad \text{mean of } \text{Bin}(n, X) \\ &= n\mathbb{E}(X) \quad \text{by linearity of expectation} \\ &= n \frac{0.6}{2} \quad \text{mean of } U(0, 0.6) \\ &= 10(0.3) \\ &= 3\end{aligned}$$

A fair coin would expect $10(0.5) = 5$ heads, whereas $\mathbb{E}(Y) = 3$ suggests that the coins produced by the product line have a smaller probability of tossing heads, e.g. have more weights on the head side. The manufacturer could fix the mint machine based on this information, if they prefer fair coins.

Also for the variance,

$$\begin{aligned}\text{Var}(Y) &= \mathbb{E}(\text{Var}(Y|X)) + \text{Var}(\mathbb{E}(Y|X)) \\ &= \mathbb{E}(nX(1 - X)) + \text{Var}(nX) \\ &= n(\mathbb{E}(X) - \mathbb{E}(X^2)) + n^2 \text{Var}(X) \quad \text{by property of mean and variance} \\ &= n(\mathbb{E}(X) - (\text{Var}(X) + \mathbb{E}(X)^2)) + n^2 \text{Var}(X) \quad \text{by variance formula} \\ &= 10(0.3 - (\frac{0.6^2}{12} + 0.3^2)) + 100 \frac{0.6^2}{12} \\ &= 10(0.3 - (0.03 + 0.09)) + 100(0.03) \\ &= 10(0.18) + 3 \\ &= 4.8\end{aligned}$$

A fair coin should have variance $(10)(0.5)(0.5) = 2.5$. That said, the produced coin has a larger variability than fair coin (why?).

The most straightforward way to determine the sum of independent distributions is multiplying their MGFs and see what the resulting MGF looks like, shown in [MGF of Sum Distribution](#). However, the formal way to determine a distribution is to find out its probability function, which involves conditional probability.

Lemma 3.6.5 (Sum Distribution)

Let X, Y be independent RVs. Let $Z = X + Y$, then the distribution of Z has characteristic function

$$\begin{cases} P(Z = z) = \sum_{x=-\infty}^{\infty} P(Y = z - x)P(X = x) & \text{if } X, Y \text{ dis.} \\ P(Z < z) = \int_{-\infty}^z P(Y < z - x)f_X(x) dx & \text{if } X, Y \text{ cts.} \end{cases}$$

Proof. dis: Conditioning on X , we have that

$$\begin{aligned} P(Z = z) &= \sum_{x=-\infty}^{\infty} P(Z = z, X = x) \\ &= \sum_{x=-\infty}^{\infty} P(Z = z|X = x)P(X = x) \quad \text{by Law of Total Probability} \\ &= \sum_{x=-\infty}^{\infty} P(\underline{X + Y}^a = z|X = x)P(X = x) \\ &= \sum_{x=-\infty}^{\infty} P(x + Y = z|X = x)P(X = x) \\ &= \sum_{x=-\infty}^{\infty} P(Y = z - x^b)P(X = x) \end{aligned}$$

cts: Still conditioning on X ,

$$\begin{aligned} P(Z < z) &= \int_{-\infty}^z P(Z < z|X = x)f_X(x) dx \\ &= \int_{-\infty}^z P(X + Y < z|X = x)f_X(x) dx \\ &= \int_{-\infty}^z P(Y < z - x)f_X(x) dx \end{aligned}$$

□

^aconvert dep. RVs Z, X into indep. Y, X by expanding Z as a function of Y

^bdrop condition since X, Y are indep.

Example 3.6.2 (Sum of Discrete Distributions)

Let $X \sim \text{Poi}(2)$ and $Y \sim \text{Bin}(10, 0.3)$ be independent. Find the distribution of $Z = X + Y$.

Solution

We have

$$P(X = x) = \frac{e^{-2}2^x}{x!} \mathbf{1}_{\{x=0,1,2,\dots\}}$$

and

$$P(Y = y) = \binom{10}{y} 0.3^y 0.7^{10-y} \mathbf{1}_{\{y=0,1,\dots,10\}}$$

The outcome of Z is discrete, thus we start with its PMF. X has sample space $\{0, 1, 2, \dots\}$ and Y has sample space

$\{0, 1, \dots, 10\}$, therefore $Z = X + Y$ has sample space $\{0, 1, 2, \dots\}$. Then we find the probability of $Z = z$ for each $z \in \{0, 1, 2, \dots\}$. Conditioning on Y ,

$$\begin{aligned}
P(Z = z) &= P(X + Y = z) \\
&= \sum_{y=0}^{10} P(X + Y = z | Y = y) P(Y = y) \\
&= \sum_{y=0}^{10} P(X = z - y) P(Y = y) \\
&= \sum_{y=0}^{10} \frac{e^{-2} 2^{z-y}}{(z-y)!} \binom{10}{y} 0.3^y 0.7^{10-y} \\
&= 0.7^{10} e^{-2} 2^z 10! \sum_{y=0}^{10} \frac{\left(\frac{3}{14}\right)^y}{(z-y)! y! (10-y)!}, z = 0, 1, 2, \dots
\end{aligned}$$

Example 3.6.3 (Sum of Continuous Distributions)

Let $X \sim \text{Exp}(1/3) \perp Y \sim \text{Uni}(1, 3)$. Find the distribution of $Z = X + Y$.

Solution

We have that

$$f_X(x) = 3e^{-3x} \mathbf{1}_{\{x>0\}} \quad f_Y(y) = \frac{1}{2} \mathbf{1}_{\{1 < y < 3\}}$$

Note that

$$F_X(x) = 1 - e^{-3x} \mathbf{1}_{\{x>0\}}$$

The support of Z is $[1, \infty)$.

Warning

The following derivation has wrong logic. Try to find out where the problem is.

$$\begin{aligned}
P(Z < z) &= \int_{-\infty}^{\infty} P(Z < z | Y = y) f_Y(y) dy \\
&= \int_1^3 P(X + Y < z | Y = y) f_Y(y) dy \\
&= \int_1^3 P(X < z - y) f_Y(y) dy \\
&= \int_1^3 (1 - e^{-3(z-y)}) \frac{1}{2} dy \\
&= 1 - \frac{1}{2} \int_1^3 e^{-3(z-y)} dy \\
&= 1 - \frac{1}{2} \frac{1}{3} e^{-3(z-y)} \Big|_1^3 \\
&= 1 - \frac{1}{6} (e^{-3(z-3)} - e^{-3(z-1)}) \\
&= 1 - \frac{1}{6} (e^9 - e^3) e^{-3z} \mathbf{1}_{\{z>1\}}
\end{aligned}$$

Or starting with the PDF,

$$\begin{aligned}
f_Z(z) &= \int_{-\infty}^{\infty} f_{Z|Y}(z|y)f_Y(y) dy \\
&= \int_1^3 f_X(z-y)f_Y(y) dy \\
&= \int_1^3 3e^{-3(z-y)} \frac{1}{2} dy \\
&= \frac{3}{2} \int_1^3 e^{-3(z-y)} dy \\
&= \frac{3}{2} \frac{1}{3} e^{-3(z-y)} \Big|_1^3 \\
&= \frac{1}{2} (e^{-3(z-3)} - e^{-3(z-1)}) \\
&= \frac{1}{2} (e^9 - e^3) e^{-3z} \mathbf{1}_{\{z>1\}}
\end{aligned}$$

Both of them are invalid probability functions.

Here goes the correct one. Note that when $z < 3$, Y cannot go beyond z . Thus we fix the logic

$$\begin{aligned}
P(Z < z) \text{ on } 1 < z < 3 &= \int_1^z P(Z < z|Y = y)f_Y(y) dy \\
&= \int_1^z P(X < z-y)f_Y(y) dy \\
&= \int_1^z (1 - e^{-3(z-y)}) \frac{1}{2} dy \\
&= \frac{z-1}{2} - \frac{1}{2} \int_1^z e^{-3(z-y)} dy \\
&= \frac{z-1}{2} - \frac{1}{2} \frac{1}{3} e^{-3(z-y)} \Big|_1^z \\
&= \frac{z-1}{2} - \frac{1}{6} (e^{-3(z-z)} - e^{-3(z-1)}) \\
&= \frac{z-1}{2} - \frac{1}{6} (1 - e^{-3(z-1)}) \\
&= \left(\frac{1}{6} e^{-3(z-1)} + \frac{z}{2} - \frac{2}{3} \right) \mathbf{1}_{\{1 < z < 3\}}
\end{aligned}$$

Note that

$$P(Z < 3) = \frac{1}{6} e^{-6} + \frac{3}{2} - \frac{2}{3} = \frac{1}{6} e^{-6} + \frac{5}{6}$$

When $z > 3$, Y can integrate all over support,

$$\begin{aligned}
P(3 < Z < z) &= \int_1^3 P(3 < Z < z | Y = y) f_Y(y) dy \\
&= \int_1^3 P(3 - y < X < z - y) f_Y(y) dy \\
&= \int_1^3 (e^{-3(3-y)} - e^{-3(z-y)}) \frac{1}{2} dy \\
&= \frac{1}{2}(e^{-9} - e^{-3z}) \int_1^3 e^{3y} dy \\
&= \frac{1}{6}(e^{-9} - e^{-3z}) e^{3y} \Big|_1^3 \\
&= \frac{1}{6}(e^{-9} - e^{-3z})(e^9 - e^3) \\
&= \left(-\frac{1}{6}(e^9 - e^3)e^{-3z} - \frac{1}{6}e^{-6} + \frac{1}{6} \right) \mathbf{1}_{\{z>3\}}
\end{aligned}$$

Therefore

$$P(Z < z) = \begin{cases} 0 & \text{if } z < 1 \\ \frac{1}{6}e^{-3(z-1)} + \frac{z}{2} - \frac{2}{3} & \text{if } 1 < z < 3 \\ 1 - \frac{1}{6}(e^9 - e^3)e^{-3z}^a & \text{if } z > 3 \end{cases}$$

The PDF is

$$f_Z(z) = \frac{1}{2}(1 - e^{-3(z-1)}) \mathbf{1}_{\{1 \leq z < 3\}} + \frac{1}{2}(e^9 - e^3)e^{-3z} \mathbf{1}_{\{z \geq 3\}}$$

^aHere is defined as $P(Z < 3) + P(3 < Z < z)$.

4 Known Distributions

4.1 Normal Distribution

Definition 4.1.1 (Normal Distribution)

A RV X following normal distribution with mean μ and variance σ^2 , denoted by $X \sim N(\mu, \sigma^2)$, has PDF

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

The standard normal distribution is

$$Z \sim N(0, 1)$$

with PDF

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right), z \in \mathbb{R}$$

and CDF

$$\Phi(z) := P(Z < z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-t^2} dt \quad (\text{Normal CDF})$$

For general $X \sim N(\mu, \sigma^2)$, the CDF can be manipulated by

$$\begin{aligned} P(X < x) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-(\frac{t-\mu}{\sigma})^2} dt \\ &\text{take } z = \frac{t-\mu}{\sigma} \text{ where } dz = \frac{1}{\sigma}dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-z^2} dz \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right) \text{ by (Normal CDF)} \end{aligned}$$

For convenience we first find $N(0, 1)$'s MGF

$$\mathcal{M}_Z(t) = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2 + tz} dz$$

Note that

$$-\frac{1}{2}z^2 + tz = -\frac{1}{2}(z^2 - 2tz + t^2 - t^2) = -\frac{1}{2}(z-t)^2 + \frac{1}{2}t^2$$

Substitute back in, and refer to [Scaled and Shifted Gaussian Integral](#),

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2 + tz} dz = \frac{e^{t^2/2}}{\sqrt{2\pi}} \underbrace{\int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-t)^2} dz}_{\sqrt{2\pi}}$$

hence

$$\mathcal{M}_Z(t) = e^{t^2/2} \quad (\star)$$

For the MGF of general $X \sim N(\mu, \sigma^2)$,

$$\begin{aligned}
\mathcal{M}_X(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2} dx \\
&\text{take } z = \frac{x-\mu}{\sigma} \text{ s.t. } dz = \frac{1}{\sigma} dx \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{t(\sigma z + \mu)} e^{-\frac{1}{2}z^2} \sigma dz = \frac{e^{\mu t}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t\sigma z} e^{-\frac{1}{2}z^2} dz \\
&= \frac{e^{\mu t}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z^2 - 2t\sigma z + t^2\sigma^2 - t^2\sigma^2)} dz = \frac{e^{\mu t + \frac{1}{2}\sigma^2 t^2}}{\sqrt{2\pi}} \underbrace{\int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-t\sigma)^2} dz}_{\sqrt{2\pi}} \\
&= e^{\mu t + \frac{1}{2}\sigma^2 t^2} \tag{Normal MGF}
\end{aligned}$$

The mean is

$$\mathbb{E}(X) = \frac{d}{dt} \Big|_{t=0} \mathcal{M}_X(t) = (\mu + \sigma^2 t) e^{\mu t + \frac{1}{2}\sigma^2 t^2} \Big|_{t=0} = \mu$$

The variance is

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \left\{ \underbrace{\sigma^2 e^{\mu t + \frac{1}{2}\sigma^2 t^2}}_{\sigma^2} + \underbrace{(\mu + \sigma^2 t)^2 e^{\mu t + \frac{1}{2}\sigma^2 t^2}}_{\mu^2} \right\} \Big|_{t=0} - \mu^2 = \sigma^2$$

For $a, b \in \mathbb{R}$, $Y = aX + b$ has MGF

$$\mathcal{M}_Y(t) = \mathbb{E}(e^{t(aX+b)}) = e^{tb} \mathbb{E}(e^{atX}) = e^{tb} e^{\mu(at) + \frac{1}{2}\sigma^2(at)^2} = e^{(a\mu+b)t + \frac{1}{2}(a^2\sigma^2)t^2} \tag{**}$$

which is the MGF of $N(a\mu + b, a^2\sigma^2)$.

Lemma 4.1.2 (Scaled and Shifted Normal Distribution)

For $X \sim N(\mu, \sigma^2)$ and $a, b \in \mathbb{R}$,

$$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

Proof. See MGF in (**). □

A quick result is that

Corollary 4.1.3 (Pivoting Normal Distribution)

Let $X \sim N(\mu, \sigma^2)$. Then

$$\frac{X - \mu}{\sigma} \sim N(0, 1)$$

Conversely, let $Z \sim N(0, 1)$. Then

$$\sigma Z + \mu \sim N(\mu, \sigma^2)$$

Recall σ is the standard deviation. Observe the shapes of $N(0, 1)$ and $N(\mu, \sigma^2)$. Changing the mean by adding a constant μ will shift the "bell" horizontally on the axis such that the mode is moved from 0 to μ , but won't affect the shape, whereas scaling the distribution by σ changes the shape of the "bell" with location unchanged. The mean μ is called *location parameter* and σ is called *scale parameter*.

Now suppose we have two **independent** normal RVs $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$. They have MGF

$$\mathcal{M}_X(t) = e^{\mu_1 t + \frac{1}{2}\sigma_1^2 t^2} \quad \mathcal{M}_Y(t) = e^{\mu_2 t + \frac{1}{2}\sigma_2^2 t^2}$$

By [MGF of Sum Distribution](#), their sum $X + Y$ has MGF

$$\mathcal{M}_X(t)\mathcal{M}_Y(t) = e^{\mu_1 t + \frac{1}{2}\sigma_1^2 t^2} e^{\mu_2 t + \frac{1}{2}\sigma_2^2 t^2} = e^{(\mu_1 + \mu_2)t + \frac{1}{2}(\sigma_1^2 + \sigma_2^2)t^2} \quad (\star\star)$$

which is the MGF of $N((\mu_1 + \mu_2), (\sigma_1^2 + \sigma_2^2))$.

Lemma 4.1.4 (Sum of Independent Normal RVs)

For independent $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$,

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Proof. See MGF in [\(★★\)](#). □

Combining [Scaled and Shifted Normal Distribution](#) and [Sum of Independent Normal RVs](#), we have

Corollary 4.1.5 (Linear Combination of Independent Normal RVs)

Let $X_1 \sim N(\mu_1, \sigma_1^2), \dots, X_n \sim N(\mu_n, \sigma_n^2)$ be independent. Then for $a_1, \dots, a_n \in \mathbb{R}, b_1, \dots, b_n \in \mathbb{R}$,

$$\sum_{i=1}^n a_i X_i + b_i \sim N\left(\sum_{i=1}^n a_i \mu_i + \sum_{i=1}^n b_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right) \quad (\star)$$

Conversely, if [\(★\)](#) holds, then $X_1 \sim N(\mu_1, \sigma_1^2), \dots, X_n \sim N(\mu_n, \sigma_n^2)$ must be independent.

4.2 Gamma Distribution

The *Gamma Function* is defined as

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt \quad (\text{Gamma Function})$$

for $z \in \mathbb{R}$ except at negative integers. Note that for positive integers $z \in \mathbb{Z}$,

$$\Gamma(z) = (z-1)!$$

which is the factorial of $z-1$.

Definition 4.2.1 (Gamma Distribution)

A RV X follows a Gamma distribution with parameters $\alpha, \beta > 0$, denoted by $X \sim \text{Gamma}(\alpha, \beta)$, has PDF

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, x > 0$$

Note that it is a valid PDF since

$$\begin{aligned} \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha-1} e^{-x/\beta} dx \\ &= \frac{\beta^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty (x/\beta)^{\alpha-1} e^{-x/\beta} dx \\ &\quad \text{take } u = x/\beta \text{ where } dx = \beta du \\ &= \frac{\beta^{\alpha-1}\beta}{\Gamma(\alpha)\beta^\alpha} \underbrace{\int_0^\infty u^{\alpha-1} e^{-u} du}_{\Gamma(\alpha)} \\ &= 1 \end{aligned} \quad (\star)$$

The *incomplete gamma function* (little gamma function) is

$$\gamma(\alpha, x) = \int_0^x e^{-t} t^{\alpha-1} dt \quad (\text{Incomplete Gamma})$$

where $\gamma(\alpha, \infty) = \Gamma(\alpha)$. Refer to (\star) , the CDF is thus

$$\begin{aligned} F(x) &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^x t^{\alpha-1} e^{-t/\beta} dt \\ &= \frac{1}{\Gamma(\alpha)} \int_0^{x/\beta} u^{\alpha-1} e^{-u} du \\ &= \frac{\gamma(\alpha, x/\beta)}{\Gamma(\alpha)} \end{aligned} \quad (\text{Gamma CDF})$$

Its MGF $\mathcal{M}_X(t)$ is

$$\begin{aligned} \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty e^{tx} \cdot x^{\alpha-1} e^{-x/\beta} dx &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha-1} e^{-(1/\beta-t)x} dx \\ &\quad \text{take } u = (1/\beta - t)x \text{ where } du = [(1 - \beta t)/\beta]dx \\ &= \frac{[\beta/(1 - \beta t)]^{1+(\alpha-1)}}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty u^{\alpha-1} e^{-u} du \\ &= \frac{1}{(1 - \beta t)^\alpha}, \quad t < 1/\beta \end{aligned} \quad (\text{Gamma MGF})$$

Note that the Gamma function requires $1/\beta - t > 0$, thus t is constrained to be less than $1/\beta$. The first moment is

$$\mathcal{M}^{(1)} = -\alpha(1 - \beta t)^{-\alpha-1}(-\beta) = \alpha\beta(1 - \beta t)^{-\alpha-1}$$

and by taking $t = 0$, the mean is thus α . The second moment is

$$\mathcal{M}^{(2)} = \alpha(\alpha + 1)\beta^2(1 - \beta t)^{-\alpha-2}$$

The variance is thus $\alpha\beta^2$. Note that its n th moment for $n \in \mathbb{Z}^+$ is

$$\mathcal{M}^{(n)} = \left(\prod_{i=0}^{n-1} \alpha + i \right) \beta^n (1 - \beta t)^{-\alpha-n}$$

which is

$$\left(\prod_{i=0}^{n-1} \alpha + i \right) \beta^n$$

by taking $t = 0$.

- 4.3 Beta Distribution**
- 4.4 Chi Squared Distribution**
- 4.5 Student t Distribution**

5 Known Multivariate Distributions

5.1 Multinomial Distribution

Just like in binomial distribution, we are performing n independently and identically distributed trials, but this time we are making the outcomes more complex. Binomial distribution, which only has two types of outcomes, is a special case of multinomial distribution, which consists of $k \geq 2$ types of different outcomes.

5.1.1 Definitions

Say the types of outcomes are numbered by $1, 2, \dots, k$ with corresponding probabilities $\theta_1, \theta_2, \dots, 1 - (\theta_1 + \dots + \theta_{k-1})$, and among the total n independent trials, there are t_1 type-1 outcomes, t_2 type-2 outcomes, ..., and $n - (t_1 + \dots + t_{k-1})$ type- k outcomes. Note that we have $k - 1$ degrees of freedom, so there are $k - 1$ variables in multinomial distributions, namely

$$\mathbf{T} = (T_1, \dots, T_k)^T, T_k = n - \sum_{i=1}^{k-1} T_i$$

with parameters

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^T, \theta_k = 1 - \sum_{i=1}^{k-1} \theta_i$$

Definition 5.1.1 (Multinomial Distribution)

A RV $T = (T_1, \dots, T_{k-1})^T$ that follows a multinomial distribution with probabilities $\theta_1, \dots, \theta_{k-1}$ such that $0 < \theta_1, \dots, \theta_{k-1} \leq 1, \sum_{i=1}^{k-1} \theta_i \leq 1$, denoted by $T \sim \text{Mult}(n, \theta_1, \dots, \theta_{k-1})$, has PMF

$$P(T = (t_1, \dots, t_{k-1})) = \frac{n!}{t_1! \cdots t_k!} \theta_1^{t_1} \cdots \theta_k^{t_k} \text{ where } t_k = n - \sum_{i=1}^{k-1} t_i, \theta_k = 1 - \sum_{i=1}^{k-1} \theta_i$$

for $0 \leq t_1, \dots, t_{k-1} \leq n$ and $\sum_{i=1}^{k-1} t_i \leq n$.

Another common way to represent it is

$$T = (T_1, \dots, T_k)^T \sim \text{Mult}(n, \theta_1, \dots, \theta_k), \sum_{i=1}^k \theta_i = 1$$

with

$$P(T = (t_1, \dots, t_k)) = \frac{n!}{t_1! \cdots t_k!} \theta_1^{t_1} \cdots \theta_k^{t_k} \mathbf{1}_{\{0 \leq t_1, \dots, t_k \leq n \text{ and } \sum_{i=1}^k t_i = n\}}$$

We break it up. The term $\theta_1^{t_1} \cdots \theta_k^{t_k}$ is the probability of the n independent trials having t_1 type-1 outcomes, ..., t_{k-1} type- $(k-1)$ outcomes and t_k type- k outcomes. However, we also need to count for which t_1 trials are outcomes 1, which t_2 trials are outcomes 2, etc. The term $\binom{n}{t_1, \dots, t_k} = \frac{n!}{t_1! \cdots t_k!}$ is the multinomial combination such that instead of choosing t objects out of the n total objects like in normal $\binom{n}{t}$, we are choosing t_1 many objects, then t_2 many, ..., until t_{k-1} and see how many ways are there in total. The derivation is interesting. First, we choose t_1 objects from the total n objects, which has $\binom{n}{t_1} = \frac{n!}{t_1!(n-t_1)!}$ ways. Then, of the rest $n - t_1$ objects, we choose t_2 objects from them, which has $\binom{n-t_1}{t_2} = \frac{(n-t_1)!}{t_2!(n-t_1-t_2)!}$ ways. Repeat until t_{k-1} , and the total number of ways is their product,

$$\binom{n}{t_1} \binom{n-t_1}{t_2} \binom{n-t_1-t_2}{t_3} \cdots = \frac{n!}{t_1!(n-t_1)!} \frac{(n-t_1)!}{t_2!(n-t_1-t_2)!} \frac{(n-t_1-t_2)!}{t_3!(n-t_1-t_2-t_3)!} \cdots = \frac{n!}{t_1! \cdots t_k!}$$

We can always adjust the parameters for special requirements. For example, if there are 10 outcomes numbered from 1 to 10, but we only care about the occurrences of numbers from 1 to 6 with corresponding probabilities $P(1), \dots, P(6)$, ignoring the rest numbers. Then this case has 7 types, namely from 1 to 6, and one extra type for the discarded numbers 7 to 10^{5.1}, with occurrence probability $P(7) + P(8) + P(9) + P(10)$. That is,

$$Mult(n, P(1), \dots, P(6)), \text{discarded type with probability } P(7) + P(8) + P(9) + P(10)$$

Example 5.1.1 (Multinomial Example)

Suppose we are rolling an unfair die 20 times independently. Let T denote the numbers rolled. The probabilities are

$$P(X = x) = \begin{cases} 0.3 & \text{if } x = 1 \\ 0.01 & \text{if } x = 2 \\ 0.4 & \text{if } x = 3 \\ 0.05 & \text{if } x = 4 \\ 0.15 & \text{if } x = 5 \\ 0.09 & \text{if } x = 6 \end{cases}$$

What is the probability of rolling two 1's, one 2, four 3's, six 4's, three 5's, four 6's?

Solution

The probability vector is

$$\theta = (0.3, 0.01, 0.4, 0.05, 0.15, 0.09)^T$$

The target vector is

$$t = (2, 1, 4, 6, 3, 4)^T$$

The probability is thus

$$P(T = t) = \frac{20!}{2!1!4!6!3!4!} 0.3^2 0.01^1 0.4^4 0.05^6 0.15^3 0.09^4 \approx 3.897 \times 10^{-8}$$

As you can see, as the sample space is huge, the probability of a single outcome is really small.

5.1.2 MGF

The MGF of multinomial distributions is

$$\mathcal{M}_T(\lambda) = \mathbb{E}(e^{\lambda^T T})$$

where $\lambda = (\lambda_1, \dots, \lambda_{k-1})^T$. Expand it,

$$\begin{aligned} \mathcal{M}_T(\lambda) &= \mathbb{E}(e^{\lambda^T T}) \\ &= \sum_{t_1+\dots+t_k=n} e^{\lambda_1 t_1 + \dots + \lambda_{k-1} t_{k-1}} \binom{n}{t_1, \dots, t_k} \theta_1^{t_1} \dots \theta_k^{t_k} \\ &= \sum_{t_1+\dots+t_k=n} \binom{n}{t_1, \dots, t_k} e^{\lambda_1 t_1} \theta_1^{t_1} \dots e^{\lambda_{k-1} t_{k-1}} \theta_{k-1}^{t_{k-1}} \theta_k^{t_k} \\ &= \sum_{t_1+\dots+t_k=n} \binom{n}{t_1, \dots, t_k} (\theta_1 e^{\lambda_1})^{t_1} \dots (\theta_{k-1} e^{\lambda_{k-1}})^{t_{k-1}} \theta_k^{t_k} \\ &= (\theta_1 e^{\lambda_1} + \dots + \theta_{k-1} e^{\lambda_{k-1}} + \theta_k)^n \text{ by Multinomial Theorem} \end{aligned}$$

^{5.1}We treat the occurrences of these numbers as one type. Since we have $k - 1$ degrees of freedom, the multinomial random vector only has $k - 1$ free variables, where exactly one of the types has no variable, a.k.a "discarded". However, any type could be the "discarded" type, depending on your choice.

5.1.3 Marginals

Now we proceed with the marginal distribution of each entry $T_i, i = 1, \dots, k$. Intuitively, if we do not care about the occurrences of other types, then it becomes a "yes or no" problem, where for each of the n trials, count once if it is type- i outcome, and skip otherwise. This is clearly the $\text{Bin}(n, \theta_i)$ distribution.

Proposition 5.1.2 (Marginal Distribution of Multinomial Distribution)

For multinomial RV $T \sim \text{Mult}(n, \theta_1, \dots, \theta_{k-1})$, the marginal distribution is $T_i \sim \text{Bin}(n, \theta_i), \forall i = 1, \dots, k-1$.

Proof. The support of T_i is $\{0, 1, \dots, n\}$. Let X_i denote the total number of outcomes other than type i .

$$\begin{aligned} P(T_i = t) &= P(T_i = t, X_i = n-t) = \frac{n!}{t!(n-t)!} \theta_i^t (1-\theta_i)^{n-t} \\ &= \binom{n}{t} \theta_i^t (1-\theta_i)^{n-t} \mathbf{1}_{\{t=0,1,\dots,n\}} \end{aligned}$$

which is the PMF of $\text{Bin}(n, \theta_i)$. \square

Proposition 5.1.3 (Sum of Multinomial Marginals)

For multinomial RV $T \sim \text{Mult}(n, \theta_1, \dots, \theta_{k-1})$, $T_i + T_j \sim \text{Bin}(n, \theta_i + \theta_j), \forall i \neq j = 1, \dots, k-1$.

Proof. We can group the outcomes of T_i, T_j as one type and treat the occurrences of others as the "discarded" type. A formal proof goes as

$$\begin{aligned} P(T_i + T_j = x) &= \sum_{t_i+t_j=x} \frac{n!}{t_1! \dots t_i! \dots t_j! \dots t_k!} \theta_1^{t_1} \dots \theta_i^{t_i} \dots \theta_j^{t_j} \dots \theta_k^{t_k}, \text{ constraint could be replaced by } \sum_{m \neq i,j} T_m = n - x \\ &= \sum_{t_i+t_j=x} \frac{n!}{t_i! t_j! \prod_{m \neq i,j} t_m!} \theta_i^{t_i} \theta_j^{t_j} \prod_{m \neq i,j} \theta_m^{t_m} \\ &= \sum_{t_i+t_j=x} \frac{n!}{x!(n-x)!} \frac{x!}{t_i! t_j!} \frac{(n-x)!}{\prod_{m \neq i,j} t_m!} \theta_i^{t_i} \theta_j^{t_j} \prod_{m \neq i,j} \theta_m^{t_m} \\ &= \sum_{t_i+t_j=x} \binom{n}{x} \frac{x!}{t_i! t_j!} \frac{(n-x)!}{\prod_{m \neq i,j} t_m!} \frac{\theta_i^{t_i} \theta_j^{t_j}}{(\theta_i + \theta_j)^x} \frac{\prod_{m \neq i,j} \theta_m^{t_m}}{(1 - \theta_i - \theta_j)^{n-x}} (\theta_i + \theta_j)^x (1 - \theta_i - \theta_j)^{n-x} \\ &= \binom{n}{x} (\theta_i + \theta_j)^x (1 - \theta_i - \theta_j)^{n-x} \underbrace{\sum_{t_i+t_j=x} \frac{x!}{t_i! t_j!} \frac{(n-x)!}{\prod_{m \neq i,j} t_m!} \left(\frac{\theta_i}{\theta_i + \theta_j} \right)^{t_i} \left(\frac{\theta_j}{\theta_i + \theta_j} \right)^{t_j} \prod_{m \neq i,j} \left(\frac{\theta_m}{1 - \theta_i - \theta_j} \right)^{t_m}}_{\text{Sum of Conditional Multinomial PMF with } n \text{ trials, } k \text{ types, given } t_i + t_j = x} \\ &= \binom{n}{x} (\theta_i + \theta_j)^x (1 - \theta_i - \theta_j)^{n-x} \mathbf{1}_{\{x=0,1,\dots,n\}} \end{aligned}$$

which is the PMF of $\text{Bin}(n, \theta_i + \theta_j)$. \square

Proposition 5.1.4 (Covariance Between Multinomial Marginals)

For marginals $T_i, T_j, i \neq j$,

$$\text{Cov}(T_i, T_j) = -np_i p_j$$

Proof. Recall Properties of Multivariate MGF, $\mathcal{M}_{T_i, \lambda_i}(T_j, \lambda_j)$ is obtained by taking $\lambda_m = 0, m = 1, \dots, k-1$

except for $m = i, j$. The joint MGF is thus

$$\mathcal{M}_{T_i, \lambda_i}(T_j, \lambda_j) = (\theta_i e^{\lambda_i} + \theta_j e^{\lambda_j} + (1 - \theta_i - \theta_j))^n$$

Note that $\mathbb{E}(T_i) = n\theta_i$, $\mathbb{E}(T_j) = n\theta_j$ since they are binomial distributions. With

$$\mathbb{E}(T_i T_j) = \frac{\partial^2}{\partial \lambda_i \partial \lambda_j} \Big|_{\lambda_1 = \lambda_2 = 0} (\theta_i e^{\lambda_i} + \theta_j e^{\lambda_j} + (1 - \theta_i - \theta_j))^n = n^2 \theta_i \theta_j - n \theta_i \theta_j$$

The covariance is thus

$$\mathbb{E}(T_i T_j) - \mathbb{E}(T_i) \mathbb{E}(T_j) = -n \theta_i \theta_j$$

□

5.1.4 Conditional Multinomial Distribution

If assumed that some marginal T_i has happened, what is the conditional distribution of T ? This can be denoted by

$$T|T_i = t_i, i = 1, \dots, k-1$$

With t_i fixed, there are only $n - t_i$ stochastic trials left, and also $k - 2$ free variables, namely $T_1, \dots, T_{i-1}, T_{i+1}, \dots, T_{k-1}$. Their marginal probabilities are now $\frac{\theta_1}{1-\theta_i}, \dots, \frac{\theta_{i-1}}{1-\theta_i}, \frac{\theta_{i+1}}{1-\theta_i}, \dots, \frac{\theta_{k-1}}{1-\theta_i}$, and the "discarded" type has probability $\frac{\theta_k}{1-\theta_i}$ since these $k-1$ probabilities need to sum to 1. The distribution is thus $Mult(n-t_i, \frac{\theta_1}{1-\theta_i}, \dots, \frac{\theta_{i-1}}{1-\theta_i}, \frac{\theta_{i+1}}{1-\theta_i}, \dots, \frac{\theta_{k-1}}{1-\theta_i})$. Using the die example seen previously, say rolling a die 20 times and we know that there are four 6's, the conditional probabilities for 1, 2, 3, 4, 5 are $\frac{0.3}{0.91}, \frac{0.01}{0.91}, \frac{0.4}{0.91}, \frac{0.05}{0.91}, \frac{0.15}{0.91}$ respectively, sharing the rest $20 - 4 = 16$ trials. Formally,

Proposition 5.1.5 (Conditional Multinomial Distribution)

For multinomial RV $T \sim Mult(n, \theta_1, \dots, \theta_{k-1})$,

$$T|T_i = t_i \sim Mult\left(n - t_i, \frac{\theta_1}{1-\theta_i}, \dots, \frac{\theta_{i-1}}{1-\theta_i}, \frac{\theta_{i+1}}{1-\theta_i}, \dots, \frac{\theta_{k-1}}{1-\theta_i}\right)$$

Proof.

$$\begin{aligned} P(T = (t_1, \dots, t_i, \dots, t_{k-1}) | T_i = t_i) &= \frac{T = (t_1, \dots, t_i, \dots, t_{k-1})}{P(T_i = t_i)} \\ &= \frac{\cancel{t_1! \dots t_{i-1}! t_{i+1}! \dots t_{k-1}!} \theta_1^{t_1} \dots \cancel{\theta_i^{t_i}} \dots \theta_{k-1}^{t_{k-1}}}{\cancel{t_i! (n-t_i)!} \theta_i^{t_i} (1-\theta_i)^{n-t_i}} \text{ by Marginal Distribution of Multinomial Distribution} \\ &= \frac{(n-t_i)!}{t_1! \dots t_{i-1}! t_{i+1}! \dots t_{k-1}!} \left(\frac{\theta_1}{1-\theta_i}\right)^{t_1} \dots \left(\frac{\theta_{i-1}}{1-\theta_i}\right)^{t_{i-1}} \left(\frac{\theta_{i+1}}{1-\theta_i}\right)^{t_{i+1}} \dots \left(\frac{\theta_{k-1}}{1-\theta_i}\right)^{t_{k-1}} \end{aligned}$$

a

b

on $0 \leq t_1, \dots, t_{i-1}, t_{i+1}, \dots, t_{k-1} \leq n$ and $t_1 + \dots + t_{i-1} + t_{i+1} + \dots + t_{k-1} \leq n$

as desired. □

Remark

It also follows that for $i \neq j$,

1. $T|T_i = t_i, T_j = t_j$ follows

$$Mult\left(n - t_i - t_j, \frac{\theta_1}{1-\theta_i - \theta_j}, \dots, \frac{\theta_{i-1}}{1-\theta_i - \theta_j}, \frac{\theta_{i+1}}{1-\theta_i - \theta_j}, \dots, \frac{\theta_{j-1}}{1-\theta_i - \theta_j}, \frac{\theta_{j+1}}{1-\theta_i - \theta_j}, \dots, \frac{\theta_{k-1}}{1-\theta_i - \theta_j}\right)$$

with $k - 3$ degrees of freedom.

2. $T|T_i + T_j = x$ for $x = 0, 1, \dots, n$ follows

$$Mult\left(n, \frac{\theta_i}{\theta_i + \theta_j}, \left\{ \frac{\theta_m}{1 - \theta_i - \theta_j} : m \in \{1, \dots, k-1\} \setminus \{i, j\} \right\} \right)$$

with $k - 2$ degrees of freedom, since

$$\begin{aligned} P(T = t|T_i + T_j = x) &= \frac{\sum_{t_i+t_j=x} \frac{n!}{t_1! \cdots t_i! \cdots t_j! \cdots t_k!} \theta_1^{t_1} \cdots \theta_i^{t_i} \cdots \theta_j^{t_j} \cdots \theta_k^{t_k}}{\frac{n!}{x!(n-x)!} (\theta_i + \theta_j)^x (1 - \theta_i - \theta_j)^{n-x}}, \quad t_i + t_j = x \\ &= \frac{x!}{t_i! t_j!} \frac{(n-x)!}{\prod_{m \neq i,j} t_m!} \left(\frac{\theta_i}{\theta_i + \theta_j} \right)^{t_i} \left(\frac{\theta_j}{\theta_i + \theta_j} \right)^{t_j} \prod_{m \neq i,j} \left(\frac{\theta_m}{1 - \theta_i - \theta_j} \right)^{t_m} \\ &\text{for } 0 \leq t_1, \dots, t_k \leq n \text{ except for } t_i, t_j \text{ and } \sum_{m \neq i,j} t_m = n - x \end{aligned}$$

^aallocating $(1 - \theta_i)^{n-t_i}$ to each $\theta_j^{t_j}, j \neq i$

^bnote that t_i is no longer a variable, since it is conditioned (fixed)

Example 5.1.2

Let $X_i \sim Poi(\lambda_i)$ for $i = 1, \dots, n$. Show that $(X_1, \dots, X_n)|X_1 + \dots + X_n = t$ follows a multinomial distribution.

Solution

Recall from MGF of Sum Poisson RVs,

$$X_1 + \dots + X_n \sim Poi(\lambda_1 + \dots + \lambda_n)$$

Let $T = \sum_{i=1}^n X_i$, we have

$$\begin{aligned} P(X_1 = x_1, \dots, X_n = x_n | T = t) &= \frac{P(X_1 = x_1, \dots, X_n = x_n, T = t)}{P(T = t)}, \quad x_1 + \dots + x_n = t \\ &= \frac{\prod_{i=1}^n e^{-\lambda_i} \lambda_i^{x_i} / x_i!}{e^{-\lambda} \lambda^t / t!}, \quad \lambda_1 + \dots + \lambda_n = \lambda \\ &= \frac{e^{-\sum_{i=1}^n \lambda_i} \prod_{i=1}^n \lambda_i^{x_i} / \prod_{i=1}^n x_i!}{e^{-\lambda} \lambda^t / t!} \\ &= \frac{t!}{\prod_{i=1}^n x_i!} \frac{\prod_{i=1}^n \lambda_i^{x_i}}{\lambda^t} = \frac{t!}{\prod_{i=1}^n x_i!} \frac{\prod_{i=1}^n \lambda_i^{x_i}}{\lambda^{\sum_{i=1}^n x_i}} \\ &= \frac{t!}{\prod_{i=1}^n x_i!} \frac{\prod_{i=1}^n \lambda_i^{x_i}}{\prod_{i=1}^n \lambda^{x_i}} = \frac{t!}{\prod_{i=1}^n x_i!} \prod_{i=1}^n \left(\frac{\lambda_i}{\lambda} \right)^{x_i} \\ &= \frac{t!}{\prod_{i=1}^n x_i!} \prod_{i=1}^n \theta_i^{x_i}, \quad \theta_i = \frac{\lambda_i}{\lambda} \end{aligned}$$

which is the PMF of $Mult(t; \theta_1, \dots, \theta_n)$. Notice that the θ 's are valid probabilities since $\theta_1 + \dots + \theta_n = 1$.

5.2 Multivariate Uniform Distribution

Suppose we have n uniform RVs

$$U_1, \dots, U_n \sim Uni(0, 1)$$

We can represent them as a random vector

$$\mathbf{U} = (U_1, \dots, U_n)^T \in Uni(0, 1)^n$$

Their joint CDF is called **Copula**, where

$$C(u_1, \dots, u_n) = P(U_1 < u_1, \dots, U_n < u_n), 0 < u_1, \dots, u_n < 1 \quad (\text{Copula})$$

If they are independent, then the copula is the product of their marginal probabilities,

$$C(u_1, \dots, u_n) = \prod_{i=1}^n P(U_i < u_i) = \prod_{i=1}^n u_i \quad (\text{Indep. Copula})$$

5.3 Multivariate Normal Distribution

5.3.1 Definitions

Just like multivariate uniform distributions, we stack marginal normal RVs $Z_1 \sim N(\mu_1, \sigma_1^2), \dots, Z_n \sim N(\mu_n, \sigma_n^2)$ (possibly co-related) in a vector $\mathbf{Z} = (Z_1, \dots, Z_n)^T \in \mathbb{R}^n$. A random vector has expectation

$$\mathbb{E}(\mathbf{Z}) = (\mathbb{E}(Z_1), \dots, \mathbb{E}(Z_n))^T$$

where in this case

$$\mathbb{E}(\mathbf{Z}) = (\mu_1, \dots, \mu_n)^T := \boldsymbol{\mu} \in \mathbb{R}^n$$

The variance $\text{Var}(\mathbf{Z})$ is the same as the covariance matrix of Z_1, \dots, Z_n , which is

$$\text{Var}(\mathbf{Z}) = [\text{Cov}(Z_i, Z_j)]_{ij} = \begin{pmatrix} \text{Var}(Z_1) & \text{Cov}(Z_1, Z_2) & \dots & \text{Cov}(Z_1, Z_n) \\ \text{Cov}(Z_2, Z_1) & \text{Var}(Z_2) & \dots & \text{Cov}(Z_2, Z_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Z_n, Z_1) & \text{Cov}(Z_n, Z_2) & \dots & \text{Var}(Z_n) \end{pmatrix} := \Sigma \in \mathbb{S}^n$$

where here for MVN, if $\rho_{ij} := \text{Corr}(Z_i, Z_j)$,

$$\text{Var}(\mathbf{Z}) = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \dots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \dots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1}\sigma_n\sigma_1 & \rho_{n2}\sigma_n\sigma_2 & \dots & \sigma_n^2 \end{pmatrix}$$

We formally define the multivariate normal distribution,

Definition 5.3.1 (Multivariate Normal Distribution)

$\mathbf{Z} = (Z_1, \dots, Z_n)^T$ follows a multivariate normal distribution, denoted by

$$\mathbf{Z} \sim MVN_n(\boldsymbol{\mu}, \Sigma)$$

such that

$$\mathbb{E}(\mathbf{Z}) = \boldsymbol{\mu} \quad \text{Var}(\mathbf{Z}) = \Sigma$$

iff

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{z} - \boldsymbol{\mu})\right), \quad \mathbf{z} = (z_1, \dots, z_n)^T \in \mathbb{R}^n$$

where $Z_1 \sim N(\mu_1, \dots, \sigma_1^2), \dots, Z_n \sim N(\mu_n, \dots, \sigma_n^2)$, and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$, $\Sigma = [\text{Cov}(Z_i, Z_j)]_{ij}^{n \times n}$.

Note that the covariance matrix Σ is defined as

$$\Sigma = \mathbb{E}((Z - \boldsymbol{\mu})(Z - \boldsymbol{\mu})^T) \stackrel{5.3}{=} \mathbb{E}(ZZ^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T$$

As well, Σ must be invertible and $|\Sigma| > 0$.

5.3.2 Stochastic Representations

The standard n -dimensional MVN RV is

$$Z \stackrel{5.4}{\sim} MVN_n(\mathbf{0}, I) \stackrel{5.5}{\sim}$$

^{5.2} \mathbb{S}^n refers to the $n \times n$ Symmetric Matrix Space, and Σ is indeed symmetric.

^{5.3}For vectors $x, y \in \mathbb{R}^n$, xy^T is called the outer product of x, y . See Outer Product of Vectors.

^{5.4}The marginals of standard MVN RV are independent.

^{5.5} $\mathbf{0}$ is a vector of n 0's, and I is the $n \times n$ identity matrix.

We have that

$$Z^T Z = \sum_{i=1}^n Z_i^2 \sim \chi^2(n) \quad (\text{MVN Inner})$$

or equivalently

$$\|Z\|^2 \sim \chi^2(n)$$

is a summation of n independent squared $N(0, 1)$ RVs, which follows $\chi^2(n)$. Like univariate normal, the standard MVN RV can be used to implement any

$$X \sim MVN(\mu, \Sigma)$$

Before approaching this, we need some more random vector techniques. For random vector $Z \in \mathbb{P}^n$ and constant $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$,

$$\mathbb{E}(AZ + b) = A\mathbb{E}(Z) + b \quad (\triangle)$$

More generally, for independent^{5.6} random matrices B, C with compatible dimensions,

$$\mathbb{E}(BC) = \mathbb{E}(B)\mathbb{E}(C) \quad (\star)$$

This can be quickly proven:

Proof. For a random matrix $B \in \mathbb{P}^{m \times n}$ and random vector $Z \in \mathbb{P}^n$ that are independent,

$$\begin{aligned} \mathbb{E}(BZ) &= \mathbb{E} \begin{bmatrix} \sum_{j=1} b_{1j} z_j \\ \vdots \\ \sum_{j=1} b_{mj} z_j \end{bmatrix} = \begin{bmatrix} \mathbb{E} \left(\sum_{j=1} b_{1j} z_j \right) \\ \vdots \\ \mathbb{E} \left(\sum_{j=1} b_{mj} z_j \right) \end{bmatrix} = \begin{bmatrix} \sum_{j=1} \mathbb{E}(b_{1j} z_j) \\ \vdots \\ \sum_{j=1} \mathbb{E}(b_{mj} z_j) \end{bmatrix} \\ &= \begin{bmatrix} \sum_{j=1} \mathbb{E}(b_{1j}) \mathbb{E}(z_j) \\ \vdots \\ \sum_{j=1} \mathbb{E}(b_{mj}) \mathbb{E}(z_j) \end{bmatrix} \quad \text{by indep. of } B, Z \\ &= \mathbb{E}(B)\mathbb{E}(Z) \end{aligned}$$

Now suppose random matrix $C = [C_1 \dots C_k] \in \mathbb{P}^{n \times k}$, where each $C_j \in \mathbb{P}^n$ is the j th column of C . Then

$$\begin{aligned} \mathbb{E}(BC) &= \mathbb{E}(B[C_1 \dots C_k]) = \mathbb{E}([BC_1 \dots BC_k]) = [\mathbb{E}(BC_1) \dots \mathbb{E}(BC_k)] \\ &= [\mathbb{E}(B)\mathbb{E}(C_1) \dots \mathbb{E}(B)\mathbb{E}(C_k)] \quad \text{by indep. and } (\star) \\ &= \mathbb{E}(B)\mathbb{E}(C) \end{aligned}$$

This also proves (\triangle) . □

We also claim that for any RV $X \sim MVN(\mu, \Sigma)$,

$$\mathbb{E}(AX + b) = A\mathbb{E}(X) + b = A\mu + b \quad \text{Var}(AX + b) = A \text{Var}(X)A^T = A\Sigma A^T$$

Proof. By (\triangle) , we directly have $\mathbb{E}(AX + b) = A\mu + b$. Next, let $Y = AX + b$, where $\mu_Y = A\mu + b$. Thus

$$\begin{aligned} \text{Var}(Y) &= \mathbb{E}((Y - \mu_Y)(Y - \mu_Y)^T) \\ &= \mathbb{E}(((AX + b) - (A\mu + b))((AX + b) - (A\mu + b))^T) \\ &= \mathbb{E}((AX - A\mu)(AX - A\mu)^T) \\ &= \mathbb{E}(A(X - \mu)(X - \mu)^T A^T) \\ &= A\mathbb{E}((X - \mu)(X - \mu)^T)A^T \\ &= A \text{Var}(X)A^T \\ &= A\Sigma A^T \end{aligned}$$

□

^{5.6}The marginals in A or B are not required to be independent.

Although we only showed the mean and variance of $AX + b$, it is true that

Proposition 5.3.2 (Stochastic Representation of MVN RVs)

For $X \sim MVN_n(\mu, \Sigma)$, and constant invertible matrix $A \in \mathbb{R}^{n \times n}$, vector $b \in \mathbb{R}^n$,

$$AX + b \sim MVN_n(A\mu + b, A\Sigma A^T)$$

The proof will be presented later with MGF. Therefore, we can build $MVN(\mu, \Sigma)$ RVs with standard MVN RVs,

Corollary 5.3.3 (Building MVN using Standard MVN)

For $X \sim MVN(\mu, \Sigma)$, we can represent it as $AZ + \mu$ where A denotes the Cholesky Factorization of Σ and $Z \sim MVN(\mathbf{0}, I)$ is the standard MVN RV, such that

$$AZ + \mu \sim X \sim MVN(\mu, \Sigma)$$

Recall that for univariate normal RVs $X_i \stackrel{ind.}{\sim} N(\mu_i, \sigma_i^2), i = 1, \dots, n$,

$$\sum_{i=1}^n \left(\underbrace{\frac{X_i - \mu_i}{\sigma_i}}_{\text{deshifting}} \right)^2 \sim \chi^2(n)$$

It is similar to deshift and descale MVN normals such that

$$X = AZ + \mu \implies Z = \underbrace{A^{-1}}_{\text{descaling}} \left(\underbrace{X - \mu}_{\text{deshifting}} \right)$$

thus

$$(A^{-1}(X - \mu))^T (A^{-1}(X - \mu)) = Z^T Z \sim \chi^2(n)$$

Proposition 5.3.4

For $X \sim MVN_n(\mu, \Sigma)$,

$$(X - \mu)^T \Sigma^{-1} (X - \mu) \sim \chi^2(n)$$

Proof. Let A denote the Cholesky factor of Σ , thus $\Sigma^{-1} = A^{-T} A^{-1}$. Decompose $X = AZ + \mu$ for some $Z \sim MVN_n(\mathbf{0}, I)$, then $Z = A^{-1}(X - \mu)$. Therefore,

$$\begin{aligned} (X - \mu)^T \Sigma^{-1} (X - \mu) &= (X - \mu)^T (A A^T)^{-1} (X - \mu) \\ &= (X - \mu)^T (A^{-T} A^{-1}) (X - \mu) \\ &= (A^{-1}(X - \mu))^T (A^{-1}(X - \mu)) \\ &= Z^T Z \sim \chi^2(n) \quad \text{by (MVN Inner)} \end{aligned}$$

□

This is a quadratic form, where

$$(X - \mu)^T \Sigma^{-1} (X - \mu) = \sum_{i=1}^n \sum_{j=1}^n \Sigma_{ij}^{-1} (X_i - \mu_i)(X_j - \mu_j) = \sum_{i=1}^n \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

5.3.3 Independence of Marginals

From [Independence Implies No Covariance](#), we know that if the marginals are independent then

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_n^2 \end{pmatrix}$$

such that Σ is a diagonal matrix, namely $\text{diag}(\sigma_1^2, \dots, \sigma_n^2)$, since the covariance between any two distinct marginals is 0. However, the converse is true for multivariate normal distributions, that said, if two normal RVs have zero covariance, then they are independent. Let $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ be a diagonal matrix, then

$$|\Sigma|^{1/2} = \prod_{i=1}^n \sigma_i$$

Thus the joint PDF is

$$f_X(x) = \frac{1}{(2\pi)^{n/2} \prod_{i=1}^n \sigma_i} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right), x \in \mathbb{R}^n$$

which is the product of n marginal normal RVs' PDFs such that

$$f_X(x) = \prod_{i=1}^n \underbrace{\frac{1}{(2\pi)^{1/2} \sigma_i} \exp\left(-\frac{1}{2} \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right)}_{f_{X_i}(x_i), i=1, \dots, n}$$

By [Independence of RVs](#), they are independent. Therefore,

Corollary 5.3.5 (Independent Normal RVs)

The marginal normal RVs of $X \sim MVN_n(\mu, \Sigma)$ are independent if and only if Σ is a diagonal matrix, i.e. the covariance between two distinct marginals is 0.

This is a special property of normal distributions, but again notice that not all distributions have this property. Zero covariance does not guarantee independence, except for normal distributions.

5.3.4 MGF

Recall that multivariate MGF with k variables is defined as

$$\mathbb{E}(e^{\mathbf{t}^T \mathbf{X}})$$

where $\mathbf{t} = (t_1, \dots, t_k)^T \in \mathbb{R}^k$, $\mathbf{X} = (X_1, \dots, X_k)$ is a random vector. Here,

$$\mathcal{M}_X(t) = \int_{\mathbb{R}^n} \exp(t^T x) \cdot \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right) dx$$

We first expand

$$-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) = -\frac{1}{2}(x^T \Sigma^{-1} x - 2x^T \Sigma^{-1} \mu + \mu^T \Sigma^{-1} \mu) = -\frac{1}{2}(x^T \Sigma^{-1} x - 2x^T \Sigma^{-1} \mu) - \frac{1}{2}\mu^T \Sigma^{-1} \mu$$

Thus we get

$$\frac{\exp\left(-\frac{1}{2}\mu^T \Sigma^{-1} \mu\right)}{(2\pi)^{n/2} |\Sigma|^{1/2}} \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}(x^T \Sigma^{-1} x - 2x^T \Sigma^{-1} \mu) + t^T x\right) dx$$

Similar to the univariate derivation, we construct a quadratic function for the exponent

$$\begin{aligned}
& -\frac{1}{2}(x^T \Sigma^{-1} x - 2x^T \Sigma^{-1} \mu) + t^T x = -\frac{1}{2}(x^T \Sigma^{-1} x - 2(x^T \Sigma^{-1} \mu + x^T t)) = -\frac{1}{2}(x^T A^{-T} A^T x - 2(x^T A^{-T} A^{-1} \mu + x^T A^{-T} A^T t)) \\
& = -\frac{1}{2} \underbrace{(x^T A^{-T} A^T x - 2(A^{-1} x)^T (A^{-1} \mu + A^T t) + (A^{-1} \mu + A^T t)^T (A^{-1} \mu + A^T t))}_{\text{quadratic function}} - (A^{-1} \mu + A^T t)^T (A^{-1} \mu + A^T t) \\
& = -\frac{1}{2}((A^{-1} x - (A^{-1} \mu + A^T t))^T (A^{-1} x - (A^{-1} \mu + A^T t)) - (A^{-1} \mu + A^T t)^T (A^{-1} \mu + A^T t)) \\
& = -\frac{1}{2} \underbrace{(A^{-1}(x - \mu) - A^T t)^T (A^{-1}(x - \mu) - A^T t)}_{\text{apply change of variable}} + \underbrace{\frac{1}{2}(\mu^T \Sigma^{-1} \mu + 2\mu^T t + t^T \Sigma t)}_{\text{pull out of integral}}
\end{aligned}$$

Set $z = A^{-1}(x - \mu)$ with $dz = A^{-1}dx$. The integral thus becomes

$$\frac{\exp(-\frac{1}{2}\mu^T \Sigma^{-1} \mu) \exp(\frac{1}{2}\mu^T \Sigma^{-1} \mu + \mu^T t + \frac{1}{2}t^T \Sigma t)}{(2\pi)^{n/2} |\Sigma|^{1/2}} \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}(z - A^T t)^T (z - A^T t)\right) |A|^{\textcolor{red}{5.7}} dz$$

Note that

$$|\Sigma| = |AA^T| = |A|^2 \implies |\Sigma|^{1/2} = |A|$$

and by multivariate Gaussian integral

$$\int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}(z - A^T t)^T (z - A^T t)\right) dz = (2\pi)^{n/2} \tag{*}$$

the integral is equal to

$$\mathcal{M}_X(t) = \exp(\mu^T t + \frac{1}{2}t^T \Sigma t) \tag{MVN MGF}$$

The 1-deg moment

$$\mathbb{E}(X) = \nabla_t \Big|_{t=0} \mathcal{M}_X(t) = (\mu + \Sigma t) \exp(\mu^T t + \frac{1}{2}t^T \Sigma t) \Big|_{t=0} = \mu$$

Note that

$$\mathbb{E}(X)_i = \mu_i, i = 1, \dots, n$$

The 2-deg moment

$$\nabla_t^{(2)} \Big|_{t=0} \mathcal{M}_X(t) = \Sigma \exp(\mu^T t + \frac{1}{2}t^T \Sigma t) + (\mu + \Sigma t)(\mu + \Sigma t)^T \exp(\mu^T t + \frac{1}{2}t^T \Sigma t) \Big|_{t=0} = \Sigma + \mu\mu^T$$

Let $H^{\textcolor{red}{5.8}} = \nabla_t^{(2)} \Big|_{t=0} \mathcal{M}_X(t)$, then

$$H_{ij} = \frac{\partial^2 \mathcal{M}_X(t)}{\partial t_i \partial t_j} \Big|_{t=0} = \mathbb{E}(X_i X_j), i, j = 1, \dots, n$$

Note that

$$\mu\mu^T = \mathbb{E}(X)\mathbb{E}(X)^T = [\mathbb{E}(X)_i \mathbb{E}(X)_j]_{ij}^{n \times n}$$

thus

$$\text{Var}(X) = H - \mu\mu^T = \Sigma$$

For example, the mutivariate MGF w.r.t. $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$ with $\text{Cov}(X, Y) = c$ is

$$\mathcal{M}_{X,Y}(t_1, t_2) = e^{(\mu_1 t_1 + \mu_2 t_2) + \frac{1}{2}(\sigma_1^2 t_1^2 + \sigma_2^2 t_2^2 + 2ct_1 t_2)}$$

^{5.7}Jacobian

^{5.8}a.k.a. *Hessian matrix*

The 1-deg moment

$$\frac{\partial \mathcal{M}_{X,Y}(t_1, t_2)}{\partial t_1} \Big|_{t_1=t_2=0} \stackrel{5.9}{=} \mu_1 \quad \frac{\partial \mathcal{M}_{X,Y}(t_1, t_2)}{\partial t_2} \Big|_{t_1=t_2=0} = \mu_2$$

The 2-deg moment $\frac{\partial^2 \mathcal{M}_{X,Y}(t_1, t_2)}{\partial t_1 \partial t_2} \Big|_{t_1=t_2=0}$ is equal to

$$c\mathcal{M}_{X,Y}(t_1, t_2) + (\mu_1 + \sigma_1^2 t_1 + ct_2)(\mu_2 + \sigma_2^2 t_2 + ct_1)\mathcal{M}_{X,Y}(t_1, t_2)$$

Messy but if we substitute in $t_1 = t_2 = 0$ we obtain

$$\frac{\partial^2 \mathcal{M}_{X,Y}(t_1, t_2)}{\partial t_1 \partial t_2} \Big|_{t_1=t_2=0} = c + \mu_1 \mu_2$$

Recall from [Properties of Multivariate MGF](#), this is $\mathbb{E}(XY)$. That explains why

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

works.

With MGF we can determine the transformation of distributions. We now prove [Stochastic Representation of MVN RVs](#). For invertible $M \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n, X \sim MVN_n(\mu, \Sigma)$, let

$$Y = MX + b$$

Its MGF is

$$\mathcal{M}_Y(t) = \mathbb{E}(e^{t^T(MX+b)}) = \mathbb{E}(e^{t^T MX + t^b}) = e^{t^T b} \mathbb{E}(e^{t^T MX})$$

by linearity of expectation. Note that by [\(MVN MGF\)](#)

$$\mathbb{E}(e^{t^T MX}) = \mathbb{E}(e^{(M^T t)^T X}) = e^{t^T M\mu + \frac{1}{2}t^T M\Sigma M^T t}$$

hence

$$\mathcal{M}_Y(t) = e^{t^T (M\mu + b) + \frac{1}{2}t^T (M\Sigma M^T)t}$$

which is the MGF of

$$Y \sim MVN_n(M\mu + b, M\Sigma M^T)$$

This completes the proof.

5.3.5 Covariance

Recall for n dimensional random vector X with mean μ_X , its variance is

$$\text{Var}(X) = \mathbb{E}((X - \mu_X)(X - \mu_X)^T)$$

Similarly, for random vectors $X, Y \in \mathbb{P}^n$ with means μ_X, μ_Y , their covariance is

$$\text{Cov}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)^T) \quad (\text{Covariance})$$

Particularly,

$$\text{Cov}(X, Y) = \mathbb{E}(XY^T) - \mu_X \mu_Y^T$$

^{5.9}can also say $\frac{d\mathcal{M}_{X,Y}(t_1, 0)}{dt_1} \Big|_{t_1=0}$

Definition 5.3.6 (MVN Covariance)

For n dimensional MVN RVs X, Y of the same dimension with means μ_X, μ_Y , their covariance is

$$\text{Cov}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)^T) = \mathbb{E}(XY^T) - \mu_X\mu_Y^T$$

Let $\sigma_{X,i} = \sqrt{\text{Var}(X_i)}$, $\sigma_{Y,i} = \sqrt{\text{Var}(Y_i)}$ for $i = 1, \dots, n$ and $\rho_{i,j} = \text{Corr}(X_i, Y_j)$. Then the covariance matrix $\text{Cov}(X, Y)$ is

$$\begin{pmatrix} \rho_{1,1}\sigma_{X,1}\sigma_{Y,1} & \rho_{1,2}\sigma_{X,1}\sigma_{Y,2} & \dots & \rho_{1,n}\sigma_{X,1}\sigma_{Y,n} \\ \rho_{2,1}\sigma_{X,2}\sigma_{Y,1} & \rho_{2,2}\sigma_{X,2}\sigma_{Y,2} & \dots & \rho_{2,n}\sigma_{X,2}\sigma_{Y,2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n,1}\sigma_{X,n}\sigma_{Y,1} & \rho_{n,2}\sigma_{X,n}\sigma_{Y,2} & \dots & \rho_{n,n}\sigma_{X,n}\sigma_{Y,n} \end{pmatrix}$$

5.3.6 Conditional MVN

Consider a simple case $(X_1, X_2) \sim MVN_2(\mu, \Sigma)$. Their joint PDF is thus

$$f(x_1, x_2) = \frac{1}{2\pi |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} \left[\left(\frac{x_1 - \mu_1}{\Sigma_{11}}\right)^2 + \left(\frac{x_2 - \mu_2}{\Sigma_{22}}\right)^2 \right]\right), x_1, x_2 \in \mathbb{R}$$

The marginal PDF is

$$\begin{aligned} f(x_1) &= \int_{-\infty}^{\infty} \frac{1}{2\pi |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} \left[\left(\frac{x_1 - \mu_1}{\Sigma_{11}}\right)^2 + \left(\frac{x_2 - \mu_2}{\Sigma_{22}}\right)^2 \right]\right) dx_2 \\ &= \frac{1}{2\pi |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} \left(\frac{x_1 - \mu_1}{\Sigma_{11}}\right)^2\right) \underbrace{\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2} \left(\frac{x_2 - \mu_2}{\Sigma_{22}}\right)^2\right) dx_2}_{(2\pi)^{1/2}\Sigma_{22}^{1/2} \text{ by Gaussian Integral}} \end{aligned}$$

Note that

$$|\Sigma| = \Sigma_{11}\Sigma_{22} - 2\Sigma_{12}$$

so

$$|\Sigma|/\Sigma_{22} = \Sigma_{11} - 2(\Sigma_{12}/\Sigma_{22})$$

and we get

$$f(x_1) = \frac{1}{(2\pi)^{1/2}(\Sigma_{11} - 2(\Sigma_{12}/\Sigma_{22}))^{1/2}} \exp\left(-\frac{1}{2} \left(\frac{x_1 - \mu_1}{\Sigma_{11}}\right)^2\right), x_1 \in \mathbb{R}$$

and similarly

$$f(x_2) = \frac{1}{(2\pi)^{1/2}(\Sigma_{22} - 2(\Sigma_{12}/\Sigma_{11}))^{1/2}} \exp\left(-\frac{1}{2} \left(\frac{x_2 - \mu_2}{\Sigma_{22}}\right)^2\right), x_2 \in \mathbb{R}$$

5.4 Complex Distributions

A complex distribution has complex sample space such that $\Omega \subseteq \mathbb{C}$. A complex number $z \in \mathbb{C}$ could be defined as

$$z = x + iy, x, y \in \mathbb{R} \quad (\text{Complex Number})$$

A complex random variable could be written as $Z = X + iY$, where X, Y are real random variables (possibly related).

Example 5.4.1

Let Z be a complex RV such that $\Re(Z) = \Im(Z) \sim N(0, 1)$. Find $P(|Z| > 3)$.

Solution

Let

$$Z = X + iX, \quad X \sim N(0, 1)$$

such that

$$|Z| = |X(1 + i)| = |X| |1 + i| = \sqrt{2} |X|$$

Thus

$$P(|Z| > 3) = P(\sqrt{2} |X| > 3) = P(|X| > 3/\sqrt{2}) = 2(1 - \Phi(3/\sqrt{2})) \approx 2(1 - 0.983) \approx 0.034$$

Example 5.4.2

A complex random variable Z is uniformly distributed in the annulus

$$O = \{z \in \mathbb{C} : a < |z| < b\}, 0 < a < b \in \mathbb{R}$$

on the complex plane. Find the CDF of Z .

Solution

Let the joint distribution (X, Y) have uniform joint probability

$$f(x, y) = c \mathbf{1}_{\{a < |z| < b\}}$$

To find c , we first calculate the area of the annulus. The outer circle with radius b has area $b^2\pi$, and the inner circle with radius a has area $a^2\pi$. Hence, the area of the annulus is $b^2\pi - a^2\pi = (b^2 - a^2)\pi$. Thus the joint probability is

$$f(x, y) = \frac{1}{(b^2 - a^2)\pi} \mathbf{1}_{\{a < |z| < b\}}$$

For simplicity we still use c to denote the uniform density. Note that the region of support is

$$a < |z| < b \implies a < \sqrt{x^2 + y^2} < b$$

Parametrize x, y using r, θ , which are the length and angle respectively,

$$x = r \cos(\theta) \quad y = r \sin(\theta)$$

The region of support is now mapped to

$$a < r < b, 0 < \theta < 2\pi$$

The Jacobian is

$$\begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos(\theta) & -r \sin(\theta) \\ \sin(\theta) & r \cos(\theta) \end{vmatrix} = r(\cos^2(\theta) + \sin^2(\theta)) = r$$

So now the joint density function is

$$f(r, \theta) = \frac{r}{(b^2 - a^2)\pi} \mathbf{1}_{\{r \in [a, b], \theta \in [0, 2\pi]\}}$$

The joint CDF is thus

$$\begin{aligned} F(r, \theta) &= \int_{-\infty}^{\theta} \int_{-\infty}^r \frac{t}{(b^2 - a^2)\pi} dt dk = \frac{1}{(b^2 - a^2)\pi} \int_0^{\theta} \int_a^r t dt dk \\ &= \frac{1}{(b^2 - a^2)\pi} \int_0^{\theta} \frac{t^2}{2} \Big|_a^r dk = \frac{1}{2(b^2 - a^2)\pi} \int_0^{\theta} r^2 - a^2 dk \\ &= \frac{\theta(r^2 - a^2)}{2(b^2 - a^2)\pi} \mathbf{1}_{\{r \in [a, b], \theta \in [0, 2\pi]\}} \end{aligned}$$

Note that it is valid since it is non-decreasing and

$$\lim_{\theta \rightarrow \infty} \lim_{r \rightarrow \infty} F(r, \theta) = F(b, 2\pi) = 1$$

6 Limiting (Asymptotic) Distributions

6.1 Convergence in Probability

6.2 Convergence in Distribution

6.3 Central Limit Theorem (CLT)

In this section, we aim to estimate the mean of some feature from the target population. For example, if we want to estimate the mean height of people in a country, we can select a small set of people as the sample and measure the mean height, which can be an estimate of the target population's mean height. We focus on how to analyze these observed information for a better estimate.

Assume that the country's people have mean height μ and variance σ^2 (both **constant** and **unknown**). To estimate them, we select a sample from the country. Let X_1, \dots, X_n be n iid RVs following some distribution, and each X_i for $i = 1, \dots, n$ denotes an individual in the sample, where in this example, it is the height of individual i . The

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (\text{Sample Mean})$$

is the sample mean. Note that it is a **random variable** whose randomness depends on X_1, \dots, X_n . It is an unbiased estimator^{6.1} of the population mean such that

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n}(n\mathbb{E}(X)) = \mu$$

The variance of \bar{X} is

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \underbrace{(n \text{Var}(X))}_{\text{by indep.}} = \frac{\sigma^2}{n}$$

Theorem 6.3.1 (Central Limit Theorem)

Let X_1, \dots, X_n be iid RVs with mean μ and variance σ^2 . Let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Then

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0, 1)$$

Another way to represent it is

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow{D} N(0, 1)$$

It states that a larger sample size will yield a better estimate. Ideally, if sample size is infinitely large we can get an exactly accurate estimate, but it is impossible in the real world.

Theorem 6.3.2 (ANOVA)

Let $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$. Then

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (\bar{X} - \mu)^2$$

^{6.1}Note: A "parameter" to be estimated is constant, whereas an "estimator" of the parameter is a random variable.

Proof.

$$\begin{aligned}
\sum_{i=1}^n (X_i - \mu)^2 &= \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu)^2 \\
&= \sum_{i=1}^n (X_i - \bar{X})^2 + (\bar{X} - \mu)^2 + 2(X_i - \bar{X})(\bar{X} - \mu) \\
&= \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (\bar{X} - \mu)^2 + \sum_{i=1}^n 2(X_i - \bar{X})(\bar{X} - \mu) \\
&= \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (\bar{X} - \mu)^2 + \cancel{2(\bar{X} - \mu)} \sum_{i=1}^n (X_i - \bar{X}) \\
&= \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (\bar{X} - \mu)^2
\end{aligned}$$

□

Remark

Note that

$$\underbrace{\frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2}}_{\sim \chi^2(n)} = \underbrace{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}}_{\sim \chi^2(n-1)} + \underbrace{\frac{\sum_{i=1}^n (\bar{X} - \mu)^2}{\sigma^2}}_{\sim \chi^2(1)} \quad (\star)$$

Proof. We indicate (\star) by

$$A = B + C$$

For A ,

$$\begin{aligned}
\underbrace{\frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2}}_{\sim \chi^2(n)} &= \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^n Z_i^2, \quad Z_1, \dots, Z_n \stackrel{i.i.d.}{\sim} N(0, 1) \\
&\sim \chi^2(n) \quad \text{by property}
\end{aligned}$$

For C ,

$$\begin{aligned}
\underbrace{\frac{\sum_{i=1}^n (\bar{X} - \mu)^2}{\sigma^2}}_{\sim \chi^2(1)} &= \sum_{i=1}^n \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 = n \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 \\
&= \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2 = Z^2, \quad Z \sim N(0, 1) \\
&\sim \chi^2(1) \quad \text{by property}
\end{aligned}$$

The middle one B is a little bit complicated. Take $Z_i = \frac{X_i - \mu}{\sigma} \sim N(0, 1)$, $i = 1, \dots, n$, and $\bar{Z} \sim N(0, 1/n)$. Then B becomes

$$\frac{\sigma^2 \sum_{i=1}^n (Z_i - \bar{Z})^2}{\sigma^2} = \sum_{i=1}^n (Z_i - \bar{Z})^2 = \sum_{i=1}^n Z_i^2 - n\bar{Z}^2$$

Note that $\sum_{i=1}^n Z_i^2 \sim \chi^2(n)$, $n\bar{Z}^2 = (\sqrt{n}\bar{Z})^2 \sim \chi^2(1)$.

□

Proposition 6.3.3 (Estimator of Sample Variance)

Let $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} F_X$, then

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

is an unbiased estimator of $\text{Var}(X) = \sigma^2$.

Proof. Note that

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X})^2 &= \sum_{i=1}^n X_i^2 - 2X_i\bar{X} + \bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - \sum_{i=1}^n 2X_i\bar{X} + \sum_{i=1}^n \bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - 2n\bar{X}^2 + \sum_{i=1}^n \bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - n\bar{X}^2 \\ \mathbb{E} \left(\sum_{i=1}^n (X_i - \bar{X})^2 \right) &= \sum_{i=1}^n \mathbb{E}(X_i^2) - n\mathbb{E}(\bar{X}^2) \text{ by linearity} \\ &= \sum_{i=1}^n (\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= n\sigma^2 + n\mu^2 - (\sigma^2 + n\mu^2) \\ &= (n - 1)\sigma^2 \end{aligned}$$

Thus

$$\mathbb{E} \left(\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} \right) = \sigma^2$$

□

From the derivation, we see that one degree of freedom is taken away by \bar{X} . A beauty of this estimator is that it can be estimated based on known information, i.e. the X_i 's and \bar{X} . We denote the convention

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} \quad (\text{Sample Var Est.})$$

to be the estimated sample variance, and notice that by **ANOVA**,

$$\frac{(n - 1)s^2}{\sigma^2} \sim \chi^2(n - 1) \quad (\text{Sample Var Distri.})$$

whose randomness comes from s^2 . Note that s is a biased estimate of σ .

$$\mathbb{E}(s^2) = \text{Var}(s) + \mathbb{E}(s)^2 \implies \mathbb{E}(s) = \sqrt{\text{Var}(s) - \mathbb{E}(s^2)}$$

Also note that $s = \sqrt{\frac{\sigma^2 W}{n-1}} = \frac{\sigma}{\sqrt{n-1}} \sqrt{W}$ where $W = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$, $\sqrt{W} \sim \chi(n-1)$.

For convenience, if we stack the samples in a vector such that

$$\mathbf{X} = (X_1, \dots, X_n)^T$$

with sample mean \bar{X} , the sample variance could be vectorized as

$$s^2 = \frac{(\mathbf{X} - \bar{\mathbf{X}})^T (\mathbf{X} - \bar{\mathbf{X}})}{n - 1} \quad (\text{Vec. Sample Var})$$

Proposition 6.3.4 (Estimating Mean with t)

For sample variance estimator s^2 ,

$$\frac{\bar{X} - \mu}{s} \sim t(n - 1)$$

Proof.

$$\begin{aligned}\frac{\bar{X} - \mu}{s} &= \frac{\frac{\bar{X} - \mu}{\sigma}}{\frac{s}{\sigma}} = \frac{Z}{\sqrt{\frac{(n-1)s^2/\sigma^2}{n-1}}} \\ &= \frac{Z}{\sqrt{\frac{W}{n-1}}}, W \sim \chi^2(n-1) \quad \text{by (Sample Var Distri.)} \\ &\sim t(n-1) \quad \text{by property}\end{aligned}$$

□

7 Maximum Likelihood Estimation

In this section we aim to estimate some unknown parameters with independently sampled observations.

7.1 Introduction

Recall the coin example 3.6.1, but we simplify the condition. Let $X \sim \text{Bern}(p)$ denote the result of tossing a coin with unknown probability p of getting a head. We flip the coin 10 times independently^{7.1} and the results are

$$1, 0, 1, 1, 0, 0, 1, 1, 1, 0$$

We want the theoretical probability of getting these results, a.k.a. the *Likelihood Function*. Since there are 6 heads and 4 tails and the trials are independent, the likelihood is

$$L(x; p) = p^6(1-p)^4$$

Again, notice that p is unknown. The likelihood $L(x; p)$ is a function on p , and since these results have happened, we assume that they have the largest probability to appear. That said, p is likely to be the number when $L(x; p)$ is maximized, namely $p = \operatorname{argmax}_p L(x; p)$. Now we aim to maximize $L(x; p)$. A common way is to find the first derivative and set to 0, and verify the optimality using second derivative test. However, most of the times it is difficult to differentiate $L(x; p)$. Note that \log is a one-to-one function and

$$a < b \implies \log(a) < \log(b)$$

That said, the position of $L(x; p)$'s optima will not change after applying \log to it, namely

$$\operatorname{argmax}_p L(x; p) = \operatorname{argmax}_p \log L(x; p)$$

but it is easier to find the derivative. Denote

$$l(x; p) = \log L(x; p)$$

and now we maximize $l(x; p)$. Recall that the maximizer $\hat{\theta}$ of $l(x; p)$ is the same as $L(x; p)$'s. In this example,

$$l(x; p) = \log(p^6(1-p)^4) = 6\log(p) + 4\log(1-p)$$

Differentiate,

$$\frac{dL}{dp} = \frac{6}{p} - \frac{4}{1-p}$$

Set to zero,

$$\frac{6}{p} - \frac{4}{1-p} = 0 \implies \hat{p} = \frac{3}{5}$$

Use second derivative to verify the maximality,

$$\frac{d^2l}{dp^2} = -\frac{6}{p^2} - \frac{4}{(1-p)^2} < 0, \forall p \in [0, 1]$$

We find that $l(x; p)$ is concave, thus the stationary point $\hat{p} = \frac{3}{5}$ is indeed the maxima. Based on these data, the probability of tossing a head is $\frac{3}{5}$, and we find out that the best estimate is number of heads divided by total number of trials.

^{7.1}All observations must be independently sampled.

7.2 Maximum Likelihood Estimate

Definition 7.2.1 (Likelihood Function and Log Likelihood Function)

Let X have probability function f depending on unknown parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^T$ as our estimate of interest. Suppose x_1, \dots, x_n are n independently sampled realizations of X , the likelihood function is

$$L(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^n f(x^{(i)}; \boldsymbol{\theta}) \quad (\text{Likelihood Function})$$

The log likelihood function is

$$l(\mathbf{x}; \boldsymbol{\theta}) = \log L(\mathbf{x}; \boldsymbol{\theta}) = \sum_{i=1}^n \log(f(x^{(i)}; \boldsymbol{\theta})) \quad (\text{Log Likeli. Fun.})$$

Lemma 7.2.2 (Invariance Property)

Let $h(\cdot)$ be a one-to-one and monotonic function. Let $g(\theta)$ be a function on θ . Then

$$\operatorname{argmax}_{\theta} g(\theta) = \operatorname{argmax}_{\theta} h(g(\theta))$$

This lemma suggests that a positive linear transformation applied to the target function will not affect the maximizer. Now we proceed with *Maximum Likelihood Estimate*,

Algorithm 7.2.3 (Maximum Likelihood Estimate (MLE))

The Score Function is defined as the first derivative of l ,

$$S(\mathbf{X}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} l(\mathbf{X}; \boldsymbol{\theta}) \in \mathbb{R}^k \quad (\text{Score})$$

The Information Function is defined as the second derivative of l ,

$$I(\mathbf{X}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}^{(2)} l(\mathbf{X}; \boldsymbol{\theta}) \in \mathbb{R}^{k \times k} \quad (\text{Information})$$

The Expected Information is

$$J(\boldsymbol{\theta}) = \mathbb{E}(I(\mathbf{X}; \boldsymbol{\theta})) \quad (\text{Expected Information})$$

Assume that L and l are sufficiently smooth. The maximum likelihood estimate of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}} = \text{MLE}(\boldsymbol{\theta})$, satisfies

$$S(\mathbf{X}; \hat{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{\theta}} l(\mathbf{X}; \hat{\boldsymbol{\theta}}) = \mathbf{0} \quad (\text{Stationary})$$

$$I(\mathbf{X}; \hat{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{\theta}}^{(2)} l(\mathbf{X}; \hat{\boldsymbol{\theta}}) \prec \mathbf{0} \quad (\text{Negative Definite})$$

Refer to [Optimization](#), there might exist multiple stationary points (such as multimodal likelihood). Always compare the stationary points and find the global maximizer. Furthermore, check boundary points if necessary. The following corollary is a direct result of [Invariance Property](#),

Corollary 7.2.4 (Invariance Property of MLE)

Let $h(\cdot)$ be a one-to-one function. Then

$$\text{MLE}(h(\boldsymbol{\theta})) = h(\text{MLE}(\boldsymbol{\theta}))$$

We look at a more general MLE example.

Example 7.2.1 (MLE for Binomial Distribution)

Suppose in a $\text{Bin}(n, p)$ experiment, there are k successes (1's). Find the MLE of Binomial distribution's success probability p .

Solution

Recall

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, \dots, n$$

The likelihood function is thus

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

Note that by **Invariance Property**, we can remove all constants and only keep the terms containing p . The log likelihood function is

$$l(k; p) = \log L(x, p) = k \log(p) + (n - k) \log(1 - p)$$

Find score and set to 0,

$$\frac{dl}{dp} = \frac{k}{p} - \frac{n - k}{1 - p} = 0$$

Solve,

$$k(1 - p) = (n - k)p \implies \hat{p} = \frac{k}{n}$$

Verify the maximality. The second derivative is

$$\frac{d^2 l}{dp^2} = -\frac{k}{p^2} - \frac{n - k}{(1 - p)^2}$$

Substitute \hat{p} in,

$$\begin{aligned} -\frac{k}{\left(\frac{k}{n}\right)^2} - \frac{n - k}{\left(\frac{n-k}{n}\right)^2} &= -\frac{n^2}{k} - \frac{n^2}{n - k} \\ &= -n^2 \left(\frac{(n - k) + k}{k(n - k)} \right) \\ &= -\frac{n^3}{k(n - k)} < 0 \end{aligned}$$

That said, $l(x; p)$ is concave at \hat{p} . Therefore $\hat{p} = \frac{k}{n}$ is indeed the MLE of p .

In the binomial example above, note that the MLE of probability of getting 0 is

$$\text{MLE}(1 - p) = \frac{n - k}{n}$$

which is the number of 0's divided by the total number of trials, similar to $\text{MLE}(p)$. In fact, we can extend this result to a more general case – the multinomial distribution,

Exercise 7.2.1 (MLE for Multinomial Distribution)

Let $(X_1, \dots, X_k)^T \sim \text{Mult}(n; p_1, \dots, p_k)$ for $X_1 + \dots + X_k = n$ and $p_1 + \dots + p_k = 1$. Show that

$$\text{MLE}(p_j) = \frac{x_j}{n}$$

where x_j denotes the number of occurrences of type j , for each $j = 1, \dots, k$.

Example 7.2.2 (MLE for Normal Distribution)

Let $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$. Find MLE(μ) and MLE(σ).

Solution

The likelihood function is

$$L(x; \mu, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}(x_i - \mu)^2/\sigma^2\right)$$

Simplify,

$$L(x; \mu, \sigma) = \cancel{\frac{1}{(2\pi)^{n/2}}} \frac{1}{\sigma^n} \exp\left(-\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2/\sigma^2\right)$$

The log likelihood is

$$l(x; \mu, \sigma) = -n \log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

Set

$$\begin{aligned} \frac{\partial}{\partial \mu} l &= \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0 \implies \hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \\ \frac{\partial}{\partial \sigma} l &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2 = 0 \implies \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 \end{aligned}$$

One can verify that they are indeed the maximum via second derivative test.

Example 7.2.3 (MLE for Multivariate Normal Distribution)

Let $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} MVN_d(\mu, \Sigma)$. Find MLE(μ) and MLE(σ).

Solution

The likelihood function

$$L(x; \mu, \sigma) = \prod_{i=1}^n \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)\right)$$

Simplify

$$L(x; \mu, \sigma) = \cancel{\frac{1}{(2\pi)^{nd/2}}} \frac{1}{|\Sigma|^{n/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)\right)$$

The log likelihood is

$$\begin{aligned} l(x; \mu, \sigma) &= -\frac{n}{2} \log(|\Sigma|) - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \\ &= -\frac{n}{2} \log(|\Sigma|) - \frac{1}{2} \left(\sum_{i=1}^n x_i^T \Sigma^{-1} x_i - 2 \sum_{i=1}^n \mu^T \Sigma^{-1} x_i + \sum_{i=1}^n \mu^T \Sigma^{-1} \mu \right) \end{aligned}$$

Take the derivative w.r.t. μ

$$\nabla_{\mu} l = \sum_{i=1}^n \Sigma^{-1} x_i - \sum_{i=1}^n \Sigma^{-1} \mu = \Sigma^{-1} \sum_{i=1}^n (x_i - \mu) = 0 \implies \hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

Note that since Σ^{-1} is non-singular, $\text{Null}(\Sigma^{-1}) = \{0\}$ such that the only solution is $\sum_{i=1}^n (x_i - \mu) = 0$. Now for Σ , take $S = \Sigma^{-1}$ for convenience. Note that

$$\nabla_S |S| = |S| S^{-T} = |S| S^{-1}$$

since S is symmetric, and thus $\nabla_S \log(|S^{-1}|) = \nabla_S \log(|S|^{-1}) = \nabla_S - \log(|S|) = -\frac{|S|S^{-1}}{|S|} = -S^{-1}$. Also note that

$$\nabla_S \sum_{i=1}^n (x_i - \mu)^T S (x_i - \mu) = \sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T$$

Therefore

$$\nabla_S l = \frac{n}{2} S^{-1} - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T = 0 \implies \hat{S}^{-1} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

In general,

$$\begin{aligned}\hat{\mu} &= \frac{1}{n} \sum_{i=1}^n x_i \\ \hat{\Sigma} &= \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})^T\end{aligned}$$

Observe that $\text{MLE}(\mu)$ is the mean of the samples and $\text{MLE}(\Sigma)$ is the mean of the samples' covariance matrices.

A Calculus

A.1 Fundamentals

Theorem A.1.1 (Fundamental Theorem of Calculus 1)

For continuous $f(t)$ and $a \in \mathbb{R}$,

$$\frac{d}{dx} F(x) = \frac{d}{dx} \int_a^x f(t) dt = f(x)$$

Theorem A.1.2 (Fundamental Theorem of Calculus 2)

For continuous $f(t)$ and $a \in \mathbb{R}$, let

$$F(x) = \int f(x) dx$$

then

$$\int_a^b f(x) dx = F(b) - F(a)$$

Definition A.1.3 (Convergence)

A series

Definition A.1.4 (Integration By Parts)

$$\int f'(x)g(x) dx = f(x)g(x) - \int f(x)g'(x) dx$$

where the definite integral

$$\int_a^b f'(x)g(x) dx = f(x)g(x)|_a^b - \int_a^b f(x)g'(x) dx$$

Theorem A.1.5 (Cauchy-Schwarz Inequality)

For some function $f(x), g(y)$ on x, y ,

$$\left(\int_a^b f(x)g(y) dx dy \right)^2 \leq \left(\int_a^b f(x)^2 dx \right) \left(\int_a^b g(y)^2 dy \right)$$

When calculating the integral of a product of two functions, observe the function that easier for integral (denoted by $f'(x)$) and the function easier for differentiation (denoted by $g(x)$), and then apply IBP on them.

Theorem A.1.6 (Gaussian Integral)

For all $x \in \mathbb{R}$,

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

Proof. Let

$$I = \int_{-\infty}^{\infty} e^{-x^2} dx$$

By changing x to y ,

$$I = \int_{-\infty}^{\infty} e^{-y^2} dy$$

Extend to a two-variate integral,

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} e^{-x^2} dx \int_{-\infty}^{\infty} e^{-y^2} dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)} dxdy \\ &\text{take } x = r \cos(\theta), y = r \sin(\theta) \text{ with Jacobian } |J| = |r| \\ &\text{therefore } x^2 + y^2 = r^2, dxdy = rdrd\theta \\ &= \int_0^{2\pi} \int_0^{\infty} re^{-r^2} dr d\theta = 2\pi \int_0^{\infty} re^{-r^2} dr \\ &= 2\pi \left(-\frac{1}{2}e^{-r^2} \right) \Big|_0^{\infty} = 2\pi \cdot \frac{1}{2} \\ &= \pi \end{aligned}$$

Therefore, ignoring the negative root since the integral should be positive,

$$I = \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

□

Corollary A.1.7 (Scaled and Shifted Gaussian Integral)

Replace x by $a(x - b)$ in the Gaussian integral,

$$\int_{-\infty}^{\infty} e^{-(a(x-b))^2} dx = \frac{\sqrt{\pi}}{a}$$

Proof. Take $z = a(x - b)$, then $dz = adx$.

$$\int_{-\infty}^{\infty} e^{-(a(x-b))^2} dx = \int_{-\infty}^{\infty} \frac{1}{a} e^{-z^2} dz = \frac{1}{a} \underbrace{\int_{-\infty}^{\infty} e^{-x^2} dx}_{\sqrt{\pi}} = \frac{\sqrt{\pi}}{a}$$

by Gaussian Integral.

□

A.2 Change of Variable

Definition A.2.1 (Jacobian)

For a bijective transformation of variables

$$x = g_1(u, v) \quad y = g_2(u, v)$$

the First-Order Derivative Matrix is defined as

$$J = \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix}$$

and Jacobian is its determinant such that

$$|J| = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$$

The Derivative Matrix can also be defined for higher dimensions, where in general for $\mathbf{x} = (x_1, \dots, x_m)^T \in \mathbb{R}^m$, $\mathbf{u} = (u_1, \dots, u_n)^T \in \mathbb{R}^n$ and $x_1 = g_1(u_1, \dots, u_n), \dots, x_m = g_m(u_1, \dots, u_n)$, the Jacobian matrix is

$$J = \frac{\partial \mathbf{x}}{\partial \mathbf{u}} = \begin{pmatrix} \frac{\partial x_1}{\partial u} \\ \vdots \\ \frac{\partial x_m}{\partial u} \end{pmatrix} = \begin{pmatrix} \frac{\partial x_1}{\partial u_1} & \cdots & \frac{\partial x_1}{\partial u_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_m}{\partial u_1} & \cdots & \frac{\partial x_m}{\partial u_n} \end{pmatrix} = \left[\frac{\partial x_i}{\partial u_j} \right]_{ij}^{m \times n} \quad (m \times n \text{ Jacobian})$$

However, determinant is only defined for square matrices, and Jacobian is meaningless for non-square J . We will see why this case fails later on. Recall that in the univariate integral

$$\int_a^b f(x) dx$$

where $f(x) = y$ is the dynamic function, the method of change of variable with $x = g(u)$ for bijective function $g(\cdot)$ is

$$\begin{aligned} \int_a^b f(x) dx &= \int_a^b f(g(u)) d(g(u)) \\ &= \int_{g^{-1}(a)}^{g^{-1}(b)} f(g(u))g'(u) du \end{aligned} \quad (\text{change of variable})$$

where

$$g'(u) = \frac{dg(u)}{du} = \frac{dx}{du}$$

is the rate of change in x with respect to u , acting as the Jacobian here. The new dynamic function is thus

$$f(g(u))g'(u) \quad \text{for } g^{-1}(a) \leq u \leq g^{-1}(b)$$

Notice that the interval of integral has also been mapped from $a \leq x \leq b$ to $g^{-1}(a) \leq u \leq g^{-1}(b)$. We are indeed integrating on the variable u over its corresponding range, resulting in an equivalent definite integral as $\int_a^b f(x) dx$.

Why does this make sense? Let $y = F(x)$ be the indefinite integral of $f(x)$,

$$y = F(x) = \int f(x) dx$$

such that

$$f(x) = \frac{dy}{dx}$$

is the rate of change in y with respect to x , that is, a change of dx in x would result in a change $dy = f(x)dx$ in y . However, since we applied the change of variable $x = g(u)$, we are instead interested in the rate of change in y with respect to u such that

$$h(u) = \frac{dy}{du} \quad \text{where } y = \int h(u) du$$

where x is totally transformed into some shape on u ^{A.1}. Again, note that it is integrated over u 's domain. In (change of variable), we can see that the new dynamic function $h(u)$ after the change of variable is

$$h(u) = f(g(u))J \quad \text{where } J = \frac{dx}{du}$$

Look at a simple example. Let

$$f(x) = x^2, x = x(u) = 2u$$

The derivative of f w.r.t. x is

$$f'_x(x) = \frac{df}{dx} = 2x \text{ and } df = f'_x(x)dx = (2x)dx$$

and the derivative of x w.r.t. u is

$$x'_u(u) = \frac{dx}{du} = 2 \text{ and } dx = x'_u(u)du = 2du$$

we have

$$f(x) = \int f'_x(x) dx = \int 2x dx$$

But we are interested in the integral with respect to u , i.e.

$$f(u) = \int f'_u(u) du \tag{△}$$

we can just replace x by the identical function on u ,

$$f(u) = \int f'_x(x) dx = \int f'_x(x(u)) x'_u(u)du \tag{★}$$

which is

$$f(u) = \int f'_x(2u) d(2u) = \int f'_x(2u) 2du = \int (2u)^2 2 du = \int \underbrace{8u^2}_{f'_u(u) \text{ in } (\triangle)} du$$

As you can see in (★), instead of directly replacing x by the identical $2u$, there is an extra $x'_u(u)$ multiplied, which is the Jacobian. We thus have

$$f'_u(u) = f'_x(x(u))x'_u(u)$$

or equivalently,

$$\frac{df}{du} = \frac{df}{dx} \frac{dx}{du} = \frac{df}{d(x(u))} \frac{dx}{du}$$

^{A.1} x should vanish after the change of variable operation

The Jacobian J is a scale factor in the **Hypervolume** (or *measure*) upon change of variable, where here in the 1-dimensional case, it is a scale factor of the **length** of the basis, namely from dx to du . Details are presented in [Change of Basis](#). In general we look at the 2-dimensional integral over region Ω_{xy} where

$$\iint_{\Omega_{xy}} f(x, y) \, dxdy, \quad x = g_1(u, v), y = g_2(u, v) \text{ for some functions } g_1, g_2 : \Omega_{uv} \rightarrow \Omega_{xy}$$

and we want to expand the integral to $\iint_{\Omega_{uv}} h(u, v) \, dudv$ with some $h(\cdot)$, just like the 1-D case. Again note that there is a bijective^{A.2} transformation between $f(x, y)$ and $h(u, v)$, i.e. for each pair of (x, y) in Ω_{xy} , there is a **UNIQUE** corresponding (u, v) in the mapped region Ω_{uv} such that $u = c_1(x, y), v = c_2(x, y)$ for some functions $c_1, c_2 : \Omega_{xy} \rightarrow \Omega_{uv}$. The functions themselves might not be linear, however, when slicing them into infinitely small increments, they are approximately linear, note that

$$\underbrace{\begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}}_J \begin{bmatrix} du & 0 \\ 0 & dv \end{bmatrix} = \begin{bmatrix} dx & 0 \\ 0 & dy \end{bmatrix} \quad (\star\star)$$

such that J is the [Change of Basis](#) matrix that maps basis du, dv to dx, dy . Let $dxdy$ be the area^{A.3} of the rectangle formed with dx, dy and $dudv$ be defined in the similar manner, then the Jacobian $|J|$ here acts as a scale factor in the area from $dudv$ to $dxdy$, such that

$$|J| \, dudv = dxdy$$

by taking the **determinant** on both sides in $(\star\star)$. That said

$$\iint_{\Omega_{xy}} f(x, y) \, dxdy = \iint_{\Omega_{uv}} \underbrace{f(g_1(u, v), g_2(u, v)) |J|}_{h(u, v)} \, dudv$$

This implies

$$f(x, y) = f(g_1(u, v), g_2(u, v)) |J| := h(u, v) \quad (*)$$

For general dimensions, let $\mathbf{x}, \mathbf{u}, J$ be defined as in ($m \times n$ [Jacobian](#)) but assuming $m = n$ and $J = \frac{\partial \mathbf{x}}{\partial \mathbf{u}}$ is non-singular. Additionally, let $\mathbf{g} = (g_1, \dots, g_n)$ such that $\mathbf{g}(\mathbf{u}) = (g_1(\mathbf{u}), \dots, g_n(\mathbf{u})) = \mathbf{x}$. Then

$$\int_{\mathbf{x}} f(\mathbf{x}) \, d\mathbf{x} = \int_{\mathbf{u}} f(\mathbf{g}(\mathbf{u})) |J| \, d\mathbf{u}$$

such that the density

$$f(\mathbf{x}) = f(\mathbf{g}(\mathbf{u})) |J| \quad (\diamond)$$

This will be useful in determining the probability density function of a transformed multivariate distribution.

^{A.2}For non-bijective transformations, the original region Ω_{xy} will be mapped to a subspace Ω_{uv} whose dimension is strictly smaller than Ω_{xy} . Thus the Jacobian which is the scale in the measure becomes 0, which is meaningless. For instance, the basis of $\Omega_{xy} = \mathbb{R}^3$'s determinant is the volume of the parallelogram formed with the basic vectors, and if it was mapped to $\Omega_{uv} = \mathbb{R}^2$ which has no volume, the scale in the volume is thus 0. Therefore we can do nothing for the integral.

^{A.3}The domain of a 3-D function, which is a 2-D plane, can be decomposed into (infinitely) many small parallelograms formed with basic vectors from some basis, say $B = \{b_1, b_2\}$ (2 vectors in this 2-D case), each with side lengths $\|b_1\|, \|b_2\|$, and $\det([b_1 \ b_2])$ is the area of each parallelogram. Here, since the basis $\left\{ \begin{bmatrix} du \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ dv \end{bmatrix} \right\}$ are principle axis (orthogonal), the parallelograms are indeed rectangles with side lengths du, dv , whose area is simply $dudy$. $f(x, y)$ measures the density at the rectangle located at (x, y) . Note that J maps $\left\{ \begin{bmatrix} du \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ dv \end{bmatrix} \right\}$ to another principle axis $\left\{ \begin{bmatrix} dx \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ dy \end{bmatrix} \right\}$, whose area is $dxdy$.

A.3 Taylor Series

For a function $f(x) \in C^n$ ^{A.4} that is not easy to be computed or operated, we wish to find a polynomial representation for it, say $P(x)$, such that

$$\begin{aligned} f(x) &= P(x) \\ f'(x) &= P'(x) \\ f''(x) &= P''(x) \\ &\dots \\ f^{(n)}(x) &= P^{(n)}(x) \end{aligned}$$

The Taylor polynomial is exactly what we want.

Definition A.3.1 (Taylor Series)

The Taylor expansion of sufficiently smooth $f : \mathbb{R} \rightarrow \mathbb{R}$ centered at $a \in \mathbb{R}$ is

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x-a)^n = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \frac{1}{6}f'''(a)(x-a)^3 + \dots$$

When the Taylor series is centered at $a = 0$, it is called Maclaurin series

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n = f(0) + f'(0)x + \frac{f''(0)}{2}x^2 + \frac{f'''(0)}{6}x^3 + \dots$$

We look at the $f(x) = e^x$'s Taylor expansion as an example. Note that

$$f^{(n)}(x) = e^x, \forall x \in \mathbb{N}$$

That said

$$f^{(n)}(0) = 1, \forall x \in \mathbb{N}$$

Substitute into the Taylor's formula and thus we get

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

Here are the Maclaurin series for significant functions.

1. $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$
2. $\sin(x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} \dots$
3. $\cos(x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} \dots$
4. $\ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} \dots$

^{A.4}i.e. $f(x)$ is n times differentiable

The x can be replaced by a function of x . For a quick insight, we prove Euler's formula via Taylor series

$$e^{ix} = \cos(x) + i \sin(x) \quad (\text{Euler's Formula})$$

Note that we can replace x with ix ,

$$e^{ix} = \sum_{n=0}^{\infty} \frac{(ix)^n}{n!} = 1 + ix + \frac{(ix)^2}{2!} + \frac{(ix)^3}{3!} + \dots = 1 + ix - \frac{x^2}{2!} - \frac{ix^3}{3!} + \frac{x^4}{4!} + \frac{ix^5}{5!} - \frac{x^6}{6!} - \frac{ix^7}{7!} \dots$$

and

$$i \sin(x) = \sum_{n=1}^{\infty} \frac{i(-1)^{n+1} x^{2n+1}}{(2n+1)!} = ix - \frac{ix^3}{3!} + \frac{ix^5}{5!} - \frac{ix^7}{7!} \dots$$

Therefore we obtain that

$$\begin{aligned} \cos(x) + i \sin(x) &= \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} \dots\right) + \left(ix - \frac{ix^3}{3!} + \frac{ix^5}{5!} - \frac{ix^7}{7!} \dots\right) \\ &= 1 + ix - \frac{x^2}{2!} - \frac{ix^3}{3!} + \frac{x^4}{4!} + \frac{ix^5}{5!} - \frac{x^6}{6!} - \frac{ix^7}{7!} \dots \\ &= e^{ix} \end{aligned}$$

We demonstrate another usage with Taylor series. Recall that

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} \dots$$

Differentiate both sides,

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + \dots$$

By the way, we see that $(1+x)(1-x+x^2-x^3+\dots) = 1$ such that $(1+x)$ is the inverse polynomial of $(1-x+x^2-x^3+\dots)$. Now replace x with $-x$,

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \dots$$

or replace with x^2 ,

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \dots$$

where we find that it matches the result of geometric series,

$$\frac{1}{1-r} = \sum_{n=0}^{\infty} r^n \quad (\text{Geometric Series})$$

which converges when $|r| < 1$. Therefore the Taylor series converges when x is in the range when $|r| < 1$. For instance,

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

that converges when $|-x| < 1 \implies |x| < 1$ ^{A.5}. As well,

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

^{A.5}a.k.a. radius of convergence, here is 1

which converges when $|-x^2| < 1 \implies x^2 < 1 \implies |x| < 1$. The Taylor series also uniquely characterizes a function. In general, when a function is complicated, we can operate on its Taylor expansion for simplicity.

The truncated Taylor series (up to step m) is

$$T_{m,a}(x) = \sum_{n=0}^m \frac{f^{(n)}(a)}{n!} (x-a)^n$$

The Taylor remainder with respect to $T_{m,a}(x)$ is

$$R_{m,a}(x) = f(x) - T_{m,a}(x)$$

where the error is defined as $|R_{m,a}(x)|$. Note that the error is $O((x-a)^{m+1})$ ^{A.6}, that said

$$f(x) = T_{m,a}(x) + O((x-a)^{m+1})$$

or equivalently,

$$f(x) = T_{m,a}(x) + o((x-a)^m)$$
^{A.7}

The Taylor expansion has advantages in various applications. For example, instead of calculating the complex functions such as $e^x, \sin(x), \cos(x)$, the computer calculates their Taylor series which is a polynomial and can be quickly computed.

Now we extend to higher dimensional spaces. The general second-degree Taylor expansion of $f \in C^3 : \mathbb{R}^n \rightarrow \mathbb{R}$ at x with sufficiently small^{A.8} increment $h \in \mathbb{R}^n$ is

$$f(x+h) = f(x) + \langle h, \nabla f(x) \rangle + \frac{1}{2} \langle h, \nabla^2 f(x)h \rangle + o(\|h\|^2) \quad (\text{Taylor Exp.})$$

Equivalently,

$$f(x+h) = f(x) + h^T \nabla f(x) + \frac{1}{2} h^T H h + o(\|h\|^2)$$

Here $\nabla^2 f(x) = H$ is the Hessian matrix of $f(x)$. A 2-d example is

$$\lim_{|h| \rightarrow 0} \lim_{|k| \rightarrow 0} f(x+h, y+k) = f(x, y) + hf_x(x, y) + kf_y(x, y) + \frac{1}{2}[h^2 f_{xx}(x, y) + 2hk f_{xy}(x, y) + k^2 f_{yy}(x, y)] + o(h^2) + o(k^2)$$

$$\text{where } f_x = \frac{\partial f}{\partial x}, f_y = \frac{\partial f}{\partial y}, f_{xy} = \frac{\partial^2 f}{\partial x \partial y}, f_{xx} = \frac{\partial^2 f}{\partial x^2}, f_{yy} = \frac{\partial^2 f}{\partial y^2}.$$

Recall that the remainder can be bounded by the last term upon truncation. As an example we consider the first-order remainder $R_1 = \frac{1}{2}h^T H h + O(\|h\|^3) = O(\|h\|^2)$. The Taylor's Theorem states that

Theorem A.3.2 (Taylor's Theorem)

For sufficiently smooth $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and small increment $h \in \mathbb{R}^n$, there exists some $\xi \in [x, x+h]$ such that

$$f(x+h) = f(x) + h^T \nabla f(x) + \frac{1}{2} h^T H(\xi) h$$

That said, the Taylor remainder (error) can be merged into the last term (since the error is bounded by the last term). For higher-order remainders, we just substitute ξ into the corresponding derivative part,

$$f(x+h) = \sum_{n=0}^m \frac{f^{(n)}(x)}{n!} h^n + \frac{f^{(m+1)}(\xi)}{(m+1)!} h^{m+1}$$

^{A.6} $f(x)$ is $O(g(h))$ iff $f(x) \leq M \cdot g(h)$ for some $M > 0$, i.e. $f(x)$ is bounded above by $Mg(h)$. Here the dominating (largest) term of the remainder is $(x-a)^{m+1}$ if $(x-a)$ is small.

^{A.7} $f(x)$ is $o(g(h))$ if $f(x)$ converges faster than $g(h)$ when approaching 0, i.e. $\lim_{h \rightarrow 0} \frac{f(h)}{g(h)} = 0$. Here it means that the remainder converges faster than the last term $(x-a)^m$ since the remainder terms have higher orders.

^{A.8} always assume that $\|h\| \rightarrow 0$

for m th order remainder, using $f : \mathbb{R} \rightarrow \mathbb{R}$ as an example.

Up to this line is enough for the course offerings. Below is not required to be understood. Refer to 2-deg [Taylor Series](#), we can extend to higher-order multivariate Taylor expansion.

Definition A.3.3 (Multivariate Taylor Series)

For sufficiently smooth $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and small increment $h \in \mathbb{R}^n$, its Taylor expansion is

$$f(x + h) = \sum_{n=0}^{\infty} \frac{1}{n!} (h^T D)^n f(x)$$

where

$$D = \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n} \right)^T$$

is the vector of partial differentiation operators.

The term $(h^T D)^n$ can be expanded using multinomial theorem. To write it explicitly, for a simple example $f : \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$\begin{aligned} f(x_1 + h_1, x_2 + h_2) &= \sum_{n=0}^{\infty} \frac{1}{n!} \underbrace{\left(h_1 \frac{\partial}{\partial x_1} x_1 + h_2 \frac{\partial}{\partial x_2} x_2 \right)^n}_{\text{expand using binomial theorem}} f(x_1, x_2) \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} \left(h_1^n \frac{\partial^n}{\partial x_1^n} + h_1^{n-1} h_2 \frac{\partial^n}{\partial x_1^{n-1} \partial x_2} + \dots + h_1 h_2^{n-1} \frac{\partial^n}{\partial x_1 \partial x_2^{n-1}} + h_2^n \frac{\partial^n}{\partial x_2^n} \right) f(x_1, x_2) \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} \left(h_1^n \frac{\partial^n f(x_1, x_2)}{\partial x_1^n} + h_1^{n-1} h_2 \frac{\partial^n f(x_1, x_2)}{\partial x_1^{n-1} \partial x_2} + \dots + h_1 h_2^{n-1} \frac{\partial^n f(x_1, x_2)}{\partial x_1 \partial x_2^{n-1}} + h_2^n \frac{\partial^n f(x_1, x_2)}{\partial x_2^n} \right) \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} \left(h_1^n f_{x_1^{(n)}}(x_1, x_2) + h_1^{n-1} h_2 f_{x_1^{(n-1)} x_2^{(1)}}(x_1, x_2) + \dots + h_1 h_2^{n-1} f_{x_1^{(1)} x_2^{(n-1)}}(x_1, x_2) + h_2^n f_{x_2^{(n)}}(x_1, x_2) \right) \end{aligned}$$

Note that the differentiation order of n th^{A.9} term is n , and is thus $O(\|h\|^n)$. For example, for the 2-d Taylor series above, the third ($O(\|h\|^3)$) term is

$$\frac{1}{3!} \left(h_1^3 f_{x_1^{(3)}}(x_1, x_2) + h_1^2 h_2 f_{x_1^{(2)} x_2^{(1)}}(x_1, x_2) + h_1 h_2^2 f_{x_1^{(1)} x_2^{(2)}}(x_1, x_2) + h_2^3 f_{x_2^{(3)}}(x_1, x_2) \right)$$

The general Taylor's Theorem is

Theorem A.3.4 (Taylor's Theorem)

For sufficiently smooth $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and small increment $h \in \mathbb{R}^n$, for $m \geq 0$, there exists $\xi \in [x, x + h]$ such that

$$f(x + h) = \sum_{n=0}^m \frac{1}{n!} (h^T D)^n f(x) + \frac{1}{(m+1)!} (h^T D)^{m+1} f(\xi)$$

A.4 Optimization

We aim to find the global optimizer of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Suppose f is sufficiently smooth. The local maximas are those points satisfying $f(\hat{x}) \geq f(x)$ for all x in the neighbor of \hat{x} . Formally speaking, there exists some open ball $B_r = \{x : \|x - \hat{x}\| < r\}$ with radius $r > 0$ centered at \hat{x} such that $f(\hat{x}) \geq f(x)$ for all $x \in B_r$. If \hat{x} satisfies

$$\nabla f(\hat{x}) = \mathbf{0} \quad \text{and} \quad \nabla^{(2)} f(\hat{x}) \prec 0 \quad (\text{opt. Test})$$

^{A.9} n counting from 0

then \hat{x} is sufficient to be a local maxima. However, the above statement is not necessary. The target function might have multiple local maximas, so ultimately we need to compare those maximas and find the global maxima. In addition, if the target function is concave, a local maxima must be the global maxima, thus any point satisfying the optimality test would be the global maxima.

We see how the optimality test works. Recall Taylor's formula,

$$f(x + h) = f(x) + \nabla f(x)h + \frac{1}{2}h^T \nabla^{(2)} f(x)h$$

Here, (non-zero) h denotes the direction^{A.10} of a small step forward starting at x , $f(x)$ denotes the current function value, and $\nabla f(x)h + \frac{1}{2}h^T \nabla^{(2)} f(x)h$ denotes the increment (or decrease) in the function value after walking the step h . Notice that the local maximas are stationary points such that $\nabla f(\hat{x}) = 0$, so $\nabla f(\hat{x})h = 0$. Therefore the change in the function value upon step h is only

$$\frac{1}{2}h^T \nabla^{(2)} f(x)h$$

This is the quadratic form associated with the Hessian matrix, and that is why the second derivative matters in optimality. Recall that \hat{x} is a local maxima iff $\nabla^{(2)} f(\hat{x})$ is negative definite, i.e.

$$h^T \nabla^{(2)} f(\hat{x})h < 0, \forall h \neq 0 \in \mathbb{R}^n$$

That said, a forward step in any direction will decrease the function value, which matches the definition of a local maxima.

^{A.10}The length of h does not matter, just to be small enough. It mainly indicates the direction.

B Linear Algebra

Definition B.0.1 (Outer Product of Vectors)

Let $x = (x_1, \dots, x_m)^T \in \mathbb{R}^m$, $y = (y_1, \dots, y_n) \in \mathbb{R}^n$. The outer product of x, y is

$$xy^T = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} (y_1 \ \dots \ y_n) = \begin{pmatrix} y_1 & \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} & \dots & y_n & \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} \end{pmatrix} = \begin{pmatrix} x_1 y_1 & x_1 y_2 & \dots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \dots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \dots & x_m y_n \end{pmatrix}$$

and similarly

$$yx^T = \begin{pmatrix} y_1 x_1 & y_1 x_2 & \dots & y_1 x_m \\ y_2 x_1 & y_2 x_2 & \dots & y_2 x_m \\ \vdots & \vdots & \ddots & \vdots \\ y_n x_1 & y_n x_2 & \dots & y_n x_m \end{pmatrix} = (xy^T)^T$$

Definition B.0.2 (Cholesky Factorization)

For $V \in \mathbb{S}^n$, its Cholesky factor, A , is a $n \times n$ matrix such that

$$V = AA^T$$

The Cholesky Factorization is only defined for symmetric matrices.

Proposition B.0.3 (Geometry of Determinant)

The determinant of a matrix formed with vectors from some basis \mathcal{B} is a geometric measure of the hypervolume in the structure of the vectors (or simply *measure*) in the corresponding vector space. In particular,

Let $\{x_1\}$ ^a be a (trivial) basis of \mathbb{R} , then $\det(x_1) = x_1$ is the **length** of the line segment between 0 and x_1 .

Let $\{v_1, v_2\}$ be a basis of \mathbb{R}^2 , then $\det([v_1 \ v_2])$ is the **area** of a parallelogram constructed with v_1, v_2 .

Let $\{w_1, w_2, w_3\}$ be the basis of \mathbb{R}^3 , then $\det([w_1 \ w_2 \ w_3])$ is the **volume** of the parallelepiped constructed with w_1, w_2, w_3 .

^aAll basic vectors must be non-zero.

Note that we didn't intend to ignore the sign of the determinant. Although typical length, area, volume should be non-negative, the sign additionally indicates the direction, e.g. negative sign means one basic vector is in the negative quadrant.

Definition B.0.4 (Change of Basis)

Let $V = \{v_1, \dots, v_n\}, W = \{w_1, \dots, w_n\}$ be two vector bases for \mathbb{R}^n . Let

$$[V] = [v_1 \ \dots \ v_n] \quad [W] = [w_1 \ \dots \ w_n]$$

be the corresponding matrix representations of V, W . Then the change of basis matrix from basis V to W is a $n \times n$ matrix $[\mathcal{C}]$ that satisfies

$$[\mathcal{C}][V] = [W]$$

Note that by property of basis, $[\mathcal{C}], [V], [W]$ are non-singular. Also for any $x \in \mathbb{R}^n$, let

$$[x]_V = [V]^{-1}x \quad [x]_W = [W]^{-1}x$$

be the coordinates^{B.1} of x with respect to bases V, W respectively, then

$$[\mathcal{C}] [x]_V = [x]_W$$

We see that the change of basis matrix \mathcal{C} is indeed an isomorphic transformation from $[x]_V$ to $[x]_W$.

^{B.1}Coordinate vectors for x are unique for their corresponding basis.

C Indices

List of Theorems

1.1.1 Definition (Sample Spaces)	3
1.1.2 Definition (Events)	3
1.1.3 Definition (Disjoint Events)	3
1.1.4 Definition (Probability Functions)	3
1.1.1 Example (Seats Selection)	3
1.1.5 Proposition (Properties of Probability Function)	4
1.1.6 Proposition (Probability of Union of Events)	4
1.1.7 Corollary (Probability of Union of Disjoint Events)	4
1.1.8 Proposition (Probability of Independent Events)	5
1.2.1 Definition (Conditional Probability)	6
1.2.1 Example (Boy or Girl?)	6
1.2.2 Theorem (Law of Total Probability)	6
1.2.2 Example (Law of Total Probability 1)	7
1.2.3 Example (Law of Total Probability 2)	8
1.2.3 Theorem (Bayes' Theorem)	8
1.2.4 Example (Infection Probability)	9
1.3.1 Definition (σ -algebra)	9
1.3.2 Definition (Probability Space)	9
1.3.3 Definition (Random Variables)	10
1.3.4 Definition (Binomial Distribution)	11
1.3.5 Definition (Poisson Distribution)	11
2.1.1 Example (Discrete Distribution Example 1)	12
2.1.1 Proposition (Properties of Probability Mass Function)	13
2.1.2 Example	13
2.1.2 Proposition (Properties of Probability Density Function)	14
2.1.3 Definition (Support)	15
2.1.4 Proposition (Properties of Cumulative Distribution Function)	15
2.1.5 Definition (Quantile Function)	15
2.2.1 Definition (Expectation)	16
2.2.2 Theorem (Expectation with TPF)	17
2.2.1 Example (Discrete Expectation)	17
2.2.2 Example (Poisson Distribution Mean)	18
2.2.3 Example (Binomial Distribution Mean)	18
2.2.4 Example (Exponential Distribution Mean)	19
2.2.3 Definition (Expectation Function)	20
2.2.4 Proposition (Linearity of Expectation)	20
2.2.5 Corollary (Linearity of Expectation)	21
2.2.6 Definition (Variance)	21
2.2.5 Example (Poisson Distribution Variance)	22
2.2.6 Example (Binomial Distribution Variance)	22
2.2.7 Example (Exponential Distribution Variance)	23
2.2.7 Proposition (Zero Variance)	24
2.2.8 Definition (Standard Deviation)	24
2.2.9 Proposition (Variance of Linearly Transformed Distribution)	25
2.3.1 Definition (Moment Generating Function (MGF))	26
2.3.2 Theorem (Generating Moments)	27
2.3.1 Example (Binomial Distribution MGF)	27

2.3.2 Example (Poisson Distribution MGF)	27
2.3.3 Theorem (Uniqueness Theorem of MGF)	28
2.3.4 Lemma (Characterizing Distributions)	28
2.4.1 Algorithm (Determine Distribution)	29
2.4.1 Example (Discrete Distribution Transformation 1)	29
2.4.2 Example (Continuous Distribution Transformation 1)	29
2.4.3 Example (Continuous Distribution Transformation 2)	29
2.4.2 Theorem (Inverse Probability Integral Transformation)	30
2.4.4 Example (Sampling Exponential Distribution)	30
3.1.1 Definition (Joint Probability Function)	31
3.1.2 Proposition (Properties of Joint Probability Function)	31
3.1.3 Definition (Marginal Probability Function)	31
3.1.4 Definition (Joint Cumulative Distribution Function)	33
3.1.5 Proposition (Properties of Joint CDF)	33
3.2.1 Definition (Independence of RVs)	34
3.2.1 Example (Independent RVs)	34
3.3.1 Definition (Expectation of Multivariate Distribution)	35
3.3.2 Definition (Expectation function of Multivariate Distribution)	35
3.3.3 Proposition (Expectation of Separable Independent Functions)	35
3.3.4 Definition (Covariance)	35
3.3.5 Proposition (Covariance of Scaled RVs)	36
3.3.6 Proposition (Independence Implies No Covariance)	36
3.3.7 Proposition (Variance of Linear Combination of RVs)	37
3.3.8 Lemma (Range of Covariance)	37
3.3.9 Definition (Correlation Coefficient)	38
3.4.1 Definition (Multivariate MGF)	39
3.4.2 Proposition (Properties of Multivariate MGF)	39
3.4.3 Lemma (MGF of Sum Distribution)	39
3.4.1 Example (MGF of Sum Bernoulli RVs)	39
3.4.2 Example (MGF of Sum Poisson RVs)	40
3.5.1 Lemma (Product and Max Distribution)	40
3.5.2 Lemma (Min Distribution)	41
3.6.1 Definition (Conditional Distribution)	43
3.6.2 Definition (Conditional Expectation)	43
3.6.3 Theorem (Law of Total Expectation)	43
3.6.4 Theorem (Law of Total Variance)	44
3.6.1 Example (Law of Total Expectation)	45
3.6.5 Lemma (Sum Distribution)	45
3.6.2 Example (Sum of Discrete Distributions)	46
3.6.3 Example (Sum of Continuous Distributions)	47
4.1.1 Definition (Normal Distribution)	50
4.1.2 Lemma (Scaled and Shifted Normal Distribution)	51
4.1.3 Corollary (Pivoting Normal Distribution)	51
4.1.4 Lemma (Sum of Independent Normal RVs)	52
4.1.5 Corollary (Linear Combination of Independent Normal RVs)	52
4.2.1 Definition (Gamma Distribution)	53
5.1.1 Definition (Multinomial Distribution)	56
5.1.1 Example (Multinomial Example)	57
5.1.2 Proposition (Marginal Distribution of Multinomial Distribution)	58
5.1.3 Proposition (Sum of Multinomial Marginals)	58
5.1.4 Proposition (Covariance Between Multinomial Marginals)	58

5.1.5 Proposition (Conditional Multinomial Distribution)	59
5.1.2 Example	60
5.3.1 Definition (Multivariate Normal Distribution)	62
5.3.2 Proposition (Stochastic Representation of MVN RVs)	64
5.3.3 Corollary (Building MVN using Standard MVN)	64
5.3.4 Proposition	64
5.3.5 Corollary (Independent Normal RVs)	65
5.3.6 Definition (MVN Covariance)	68
5.4.1 Example	69
5.4.2 Example	69
6.3.1 Theorem (Central Limit Theorem)	72
6.3.2 Theorem (ANOVA)	72
6.3.3 Proposition (Estimator of Sample Variance)	73
6.3.4 Proposition (Estimating Mean with t)	75
7.2.1 Definition (Likelihood Function and Log Likelihood Function)	77
7.2.2 Lemma (Invariance Property)	77
7.2.3 Algorithm (Maximum Likelihood Estimate (MLE))	77
7.2.4 Corollary (Invariance Property of MLE)	77
7.2.1 Example (MLE for Binomial Distribution)	78
7.2.2 Example (MLE for Normal Distribution)	79
7.2.3 Example (MLE for Multivariate Normal Distribution)	79
A.1.1 Theorem (Fundamental Theorem of Calculus 1)	81
A.1.2 Theorem (Fundamental Theorem of Calculus 2)	81
A.1.3 Definition (Convergence)	81
A.1.4 Definition (Integration By Parts)	81
A.1.5 Theorem (Cauchy-Schwarz Inequality)	81
A.1.6 Theorem (Gaussian Integral)	81
A.1.7 Corollary (Scaled and Shifted Gaussian Integral)	82
A.2.1 Definition (Jacobian)	83
A.3.1 Definition (Taylor Series)	86
A.3.2 Theorem (Taylor's Theorem)	88
A.3.3 Definition (Multivariate Taylor Series)	89
A.3.4 Theorem (Taylor's Theorem)	89
B.0.1 Definition (Outer Product of Vectors)	91
B.0.2 Definition (Cholesky Factorization)	91
B.0.3 Proposition (Geometry of Determinant)	91
B.0.4 Definition (Change of Basis)	91

List of Figures

2.1 <i>Bin(5, 0.5)</i> Probability Functions	13
2.2 Same Mean Different Spread	21
2.3 100 <i>Poi(2)</i> samples	24
2.4 scaled 100 <i>Poi(2)</i> samples	25
3.1 RVs with covariance 3.43	36

List of Tables

3.1.1 Weather Probability	31
3.1.2 Weather Probability	32

D References