GU4001: Into to Prob and Stat
Class Notes

Xiao xy2437

January 18, 2022

Contents

1 Logistics **3**

2 Introduction **4**

3 Basics **5**
3.1 Multiplication Rule 6
3.2 Set Theory Tools 7
3.3 Redefinition of Probability 10

4 Conditional Probability **13**
4.1 Rules for Conditional Probability 15
4.2 Independence 18

5 Discrete Random Variables **21**
5.1 Probability Mass Function 23
5.2 Cumulative Distribution Function 24
5.3 Mean and Variance 26
5.3.1 Mean of a Random Variable 26
5.3.2 LOTUS 29
5.3.3 Variance 32

6 Continuous Random Variables **34**
6.1 CDF for Continuous RV 34
6.2 Probability Density Function 37
6.2.1 Uniform Random Variable 39
6.2.2 Gaussian(Normal) Random Variable 40
6.2.3 Exponential Random Variable 43
6.3 Mean and Variance 44
6.3.1 Mean 44
6.3.2 Variance 47
6.3.3 Mean and Variance for Common Distributions 47

7 Conditional, Joint, and Marginal **52**
7.1 Conditional Distribution 52
7.2 Joint for Discrete RV 54
7.2.1 Joint Distribution for Discrete RV 54
7.2.2 Joint CDF for Discrete RV 56
7.3 Marginals for Discrete RV 58
7.3.1 Marginal PMF for Discrete RV 58
7.3.2 Marginal CDF for Discrete RV 59
7.4 Independence of Discrete RV 60
7.5 Joint for Continuous RV 62
7.5.1 Joint PDF for Continuous RV 62
7.5.2 Joint CDF for Continuous RV 63
7.6 Independence of Continuous RV 64
7.7 Conditional for Continuous RV 67

8 Covariance and Correlation **69**
8.1 Covariance 69
8.2 Correlation Coefficient 77
8.3 Covariance v.s. Correlation 80

9 Multiple Random Variables **82**
9.1 Independence 83

10 Parameter Estimation **84**
10.1 Method of Moments 86
10.2 Maximum Likelihood Estimation 89

11 Linearity of Expected Value **93**

12 Review **94**
12.1 Set Laws 94

1 Logistics

Overall:

- all material on **google site** instead of Canva
- textbooks are not required, but the first one on the list is recommended

Grading:

- details see google site
- Midterm is take-home, harder
- Final is in-class easier

2 Introduction

First, let us consider some examples.

***Example*: Birthday Paradox**

Question: what is the probability that, for n people in a room, no two are having the same birthday?

Solution: Note that this is the same as calculating the probability of everyone having a different birthday. This means that:

- if we have 2 people, then probability is 364/365
- if we have 3 people, then probability is $(364/365) * (363/365)$
- ...

Therefore, if we have n people, the probability is:

$$1 \cdot \frac{364}{365} \cdot ... \cdot \frac{365-(n-1)}{365} = \frac{365!}{365^n(365-n)!}$$

which if you evaluate for $n = 23$ people, you get something like 50%! (A quick check of this would to be thinking that, if we have 366 people, then by the pigeon-hole principle, the probability has to be 1).

Some brief history about probability would also be interesting.

Probability was started mainly by the **rich gamblers**, as they have the time and effort to develop theories related to probability.

- Girolamo Cardano: a famous physician addicted to gambling, and invented one of the most influential foundation of it.
- then it is advanced by numerous others, such as the famous work of "unfinished game".

3 Basics

Now we consider some basic principles of it.

Definition 3.1. *A **random event** is an event with a probability of occurrence determined by some probability distribution. In other words, you do not know in advance if it will happen or not.*

Definition 3.2. ***Sample Space** of a random event/experiment is the set of all possible outcomes of this experiment.*

Example

For tossing a coin twice, the sample space $\mathcal{S} = \{HH, HT, TH, TT\}$.

Definition 3.3. *An event A is a subset of sample space $\mathcal{S}$. After performing an experiment we say that an **event** A **has happened** if the outcome is **in** A.*

Example

If you are rolling a die, and we want to find out the case if we get an even number.

In this case, we know $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, and the event we are looking for is $A = \{2, 4, 6\}$.

Then if we got the outcome of 2, we can say that A has happened.

Definition 3.4. ***Cardano's definition of probability**, which is partially correct: Let A be an event for a random experiment with a finite sample space $\mathcal{S}$, the probability of A happening is defined as:*

$$P_{cardano}(A) = \frac{|A|}{|\mathcal{S}|}$$

However, note that we are assuming here that each outcome has an **equal probability of happening**.

Example

Suppose we are rolling dice again, and we want to calculate the probability of getting at least one 2 or one 3 for two tosses.

We know then the sample space is simply:

$$|\mathcal{S}| = 6 \times 6 = 36$$

And the event we are looking at is:

$$|A| = 36 - (4 \times 4) = 20$$

which basically is the number of outcomes that we got no 2 and no 3

in two tosses.

Therefore, according to Cardano, we get:

$$P_{card}(A) = \frac{|A|}{|\mathcal{S}|} = \frac{20}{36}$$

Note:

Question: What does the probability of something mean? The future outcome is always determined (by the god), so probability as a number does not make sense?

Answer: One way of interpreting probability would be the **frequency** of an event happening if I **repeat the experiment for a lot of times**. Consider the case of dice roll, if I record the frequency

$$\frac{\#\text{ times event happened}}{\#\text{ experiment}}$$

of getting a 1 over time, you will see something like the figure below.

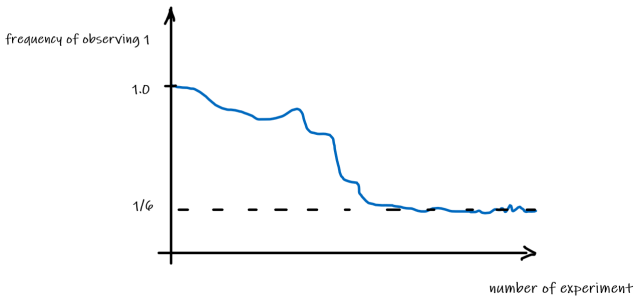


Figure 1: Frequency of observing 1 in dice rolls

In the end, this is related to the idea of **law of large numbers**, which basically discusses the idea that **random events become non-random, computable quantity** if we take a large amount of experiments.

3.1 Multiplication Rule

Consider the case when $\mathcal{S}$ is large in size. Then instead of counting, you can use **permutation** and **combination** to compute the size quickly.

Example

Consider the case when you have 10 marbles in a jar, each having a different color. What is the sample size when you can choose four balls **with replacement**?

The idea is to consider a **single experiment**, then the output should look like:

$$(\text{green}, \text{red}, \text{red}, \text{blue})$$

In particular, you can have any of the following **placeholders**:

$$(\_, \_, \_, \_)$$

Since each can take values of 10 possible ways, you have in the end:

$$10 \times 10 \times 10 \times 10$$

Therefore, we have the following lemma.

Lemma 3.1. *Consider a compound experiment, made of $m \geq 2$ experiments. Suppose each experiment has n_i outcomes. Then the compound experiment has:*

$$n_1 \times n_2 \times ... \times n_m$$

possible outcomes.

Therefore, we have the following useful rules.

Definition 3.5 (Permutation)**.** *Consider choosing k objects out of n objects without replacement. If the order of the object matters, then we have:*

$$\#\ ways = n \times (n-1) \times ... \times (n-k+1) = \frac{n!}{(n-k)!}$$

for the total number of arrangements.

Note:

note that here we can think about it in terms of placeholders as well, since basically each outcome is a k-tuple, we then just need to consider the number of cases for each placeholder in the k-tuple.

Definition 3.6 (Combination)**.** *If the order matters in the above example, then since **each k-tuple/outcome** will be repeated by $k!$ times, we have*

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

for the total number of arrangements.

3.2 Set Theory Tools

Another useful tool would be set theory. Consider some sample space $\mathcal{S}$ and some event A. Some **set identities** you already know would be:

- $A \cup A^c = \mathcal{S}$
- $|A \cup B| = |A| + |B| - |A \cap B|$

***Example*: Back to Birthday Paradox**

Consider the birthday paradox we introduced before: we have n students in a class, what is the probability that at **least two of them** have the same birthday?

Using what we knew, we first consider an outcome of a experiment:

$$(1, 45, 78, \_, ..., \_)$$

where each placeholder puts the birthday in the range of $[1, 365]$.

Therefore, we first now that the size of our sample space is:

$$|\mathcal{S}| = 365^n$$

Now, for the **event A we want**, it should look something like:

$$(1, 1, 4, 8, ..., \_) \text{ or } (1, 2, 2, 2, 8, ..., \_)$$

So we can define:

$A = \{\text{all the n-tuples in which at least two of the numbers are the same}\}$

This looks hard, but we know $A + A^c = \mathcal{S}$, and notice that:

$$A^c = \{\text{all the n-tuples contains unique numbers}\}$$

This is easy, since we just need to have different numbers in the n-placeholders. Therefore, we have:

$$A^c = {}^{365}P_n = \frac{365!}{(365-n)!}$$

and note that **order matters here**, so we are having a permutation instead of a combination.

Finally, we can compute what we wanted:

$$|A| = |\mathcal{S}| - |A^c| = 365^n - \frac{365!}{(365-n)!}$$

and the probability is:

$$P_{\text{card}}(A) = \frac{|A|}{|S|} = 1 - \frac{365!}{(365-n)!(365)^n}$$

assuming $n \leq 365$. For $n > 365$, then by pigeon hole principle, there must exists two people having the same birthday.

Note:

An implicit assumption made above is that we are assuming that every birthday are **equally likely**. That is also way we have noted it is the Cardano's probability. In fact, babies are more often born in July to Sep.

- this means that the actual probabilit is **higher**. (e.g. think of the extreme case that 0 people are born in Jan. Then obviously the chances of same birthday is higher).

Example

Question: I have n distinguishable marbles of different color in a jar, and a I draw k marbles with replacement (order matters). What is the probability of observing at least one red and **not any blue**?

Solution: One example of the event happening would be:

$$(\text{red}, \text{red}, \text{green}, ...)$$

Again, you notice that this is hard. So we can consider:

$$\begin{cases} A = \{\text{seen at least one red}\} \\ B = \{\text{seen at least one blue}\} \end{cases}$$

Therefore, we want to calculate:

$$P_{\text{card}}(A \cap B^c)$$

The easy thing is B^c or A^c

$$|B^c| = |A^c| = (n-1) \times (n-1) \times (n-1) \times ...$$

since we are doing it **with replacement**. One trick you will notice is that:

$$\begin{aligned} (A \cap B^c) \cup (A^c \cap B^c) &= B^c \\ |A \cap B^c| &= |B^c| - |A^c \cap B^c| \end{aligned}$$

this is useful since we know already A^c, B^c, and the computation of $A^c \cap B^c = \{\text{not seen red or blue}\}$:

$$|A^c \cap B^c| = (n-2)^k$$

Therefore, you finally get:

$$|A \cap B^c| = (n-1)^k - (n-2)^k$$

and now calculating the probability is trivial.

3.3 Redefinition of Probability

Now, we see problems of Cardano's definition (**recall that it assumed that each outcome is equally likely to happen**).

For instance, assume that you have an **unfair coin**, then the **sample space remains the same**, but the probability of getting an head **will not be** $\frac{1}{2}$.

***Example*: Unfair Coin**

Consider now having an unfair coin, such that there is still either a head or tail, but that, after tossing 10000 times:

- you get 7000 tails
- you get 3000 heads

Then, we define that:

$$\mathbb{P}(\{T\}) = \frac{7000}{10000} = 0.7$$

$$\mathbb{P}(\{H\}) = \frac{3000}{10000} = 0.3$$

Therefore, the idea is that actual **outcomes/events might not be equally possible**. If this is the case, then usually those probability can only be told in advance or tested empirically, instead of using $\mathbb{P}_{\text{card}}$. Yet, there are still **rules that we can still use**.

Theorem 3.2 (Axiom of Probability)**.** *Suppose A and B are disjoint events such that $A \cap B = \emptyset$, then:*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$$

Therefore, if you know $\mathbb{P}(A)$ and $\mathbb{P}(B)$, then you are done (which could be unbalanced).

Therefore, we consider the following definition of probability revised from the above phenomenon.

Definition 3.7 (Semi-Formal Definitino of Probability)**.** *A probability space consists of a sample space $\mathcal{S}$ and a probability* ***function*** *$\mathbb{P}$, which takes an event $A \subseteq \mathcal{S}$ as input and returns a number between $[0, 1]$ as output. The function $\mathbb{P}$ satisfy the following two axioms:*

1. $\mathbb{P}(\mathcal{S}) = 1$
2. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$, *if* $A \cap B = \emptyset$

However, you will notice that this is not entirely true if the sample space is infinite and your event size is also infinite.

***Example*: Sum of Two Dice**

Question: Consider two dices, where you want to roll both of them and compute the sum. Therefore, you obviously know that:

$$\mathcal{S} = \{2, 3, 4, ..., 12\}$$

Then, what is the probability of observing 7?

Solution: The idea here is that each event is not equally probable of happening in. There are at least two ways of computing this:

1. Consider directly the size of event. You will get it is 6. Then you can divide it by the size of sample space of 2 tuples, which has size of 36.
2. Or you can consider a matrix as your sample space:

$$\begin{bmatrix} (1,1) & (1,2) & ... & (1,6) \\ (2,1) & (2,2) & \dots & (2,6) \\ \vdots & \vdots & \vdots & \vdots \\ ... & ... & ... & (6,6) \end{bmatrix}$$

   Then you just count the diagonal.

Additionally, if you were to compute $\mathbb{P}(\{2, 7\})$, we can easily do without re-computation:

$$\mathbb{P}(\{2, 7\}) = \frac{1}{36} + \frac{6}{36} = \frac{7}{36}$$

Due to the axiom.

Now, we will show more on how powerful those axioms are.

Lemma 3.3. *If $A \subseteq \mathcal{S}$ is an event, then:*

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

note that this is true in general, regardless of Cardano's error.

Proof. Using the axiom 2, we know that $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$. Then, we know that since:

$$A \cup A^c = S$$

Then using axiom 1:

$$P(A \cup A^c) = 1$$

Now, by definition $A \cap A^c = \emptyset$, so we get:

$$P(A \cup A^c) = P(A) + P(A^c) = 1$$

□

Lemma 3.4.

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

which is true in general as well.

Proof. First, we notice that $(A \cap B)$ is disjoint with $(A \cap B^c)$. Therefore:

$$P((A \cap B) \cup (A \cap B^c)) = P(A \cap B) + P(A \cap B^c)$$

Then we need to factor out (using distributive law) such that:

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

because $A \cap (B \cup B^c) = A \cap \mathcal{S} = A$.
Finally, combing the above, we finishes the proof. □

Lemma 3.5. *For any two event A, B:*

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

Proof. First we know that $A \cup B = A \cup (B \cap A^c)$, so that A and $B \cap A^c$ are disjoint events. Then:

$$P(A \cup B) = P(A) + P(B \cap A^c)$$

From Lemma 3.4 we know that:

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

Combining the two we obtain the result. □

4 Conditional Probability

Consider the presidential election in 2016, Trump vs Clinton. We knew that before, the overall estimate of electro votes would be:

- $\mathbb{P}(\text{Clinton Wins}) = 0.66$
- $\mathbb{P}(\text{Trump Wins}) = 0.34$

But we know that some **states matter a lot** in the election. Consider this probability:

$$\mathbb{P}\left(\text{Clinton Winning} \mid \text{Trump wins Florida and Iowa}\right)$$

which means that **given that** Trump wins Florida and Iowa, what is the probability of Clinton still winning? We now have **new evidence**, so the probability would have changed.

Definition 4.1 (Conditional Probability)**.** *Suppose we are interested in event A, and we know that B has happened. Then it means that:*

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

If you have multiple events B_1, B_2, then it is still similar:

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

This is best understood by diagrams and Cardano's definition of probability

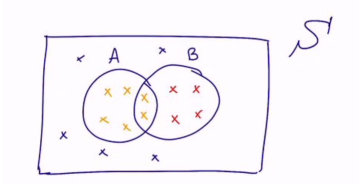


Figure 2: Conditional Probability Example

Assuming that each outcome is equally likely to happen. Then we know that:

$$P_{card}(A) = \frac{6}{15}, \quad P_{card}(B) = \frac{6}{15}$$

But know assume that B **has happened**. Then our **sample space becomes all the B events**. Therefore:

$$P_{card} = \frac{2}{6} = \frac{|A \cap B|}{|B|}$$

Lastly, we know:

$$P_{card}(A|B) = \frac{|A \cap B| \,/\, |S|}{|B| / |S|} = \frac{P_{card}(A \cap B)}{P_{card}(B)}$$

which is our previous definition of conditional probability.

Note:

The meaning of conditional probability $A|B$ is similar to that of a normal probability: the number of event satisfying A, when there is a large number of experiments done but you **only look at events that satisfied** B.

***Example*: Conditional Prob**

Consider two cards drawn at random, one at a time, without replacement. Let:

$$\begin{cases} A = \text{first card being a heart} \\ B = \text{second card being a heart} \end{cases}$$

Question: What is the $P(B|A)$?

Solution: By definition, we know that:

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

We know that the simple stuff:

$$P(A) = \frac{1}{4}$$

But now, we need to care a little bit for $P(A \cap B)$. First, each outcome here is a 2-tuple, therefore, the sample space here is 52×51. Then, if we want to choose two hearts, then 13×12 ways. Hence:

$$\frac{P(A \cap B)}{P(A)} = \frac{13 \times 12/52 \times 51}{1/4} = \frac{12}{51}$$

Alternatively, this can be actually computed directly:

1. Given that A happened, the sample space is now 51 cards, since you have picked a (heart) card
2. Then, to get B happen, you can only choose 12 ways, since there are only 12 hearts left.
3. Therefore, we get 12/51

***Example*: Baby Birth**

A family has two kids, the elder one a girl. What is the chance of the second kid being a girl?

Solution: Given that the elder one being a girl, we have the following **new sample space**:

$$\mathcal{S}_B = \{(g, g), (b, g)\}$$

where the elder one is in the second position. Therefore, the probability is:

$$P(A|B) = \frac{1}{2}$$

4.1 Rules for Conditional Probability

There are two useful rules for calculating conditional probabilities.

Theorem 4.1 (Bayes Rule)**.** *If A and B are two events, and $P(A) > 0$ and $P(B) > 0$, then:*

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

Proof for Bayes Rule. We know that:

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

but we also know that:

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

Therefore, substituting the above into the first equation, we get that:

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

□

Before the next theorem, we need to know what is means to be a **partition of sample space**. This means that:

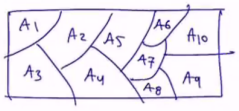


Figure 3: Partition of Sample Space

Definition 4.2 (Paritition of Sample Space)**.** *The event $A_1, A_2, ..., A_n$ are called a partition of sample space if:*

- $A_1 \cup A_2 ... \cup A_n = \mathcal{S}$
- $A_i \cap A_j = \emptyset, \forall i \neq j$

Theorem 4.2 (Law of Total Probability)**.** *Let $A_1, A_2, ..., A_n$ be partition of the sample space $\mathcal{S}$ with $P(A_i) > 0, \forall i$. Then, if we consider another event B:*

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Again the proof is straight forward. Graphically:

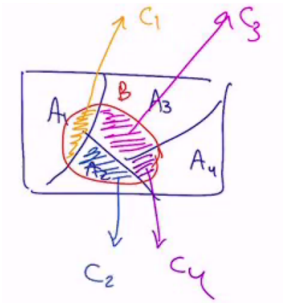


Figure 4: Example of Law of Total Probability

Proof. Define $C_i = A_i \cap B$. We notice in the diagram above, C_i are all disjoint, because A_i are all disjoint by definition. Since we know that:

$$C_1 \cup C_2 \cup C_3 ... \cup C_n = B$$

Therefore, we get:

$$\begin{aligned} P(B) &= P(C_1 \cup C_2 ... \cup C_n) \\ &= P(C_1) + P(C_2) + ... + P(C_n) \\ &= P(A_1 \cap B) + P(A_2 \cap B) + ... + P(A_n \cap B) \\ &= P(B|A_1)P(A) + ... + P(B|A_n)P(A_n) \\ &= \sum_{i=1}^{n} P(B|A_i)P(A_i) \end{aligned}$$

where the second equality comes from C_i being disjoint from each other. □

***Example*: Rare Disease Paradox**

Suppose only one person out of 1000 is affected by a disease A. Now, suppose we have a very good test for this. This means that:

$$\begin{cases} P(\text{Test Positive} \mid \text{Not have A}) = 0.01 \\ P(\text{Test Negative} \mid \text{Have A}) = 0.01 \end{cases}$$

which is basically False Positive and False Negative.

Question: Suppose a person is tested positive. What is the probability that the person actually has the disease?

Solution: Let the even of having the disease being E, and let the event of being tested positive P. Then we need to compute $\mathbb{P}(E|P)$. Using Bayes Rule:

$$\mathbb{P}(E|P) = \frac{\mathbb{P}(P|E)\mathbb{P}(E)}{\mathbb{P}(P)}$$

we knew already that:

- $\mathbb{P}(E) = 0.001$ of actually have the disease
- $\mathbb{P}(P|E) = 1 - 0.01 = 0.99$
- $\mathbb{P}(P)$ we don't know yet

But notice that we know:

$$\mathbb{P}(P) = \mathbb{P}(P|E)\mathbb{P}(E) + \mathbb{P}(P|E^c)P(E^c)$$

which is the law of probability. This is now computable, being:

$$\mathbb{P}(P) = 0.99 \cdot 0.001 + 0.01 \cdot 0.999 \approx 0.01098$$

Therefore, our answer is:

$$\mathbb{P}(E|P) = \frac{0.99 \cdot 0.001}{0.01098} \approx 0.09$$

Therefore, even if the test is very accurate, the chance of you actually having the disease is still quite low.

If you think about this, suppose we have 10000 people. Then we know that:

- there will be 10 people who actually has diseases
- in total 100 people will be tested positive

This means that you will make **at least 90 errors!**

***Example*: Simpson's Paradox**

Suppose a person wants to remove the kidney's stone in your body. You know two doctors, who are different at skills with large/small kidney stones.

Within 100 operations, the success rates look like

$$\begin{cases} \text{Doctor A: } 65/85 \text{ Large}, & 15/15 \text{ Small} \\ \text{Doctor B: } 2/10 \text{ Large}, & 80/90 \text{ Small} \end{cases}$$

note that Doctor B has 82/100 for success overall (wrong calculation anyway), though it is not representative of him being a better doctor. This is because the probability of having a large stone or a small stone is missing.

The upshot here is that, often the result is the overall probability. This means that doctors will **reject hard operations** so that their rating will be high.

4.2 Independence

Definition 4.3 (Independence of Two Events)**.** *Two evens A, B are represented with $P(A), P(B)$. We say that the two events are independent if and only if:*

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

Lemma 4.3. *If A, B are independent, then:*

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

So even given B has happened, the probability of A happening does not change.

Proof. Basically, the idea is that we know the two events are independent. So:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A)P(B)}{P(B)} \\ &= P(A) \end{aligned}$$

□

***Example*: Presidential Election**

Let the event A being Biden winning, and event B be very heavy rain on the day of election in NYC.

Then to consider if they are independent, consider:

$$P(A|B), \quad P(A)$$

we can see that would be different, since if it is raining, then people in NYC might not be able to vote, then Biden's win in NY will be affected.

There are also other examples. In reality, two events could arguably never be **completely independent**, but often they are more of less independent. In those cases, we just assume that for simplicity they are independent.

Definition 4.4 (Multiple Independence)**.** *Suppose we have three events, A, B, C. They are independent if an only if:*

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

AND that:

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

***Example*: Three Events**

Consider three events, A=Me getting COVID, B raining in NYC, and C=I bring my umbrella to office.

Are those events independent?

Solution: A and B is arguably independent. A and C are also arguably independent. But are B, C independent? We needed:

$$P(B \cap C) \stackrel{?}{=} P(B)P(C)$$

This is NOT true. I am more likely to bring umbrella to the office if it is raining.

So in general, if we have $A_1, A_2, ..., A_n$ are independent, then **all** the conditions below needs to be satisfied:

$$\begin{cases} P(A_i \cap A_j) = P(A_i)P(A_j), & \forall i, j \\ P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k), & \forall i, j, k \\ \vdots \\ P(A_i \cap A_j \cap ... \cap A_n) = P(A_i)P(A_j)...P(A_n) \end{cases}$$

Lemma 4.4. *Let A, B, C are independent. Then:*

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

Proof. Since we know A, B, C are independent, then

$$\begin{aligned} P(A|B \cap C) &= \frac{P(A \cap B \cap C)}{P(B \cap C)} \\ &= \frac{P(A)P(B)P(C)}{P(B)P(C)} \\ &= P(A) \end{aligned}$$

□

Now, consider a case of A, B are independent given C has happened. For example, A=me getting COVID, B=someone in the class getting COVID. C=we have mask mandate. Then A, B could be independent if we assume that wearing a mask can stop spreading COVID to people in class.

Definition 4.5 (Conditional Independence)**.** *A, B are independent given C if an only if:*

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

Alternatively, this is equivalent to:

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

proof can be found on Wikipedia on Conditional Independence.

***Example*: Trial of Sally Clark**

She has a baby, who died 11 weeks after birth (sad) due to some unknown reasons. Then, she has a second baby, who died 8 weeks after birth. Again, no idea why. So something must went awry. One guy believed Sally Clark killed the baby.

We know that the chance of one baby with sudden death is 1/8500. One guy said then:

$$P(\text{Both Died}) = \frac{1}{8500} \times \frac{1}{8500} = \frac{1}{73000000}$$

Is this correct?

Solution: Let the first baby died suddenly be event A. Let the second death be event B. Then the guy is claiming that A, B are **independent**. This would very likely **not be true**, because this could be some genetic issues. Then given that A has happened, the chance of Sally has some genetic issue is high, so then B is also likely to happen.

Additionally, let C=Sally being innocent. The guy is calculating:

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

But what we **should calculate** is:

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

instead. So even if $P(A \cap B)$ is rare, but $P(C|A \cap B)$ could be large.

5 Discrete Random Variables

Before, we have a sample space $\mathcal{S}$, and we were thinking about a **particular experiment** A and compute $P(A)$. But we want to be more generic for any A.

Consider coin tosses, where you get a head or a tail. Then a random variable X could be used to represent either getting a head $X = 1$ or a tail $X = 0$. In essence, we are representing **different events with different numbers**, and those numbers all has some associated probability of happening.

Example

Suppose we are tossing the coin n times, and here we consider $n = 1000$. Then suppose we let X be the **number of heads** we got. Then, something like:

$$\begin{cases} P(X = 0) = P(\text{getting 0 heads}) \\ P(X = 1) = P(\text{getting 1 heads}) \\ \dots \end{cases}$$

so basically we are **mapping $X \in \mathbb{N} \to$ some events**. Sin the random variable X is basically mapping a number to some events.

Definition 5.1 (Random Variable)**.** *A random variable is a function from the sample space to numbers.*

$$X : \mathcal{S} \to \mathbb{R}$$

*Often, random variables are shown in **capital letters**, using symbols like X, Y, W, etc.*

***Example*: Random Variable and Coin Toss**

Let us have a coin which is $P(\{H\}) = p$, and $P(\{T\}) = 1 - p$. And we define the random variable to be:

$$\begin{cases} X(H) \to 1 \\ X(T) \to 0 \end{cases}$$

where X is the random variable. Then obviously:

$$\begin{cases} P(X = 1) = p \\ P(X = 0) = 1 - p \end{cases}$$

and $P(0 \leq X \leq 1) = 1$ in this case.

This form of random variable is also called a Bernoulli random variable $X \sim \text{Bernoulli}(p)$.

The key take away of a random variable (e.g. X) is that X =? is random.

***Example*: Binomial Random Variable**

Consider tossing a coin n times, then we basically have the sample space being:

$$\mathcal{S} = \{(H, H, \dots, H), (H, H, ..., H, T), ..., (T, T, ..., T)\}$$

Then we can define random variable Y to map those events:

$$\begin{cases} Y\left(\{(H, H, ..., H)\}\right) = n \\ Y\left(\{(H, H, ..., T), (H, H, ..., T, H), ...\}\right) = n - 1 \\ ... \\ Y\left(\{(T, T, ..., T)\}\right) = 0 \end{cases}$$

Therefore, $Y : \mathcal{S} \to \{0, 1, 2, ...n\}$ represents how many heads you get in n tosses $\in \mathcal{S}$. Notice now that each random variable has:

$$Y = k \to {}^nC_k \text{ events}$$

And notice that:

$$\sum_{k=0}^{n} {}^nC_k = 2^n$$

which was proven before using various different methods. Therefore, we know that, using Cardano's definition:

$$\mathbb{P}(Y = 0) = \frac{1}{2^n}$$

but we can do better using independence:

$$\begin{aligned} \mathbb{P}(Y = 0) &= \mathbb{P}(\text{first coin toss} = \text{T})\mathbb{P}(\text{second coin toss} = \text{T}) \\ &\quad ...\mathbb{P}(\text{first coin toss} = \text{T}) \\ &= \left(\frac{1}{2}\right)^n \end{aligned}$$

Now consider $\mathbb{P}(Y = 1)$. Then, again being more careful (not using Cardano's probability), we have:

$$\begin{aligned} \mathbb{P}(Y = 1) &= P\left((H, T, ..., T) \cup (T, H, T, ..., T)...\right) \\ &= \mathbb{P}\left((H, T, ..., T)\right) + \mathbb{P}\left((T, H, T, ..., T) + ...\right. \\ &= n\left(\frac{1}{2}\right)^n \end{aligned}$$

using the same **formal approach**, we get:

$$\mathbb{P}(Y = k) = {}^nC_k \left(\frac{1}{2}\right)^n$$

and this would be a **binomial random variable**, such that we can even say if $\mathbb{P}(\{H\}) = p$ is biased, then:

$$\mathbb{P}(Y = k) = {}^{n}C_k \cdot p^k (1-p)^{n-k} \iff Y \sim \text{Binomial}(n, p)$$

5.1 Probability Mass Function

Definition 5.2 (Probability Mass Function)**.** *A probability mass function (pmf) of a discrete random variable $X : \mathcal{S} \to O$ is deifned as the function that maps every $a \in O$ to $P(X = a)$.*

$$\mathbb{P}_X(a) = \mathbb{P}(X = a)$$

where the subscript is useful when we have many random variables.

In other words, we are mapping every event to a probability. Basically PMF tells you probability of any event.

A quick example would be to consider binomial random variable, such that we consider $X : \mathcal{S} \to \{0, 1, ..., n\} = O$, and $a \in \{0, 1, ..., n\}$ such that the probability mass function would be:

$$\mathbb{P}_X(a) = \mathbb{P}(X = a) = {}^{n}C_a \cdot p^a (1-p)^{n-a}$$

is basically a function that maps any a to a probability. (Note that if you see $X \sim \text{Bin}(p)$, it means $n = 1$ is assumed.) Graphically:

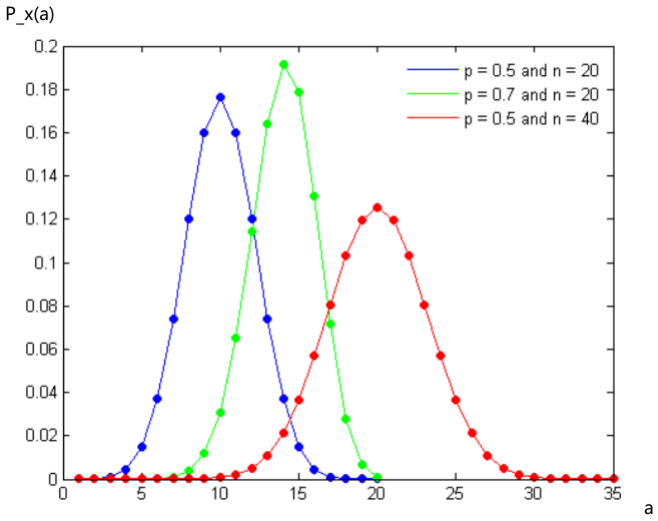


Figure 5: Binomial $X \sim \text{Bin}(n, p)$ Probability Mass Function

For Bernoulli random variable X, we would have:

$$\begin{cases} \mathbb{P}_X(0) = 1 - p \\ \mathbb{P}_X(1) = p \end{cases}$$

being the probability mass function. Moreover, we could plot this $\mathbb{P}_X(a)$ in Figure 6.

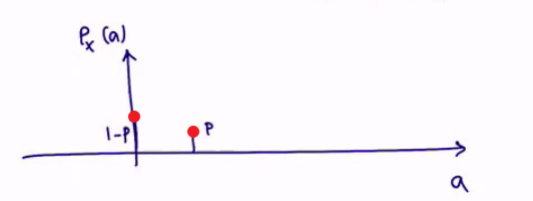


Figure 6: Bernoulli $X \sim \text{Bern}(p)$ Probability Mass Function

***Example*: Geometric Random Variable**

This is another common random variable in application. Consider that we are in downtown NYC. Each year, the probability of flooding is p. Let X being the random variable for the number of years before we see the first flood (from now).

Then, for instance, $P(X = 1)$ means the probability that your house is flooded after one year. Therefore:

$$\begin{cases} P_X(1) = P(X = 1) = p \\ P_X(2) = P(X = 2) = (1-p)p \end{cases}$$

where here we **assumed independence**

$$\begin{aligned} P_X(2) &= P(X = 2) \\ &= P(\text{no flood first year} \cap \text{yes flood second year}) \\ &= P(\text{no flood first year})P(\text{yes flood second year}) \\ &= (1-p) \cdot p \end{aligned}$$

Then, we can continue defining the PMF easily:

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

Graphically, the geometric random variable looks like Figure 7.

5.2 Cumulative Distribution Function

Definition 5.3 (CDF)**.** *The cumulative distribution function (CDF) of a random variable $X : \mathcal{S} \to O$ is defined as the function $F_x(a)$:*

$$F_x(a) = \mathbb{P}(X \leq a) = \sum_{i \in O : i \leq a} P_X(i)$$

for a being any arbitrary real number, and assuming we have a discrete random variable.

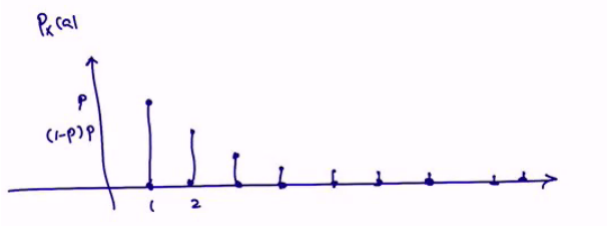


Figure 7: Geometric Random Variable PMF

A quick example would again be the Bernoulli random variable $X \sim \text{Bern}(p)$. Then the comparison of PMF and CDF looks like Figure 8.

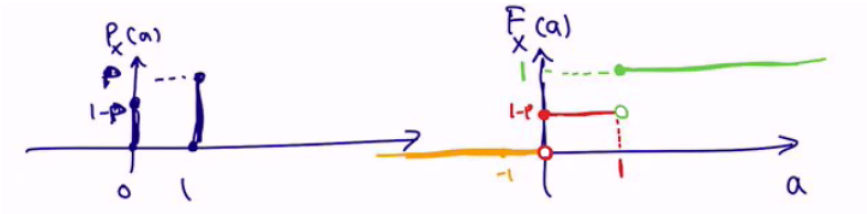


Figure 8: Bernoulli PMF on the left vs Bernoulli CDF on the right

notice that:

- there could be **jump discontinuity** in CDF, and it happens when the random variable being the jump discontinuity. (In fact, CDF should be **right continuous**.)
- every time when we have a PMF $P_X(a)$, we can find out its CDF $F_X(a)$.
- every time when we have a CDF $F_X(a)$, we can find out its PDF $P_X(a)$. For example, to get $P_X(1)$ from CDF:

$$\begin{aligned} F_X(1) &= 1 \\ P(X=1) + P(X<1) &= 1 \\ P(X=1) + F_X(0) &= 1 \\ P(X=1) + (1-p) &= 1 \\ P(X=1) &= p \end{aligned}$$

  where technically we could have used $P(X<1) = F_X(0) = F_X(0.1) = F_X(0.99) = ...$ in this example. Formally:

$$P_X(1) = F_X(1) - \lim_{a \to 1_-} F_X(a)$$

- for valid CDF:

$$\lim_{x \to -\infty} F(X=x) = 0, \quad \lim_{x \to \infty} F(X=x) = 1$$

- for **discrete random variable**, CDF gives the same amount of information as PMF. So mathematically, they are redundant.

5.3 Mean and Variance

Consider that a given distribution looks like:

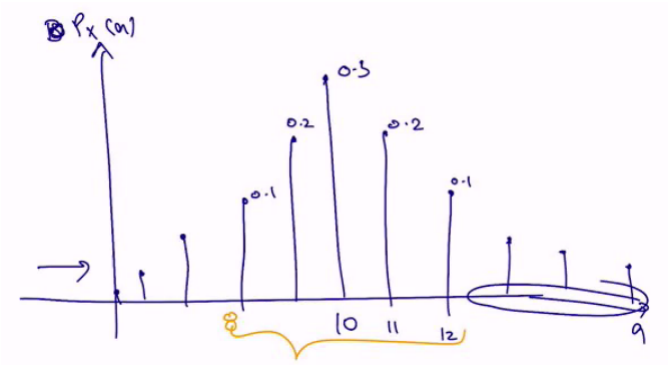


Figure 9: Center of Mass of Some Distribution

where what is important is now the **center of mass** of the distribution (in yellow), which can be computed **without knowing actual distribution function** of that random variable.

- this is useful when you don't necessarily know what distribution function your random variable is. In reality, this is always the case
- The upshot is that usually you need to **figure out the probability mass function** given some data.

Then, some useful quantity you might want to find out the center of mass is:

1. **mean** of the distribution ("center" of center of mass)
2. **standard deviation**/variance of the distribution ("width" of center of mass)

5.3.1 Mean of a Random Variable

Consider four numbers, $1, 3, 3, 5, 5, 5, 7, 7, 9$. We can compute the average which is:

$$\frac{1 + 2 \cdot 3 + 3 \cdot 5 + 2 \cdot 7 + 9}{9} = 5$$

Notice that we used "frequency $\cdot$ values ". This is essentially the same concept in random variable. Consider now a random variable $X : \mathcal{S} \to \{1, 2, 3, 4, 5, 6\}$, which is a biased die. And you are given:

$$\begin{cases} P_x(1) = P_X(6) = 1/12 \\ P_x(2) = P_X(5) = 2/12 \\ P_x(3) = P_X(4) = 3/12 \end{cases}$$

Then, I can **throw the die many times**, and we might get for X_ith experiment:

$$X_1 = 1, X_2 = 5, ..., X_{1000} = 2$$

Then, we can take the average of those numbers to be:

$$\text{Average} = \frac{\sum_{i=1}^{1000} X_i}{1000} = \frac{\#\text{ of 1s} \cdot 1 + ... + \#\text{ of 6s} \cdot 6}{1000}$$

But from the weak law of large numbers, we also know that:

$$\begin{cases} \#\text{ of 1s} \approx P_X(1) \cdot 1000 \\ \#\text{ of 2s} \approx P_X(2) \cdot 1000 \\ ... \\ \#\text{ of 3s} \approx P_X(3) \cdot 1000 \end{cases}$$

Therefore, you get:

$$\begin{aligned} \text{Average} &\approx \frac{1}{12} \cdot 1 + \frac{2}{12} \cdot 2 + ... + \frac{1}{12} \cdot 6 \\ &= \sum_{x=1}^{6} P(X = x) \cdot x \end{aligned}$$

This is essentially what it means to have a mean of a random variable.

Definition 5.4 (Mean of Random Variable)**.** *Let X denote a* ***discrete random variable*** *with PMF of $P_X(a)$. Then, the mean of this random variable is defined as:*

$$\mathbb{E}(X) = \sum_{a \in O} a \cdot P_X(a)$$

for $X : \mathcal{S} \to O$. This $\mathbb{E}$ is also called the ***expected value****.*

From the above example, it is clear that mean/expectation of a random variable is useful since:

- if you do the experiment for a large amount of times, the average of what you get is the mean (**weak law of large numbers**).
- this is very closely related to the center of mass of a PMF (see later).

***Example*: Mean of Bernoulli Random Variable**

Consider a Bernoulli random variable, with $X \sim \text{Bern}(p)$. Then the mean/expected value of X is:

$$\mathbb{E}(X) = \sum_{a=\{0,1\}} a \cdot P_X(a) = 0 + p = p$$

Other easy examples are on the lecture notes.

***Example*: Mean of Binomrial Random Variable**

Now, we have $X \sim \text{Bin}(n, p)$. Then, the mean is:

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{a=0}^{n} a \cdot P_X(a) \\
&= \sum_{a=0}^{n} a \binom{n}{a} p^a (1-p)^{n-a} \\
&= \sum_{a=1}^{n} a \binom{n}{a} p^a (1-p)^{n-a} \\
&= \sum_{a=1}^{n} a \frac{n!}{a!(n-a)!} p^a (1-p)^{n-a} \\
&= \sum_{a=1}^{n} \frac{n!}{(a-1)!(n-a)!} p^a (1-p)^{n-a} \\
&= np \sum_{a=1}^{n} \frac{(n-1)!}{(a-1)!(n-a)!} p^{a-1} (1-p)^{n-a} \\
&= np \sum_{b=0}^{n-1} \frac{(n-1)!}{b!(n-1-b)!} p^b (1-p)^{n-1-b} \\
&= np \sum_{b=0}^{n-1} \binom{n-1}{b} p^b (1-p)^{n-1-b} \\
&= np \sum_{b=0}^{n-1} P_Y(b), \quad Y \sim \text{Bin}(n-1, p) \\
&= np
\end{aligned}$$

where the third equality comes from the fact that $a = 0$ makes that term 0; the seventh equality used $b \equiv a - 1$; the second last equality uses the fact that this is $Y \sim \text{Bin}(n-1, p)$; the last equality comes from the law of total probability.

Yet one "problem" with expected value is your estimation of mean using $P_X(a)$ went wrong. For instance, an insurance company might expect average cost of a college student to be $1,500$. However, this is assuming things are "normal". What if there is a pandemic happened? Then this calculation would not work if you didn't take that into account.

5.3.2 LOTUS

Before going to the definition of variance, we need to know the Law of Unconscious Statistician (LOTUS).

Consider X being the square foot of a random apartment in the city, and assume that the price of the apartment **only dependents on** the size of the apartment. Let us now denote the price to be $Y = g(X)$.

Now, the question is, **if we knoin the function** g:

- can we calculate PMF of Y if we have the PMF of X?
- can we calculate $\mathbb{E}(Y)$ given the information that we have about X?

The answer to both is yes. We will first see some examples.

***Example*: Calculate PMF of Y**

Consider the case that $X : \mathcal{S} \to \{-3, -2, -1, 1, 2, 3\}$. And suppose that the distribution is uniform, so that:

$$P_X(a) = \frac{1}{6}, \quad \forall a \in \{-3 - 2 - 1, 1, 2, 3\}$$

So we are given the PMF of X. From here, we can calculate:

$$\mathbb{E}(X) = \sum_{a=-3}^{+3} aP_X(a) = 0$$

Now, suppose the new random variable Y is defined to be $Y = X^2$. Then, this means that $Y : \mathcal{S} \to \{1, 4, 9\}$. Then we know that:

$$\begin{aligned} P_y(1) = \mathbb{P}(Y = 1) &= \mathbb{P}(X = -1 \text{ or } X = 1) \\ &= \frac{1}{6} + \frac{1}{6} \\ &= \frac{1}{3} \end{aligned}$$

The same story goes for $P_y(4)$ and $P_y(9)$.

Therefore, we figured out the PMF of Y, which means we can figure out the expected value as well. So, we have figured out everything about Y using information of X and $g(X) = X^2$.

Graphically, what happened is:

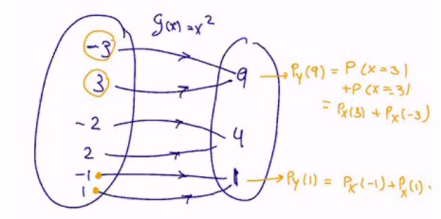


Figure 10: Example of Calculating PMF of Y

In a more general setting, the function $g(X)$ will be mapping random variables in X to another space:

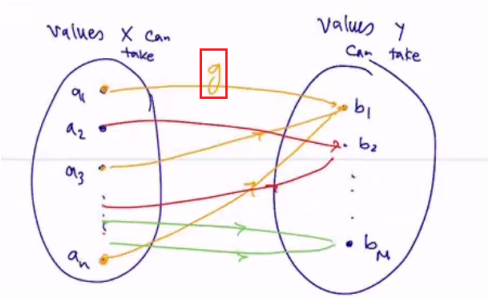


Figure 11: General Case of Calculating PMF of Y

where in Figure 11, for instance, we know that $P_Y(b_1) = P_X(a_1) + P_X(a_2) + P_X(a_n)$. More generally:

$$P_Y(b) = \sum_{a:g(a)=b} P_X(a)$$

so basically we are adding up all the $P_X(a)$ for all a that are mapped to the specific b. Note that there are certain restrictions on $g(X)$:

1. $g(X)$ must have mapped **every event in** $X : \mathcal{S} \to O$. (Otherwise PMF of Y won't add up to 1)
2. $g(X)$ is a many to one function.
3. $g(X)$ is cannot be a one to many function.

Hence, we can also calculate the expected value easily:

$$\mathbb{E}(Y) = \sum_{a \in Q} aP_Y(a)$$

but it turns out there is a simpler trick.

Lemma 5.1 (LOTUS)**.** *Let $Y = g(X)$. Then, we have:*

$$\mathbb{E}(Y) = \sum_{a \in O} g(a)P_X(a)$$

note that LOTUS stands for law of the unconscious statistician, i.e. this definition is memorized by almost all statistician. This is most always used.

Proof. We know that:

$$\begin{aligned}
\mathbb{E}(Y) &= \sum_{b\in Q} bP_Y(b) \\
&= \sum_{b\in Q} b \sum_{a:g(a)=b} P_X(a) \\
&= \sum_{b\in Q} \sum_{a:g(a)=b} bP_X(a) \\
&= \sum_{b\in Q} \sum_{a:g(a)=b} g(a)P_X(a) \\
&= \sum_{a\in O} g(a)P_X(a)
\end{aligned}$$

where the last equality comes from the fact that eventually all a will be iterated/summed since $g(X)$ is a many-to-one function for all the input space of X. □

Now, we talk about how to think about variance. Basically, think of variance as a measure of **how close/concentrated is my data towards the mean**. Therefore, suppose we have:

$$\begin{cases} \{x_1, x_2, x_3, ..., x_{10000}\}, & \text{my data} \\ \frac{1}{10000}\sum_i X_i \approx \mathbb{E}(X) = \mu \end{cases}$$

To measure how far away, we consider:

$$\{x_1 - \mu, x_2 - \mu, ..., x_{10000} - \mu\}$$

Now, we can either compute the mean directly form it, for we square it. In fact:

- if we take mean directly:

  $$\frac{1}{10000}\sum_i (x_i - \mu) = \frac{1}{10000}\left(\sum_i x_i - \sum_i \mu\right) = \mu - \mu = 0$$

  which is useless

- if we take square, it then works because:

  $$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \approx \mathbb{E}(Y_i) \equiv \text{Var}(X)$$

  where we defined the new random variable $Y \equiv X - \mu$. Therefore, we then get that:

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

5.3.3 Variance

Now, we can look at variance.

Definition 5.5 (Variance and Standard Deviation)**.** *Therefore, variance for a discrete random variable X is:*

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

Additionally, the standard deviation is defined as:

$$\sigma(X) = \sqrt{Var[X]}$$

Proof. We know that:

$$\begin{aligned}
\text{Var}[X] &= \mathbb{E}[(X-\mu)^2] \\
&= \sum_{a \in O} (a-\mu)^2 P_X(a) \\
&= \sum_{a \in O} (a^2 - 2a\mu + \mu^2) P_X(a) \\
&= \sum_{a \in O} a^2 P_X(a) + \mu^2 \sum_{a \in O} P_X(a) - 2 \sum_{a \in O} a\mu P_X(a) \\
&= \mathbb{E}[X^2] + \mu^2 - 2\mu \sum_{a \in O} a P_X(a) \\
&= \mathbb{E}[X^2] + \mu^2 - 2\mu \cdot \mu \\
&= \mathbb{E}[X^2] - \mu^2
\end{aligned}$$

which is want we had in the definition. □

***Example*: Variance of Bernoulli Distribution**

Consider $X \sim \text{Bern}(p)$. Then, we want to compute the variance.

Solution: Since we know that:

$$\text{Var}[X] = \mathbb{E}[X^2] - \mu^2$$

We know μ which is:

$$\mathbb{E}[X] = \mu = \sum_{a \in O} a P_X(a) = 0 + 1 \cdot p = p$$

Similarly:

$$\mathbb{E}[X^2] = \sum_{a \in O} a^2 P_X(a) = 0 + 1^2 \cdot p = p$$

Therefore, variance is:

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

Intuitively, variance for Bernoulli makes sense because if you have $p = 1$ or $p = 0$, then variance would be 0.

For another example, consider rolling dice.

***Example*: Variance of Rolling Dice**

For a dice, let X denote the outcome from 1 to 6. And let the probability be uniform. What is the variance?

Solution: Again use the same formula:

$$\text{Var}[X] = \mathbb{E}[X^2] - \mu^2$$

The mean is trivially $\mu = 3.5$. The other quantity needs to be computed:

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{a=1}^{6} a^2 P_X(a) \\ &= 1 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + ... + 36 \cdot \frac{1}{6} \\ &= \frac{91}{6}\end{aligned}$$

Hence, the variance is:

$$\text{Var}[X] = \frac{91}{6} - (\frac{7}{2})^2 = \frac{35}{12} \approx 3$$

which is basically the width of the entire data distribution.

Graphically, again standard variance (sqrt of variance) is related to the spread of the **center of mass** region.

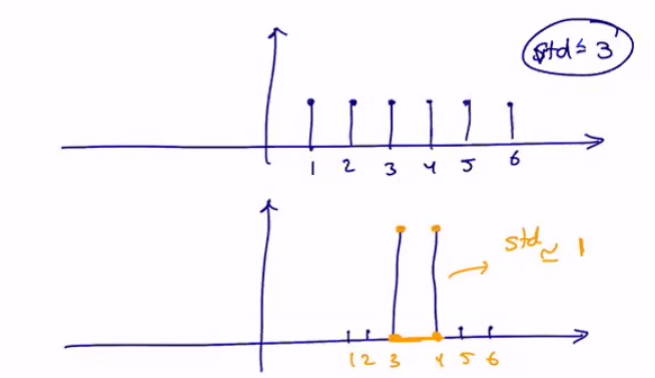


Figure 12: Dice Roll Standard Deviation

6 Continuous Random Variables

First let us begin by some examples. Some simple examples will continues random variable would be helpful would be:

- generating random numbers in $\mathbb{R}$
- voltage that your cell phone antenna receives
- etc.

Note that you technically **can discretize** them, but it is just easier to use continuous random variables in this case.

Discrete	Continuous	Comment
$P_X(a) = P(X = a)$	$P_X(a) = 0$	Because the outcome is in real, which has infinite possibility
$F_X(a) = P(X \leq a)$	$F_X(a) = P(X \leq a)$	Is meaningful
$\sum_{a \in O} P_X(a) = 1$	$\int_{-\infty}^{\infty} f_X(a)\, da = 1$	Law of Total Probability
$E[X] = \sum_{a \in O} aP_X(a)$	$E[X] = \int_{-\infty}^{\infty} a \cdot f_X(a)\, da$	Expected Value/Mean
$\mathbb{E}[Y] = \sum_{a \in O} g(a)P_X(a)$	$\mathbb{E}[Y] = \int_{-\infty}^{\infty} g(a) f_X(a)\, da$	LOTUS for $Y = g(x)$

where it is important to notice that, for continuous variable X:

- $P_X(a) = 0$. Consider estimating the train arrival. Then $P(X = 10.001010) = 0$ since the train could have arrived in on more millisecond, one more microsecond, ten more nanosecond, and etc. There is an infinite possibility for that.
- the CDF for continues variable is actually meaningful. Therefore, we will start with this for continuous variable.
- Basically the analogy of PMF in continuous variable is the PDF $f_X(a)$, but that their meaning is quite different.

Last but not least, we will only consider that $F_X(a)$ is defined for $a \in \mathbb{R}$ for the entire real space.

6.1 CDF for Continuous RV

In fact, since for continues variable X, $P_X(a) = 0$ for any a, so there is **no jump in the CDF**.

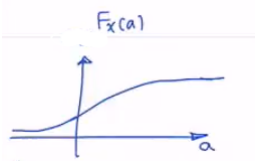


Figure 13: CDF of a Continuous Random Variable

Lemma 6.1. *Let $F_X(a)$ denote the CDF of a continuous random variable. Then, the following holds true:*

- $F_X(a)$ *is a* ***non-decreasing*** *function*
- $\lim_{a \to -\infty} F_X(a) = 0$
- $\lim_{a \to +\infty} F_X(a) = 1$
- *if* $a \leq b$, *then* $P(X \leq a) \leq P(X \leq b)$

For example, consider estimating the time that you have to wait for the train to arrive.

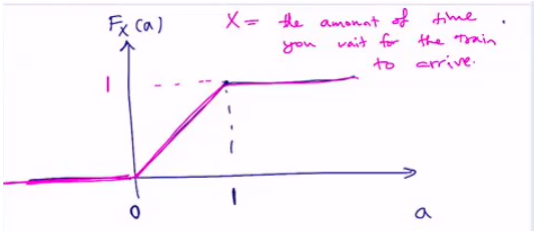


Figure 14: Waiting for Train Arrival Example

note that this is a valid CDF, because

- all three property holds
- It does **not have to be smooth**. You just need no jump discontinuity.

***Example*: Train Arrivial**

Based on the graph above, we can then compute several quantities such as:

- $P(X \leq 0.5) = F_X(0.5) = 0.5$ reading from graph
- what about $P(0.25 < X \leq 0.5)$?
- what about $P(X \geq 0.5)$?

Solution: We can compute using the fact that $P_X(a) = 0$, so that:

$$P(0.25 < X \leq 0.5) = P(0.25 \leq X \leq 0.5)$$

Therefore:

$$\begin{aligned} P(0.25 \leq X \leq 0.5) &= F_X(0.5) - F_X(0.25) \\ &= 0.5 - 0.25 = 0.25 \end{aligned}$$

And the second question:

$$\begin{aligned} P(X \geq 0.5) &= P(X > 0.5) \\ &= 1 - F_X(0.5) \\ &= 1 - 0.5 = 0.5 \end{aligned}$$

6.2 Probability Density Function

Basically the concept that we **can use** instead of probability mass function, but not **exactly equivalent**.

Definition 6.1 (PDF)**.** *Let X denote a continues variable with CDF of $F_X(a)$. Then the PDF of X is defined as:*

$$f_X(a) \equiv \frac{dF_X(a)}{da}$$

Before going into details, let us consider some examples

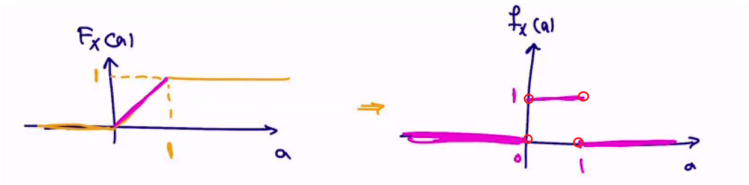


Figure 15: Example PDF of X Given CDF

notice that:

- technically derivative at $a = 0, a = 1$ is **undefined**. However, that does not matter because PDF of a point is always meaningless
- therefore, sometimes you can just assign this the average of the jump discontinuity

Lemma 6.2 (Property of PDF)**.** *The following properties of PDF $f_X(a)$ holds true:*

- $f_X(a) \geq 0, \forall a$*, since CDF is non-decreasing.*
- $\int_{-\infty}^{\infty} f_X(y)\, dy = 1$*. This is derived form the fundamental theorem of calculus:*

$$\int_a^b f_X(y)\, dy = F_X(b) - F_X(a)$$
$$\int_{-\infty}^{\infty} f_X(y)\, dy = \lim_{b\to\infty} F_X(b) - \lim_{a\to-\infty} F_X(a) = 1 - 0 = 1$$

Now, we can talk about how to **understand** what PDF is.

Consider the probability of X in a small neighborhood of c:

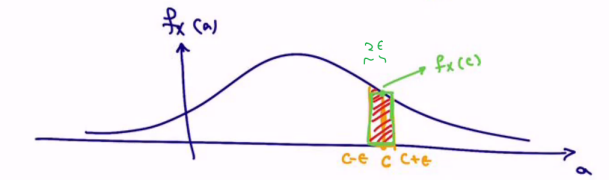


Figure 16: Understanding PDF of X

Then we know that:

$$\begin{aligned} P(c-\epsilon \leq X \leq c+\epsilon) &= \int_{c-\epsilon}^{c+\epsilon} f_X(y)\,dy \\ &\approx 2\epsilon \cdot f_X(c) \end{aligned}$$

which means that fundamentally it is probability = volume $\times$ probability density. Also, since volumes of the neighborhood is always positive, and probability is always positive, so PDF is always positive.

6.2.1 Uniform Random Variable

Now, reconsider the following distribution

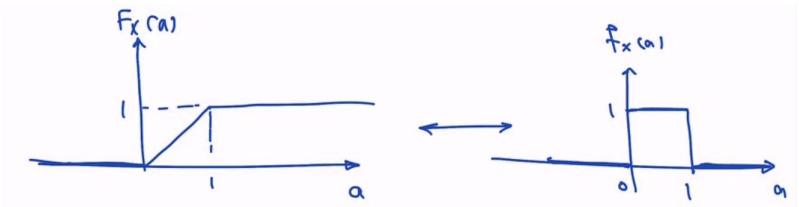


Figure 17: Uniform Distribution $X \sim$ Unif(0,1)

***Example*: Uniform RVs**

Consider $X \sim \text{Unif}(0, 1)$ in the figure above. Then we can compute that:

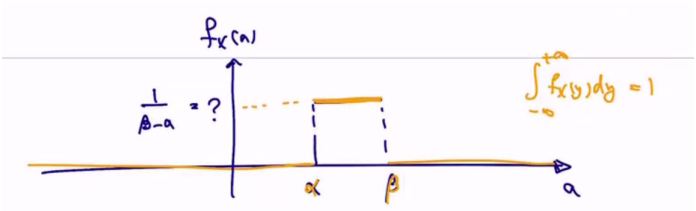
$$P(0.5 \leq X \leq 0.75) = \int_{0.5}^{0.75} 1 \, dy = 0.25$$

And a bit more interestingly:

$$\begin{aligned} P(X \geq 0.75) &= \int_{0.75}^{\infty} f_X(y) \, dy \\ &= \int_{0.75}^{1} f_X(y) \, dy + \int_{1}^{\infty} f_X(y) \, dy \\ &= 0.25 + 0 = 0.25 \end{aligned}$$

This is useful when the chance of an event happening between some region is **equally likely to happen**. (e.g. your pizza could come equally likely in 1min, 2min, 2.3min, etc.)

Definition 6.2 (Uniform PDF)**.** *The PDF of uniform distribution is defined generally as* $X \sim$ *Unif*(α, β) *looks like:*



obviously, this is because we need to make sure that $\int_{-\infty}^{\infty} f_X(y) \, dy = 1$.

6.2.2 Gaussian(Normal) Random Variable

This is by far the most often used continuous distribution in real life.

Definition 6.3 (Normal PDF)**.** *Alternatively, this is also called the Gaussian Variable. This distribution is* $X \sim N(\mu, \sigma^2)$*:*

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

where mean/maximum of the PDF happens at $a = \mu$*, and the width is defined by* σ^2*. Graphically:*

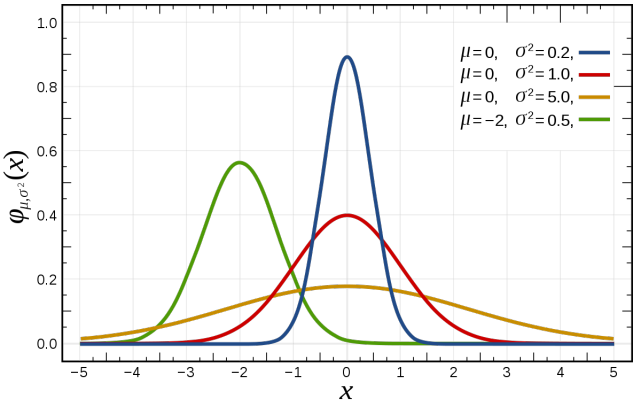
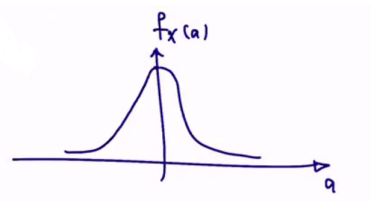


Figure 18: General Gaussian Variable

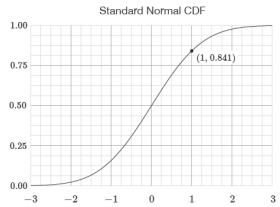
There is also a **standard Gaussian distribution**, which is defined as:

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

this is the **standard version** because $\mu = 0, \sigma^2 = 1$. This graphically will look like



(a) PDF



(b) CDF

Figure 19: Standard Gaussian Distribution

Note:

The CDF of the Gaussian doesn't have an analytical form, you basically cannot do much other than:

$$F_X(a) = \int_{\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy$$

Theorem 6.3 (Central Limit Theorem)**.** *This is basically why Gaussian distribution is important and appears often in real life. Suppose I generated n random numbers $x_1, x_2, ..., x_n$ from* ***a*** *certain distribution D (can be any distribution). Then, consider the* ***average of those numbers****:*

$$S_n = \frac{1}{n}\sum_{i=1}^{n} x_i \approx \mathbb{E}[X]$$

Notice that for the S_n :

1. *The above holds due to weak law of large numbers*
2. *Since S_n is sum of random number, it is* ***a random number*** *itself!*

Therefore, if we consider the ***distribution of*** *S_n, we will see that:*

$$S_n \sim N(\mu, \sigma^2)$$

for $\mu = \mathbb{E}[X]$ and $\sigma^2 = Var[X]/n$

Graphically, this is what happens:

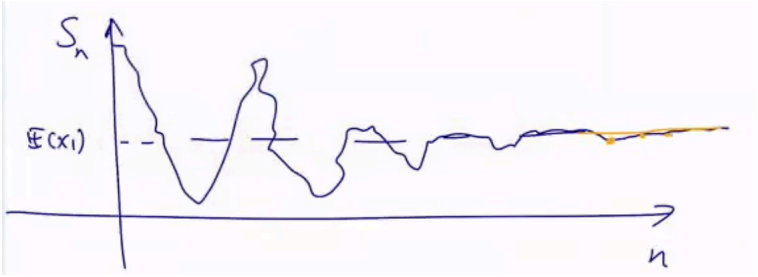


Figure 20: CLT

***Example*: Student Exam Grades**

Suppose you want to model student exam grades Z, and you think it can be modelled by averaging the following features (which is assumed to come from the same distribution, or at least weakly dependent, which is from a more general CLT):

1. Their IQ level
2. How well did the student learn algebra
3. How well did the student learn calculus
4. How well can they manage the stress during exam

Then, for each student, we gather x_1, x_2, x_3, x_4, and we say their exam grade is $Z = \frac{1}{4}\sum_i^4 x_i$. Then, from the central limit theorem, student's score will be a normal distribution.

Perhaps the more important application is with phone signal "filtering". Notice that when we are doing a phone call, we want to transmit sound wave X, but there will be lots of noises $Z_1, ..., Z_{100}$ that will also be picked up. How can your phone know which one is your voice?

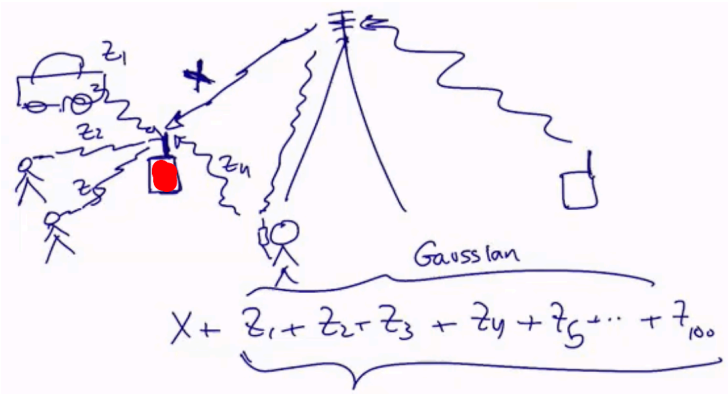


Figure 21: CLT Example with Phone Signal

basically from the CLT we know that:

- $Z_1 + ... + Z_{100}$ will be Gaussian distributed
- then we can remove that contribution (by some technique presented later)

6.2.3 Exponential Random Variable

Definition 6.4 (Exponential PDF)**.** *The exponential random variable* $X \sim Exp(\lambda)$ *is defined by:*

$$f_X(a) = \begin{cases} \lambda e^{-\lambda a}, & a \geq 0 \\ 0, & otherwise \end{cases}$$

Graphically, this looks like:

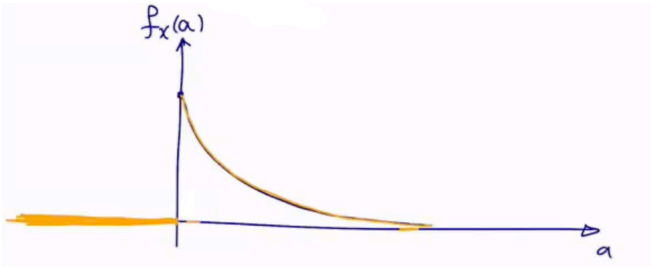


Figure 22: Exponential Random Variable Example

We can further compute the CDF of this distribution:

$$\begin{aligned} F_X(b) &= \mathbb{P}(X \leq b) \\ &= \int_0^b f_X(a)\, da \\ &= \int_0^b \lambda e^{-\lambda a}\, da \\ &= -e^{-\lambda a}\Big|_0^b \\ &= 1 - e^{-\lambda b} \end{aligned}$$

Graphically, this looks like:

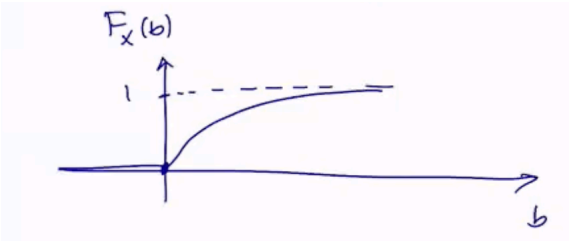


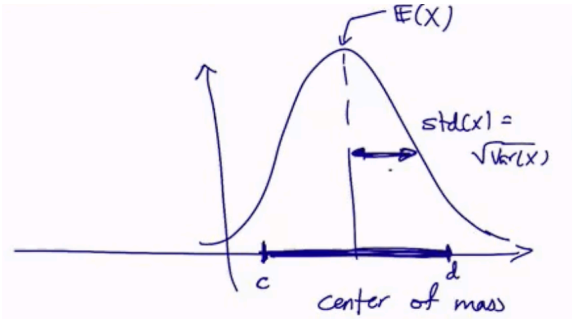
Figure 23: CDF for Exponential Random Variable

Note:

> The **unsmoothness in CDF** comes exactly from the **discontinuity in PDF**!

6.3 Mean and Variance

Basically the idea here is the same as the PMF of a discrete random variable, where we want to **measure the center of mass** of some **given, unknown PDF**:



where here, we basically want that $P(c \leq X \leq d) \approx 0.68$.

6.3.1 Mean

Definition 6.5 (Mean of Continous RV)**.** *The mean of continuous variable X is defined as:*

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} y f_X(y)\, dy$$

which is basically the analogy of expected value of discrete RV.

Intuitively, you might think variance would be somehting like:

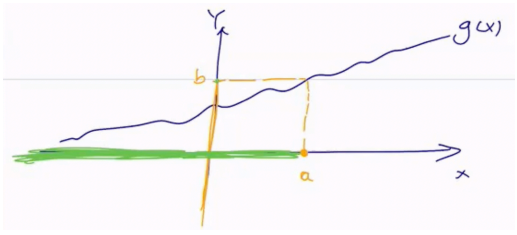
$$\text{Var}[X] = \mathbb{E}[(X - \mu)^2]$$

Then we are back to the same problem as discrete RV: we want to know the distribution of Y with $Y = g(X)$ for some distribution of X (i.e. expressed as $f_X(a)$ or $F_X(a)$). For continuous random variable. Consider:

$$F_y(b) = P(Y \leq b)$$

- note that we could do this by PDF as well, but usually **thinking in CDF** would be easier for continuous random variable.

For simplicity, assume that $g(X)$ looks like:



Then, we basically know:

$$P(Y \leq b) = P(X \leq a) = F_X(a)$$

Example

Given $X \sim \text{Exp}(1)$, we know that:

$$f_\lambda(a) \begin{cases} e^{-a}, & a \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Suppose that $Y = \sqrt{X}$. Then what is the PDF of Y?

Solution: Basically write out the definition:

$$\begin{aligned} F_Y(b) &= \mathbb{P}(Y \leq b) \\ &= \mathbb{P}(\sqrt{X} \leq b) \\ &= \mathbb{P}(0 \leq X \leq b^2) \\ &= \int_0^{b^2} e^{-a}\, da \\ &= 1 - e^{-b^2} \end{aligned}$$

Finally, taking the derivative to get the PDF:

$$f_Y(a) = \frac{d}{db} F_Y(b) = 2be^{-b^2}$$

Hence we actually get (**notice that $Y < 0$ has $F_Y(a) = 0$**). Therefore, you get:

$$f_Y(a) = \begin{cases} 2be^{-b^2}, & b \geq 0 \\ 0, & b < 0 \end{cases}$$

(try it from PDF directly, which is tricky)

For another example, consider that we have $X \sim \text{Unif}[0, 1]$, and the $Y = g(X)$ is a saw tooth function:

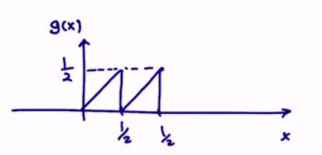


Figure 24: Saw Tooth Function of $Y = g(X)$

***Example*: Uniform to Saw Tooth**

We want to figure out the distribution of Y. First, observe that:

- $b < 0$, then $F_y(b) = 0$ since this correspond to $X < 0$.

- $0 \leq b \leq \frac{1}{2}$, then:

$$\begin{aligned} F_Y(b) &= \mathbb{P}(Y \leq b) \\ &= \mathbb{P}(0 \leq X \leq b) + \mathbb{P}(\frac{1}{2} \leq X \leq \frac{1}{2} + b) \\ &= \int_0^b f_X(a)\, da + \int_{\frac{1}{2}}^{\frac{1}{2}+b} f_X(a)\, da \\ &= \int_0^b da + \int_{\frac{1}{2}}^{\frac{1}{2}+b} da \\ &= b + b = 2b \end{aligned}$$

  for visually how we got there, check out Figure 25.

- lastly, for $b > \frac{1}{2}$, then $F_Y(b) = \mathbb{P}(Y \leq b) = 1$.

Notice that this is exactly a **uniform distribution**, where basically $Y \sim \text{Unif}\left[0, \frac{1}{2}\right]$

Graphically, we basically computed $\mathbb{P}(Y \leq b)$ being:

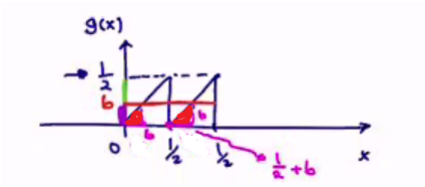


Figure 25: Computing $\mathbb{P}(Y \leq b)$

basically $\mathbb{P}(Y \leq b)$ would be the **two red areas**.

Note:

Up to this point, we basically see that we can **brute force** calculate $\mathbb{E}(Y)$ by:

- computing the $F_y(b)$
- compute $\mathbb{E}[Y]$ using $f_Y(b) = dF_y(b)/db$

But there is a away again to compute directly from distribution of X.

Theorem 6.4 (LOTUS with Continuous RV)**.** *If X is a continuous random variable $Y = g(X)$, then:*

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} g(a) f_X(a) da$$

which is basically analogous with the discrete case.

6.3.2 Variance

Definition 6.6 (Variance of Continuous RV)**.** *Therefore, similar to the discrete case, we have:*

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

from which we could compute directly by letting $Y = g(X) = (X-\mu)^2$ *and use LOTUS.*

Lemma 6.5. *In the continuous case, this still hold.*

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

Proof. Here the bounds are all from negative infinity to positive infinite, so it is omitted here.

$$\begin{aligned}
\text{Var}[X] &= \mathbb{E}[(X-\mu)^2] \\
&= \int (a-\mu)^2 f_X(a) da \\
&= \int (a^2 - 2\mu a + \mu^2) f_X(a) da \\
&= \int a^2 f_X(a) da - 2\mu \int a f_X(a) da + \mu^2 \int f_X(a) da \\
&= \int a^2 f_X(a) da - 2\mu \cdot \mu + \mu^2 \\
&= \int a^2 f_X(a) da - \mu^2 \\
&= \mathbb{E}(X^2) - \mu^2
\end{aligned}$$

which completes the proof. □

6.3.3 Mean and Variance for Common Distributions

Here we provide example computations for some of the common distributions with continuous random variables.

***Example*: Mean and Variance of Uniform Distribution**

Consider a uniform distribution $X \sim \text{Unif}[0,1]$. Calculating the mean is pretty easy, basically:

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} a f_X(a)\, da \\
&= \int_0^1 a\, da \\
&= \frac{1}{2}
\end{aligned}$$

Then, computing the variance, we can first compute $\mathbb{E}[X^2]$ using LO-

TUS:

$$\begin{aligned}\mathbb{E}[X^2] &= \int_{-\infty}^{\infty} a^2 f_X(a)\, da \\ &= \int_0^1 a^2\, da \\ &= \frac{1}{3}\end{aligned}$$

So we get the variance by:

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

***Example*: Mean and Variance of Exp(λ)**

Consider now we have an exponential random variable. So we have:

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

Now, if we compute the mean:

$$\begin{aligned}\mathbb{E}[X] &= \int_{-\infty}^{\infty} a f_X(a)\, da \\ &= \lambda \int_0^{\infty} a e^{-\lambda a}\, da\end{aligned}$$

which now we need to use integration by parts. Since we know:

$$\int_a^b f(x)g'(x)\, dx = [f(x)g(x)]_a^b - \int_a^b f'(x)g(x)\, dx$$

so we can let $a = f(x)$, and $g'(x) = e^{-\lambda a}$. Then we have:

$$\begin{aligned}\lambda \int_{-\infty}^{\infty} a e^{-\lambda a}\, da &= \left[\lambda \frac{a e^{-\lambda a}}{\lambda}\right]_0^{\infty} - \int_0^{\infty} 1 \cdot \lambda \cdot \frac{-1}{\lambda} e^{-\lambda a}\, da \\ &= 0 + \int_0^{\infty} e^{-\lambda a}\, da \\ &= \frac{1}{\lambda}\end{aligned}$$

where the first term goes to zero because a polynomial grows **slower** than exponential, which can be shown using **L'Hopital's rule**:

$$\lim_{a\to\infty} a e^{-\lambda a} = \lim_{a\to\infty} \frac{a}{e^{\lambda a}} = \lim_{a\to\infty} \frac{1}{\lambda e^{\lambda a}} = 0$$

Last but not least, we compute the **variance**. Since we know that

$\text{Var}[X] = \mathbb{E}[X^2] - \mu^2$, we just need to compute, using LOTUS:

$$\begin{aligned}\mathbb{E}[X^2] &= \int_{-\infty}^{\infty} a^2 f_X(a)\, da \\ &= \int_0^{\infty} a^2 \lambda e^{-\lambda a}\, da\end{aligned}$$

Again, doing integration by parts by taking $a = f(x)$, $g'(a) = e^{-\lambda a}$, we get:

$$\begin{aligned}\mathbb{E}[X^2] &= \left[a^2 \frac{-1}{\lambda} e^{-\lambda a}\right]_0^{\infty} + 2\int_0^{\infty} a\lambda \frac{-1}{\lambda} e^{-\lambda a}\, da \\ &= 0 + 2\frac{1}{\lambda^2} \\ &= \frac{2}{\lambda^2}\end{aligned}$$

note that the first term of the second equality comes from again taking L'Hopital's rule **twice**, computing the ratio between a^2 and $e^{\lambda a}$, and the second term comes from using the result from $\mathbb{E}[X]$ calculation above.

Therefore, the variance is:

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

Note:

There are several observations for the exponential variable.

- the mean is $1/\lambda$, which means that if λ becomes large, your data will concentrate more towards 0.
- This makes sense since standard deviation is also $1/\lambda$.
- Exponential random variable is very related to the **geometric random variable in discrete**, by simply discretising the continuous variable to discrete buckets.

Last but not least, we consider the Normal Distribution of $X \sim N(\mu, \sigma^2)$.

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

Before we go into compute the mean and variance, it is useful to know the following quantities:

- the formal proof can be found online and in professor's notes

$$\int_{-\infty}^{\infty} f_X(a)\, da = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(a-\mu)^2/(2\sigma^2)}\, da = 1$$

- and that:
$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} a e^{-a^2/(2\sigma^2)}\, da = 0$$
this can be seen geometrically since this function is **odd**, but it can also be **formally computed** by letting $h(a) = e^{-a^2/(2\sigma^2)}$, which is useful because its derivative has $ae^{-a^2/(2\sigma^2)}$, which is basically **integration by substitution**.

Now, we can compute the mean and variance of a Gaussian.

***Example*: Mean and Variance of Normal Distribution**

First we compute:
$$\mathbb{E}[X] = \int_{-\infty}^{\infty} a \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(a-\mu)^2}{2\sigma^2}}\, da$$
here, we can let $b = a - \mu$ so that we get the same form of our second quantity computed in above. So we get:
$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} (b+\mu) \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{b^2}{2\sigma^2}}\, db \\
&= \int_{-\infty}^{\infty} b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{b^2}{2\sigma^2}}\, db + \mu \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{b^2}{2\sigma^2}}\, db \\
&= 0 + \mu \cdot 1 \\
&= \mu
\end{aligned}$$
where the third equality comes from using the two quantities we computed prior to our example. The result makes sense by the definition of Normal Distribution.

Lastly, we compute the variance, which here we do it directly using LOTUS:
$$\begin{aligned}
\mathrm{Var}[X] &= \int_{-\infty}^{\infty} (a-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(a-\mu)^2}{2\sigma^2}}\, da \\
&= \int_{-\infty}^{\infty} b^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{b^2}{2\sigma^2}}\, db \\
&= \int_{-\infty}^{\infty} b \frac{1}{\sqrt{2\pi}} \frac{b}{\sigma} e^{-\frac{b^2}{2\sigma^2}}\, db
\end{aligned}$$
Here we used the trick of putting b^2 **into two terms**, which is a common **trick for computing with Normal Distribution**.

Now, we notice that we can let $g(b) = \sigma e^{-b^2/(2\sigma^2)}$, so then:
$$g'(b) = -\frac{b}{\sigma} e^{-\frac{b^2}{2\sigma^2}}$$

so now we can use **integration by parts**.

$$\begin{aligned}
\mathrm{Var}[X] &= \int_{-\infty}^{\infty} b \frac{1}{\sqrt{2\pi}} \frac{b}{\sigma} e^{-\frac{b^2}{2\sigma^2}}\, db \\
&= \frac{1}{\sqrt{2\pi}} \left[\left(-b\sigma e^{-\frac{b^2}{2\sigma^2}} \right)_{-\infty}^{\infty} - \int_{-\infty}^{\infty} + \sigma e^{\frac{b^2}{2\sigma^2}}\, db \right] \\
&= 0 + \sigma^2 \left(\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{b^2}{2\sigma^2}}\, db \right) \\
&= \sigma^2
\end{aligned}$$

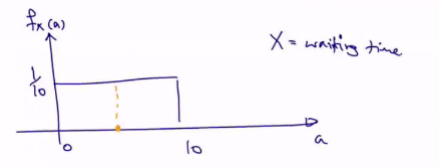
since basically the integral in the bracket in the third equality is the integral over the normal distribution itself.

7 Conditional, Joint, and Marginal

Here, basically we are extending the simple concept of single RV with single distribution to other more interesting situations.

7.1 Conditional Distribution

Consider that your are waiting for a train, and the waiting time X is modelled as follows:



But what if you have **already waited for 5 minutes**, consider:

$$\begin{aligned}\mathbb{P}(X \geq 8 | X \geq 5) &= \frac{\mathbb{P}(X \geq 8 \cap X \geq 5)}{P(X \geq 5)} \\ &= \frac{P(X \geq 8)}{P(X \geq 5)} \\ &= \frac{\int_8^{10} f_X(a)\, da}{\int_5^{10} f_X(a)\, da} \\ &= \frac{2}{5}\end{aligned}$$

But what if you want to be **more general**, consider:

$$P(X \leq a | X \geq 5) = \frac{P(X \leq a, X \geq 5)}{P(X \geq 5)}$$

Note:

> Note that this **doesn't** work if you try this technique for computing $f_{X|A}(a|A)$.

then we need to be careful because there are actually three cases:

$$P(X \leq a | X \geq 5) = \begin{cases} 0, & a \leq 5 \\ P(5 \leq X \leq a)/P(X \geq 5) = (a-5)/5, & 5 \leq a \leq 10 \\ 1, & a \geq 10 \end{cases}$$

where the third equality comes from that for $a \geq 10$, $P(5 \leq X \leq a) = P(5 \leq X)$.

This hints at the definition of **conditional CDF**:

Definition 7.1 (Conditional CDF)**.** *Consider an event A (e.g. you have already waited for 5 minutes) that has already happened, then conditional CDF of a variable X is defined as:*

$$F_{X|A}(a|A) = \mathbb{P}(X \leq a | A)$$

Definition 7.2 (Conditional PDF)**.** *Then the definition of conditional PDF is then:*

$$f_{X|A}(a|A) = \frac{dF_{X|A}(a|A)}{da}$$

so, for example, in the train case above, the conditional PDF is:

$$f_{X|A}(a|A) = \begin{cases} 0, & a < 5 \\ 1/5, & 5 \leq a \leq 10 \\ 0, & a > 10 \end{cases}$$

notice that integrating this sums up to 1. So itself is like a "new" distribution!

Note:

> Though there is a way to compute conditional PDF directly from a PDF, but it is much more involved, so it is suggested that you do it from CDF.

Definition 7.3 (Conditional Expectation)**.** *Then similarly, we can define expectation:*

$$\mathbb{E}[X|A] = \int_{-\infty}^{\infty} a f_{X|A}(a|A)\, da$$

Again, in the train example above:

$$\mathbb{E}[X|A] = \int_{5}^{10} a \cdot \frac{1}{5}\, da = 7.5$$

which makes sense because you have already waited for 5 minutes.

The take-away message here is that you basically obtain **a "new" distribution** based on some condition. Then it follows that you can define other stuff such as:

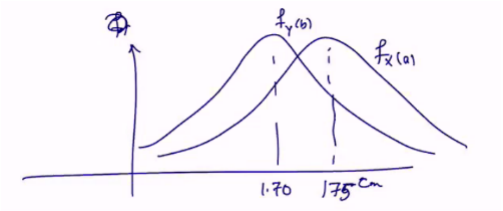
- variance:

  $$\mathrm{Var}[X|A] = \mathbb{E}[(X - \mu)^2|A]$$

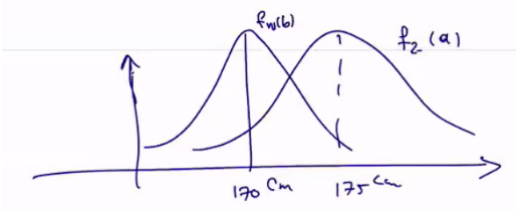
  where the mean is the conditional mean $\mu = \mathbb{E}[X|A]$
- etc.

7.2 Joint for Discrete RV

Consider now, we want to measure **two things at a time**: X being the height of the husband, and Y being the height of the wife (in the same family). Suppose that their distribution is Gaussian:



But consider **another measurement**: Z is the height of a husband, but W is the height of wife **from a different family**. Then technically, the distribution of the variables look like:



which basically looks the same. But consider that if you picked a husband of 2 meters, then:

- in the first case, you should expect the wife to have a **similar height**, around 1.9 to 2.1 meters
- in the second cases, it is completely "random"/Gaussian.

Take-away message: this dependency is **not captured** in the PDF of the variable. This **dependency of RV** is captured in the **joint distribution** (if your RV is dependent).

- if your RV is independent, then maybe joint distribution may not be as helpful

7.2.1 Joint Distribution for Discrete RV

Definition 7.4 (Joint PMF)**.** *The joint PMF of random variable of X, Y is defined as:*

$$P_{XY}(a, b) = \mathbb{P}(X = a, Y = b)$$

***Example*: Independent RVs**

Consider both X, Y being a Bernoulli random variable, both with $p =$

0.5. Additionally, let them be independent. Then:

$$\begin{cases} P_{XY}(0,0) &= 1/4 \\ P_{XY}(0,1) &= 1/4 \\ P_{XY}(1,0) &= 1/4 \\ P_{XY}(1,1) &= 1/4 \end{cases}$$

here it basically we are imagining **two independent coin tosses**. So X, Y are independent, which is like the second example of husband and wife.

However, the more interesting quantities to compute would be:

- consider:

$$\begin{aligned} \mathbb{P}(X=1) &= \mathbb{P}((X,Y)=(1,0) \text{ or } (X,Y)=(1,1)) \\ &= P_{XY}(1,0) + P_{XY}(1,1) \\ &= \frac{1}{2} \end{aligned}$$

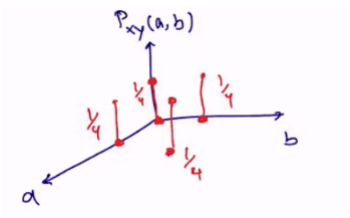
  Basically we are seeing **more info in joint distribution**, that we can revert back to PMF of one variable

- consider:

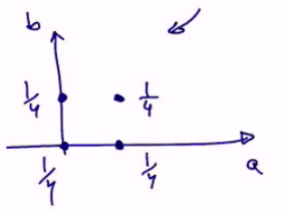
$$\sum_{a=0}^{1}\sum_{b=0}^{1} P_{XY}(a,b) = 1$$

  which makes sense since we are covering over all the probabilities.

Graphically, the joint distribution looks like:



(a) More visual



(b) More Popular for Discrete RV

Then, from the above example, we get some intuition on how joint distribution works.

Lemma 7.1. *The joint PDF satisfies:*

- $P_{XY}(a,b) \geq 0$
- $\sum_a \sum_b P_{XY}(a,b) = 1$

which is true no matter if X, Y are dependent or not.

Now, let us consider a case that **two variables are dependent**. Consider the case that:

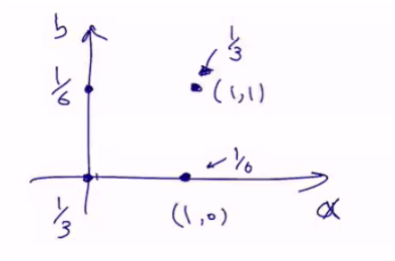
$$X = \begin{cases} 1, \text{if husband is higher than 180cm} \\ 0, \text{otherwise} \end{cases}$$

$$Y = \begin{cases} 1, \text{if wife is higher than 170cm} \\ 0, \text{otherwise} \end{cases}$$

where the height of husband and wife comes from the Gaussian we defined in the beginning of the example. In fact, this random variable X, Y is now **dependent** because if you consider:

- $P(Y = 1|X = 1) > P(Y = 0|X = 1)$ since we expect wives to be similar height of husband.

Therefore, for dependent RV, we would see something like:



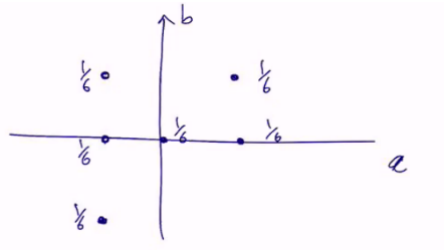
notice that **they are not uniform**!

7.2.2 Joint CDF for Discrete RV

Definition 7.5 (Joint CDF)**.** *The joint CDF of X, Y is defined as:*

$$F_{XY}(a, b) = \mathbb{P}(X \leq a, Y \leq b)$$

Consider the case that we are given the joint PMF being:



Then, some quantities to compute would be:

- consider

$$\begin{aligned} P(Y=0) &= P((X,Y) = \{(-1,0),...,(1,0)\} \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \\ &= \frac{1}{2} \end{aligned}$$

- Now consider the CDF:

$$\begin{aligned} F_{XY}(0,0) &= P(X \leq 0, Y \leq 0) \\ &= P((X,Y) = \{(0,0),(-1,0),(-1,-1)\}) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \\ &= \frac{1}{2} \end{aligned}$$

  so in the discrete case it is just adding up a bunch of values from the known PMF.

7.3 Marginals for Discrete RV

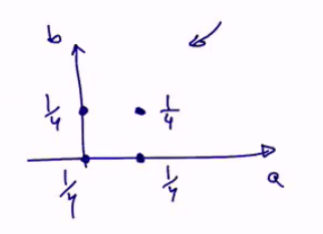
Now, let us consider the marginals for discrete variables.

7.3.1 Marginal PMF for Discrete RV

Marginals only make sense when you have a **joint distribution**. Basically, given some distribution $P_{xy}(a, b)$:

$$\text{Marginals} = \begin{cases} P_X(a) = P(X = a) \\ P_Y(b) = P(Y = b) \end{cases}$$

Again, lets consider the simple example:



Example

From the same example above, we can calculate $P_X(a)$ by:

$$P_X(0) = P(X = 0, Y = 0) + P(X = 0, Y = 1) = \frac{1}{2}$$
$$P_X(1) = P(X = 1, Y = 0) + P(X = 1, Y = 1) = \frac{1}{2}$$

which basically gives you:

$$P_X(a) = \begin{cases} 1/2, & a = 0 \\ 1/2, & a = 1 \end{cases}$$

The **take-away message** is that given a joint, you can always get the individual marginal distribution from it. (If you are given marginals, then you **cannot** reconstruct the joint since dependency information would be lost)

Therefore, we can have the formula in general for discrete variables:

Definition 7.6 (Marginal for DRV)**.** *Given a joint PMF $P_{xy}(a, b)$, then:*

$$P_X(a) = \sum_{b=-\infty}^{\infty} P_{XY}(a, b)$$
$$P_Y(b) = \sum_{a=-\infty}^{\infty} P_{XY}(a, b)$$

so basically if you want to calculate $P_X(a)$, fix a and add up all possibilities of b, and vice versa.

7.3.2 Marginal CDF for Discrete RV

Given some joint distribution $P_{X,Y}(a,b)$, we basically have the marginal CDF being the CDF for the marginal PDF:

$$F_X(a) = P(X \leq a)$$
$$F_Y(b) = P(Y \leq b)$$

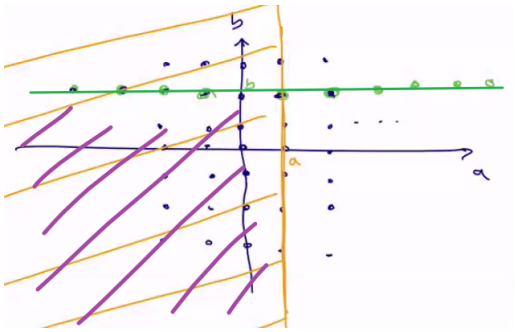
where $P(X = a), P(Y = b)$ are the marginal PMF.

However, alternatively from getting them from the marginal PDF, we could actually get it from the joint CDF.

Definition 7.7 (Marginal CDF for DRV)**.** *Given joint CDF $F_{XY}(a,b)$:*

$$F_X(a) = P(X \leq a) = \lim_{b\to\infty} F_{XY}(a,b)$$
$$F_Y(b) = P(Y \leq b) = \lim_{a\to\infty} F_{XY}(a,b)$$

Graphically, what is happening is we know that $F_{XY}(a,b)$ is the purple area here:



then, what we need to do is:

- $F_X(a)$ is the yellow area. To get the yellow area from the purple, we just need to do:
$$F_X(a) = \lim_{b\to\infty} F_{XY}(a,b)$$

7.4 Independence of Discrete RV

Definition 7.8. *Two RV X, Y are independent iff the following is true:*

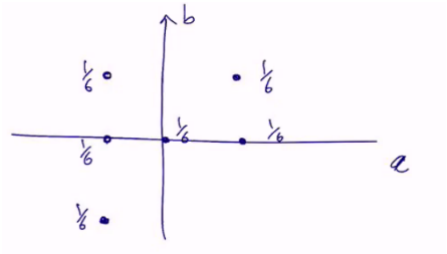
$$P_{XY}(a, b) = P_X(a)P_Y(b)$$

In other words, for ***every pair of*** *a, b:*

$$P(X = a, Y = b) = P(X = a)P(Y = b)$$

Intuitively, knowing X should **not** give you any more nformation on Y than $P_Y(b)$ if they are independent.

Consider the following joint distribution:

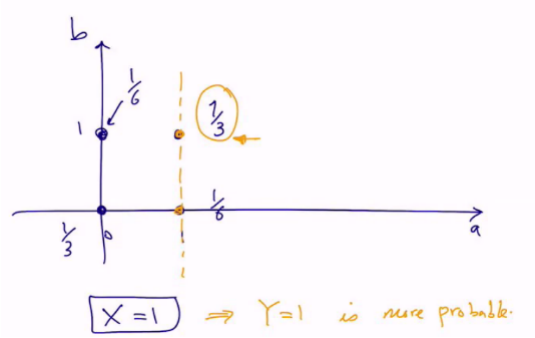


the above joint is **dependent**, which can be seen in many ways:

- given $X = 0$, then $P(Y = 0|X = 0) = 1$! So basically it tells you $Y = 0$ which is some additional information.
- you can also do it mathematically, since we know $P_X(0) = 1/6$, and $P_Y(0) = 1/2$, then:

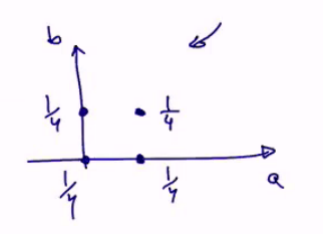
$$P_{XY}(0, 0) = \frac{1}{6} \neq \frac{1}{6} \cdot \frac{1}{2} = P_X(0)P_Y(0)$$

Another quick example of dependence is:



Hence, this one is also dependent RVs, since you suddenly got more information on what Y will be when given $X = 1$.

Now, let us see an example that is independent

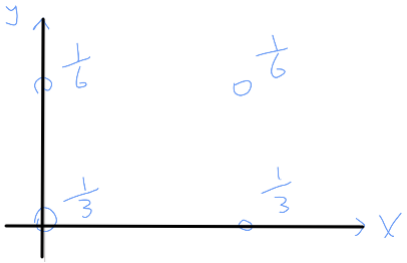


The key idea/check is that, when given an $X = a_1$, the probability of finding $Y = b, \forall b$ should be the **same as** $P_Y(b)$ (i.e. the marginal):

$$P(Y|X = a) = P(Y)$$

(which can be derived from the above definition.

Therefore, the following joint distribution would also work to be independent:



where if we compute the marginals, you will get:

- $P_X(a) = 1/2$
- then for Y:

$$P_Y(b) = \begin{cases} \frac{2}{3}, & b = 0 \\ \frac{1}{3}, & b = 1 \end{cases}$$

- but still if you check on any value:

$$P_{XY}(a, b) = P_X(a)P_Y(b)$$

the upshot is that from the graph, if you take any slice $X = a$, the probability of getting y always has the same ratio of $2 : 1$ as given in the marginal of Y. Hence we have independence.

7.5 Joint for Continuous RV

Before going into the definitions, lets us see the comparison again.

- if we go with $P(X = a, Y = b)$ for continuous RV, it is meaningless as it is just 0
- Hence we start with $F_{xy}(a, b)$ being the CDF of continuous RV

Discrete	Continuous	Comment
$P_{XY}(a, b)$	$P_{XY}(a, b) = 0$	Because the outcome is in real, which has infinite possibility
$F_{XY}(a, b)$	$F_{XY}(a, b) = P(X \leq a, Y \leq b)$	Is meaningful

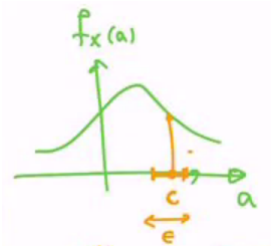
7.5.1 Joint PDF for Continuous RV

In particular, we will defined the **PDF for continuous RV** be like:

Definition 7.9 (Joint PDF)**.** *Basically it is best to be computed from CDF:*

$$f_{XY}(a, b) = \frac{\partial^2}{\partial a \partial b} F_{XY}(a, b)$$

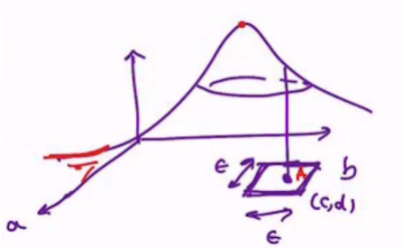
Recall that we interpreted the PDF in 1-D by thinking about its neighborhood:



then:

$$P(c - \frac{\epsilon}{2} \leq x \leq c + \frac{\epsilon}{2}) \approx P_X(c) \cdot \epsilon$$

Then, in the joint case:



In the joint case, we just have:

$$P_{XY}((a, b) \in A) \approx f_{XY}(c, d) \cdot \epsilon^2$$

Lemma 7.2 (Properties of Joint PDF)**.** *Below are true for any continuous joint PDF:*

- $f_{XY}(a,b) \geq 0$ *(if it were negative, then some small neighborhood could have negative probability)*
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(a,b)\, da\, db = 1$

Lemma 7.3 (Marginal from Joint PDF)**.** *Given a joint PDF $f_{XY}(a,b)$, then:*

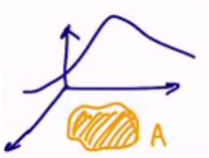
$$f_X(a) = \int_{-\infty}^{\infty} f_{XY}(a,b)\, db$$
$$f_X(b) = \int_{-\infty}^{\infty} f_{XY}(a,b)\, da$$

Lemma 7.4 (Calculating Probability)**.** *For a given area $A \in \mathbb{R}^2$:*

$$P((x,y) \in A) = \int \int_{(a,b)\in A} f_{XY}(a,b)\, dadb$$

which basically it integrating the distribution over the area:

Graphically we are just doing:



7.5.2 Joint CDF for Continuous RV

Basically given a PDF $f_{XY}(a,b)$, we can compute the PDF.

Definition 7.10 (CDF for Contionus RV)**.** *Given a PDF:*

$$F_{XY}(a,b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f_{XY}(c,e)\, dcde$$

7.6 Independence of Continuous RV

In the discrete case, independence is marked by:

$$P_{XY}(a,b) = P_X(a)P_Y(b)$$

which is meaningful because $P_X(a) = P(X = a)$ are probabilities for events in the discrete case.

In the continuous case, we have the following definition

Definition 7.11 (Independence of Continuous RV)**.** *In the continuous world:*

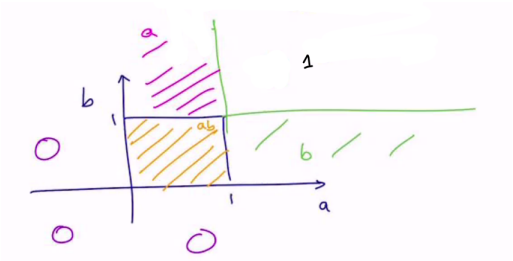
$$f_{XY}(a,b) = f_X(a)f_Y(b)$$

whose intuition comes from that:

$$F_{XY}(a,b) = F_X(a)F_Y(b)$$

which is more meaningful because it means $P(X \le a, Y \le b) = P(X \le a)P(Y \le b)$*, but both are correct/equivalent definitions.*

Below we discuss an example with the following CDF:



Example

Suppose the joint CDF of X, Y are given by (graph above):

$$F_{XY}(a,b) = \begin{cases} ab, & 0 \le a \le 1, 0 \le b \le 1 \\ a, & 0 \le a \le 1, 1 < b \\ b, & 0 \le b \le 1, 1 < a \\ 1, & 1 < b, 1 < a \\ 0, & a \le 0 \text{ or } b \le 0 \end{cases}$$

To calculate the PDF, using the formula that:

$$f_{XY}(a,b) = \frac{\partial^2}{\partial a \partial b} F_{XY}(a,b)$$

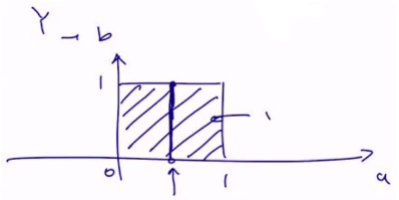
We know that:

$$f_{XY}(a,b) = \begin{cases} 1, & 0 \le a \le 1, 0 \le b \le 1 \\ 0, & 0 \le a \le 1, 1 < b \\ 0, & 0 \le b \le 1, 1 < a \\ 0, & 1 < b, 1 < a \\ 0, & a \le 0 \text{ or } b \le 0 \end{cases}$$

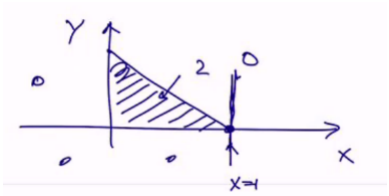
notice that the PDF we got:

- $f_{XY}(a,b) > 0$
- integrates to 1 (as a sanity check).

Notice that the above PDF we got is basically a uniform distribution, and we can think about independence:



(c) Square Uniform



(d) Triangle Uniform

To prove the independent in (a):

- intuitively, knowing any value of $X = a$ doesn't give you any information for Y
- consider showing that:

  $$f_{XY}(a,b) = f_X(a) f_Y(b)$$

  then, we need to compute $f_X(a), f_Y(b)$:

  $$f_X(a) = \begin{cases} \int_{-\infty}^{\infty} f_{XY}(a,b)\, db = 0, & a < 0, a > 1 \\ \int_{-\infty}^{\infty} f_{XY}(a,b)\, db = 1, & 0 \le a \le 1 \end{cases}$$

  which is actually a uniform distribution, $X \sim \text{Unif}(0,1)$. Similarly, it can be shown that $Y \sim \text{Unif}(0,1)$.

  Finally:

  $$f_{XY}(a,b) = \begin{cases} 1 = f_X(a) f_Y(b), & 0 \le a \le 1, 0 \le b \le 1 \\ 0 = f_X(a) f_Y(b), & \text{otherwise} \end{cases}$$

  Hence X, Y are independent.

The dependence in (b):

- intuitively, the higher the X the less choice of Y you have. So this must be dependent.
- First, make sure that the integral of the distribution is 1:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(a,b)\,da\,db = \int_{0}^{1}\int_{0}^{1-b} 2\,dadb = 1$$

- Then, for an exercise, consider $P(X \geq 0.5)$ directly from joint:

$$\begin{aligned} P(X \geq 0.5) &= \int_{0}^{0.5}\int_{0.5}^{1-b} 2\,dadb \\ &= \int_{0.5}^{1}\int_{0}^{1-a} 2\,dbda \\ &= 0.25 \end{aligned}$$

- additionally, consider $f_X(a)$, for $0 \leq a \leq 1$:

$$\begin{aligned} f_X(a) &= \int_{-\infty}^{\infty} f_{XY}(a,b)\,db \\ &= \int_{0}^{1-a} 2\,db \\ &= 2(1-a) \end{aligned}$$

  again notice that the line in the triangle is signified by $y = 1 - x \to b = 1 - a$.
- Lastly, I want to show that X, Y are **dependent** in this case, for $a = 0.75, b = 0.75$:

$$\begin{aligned} f_{XY}(0.75, 0.75) &= 0 \\ f_X(0.75) f_Y(0.75) &= 0.5^2 \neq 0 \end{aligned}$$

  So they are dependent.

7.7 Conditional for Continuous RV

Recall that we defined the conditional PDF from CDF:

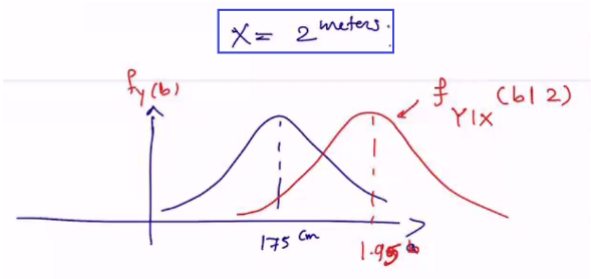
$$f_{X|A}(a|A) = \frac{d}{da} F_{X|A}(a|A)$$

where $F_{X|A}(a|A) = P(X \leq a|A)$. But what if I want to condition on **another variable**?

For example, consider that:

- X is the height of husband
- Y is the height of wife

Suppose that the height of a wife is normally distributed around $\mu = 1.75$m. Then, given that $X = 2$m, we expect the distribution to change since the height of couples usually match up



where basically the conditional distribution is different from the marginal.

Definition 7.12 (Conditional for Joint)**.** *Given a Joint distribution $P_{WZ}(a, b)$:*

$$P_{W|Z}(a|b) = \frac{P_{WZ}(a, b)}{P_Z(b)}$$

which is basically the joint divided the marginal. This is true for continuous RV as well:

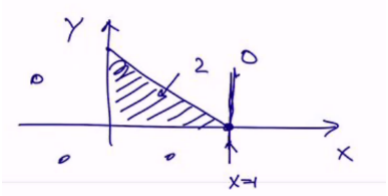
$$f_{W|Z}(a|b) = \frac{f_{WZ}(a, b)}{f_Z(b)}$$

This formula makes sense because if you consider a random value for Z, then:

$$\begin{aligned} P_{W|Z}(a|b = 1) &= P(W = a|Z = 1) \\ &= \frac{P(W = a \cap Z = 1)}{P(Z = 1)} \\ &= \frac{P_{WZ}(a, 1)}{P_Z(1)} \end{aligned}$$

which is the same formula given above. But what if you are at $f_Z(b = 1) = 0$? Then won't you be dividing by 0? The remedy is that if $f_Z(b = 1) = 0$, then considering $f_{W|Z}(a|b = 1)$ doesn't make sense as the event $b = 1$ could never happen.

Now, let us compute with an example. Let the distribution be:



Example

Let the joint be given by the diagram above. Then consider:

$$f_{Y|X}(y|0.5)$$

Intuitively, this should be uniform as it is just a uniform slice at $X = 0.5$. First, we consider $0 < y < 0.5$:

$$\begin{aligned} f_{Y|X}(y|0.5) &= \frac{f_{XY}(y, 0.5)}{f_X(0.5)} \\ &= \frac{2}{2(1-0.5)} \\ &= 2 \end{aligned}$$

where the formula of marginal used in the second equality comes from our result on the previous section.

Then, for $y > 0.5$ or $y < 0$, given that $X = 0.5$, we just have:

$$f_{Y|X}(y|0.5) = 0$$

Therefore, we do get a uniform distribution:

$$f_{Y|X}(y|0.5) = \begin{cases} 2, & 0 \leq y \leq 0.5 \\ 0, & \text{otherwise} \end{cases}$$

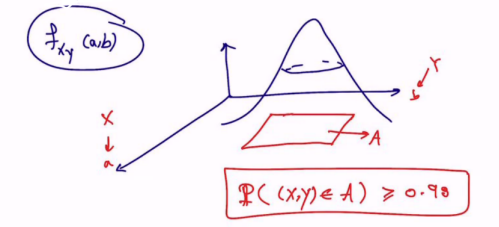
8 Covariance and Correlation

The idea is to think about the analog of variance in 1-D (for 1 random variable) for a joint distribution.

- recall that for single random variable, standard deviation was about the width of the center of mass (mean is the center of center of mass)

8.1 Covariance

Recall the concept of center of mass of a distribution we had before. For example, given some joint $f_{XY}(a, b)$:



where:

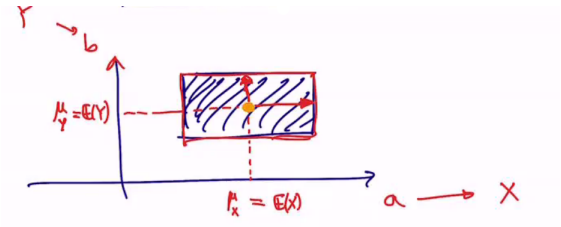
- the center of mass is the region A
- then, mean would be the center of the region A
- but now, **how do we define standard deviation/variance** (analog with PDF of 1 random variable)

To get some sense on what we should do, let us start with **examples with uniform distribution** $f_{XY}(a, b)$, with some given center of mass region:

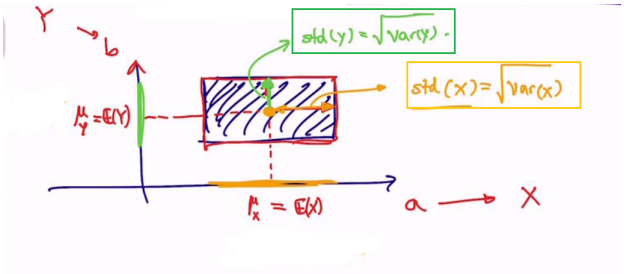
- center of "center of mass mass" would intuitively be specified by:

$$(\mu_x, \mu_y)$$

  Graphically, let the squared region is the center of mass:

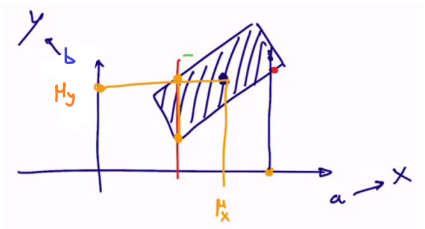


- **standard deviation**/width of the center of mass would be



- notice that in the above case, since we assumed a uniform distribution, X, Y are **independent**. Therefore we **didn't need covariance** to tell us any information about the center of mass.

However, suppose we still have a uniform distribution and a given center of mass region, but now X, Y are **dependent**:



so the problem here is how do we **convey information about the tilt** of the center of mass?

- the formal information:

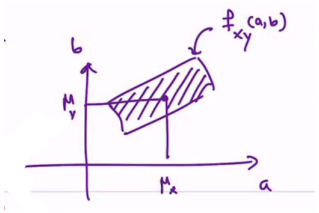
$$(\sigma_x, \sigma_y)$$

  does not work since it only tells us the "width and height" of an untilted region.

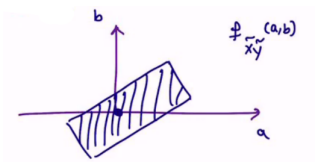
- But consider the following definition of two new random variable:

$$\tilde{X}, \tilde{Y} \equiv X - \mu_x, Y - \mu_y$$

  Graphically, this only shifts the center of mass to the origin:



(e) Original



(f) Shifted

Now, the idea is that we want to know the **quantity** γ, for the line through

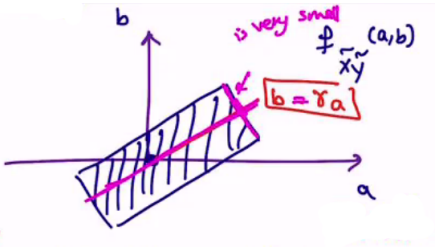


Figure 26: Covariance and Tilt of Center of Mass

the center of mass would be governed by $b = \gamma a$ as shown in Figure 26. It turns out that we can **find out γ by considering Cov(X, Y)**:

$$\text{Cov}(X, Y) = \mathbb{E}[\tilde{X}\tilde{Y}]$$

By imagining the center of mass being a "thin rectangle" such that:

$$\tilde{Y} \approx \gamma \tilde{X}$$

Then we get from computing covariance:

$$\begin{aligned}
\text{Cov}(X, Y) &= \mathbb{E}[\tilde{X}\tilde{Y}] \\
&\approx \mathbb{E}[\gamma \tilde{X}^2] \\
&= \gamma \mathbb{E}[\tilde{X}^2] \\
&= \gamma \mathbb{E}[(X - \mu_x)^2] \\
&= \gamma \text{Var}[X]
\end{aligned}$$

Therefore, if we know Var$[X]$, we get the **tilt from covariance**!

Note:

The above shows us the intuition behind covariance. Interestingly, the following should also make sense:

- **covariance of independent variable is 0**, which means **there is no tilt** in the distribution (makes sense as having a tilted distribution means dependency of random variables).

  In fact, consider 27(a), let Z, W be independent random variables with **center of mass at origin** (WLOG), and we are given their join $f_{ZW}(a, b)$. Then:

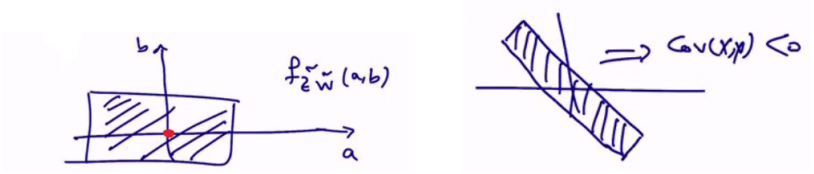
$$\begin{aligned}
\mathbb{E}[\tilde{Z}\tilde{W}] = \mathbb{E}[ZW] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} ab f_{ZW}(a, b)\, da db \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} ab f_Z(a) f_W(b)\, da db \\
&= \int_{-\infty}^{\infty} a f_Z(a)\, da \int_{-\infty}^{\infty} b f_W(b)\, db \\
&= 0 \cdot 0 = 0
\end{aligned}$$

Conclusion: if the two random variables are independent, the covariance will be zero. However, a covariance of zero does not necessarily mean that the variables are independent. An example would be Figure 28.

- the covariance **can be negative**, if you have something like 27(b).
- covariance only tells you about the tilt, to know the width/height of the center of mass you **still need Var$[X]$, Var$[Y]$**. However, often they are included in the "covariance matrix", because:

$$\text{Cov}[X, X] = \text{Var}[X]$$

which are the diagonal entries in the covariance matrix.



(a) Zero Covariance (b) Negative Covariance

Figure 27: Other Covariance Examples

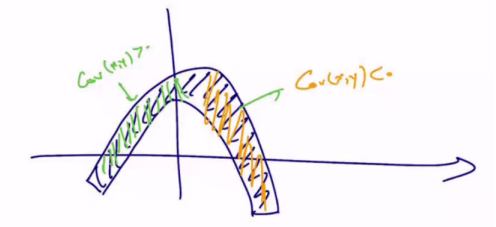


Figure 28: Cov[X,Y] = 0 but X, Y are dependent

Now let us div into covariance in a more formal manner.

Definition 8.1 (Covariance)**.** *Suppose* $\mathbb{E}[X] = \mu_x$*, and* $\mathbb{E}[Y] = \mu_y$*. Then covariance of* X *and* Y *is defined by:*

$$Cov[X, Y] = \mathbb{E}\left[(X - \mu_x)(Y - \mu_y)\right]$$

Now, the question is **how do we compute the covariance** given some joint distribution $f_{XY}(a, b)$?

1. Brute force way:
   (a) We can let:
   $$Z \equiv (X - \mu_X)(Y - \mu_Y)$$

(b) compute the PDF of Z

(c) compute:

$$\mathbb{E}[Z] = \int a f_Z(a) da$$

this would be quite painful to compute. But recall that with 1 variable, we were able to simplify step b) and c) with LOTUS.

2. Simplification using LOTUS in higher dimension

   (a) we can compute directly using LOTUS:

$$\mathbb{E}\left[(X - \mu_x)(Y - \mu_y)\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (a - \mu_x)(b - \mu_y) f_{XY}(a, b)\, da\, db$$

3. Use the Lemma 8.1.

   (a) we just need to compute the quantity below using LOTUS:

$$\text{Cov}[X, Y] = \mathbb{E}[XY] - \mu_x \mu_y$$

Definition 8.2 (Two Dimensional LOTUS)**.** *Suppose that $Z = g(X, Y)$ is a random variable. Then:*

$$\mathbb{E}[Z] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(a, b) f_{XY}(a, b)\, dadb$$

or for a discrete random variable:

$$\mathbb{E}[Z] = \sum_{-\infty}^{\infty} \sum_{-\infty}^{\infty} g(a, b) P_{XY}(a, b)$$

(recall that for 1 random variable LOTUS, we had $f_X(a) \to f_{XY}(a, b)$ instead and we were only integrating/summing over one variable)

Lemma 8.1. *The covariance can also be simplified in a similar manner as variance:*

$$\begin{aligned} Cov[X, Y] &= \mathbb{E}\left[(X - \mu_x)(Y - \mu_y)\right] \\ &= \mathbb{E}[XY] - \mu_x \mu_y \end{aligned}$$

Proof. For continues random variable:

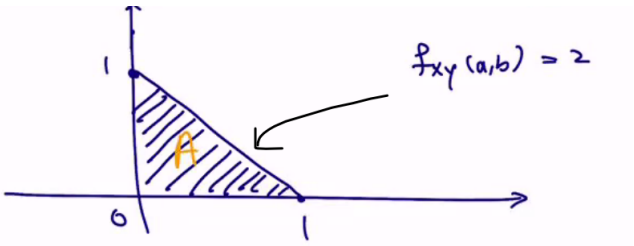
$$\begin{aligned}
\mathrm{Cov}[X,Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(a-\mu_x)(b-\mu_y)f_{XY}(a,b)\,dadb \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}abf_{XY}(a,b)\,dadb - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mu_x bf_{XY}(a,b)\,dadb \\
&\quad - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}a\mu_y f_{XY}(a,b)\,dadb + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mu_x\mu_y f_{XY}(a,b)\,dadb \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}abf_{XY}(a,b)\,dadb - \mu_x\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}bf_{XY}(a,b)\,dadb \\
&\quad - \mu_y\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}af_{XY}(a,b)\,dadb + \mu_x\mu_y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}abf_{XY}(a,b)\,dadb - \mu_x\int_{-\infty}^{\infty}bf_Y(b)\,db \\
&\quad - \mu_y\int_{-\infty}^{\infty}af_X(a)\,da + \mu_x\mu_y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}abf_{XY}(a,b)\,dadb - \mu_x\mu_y - \mu_y\mu_x + \mu_x\mu_y \\
&= \mathbb{E}[XY] - \mu_x\mu_y
\end{aligned}$$

where we used the fact that:

- on the third step:
$$f_X(a) = \int_{-\infty}^{\infty} f_{XY}(a,b)\,db$$
and the same for $f_Y(b)$.
- the last step used LOTUS in reverse:
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}abf_{XY}(a,b)\,dadb = \mathbb{E}[XY]$$

□

Let us compute covariance looking at some example



Example

Consider the joint distribution above. We want to compute the covariance of X, Y.

Intuitively, this should be negative because the larger the X value, the more likely that Y is small, i.e. $\gamma < 0$.

We want to compute this using

$$\mathrm{Cov}[X, Y] = \mathbb{E}[XY] - \mu_x \mu_y$$

1. firs let us compute $\mathbb{E}[XY]$:

$$\begin{aligned}
\mathbb{E}[XY] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} ab f_{XY}(a, b)\, da db \\
&= \int_0^1 \int_0^{1-b} ab \cdot 2\, da\, db \\
&= \int_0^1 b \left[a^2\right]_0^{1-b}\, db \\
&= \int_0^1 b(1-b)^2\, db \\
&= \frac{1}{12}
\end{aligned}$$

   where the integral bound in step 2 is done by thinking about:

$$\int_0^1 \text{horizontal slices}\, db$$

2. Then we need to compute $\mathbb{E}[X] = \mu_x$. Here we use the calculation done before that:

$$f_X(a) = 2(1-a), \quad 0 \le a \le 1$$

   so:

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} a f_X(a)\, da \\
&= \int_0^1 2a(1-a)\, da \\
&= \frac{1}{3}
\end{aligned}$$

3. Due to symmetry:

$$\mathbb{E}[Y] = \frac{1}{3}$$

   since swapping X, Y makes no difference in this distribution.

4. Lastly, we compute the **covariance**:

$$\begin{aligned}\mathrm{Cov}[X, Y] &= \mathbb{E}[XY] - \mu_x \mu_y \\ &= \frac{1}{12} - \frac{1}{3} \cdot \frac{1}{3} \\ &= -\frac{1}{36}\end{aligned}$$

8.2 Correlation Coefficient

In short, correlation coeffcient is a function of covariance and the individual variance. The aim of this is to give some **intuition** on the concentration/strength of correlation between X, Y. (e.g. you will see $\rho_{XY} = 1$ if $Y = \gamma X$ exactly, i.e. strength of correlation at max)

Definition 8.3 (Correlation Coefficient)**.** *This is a better tool to* ***understand the magnitude of tilt****, by "normalizing" the covariance.*

$$\rho_{XY} = \frac{Cov[X, Y]}{\sqrt{Var[X]}\sqrt{Var[Y]}}$$

Lemma 8.2. *This quantity is better for understating the tilt, because:*

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

Note:

- Because correlation coefficient is normalized, it could be that , for joint X, Y and joint Z, W:

  $$\text{Cov}[X, Y] > \text{Cov}[Z, W]$$

  but that:

  $$\rho_{XY} < \rho_{ZW}$$

Now, you will see that correlation coefficient essentially computes the **spread-/correlation** of the two random variables. Consider the case that:

$$Y = \gamma X$$

and let $\mathbb{E}[X] = 0$ be centered at the origin.

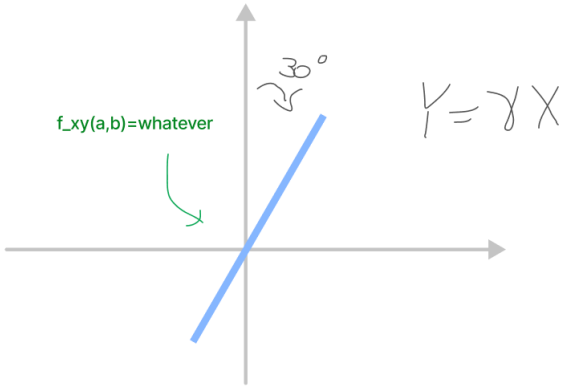


Figure 29: Case when $\rho_{XY} = 1$

For **whatever distribution X has**, we see that:

$$\begin{aligned}\rho_{XY} &= \frac{\text{Cov}[X,Y]}{\sqrt{\text{Var}[X]}\sqrt{\text{Var}[Y]}} \\ &= \frac{\gamma\mathbb{E}[X^2]}{\sqrt{\mathbb{E}[X^2]}\sqrt{\gamma^2\mathbb{E}[X^2]}} \\ &= 1\end{aligned}$$

which is the **largest you can get!** (the second equality comes from that $\text{Var}[Y] = \mathbb{E}[Y^2] = \mathbb{E}[\gamma^2 X^2] = \gamma^2\mathbb{E}[X^2]$.

Note:

An example whose shape look like Figure 29 would be to consider $X \sim \text{Unif}[-1,1]$, and $Y = 2X$ for instance. Then, it will be easy to see that:

$$Y \sim \text{Unif}[-2,2]$$

But their joint is tilted (i.e. the marginals does not give dependency information). This example will be complicated because we consider:

$$\begin{aligned}F_{XY}(a,b) &= F(X \le a, Y \le b) \\ &= F(X \le a, X \le b/2)\end{aligned}$$

Then, this region is basically the same as:

$$F(X \le a, X \le b/2) = F(X \le \min(a, b/2))$$

But then if you compute:

$$f_{XY}(a,b) = \frac{\partial^2}{\partial a \partial b} F_{XY}(a,b)$$

You will find strange things as Y is "not a random variable" given X,i.e. we are calculating joint of one variable X essentially.

Now, consider this **special** case:

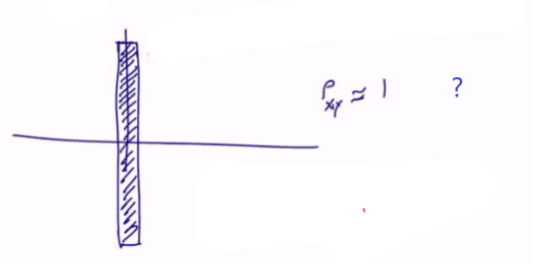


Figure 30: Special Case of γ

here in Figure 30, we noticed that our conclusion $\rho_{XY} = 1$ doesn't work as the look independent. The reason is because:

- recall that the definition of correlation is:

$$\rho_{XY} = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]}\sqrt{\text{Var}[Y]}}$$

  But in this case, $\text{Var}[X] = 0$, which also caused $\text{Cov}[X, Y] = 0$. Now, calculating correlation doesn't make much sense anymore because essentially X **is no longer a random variable**

- the same argument goes with covariance, as the result doesn't make sense if $\text{Var}[X] = 0$, meaning there is no randomness in one of the variable!

8.3 Covariance v.s. Correlation

The graphical interpretation we had so far is that, if we have $Y \approx \gamma X$ for some distribution X:

- $\text{Cov}[X, Y] = \gamma \mathbb{E}[X^2] = \gamma \text{Var}[X]$ if $\mathbb{E}[X] = 0$
- $\rho_{XY} = 1$

Now, we consider some possibilities of manipulation

- two ways of increasing covariance. Since covariance is:

  $$\text{Cov}[X, Y] = \gamma \mathbb{E}[X^2]$$

  in this case, then:

  1. Increase $\mathbb{E}[X^2]$ while fixing γ.

     For example, consider using $X \sim \text{Unif}[-1, 1]$. Then, we can calculate:

     $$\gamma \mathbb{E}[X^2] = \gamma \int_{-1}^{1} \frac{a^2}{2}\, da = \frac{\gamma}{3}$$

     Therefore, we got:

     $$\text{Cov}[X, Y] = \frac{\gamma}{3}$$

     However, if we then consider $X \sim \text{Unif}[-10, 10]$, then we can calculate:

     $$\gamma \mathbb{E}[X^2] = \gamma \int_{-10}^{10} \frac{a^{20}}{2}\, da = \gamma \frac{2000}{60} = \frac{100}{3}\gamma$$

     Therefore, now with the same γ we got:

     $$\text{Cov}[X, Y] = \frac{100\gamma}{3}$$

     which is larger.

  2. Increase γ while fixing $\text{Var}[X]$.

     Essentially, consider fixing $X \sim \text{Unif}[-1, 1]$. Then:

     $$\text{Var}[X] = \mathbb{E}[X^2] = \int_{-1}^{1} a^2\, da = \frac{2}{3}$$

     But then consider:

     $$\text{Var}[Y] = \gamma^2 \mathbb{E}[X^2] = \gamma^2 \frac{2}{3}$$

     or in other words:

     $$Y \sim \text{Unif}[-\gamma, \gamma]$$

     Therefore, we can increase γ, but we will stretch the distribution for Y as well.

- Determining tilt by just looking at the spread, if $Y = \gamma X$ holds.

  Essentially now you need to tilt the center of mass **while making variance constant**. For example, consider the case that we enforce:

  $$\begin{cases} \text{Var}[X] = 1 \\ \text{Var}[Y] = 1 \end{cases}$$

  Then, for centered distribution, and that $Y = \gamma X$, this means that:

  $$\text{Var}[Y] = \mathbb{E}[\gamma^2 X^2] = \gamma^2 \mathbb{E}[X^2] = 1$$

  so we get $\gamma = 1$ **as this is symmetric.**

  But if we consider

  $$\begin{cases} \text{Var}[X] = 2 \\ \text{Var}[Y] = 1/2 \end{cases}$$

  Then you will get that:

  $$\gamma^2 \mathbb{E}[X^2] = \frac{1}{2}$$

  so that $\gamma = 1/2$ is the "maximum".

- Fixing correlation $\rho_{XY} = 1$ while increasing covariance.

  Now, from the discussion above, you can easily change covariance without affecting even γ by changing $\mathbb{E}[X^2]$.

  Alternatively, we can just increase γ while keeping $\mathbb{E}[X^2]$ fixed. Then $\rho_{XY} = 1$ stil holds as long as $Y = \gamma X$, but the only thing side-effect is that we will increase the variance of Y:

  $$\text{Var}[Y] = \gamma^2 \mathbb{E}[X^2]$$

  (again, assuming distributions centered at origin)

9 Multiple Random Variables

Now we just extend our discussion from 2 Random Variable to multiple random variables, which is suitable in real life as we have so many features to observe!

Consider n random variables $X_1, X_2, ..., X_n$. Basically we then have

Definition 9.1 (PDF/PMF)**.** *Basically the extension of 2 variable such that:*

- *discrete case:*

$$P_{X_1...X_n}(a_1, ..., a_n) = P(X_1 = a_1, ..., X_n = a_n)$$

- *continuous case:*

$$f_{X_1...X_n}(a_1, ..., a_n) = \frac{\partial^n F_{X_1...X_n}(a_1, ..., a_n)}{\partial a_1 ... \partial a_n}$$

  where basically we still need the help of CDF.

Definition 9.2 (CDF)**.** *The same as 2 variable case, but now discrete and continuous case has the same equation:*

$$F_{X_1...X_n} = P(X_1 \leq a_1, ..., X_n \leq a_n)$$

the only difference is that:

- *discrete case: use summation to compute the CDF from PMF*
- *contionous case: use integral to compute the CDF from PDF*

Definition 9.3 (Computing some Event A)**.** *Given the multivarate PDF, if you want to calculate the probability that A happened:*

- *discrete case*

$$P((x_1, x_2, ..., x_n) \in A) = \sum_{(a_1,...,a_n)\in A} P_{X_1...X_n}(a_1, ..., a_n)$$

- *continuous case:*

$$P((x_1, x_2, ..., x_n) \in A) = \int_{(a_1,...,a_n)\in A} f_{X_1...X_n}(a_1, ..., a_n) d\vec{a}$$

Those concepts mentioned above are basically the analogous ones we had in 1-D and 2-D. The only thing we are going to discuss which is very useful is their independence.

9.1 Independence

Recall that for two random variables X, Y:

- discrete case:
$$P_{XY}(a,b) = P_X(a)P_Y(b)$$
- continuous case:
$$f_{XY}(a,b) = f_X(a)f_Y(b)$$

Then, in the case of n variables:

- discrete case:
$$P_{X_1...X_n}(a_1,...,a_n) = P_{X_1}(a_1)...P_{X_n}(a_n)$$
- continues case:
$$f_{X_1...X_n}(a_1,...,a_n) = f_{X_1}(a_1)...f_{X_n}(a_n)$$

which is basically the analogous, but they are **very useful and critical** to use in reality.

***Example*: Multivariate Gaussian with Diagonal Covariance**

Let us consider the multivariate Gaussian $X \sim N(\vec{\mu}, \Sigma)$, where $X \in \mathbb{R}^d$ basically contains d **random variables**. Then, the joint of all d variables is:

$$\begin{aligned} &P(\vec{X} = \vec{a}) \\ &= P(X_1 = a_1, ..., X_d = a_d) \\ &= \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\vec{a} - \vec{\mu}^T \Sigma^{-1}(\vec{a} - \vec{\mu})\right) \end{aligned}$$

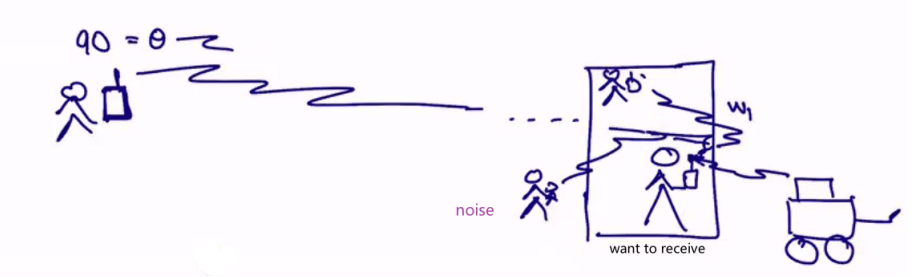
But now, since the covariance is diagonal, this means each variable $X_i, X_j, i \neq j$ is **independent with each other** (note that in general, zero covariance does not imply independence, but it is true in this special case). Therefore, this means that:

$$P(\vec{X} = \vec{a}) = N(x_1; \mu_1, \sigma_1^2) \cdot N(x_2; \mu_2, \sigma_2^2) \cdot ... \cdot N(x_d; \mu_d, \sigma_d^2)$$

which gives **product of univariate Gaussians!**

10 Parameter Estimation

Consider the problem of receiving a number that I have sent to your phone. Let that number be θ. But actually they are many noises that you are going to pick up



Therefore, let there be 50 noises, and let what you have received to be $y = 65$:

$$y = \theta + w_1 + w_2 + ... + w_{50}$$

But we can assume that those noise are independent, and random, hence:

$$y = \theta + z$$

and we **assume that**:

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

Then, we know $y \sim N(\theta, 1)$. However, now the key problem is that if we are given **multiple data** of y, each with some random noise, you eventually get:

$$(y_1 = 65), (y_2 = 75), ..., (y_{50} = 102)$$

Now, we can formula the question as, given $y_i = \theta + z_i$, for $z_i \sim N(0, 1)$. What is the **best estimate for** θ?

- you will be tempted to do MLE, which basically does:

  $$\arg\max_{\theta} P[(y_1, y_2, ..., y_n)|\theta]$$

  and using IID assumption and computing for stationary points, you will get:

  $$\hat{\theta} = \frac{1}{n}\sum_i Y_i$$

  for the fact that $Y \sim N(\theta, 1)$

- But you will soon see that there is another method, called **method of moments**, that will basically give you the same result.

***Example*: Biased Coin**

Given some coin flips, and that the coin itself is biased. How do you find out the bias? In other words, let each toss $X_i \in \{0, 1\}$ be what I have observed:

$$X_1, X_2, ..., X_n \sim \text{Bern}(p, n)$$

Again, once n is known you can estimate p either using a MLE or a

Method of Moment.

Definition 10.1 (Estimation Problem)**.** *Suppose we have* ***observations*** $X_1, X_2, ..., X_n$ *that are IID and have some distribution* $f_X(a;\theta)$*, where* θ *is* ***parameter of interest****. Then, the problem of parameter estimation is to find a good estimate* $\hat{\theta}$ *that is close to* θ*.*

- *e.g.* $f_X(a;\theta)$ *could be* $N(x;0,1)$*, which means the* $\theta = (0,1)$*, and* x *is our observations.*

Here, we show **two general methods** for estimating $\hat{\theta}$:

1. method of moments
2. maximum likelihood estimation (used above)

10.1 Method of Moments

Given some observations and the distribution:

$$X_1, X_2, ..., X_n \sim f_X(a; \theta)$$

for some observations X_i and an **unknown parameter** θ.

1. First, consider:

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X_1] \\ &= \int_{-\infty}^{\infty} a f_X(a;\theta)\, da \\ &= g(\theta)\end{aligned}$$

   for some function g. (This will always be true that the **expected value will depend on** θ).

   - e.g. if $f_X(a;\theta) = \text{Exp}(\theta)$, then $g(\theta) = 1/\theta$

2. By the weak law of large numbers, we have the estimate for the expected value:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \approx \mathbb{E}[X]$$

3. Then, we basically get:

$$\theta \approx g^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \equiv \hat{\theta}$$

   which is (a simple version of) method of moments.

Example

Consider given 100 data points:

$$X_1, X_2, ..., X_{100} \sim \text{Exp}(\theta)$$

for some unknown θ. Then recall that:

$$f_X(a;\theta) = \begin{cases}\theta e^{-\theta a}, & a \geq 0 \\ 0, & \text{otherwise}\end{cases}$$

Then, using **method of moments**:

1. first we know that:

$$\mathbb{E}[X] = \mathbb{E}[X_1] = \frac{1}{\theta}$$

2. then, our current estimate for expected value is:

$$\frac{1}{100}\sum_{i} X_i \approx \frac{1}{\theta}$$

3. Then:
$$\hat{\theta} = \frac{1}{\frac{1}{100}\sum_i X_i} = \frac{100}{\sum_i X_i}$$

***Example*: Back to Cell Phone**

Now, again given some signal received y, such that $y = \theta + z$ for z being some noise. Now, suppose that we only know:

$$Z \sim N(0, \sigma^2)$$

where σ is **now unknown**. Then, we know:

$$Y_1, Y_2, ..., Y_{50} \sim N(\theta, \sigma^2)$$

How do we obtain estimation of **both** θ, σ**!**

Using method of moments, we can only obtain θ:

- obtaining θ is easy as before:
  1. $\mathbb{E}[Y_1] = \theta$
  2. then since:
  $$\frac{1}{50}\sum_{i=1}^{n} y_i \approx \mathbb{E}[Y_1]$$
  3. Finally:
  $$\hat{\theta} = \frac{1}{50}\sum_{i=1}^{n} y_i$$
- now, to obtain σ^2, we consider **similar idea as before**:
  1. consider that:
  $$\mathbb{E}[Y_1^2] = \theta^2 + \sigma^2$$
  2. Then, we know that:
  $$\frac{1}{50}\sum_{i=1}^{n} Y_i^2 \approx \mathbb{E}[Y_1^2]$$
  3. Finally, we can then:
  $$\sigma^2 \approx \frac{1}{50}\sum_{i=1}^{n} Y_i^2 \approx \mathbb{E}[Y_1^2] - \frac{1}{50}\sum_{i=1}^{n} y_i = \hat{\sigma^2}$$

Now, we generalize the idea that if we have multiple parameters to estimate:

$$X_1, X_2, ..., X_n \sim f_X(a; \theta_1, \theta_2)$$

But now, you can do:

1. Now we need two equations, but we can do that easily:
   - $\mathbb{E}[X] = g_1(\theta_1, \theta_2)$
   - $\mathbb{E}[X^2] = g_2(\theta_1, \theta_2)$
2. now we can estimate them by:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \approx \mathbb{E}[X] = g_1(\theta_1, \theta_2)$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \approx \mathbb{E}[X^2] = g_2(\theta_1, \theta_2)$$

3. then the parameters $\hat{\theta}_1, \hat{\theta}_2$ can be estimated by:

$$g_1(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$g_2(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

   where sometimes you will see difficulty to invert those functions to actually obtain expression of $\hat{\theta}_1, \hat{\theta}_2$.

Note:

Therefore, the generalized version of method of moments is easily that:

$$X_1, X_2, ..., X_n \sim f_X(a; \theta_1, \theta_2, ..., \theta_k)$$

for k unknown parameters. Then essentially you try to compute:

$$\mathbb{E}[X^1], \mathbb{E}[X^2], ..., \mathbb{E}[X^k]$$

with k equations, each estimated by:

$$\frac{1}{n}\sum_i X_i, \frac{1}{n}\sum_i X_i^2, ..., \frac{1}{n}\sum_i X_i^k$$

However, in reality for multiple parameters, this method is **not often used** because inverting some of the functions $g_i(\theta_1, ..., \theta_k)$ is very complicated, which is why in Machine Learning we often use the other technique of Maximum Likelihood Estimation (MLE).

10.2 Maximum Likelihood Estimation

Now, given some data points $X_1, X_2, ..., X_n$ being drawn IID from some distribution $f_X(a; \vec{\theta})$, for $\vec{\theta} \in \mathbb{R}^k$ if you have k parameters to estimate. Then, essentially the idea is I want to estimate:

$$\mathcal{L}[\vec{\theta}|(X_1, ..., X_n)] = P[(X_1, ..., X_n)|\vec{\theta}]$$

Then using IID assumption, I can calculate the RHS term to be a product. Then, the key idea is to **find θ that maximizes $\mathcal{L}$** (or more often, the log of the likelihood).

***Example*: MLE Coin Toss**

Let us have $X_1, X_2, ..., X_n$ be n coin tosses you have performed, with a biased coin such that we **know**:

$$X \sim \text{Bern}(\theta)$$

Then, consider estimating the likelihood of θ:

$$\begin{aligned}\mathcal{L}[\theta|(X_1, ..., X_n)] &= P[(X_1, ..., X_n)|\theta] \\ &= \prod_{i=1}^{n} P[X_i|\theta] \\ &= \theta^{n_1}(1-\theta)^{n_0}\end{aligned}$$

where notice that:

- the second equality used the fact that, given a joint:

  $$P[X_1, X_2, ..., X_n|\theta]$$

  since we know that X_i are **independent draws/variables**, then:

  $$P[X_1, X_2, ..., X_n|\theta] = P[X_1] \cdot P[X_2] \cdot ... \cdot P[X_n]$$

  which is what we used.

- in this example, $n_1 = \sum_n X_i$ is the number of 1s I got, and $n_0 = n - n_1$ is the number of 0s I got.

Instead of computing the likelihood which is a product of terms:

$$\frac{d\mathcal{L}}{d\theta} = 0$$

We can use **log likelihood instead** because log are monotonic increasing. Then, this can **simplify the calculation** and still gives the correct

result as:

$$\frac{d}{d\theta}\log\left(\mathcal{L}[\theta|X]\right) = \frac{d}{d\theta}\left(n_1\log\theta + (n-n_1)\log\theta\right)$$
$$= \frac{n_1}{\theta} - \frac{n-n_1}{1-\theta}$$

Then to solve for the stationary point (**assuming it is a concave function**, so only one global maximum), then:

$$\hat{\theta} \leftarrow \theta \text{ such that } \frac{d}{d\theta}\log\left(\mathcal{L}[\theta|X]\right) = 0$$

(a sanity check, if $n_1 = n_2 = 50$, then the above will spit out $\hat{\theta} = 1/2$.

Note:

To solve for $\max \mathcal{L}[\theta, X]$, sometimes you **cannot** take derivatives (e.g. there an indicator function inside). In those cases, it is often that you can find the maximum by inspection.
Sometimes the MLE will have a different result than Method of Moments.

***Example*: MLE with Gaussian**

Now, given n samples, i.e. n random variables from the **same distribution** (IID):

$$X_1, ..., X_n \sim \mathcal{N}(\theta, 1)$$

where for simplicity we gave you $\sigma^2 = 1$.

Then, to compute the likelihood:

$$\mathcal{L}[\theta|X_1, ..., X_n] = f_{X_1,...,X_n}[X_1, ..., X_n|\theta]$$
$$= \prod_{i=1}^{n} f_{X_i}[x_i|\theta]$$

where each $f_{X_i}[x_i|\theta]$ is a Gaussian! So we know that:

$$f_{X_i}[x_i|\theta] = \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2}$$

for each **known data point** x_i. Therefore, again the **only unknown is** θ. To maximize the product, we can again use **log likelihood**:

$$\log\mathcal{L}[\theta|X_1, ...X_n] = n\log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2$$

which then the solution for **solving for stationary point of** θ is then:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Lastly you can do as an exercise that the MLE estimate for data from Exp$[\theta]$ will have:

$$\hat{\theta} = \frac{n}{\sum\limits_{n} x_i}$$

***Example*: MLE for Uniform Distrib**

This example is interesting because it involves indicator functions, so you cannot easily take derivatives and solve for maximum likelihood.

Consider n samples taken from Unif$[0, \theta]$, so:

$$X_1, ..., X_n \sim \text{Unif}[0, \theta]$$

Now, we want to find $\hat{\theta}$ using MLE.

Solution: First we do standard procedure of figuring out the likelihood:

$$\begin{aligned}
\mathcal{L}[\theta|X] &= P[X_1, ..., X_n|\theta] \\
&= \prod_{i=1}^{n} P[X_i|\theta] \\
&= \prod_{i=1}^{n} \frac{1}{\theta} \cdot \mathbb{1}\{0 \leq X_i \leq \theta\} \\
&= \frac{1}{\theta^n} \mathbb{1}\{0 \leq X_1 \leq \theta\} \cdot ... \cdot \mathbb{1}\{0 \leq X_n \leq \theta\} \\
&= \frac{1}{\theta^n} \mathbb{1}\{0 \leq X_1 \leq \theta, ..., 0 \leq X_n \leq \theta\}
\end{aligned}$$

the indicator in the third equality basically comes from the fact that our uniform distribution is zero outside the range.

Now, we can order the element $X_1, ..., X_n$ such that $X_1 = \min(X_1, ..., X_n)$ and $X_n = \max(X_1, ..., X_n)$. This is useful since the indicator can then be simplified as:

$$\mathbb{1}\{0 \leq X_1 \leq \theta, ..., 0 \leq X_n \leq \theta\} = \mathbb{1}\{0 \leq X_1, X_n \leq \theta\}$$

Hence the likelihood becomes:

$$\mathcal{L}[\theta|X] = \frac{1}{\theta^n} \mathbb{1}\{0 \leq X_1, X_n \leq \theta\}$$

which is largest if θ is small. But since the smallest we can use is $X_n = \theta$ (otherwise indicator gives 0), so we conclude with:

$$\arg\max_{\theta} \mathcal{L}[\theta|X] = X_n = \max\{X_1, ..., X_n\}$$

Hence:

$$\hat{\theta} = \max\{X_1, ..., X_n\}$$

MLE can also be easily generalized with multiple parameters to solve. Essentially, consider some distribution with k parameters:

$$X_1, ..., X_n \sim f_X(X; \vec{\theta}), \quad \vec{\theta} \in \mathbb{R}^k$$

Then, essentially you will get:

$$\mathcal{L}[\vec{\theta}|X_1, ..., X_n] = \prod_{i=1}^{n} f_{X_i}[x_i|\vec{\theta}]$$

Notice that now, we have a $\vec{\theta}$ vector as input. Then, considering log likelihood and taking **partial derivatives** to solve for $\vec{\theta} = [\theta_1, ..., \theta_k]^T$:

$$\begin{aligned}
\frac{\partial}{\partial \theta_1} \log \mathcal{L} &= 0 \\
\frac{\partial}{\partial \theta_2} \log \mathcal{L} &= 0 \\
&\vdots \\
\frac{\partial}{\partial \theta_k} \log \mathcal{L} &= 0
\end{aligned}$$

to obtain k equations with k unknowns and solve it. (Sometimes you cannot take derivatives as the previous example shows. Then you need to analyze it a different way).

11 Linearity of Expected Value

The most important theorem is basically this:

Theorem 11.1 (Linearity of Expectation)**.** *For any n random variables $X_1, X_2, ..X_n$, which **may or may not be independent**, and with $c_1, c_2, ..., c_n$ be n coefficients, the following holds true:*

$$\mathbb{E}\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i \mathbb{E}[X_i]$$

This is so useful as a common example would be:

Corollary 11.1.1. *Let X, Y be two random variable which may or may not be independent. Then, from the above theorem:*

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

The proof for the theorem can be done once you prove the above corollary, and then using induction to prove for n random variables.

Theorem 11.2 (Linearity of Variance)**.** *Sadly, for variance we need all the variables are **independent**. So that given n independent random variables $X_1, ..., X_n$ with $c_1, ..., c_n$ coefficients:*

$$Var\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i^2 \, Var[X_i]$$

Corollary 11.2.1. *Then an easy example would be to consider X, Y being independent, then:*

$$Var[c_1 X + c_2 Y] = c_1^2 \, Var[X] + c_2^2 \, Var[Y]$$

The corollary can be proved by using the fact that $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, and then using induction for n independent random variables.

12 Review

This section contains some preliminary knowledge you should keep in mind.

12.1 Set Laws

All the laws below are true regardless of symmetry of event, since they are set laws. They can be visualized easily using Venn Diagram.

Theorem 12.1 (Associative Law)**.**

$$(A \cup B) \cup C = A \cup (B \cup C)$$
$$(A \cap B) \cap C = A \cap (B \cap C)$$

Theorem 12.2 (Distributive Law)**.**

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$