

Probability Theory II

CS771: Introduction to Machine Learning

Purushottam Kar

Announcements

2

Assignment 1 deadline extended: Sat 07 Sept 2019, 9:59 PM IST

Applies to both PDF and code submission

Quiz: August 30 (Friday), 6PM, **L20 – same as quiz 1**

Assigned seating – don't be late (will waste time finding your seat)

Syllabus is till whatever we cover today i.e. Aug 28 (Wed)

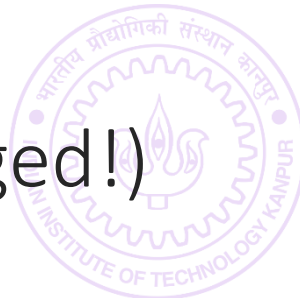
*Bring your **institute ID card** with you – will lose time if you forget*

*Bring a **pencil, pen, eraser, sharpener** with you – we wont provide!*

*Answers to be written on question paper itself. If you write with pen and make a mistake, no extra paper. Final answer **must be in pen***

***Auditors cannot appear** for quiz – please come to L20 at ~ 6:40PM*

Doubt clearing session: Aug 29 (Thu), 6PM **KD101** (venue changed!)



Recap of Last Lecture

3

Probability as Proportions (to get started and develop intuition)

Probability Mass Function (gives prob. of an r.v. taking some value)

Joint Probability (gives prob. of two or more r.v.s taking blah values)

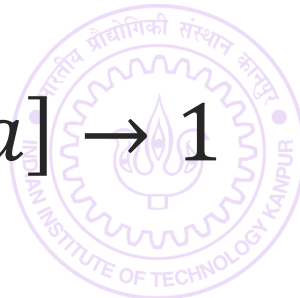
Marginal Probability (gives prob. of a subset of r.v.s taking blah values)

Conditional Probability (gives prob. of one r.v. taking blah value given that we know that another r.v. already taken some value)

Joint/Marginal/Conditional are all proper prob dist and have PMFs

Rules of Probability: Sum, Product, Chain, Bayes Theorem

Use of probability in RecSys: find a so that $\mathbb{P}[P \neq -1 \mid x, y, z, a] \rightarrow 1$



Creating Events from Random Variables?? 4

In the RecSys example, we saw the expression $\mathbb{P}[P \neq -1 \mid x, y, z, a]$

Here $\{P \neq -1\}$ is an event (that of the user purchasing something) and we are interested in the probability of this event given x, y, z, a

Recall: events are merely a description of interesting facts about an outcome

Similarly $\{X = 1 \wedge Y = 2\}$ is also an event (that the ball we picked is green and has the number 1 stamped on it)

$\{X = 1\}$ is also an event (that the ball has the number 1 stamped on it)

$\{Y = 2\}$ is also an event (that the ball is green colored)

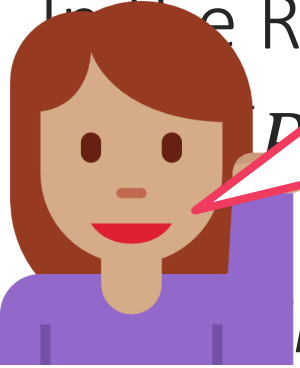
Given an event, it may happen on certain outcomes, not happen on others

e.g. if $\{Y = 2\}$ then this event will not have taken place if we pick a blue ball

Thus, given any collection of random variables, we can create events out of them and ask interesting questions about the r.v.s



Creating Events from Random Variables?? 5



We can also have events like $\{X < x_0\}$ (that the random variable takes a value less than x_0). We define $\mathbb{P}[X < x_0] \triangleq \sum_{\omega: X(\omega) < x_0} p_{\omega}$.
Similarly, $\{X \leq x_0\}$, $\{X \geq x_0\}$ are also valid events

are interested in the probability of this event given x, y, z, a
!: events are merely a description of interesting facts about an outcome

Similarly $\{X = 1 \wedge Y = 2\}$ is also an event (that the ball we picked is green and has the number 1 stamped on it)

$\{X = 1\}$ is also an event (that the ball has the number 1 stamped on it)

$\{Y = 2\}$ is also an event (that the ball is green colored)

*Given an event, it may happen on certain outcomes, not happen on others
e.g. if $\{Y = 2\}$ then this event will not have taken place if we pick a blue ball*

Thus, given any collection of random variables, we can create events out of them and ask interesting questions about the r.v.s



Event Calculus

6

Let A, B be events (possibly defined using same/different r.v.s etc)

$\neg A$ is also an event

Called the negation or complement of A (A did not happen)

$A \cup B$ is also an event

Called the union of the two events (either A happened or B happened or both happened)

$A \cap B$ is also an event

Called the intersection of the two events (both A and B happened)

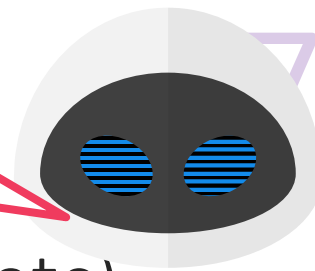
De Morgan's Laws: for any two events A, B , we must have

$$\neg(A \cup B) = \neg A \cap \neg B \text{ as well as } \neg(A \cap B) = \neg A \cup \neg B$$



Event Calculus

The term *calculus* in general, means a system of rules – no differentiation or integration going on here 😊

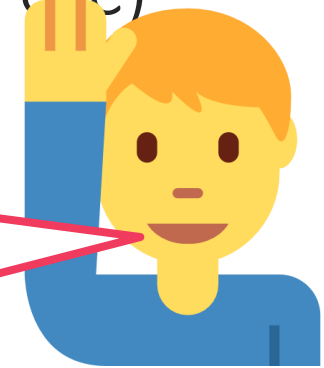


Let A, B be events (possibly defined using same/different r.v.s etc)

$\neg(A \cup B) = \neg A \cap \neg B$ tell us that saying

"It is not the case that either A happened or B happened or both happened"
is just a funny way of saying

"A did not happen and B did not happen i.e. neither happened"

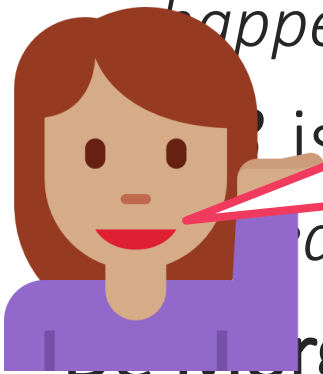


Called the
happened

$\neg(A \cap B) = \neg A \cup \neg B$ tell us that saying

"It is not the case that both A and B happened"
is just another way of saying

"Either A did not happen or B did not happen or both did not happen"

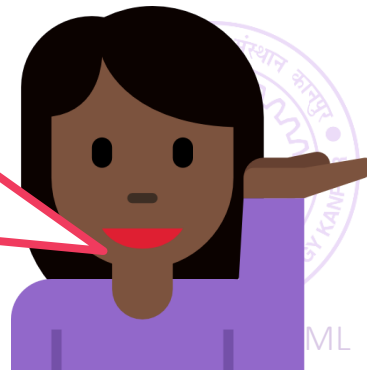


and the intersection of the two events (both A and B happened)

De Morgan's Law

$$\neg(A \cup B) = \neg A \cap \neg B$$

This means we can be more creative in defining events,
e.g. $\{X \leq x_0 \wedge X \geq x_1\}$ (also written as $\{x_1 \leq X \leq x_0\}$)
or else $\{x_1 \leq X \leq x_0\} \vee \{x_3 \leq X \leq x_2\}$



Event Calculus

8

Example: let $A \equiv \{X = 2\}$, $B \equiv \{Y = 1\}$ be events

$\neg A$ is also an event (number on ball is something other than 2)

$\neg B$ is also an event (the colour of the ball is something other than red)

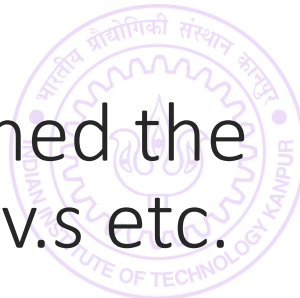
$A \cup B$ is also an event (either I have a red ball or a ball with number 2 written on it or else a red ball with number 2 written on it)

$A \cap B$ is also an event (I have a red ball and the number on it is 2)

De Morgan's Laws: for any two events A, B , we must have

$$\neg(A \cup B) = \neg A \cap \neg B \text{ as well as } \neg(A \cap B) = \neg A \cup \neg B$$

Caution: De Morgan's Laws **always** hold no matter how we defined the events. They do not require events to be defined on separate r.v.s etc.



Rules of Probability for Complex Events

9

Suppose we know probability of two events $\mathbb{P}[A], \mathbb{P}[B]$

Complement Rule: $\mathbb{P}[\neg A] = 1 - \mathbb{P}[A]$

Union Rule: $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$

These rules can be proved using a similar proof technique that we used for joint/marginal probability derivations in the last lecture

The above allow us to calculate more interesting probabilities

$$\mathbb{P}[X = 1 \vee Y = 2] = \mathbb{P}[X = 1] + \mathbb{P}[Y = 2] - \mathbb{P}[X = 1 \wedge Y = 2]$$

We used only the marginal and the joint probability distributions

$$\mathbb{P}[X \neq 1] = 1 - \mathbb{P}[X = 1]$$

The above rules hold even for conditional probability



We can derive an “**intersection rule**” using de-Morgan’s laws and these rules

$$\begin{aligned}\mathbb{P}[A \cap B] &= \mathbb{P}[\neg(\neg(A \cap B))] = 1 - \mathbb{P}[\neg(A \cap B)] = 1 - \mathbb{P}[\neg A \cup \neg B] \\ &= \mathbb{P}[A] + \mathbb{P}[B] + \mathbb{P}[\neg A \cap \neg B] - 1 \text{ (apply union and complement rules)}\end{aligned}$$

Suppose we know probability of two events X and Y

Complement Rule: $\mathbb{P}[\neg A] = 1 - \mathbb{P}[A]$

Union Rule: $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$

These rules can be proved using a similar proof technique that we used for joint/marginal probability derivations in the last lecture

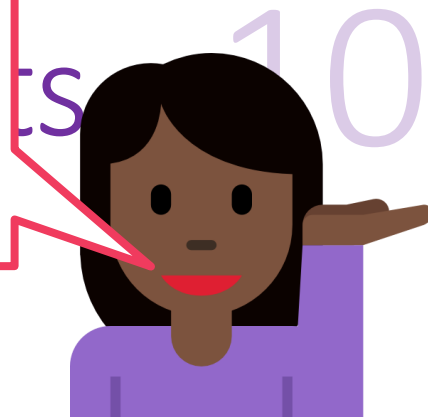
The above allow us to calculate more interesting probabilities

$$\mathbb{P}[X = 1 \vee Y = 2] = \mathbb{P}[X = 1] + \mathbb{P}[Y = 2] - \mathbb{P}[X = 1 \wedge Y = 2]$$

We used only the marginal and the joint probability distributions

$$\mathbb{P}[X \neq 1] = 1 - \mathbb{P}[X = 1]$$

The above rules hold even for conditional probability



Rules of Conditional Probability

11

Let C denote an event

For example $C \equiv \{X = x_0 \wedge Y = y_0 \wedge Z = z_0 \wedge A = a_0\}$

A user with gender x_0 , age y_0 spent z_0 sec on website and was shown ad a_0

Let A, B denote two events

*A, B may be related/unrelated to C – does not matter – the rules **always** hold*

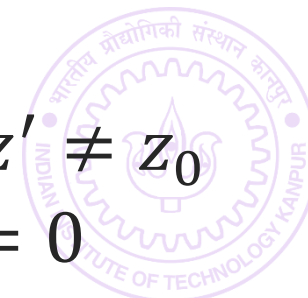
Conditional Complement Rule: $\mathbb{P}[\neg A \mid C] = 1 - \mathbb{P}[A \mid C]$

Conditional Union Rule: $\mathbb{P}[A \cup B \mid C] = \mathbb{P}[A \mid C] + \mathbb{P}[B \mid C] - \mathbb{P}[A \cap B \mid C]$

Conditional Implication Rule: If $C \Rightarrow A$, then $\mathbb{P}[A \mid C] = 1$. On the other hand, if $C \Rightarrow \neg B$, then $\mathbb{P}[B \mid C] = 0$

Example: C defined as above and $A \equiv \{X = x_0\}$ and $B \equiv \{Z = z'\}, z' \neq z_0$

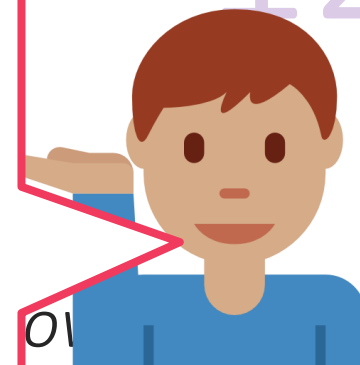
In the above cases, we will indeed have $\mathbb{P}[A \mid C] = 1$ and $\mathbb{P}[B \mid C] = 0$



In fact, Bayes Theorem applies to events just as well i.e. if A, B are events, then

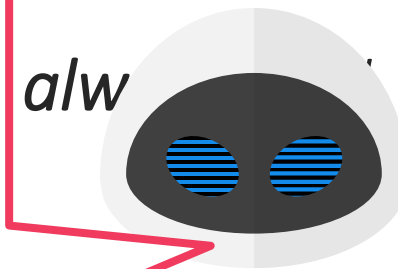
$$\mathbb{P}[A | B] = \frac{\mathbb{P}[B | A] \cdot \mathbb{P}[A]}{\mathbb{P}[B]}$$

Proof: create a random variable for each event e.g. $M = 1$ if event A happens and $M = 0$ if A does not happen. Similarly define a random variable N to tell us whether B happened or not. Now use the Bayes theorem on M, N – done!



The random variables M, N we defined above to tell us whether some event happened or not are called *indicator random variables* since they *indicate* whether an event took place (in which case the r.v. takes value 1) or not (in which case the r.v. takes value 0). **Notation:** $M = \mathbb{I}\{A\}, N = \mathbb{I}\{B\}$.

In general, $\mathbb{I}\{\text{blah}\} = 1$ if blah is true else 0



$$\mathbb{P}[A \cap B | C]$$

Conditional Implication Rule: If $C \Rightarrow A$, then $\mathbb{P}[A | C] = 1$. On the other hand, if $C \Rightarrow \neg B$, then $\mathbb{P}[B | C] = 0$.

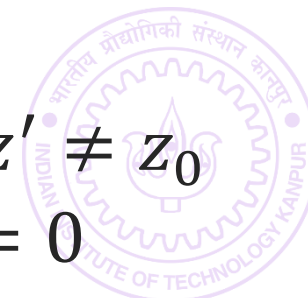
Example: C defined as above and

In the above cases, we will indeed have

Caution: not a standard rule you would find in textbooks

$$\{Z = z'\}, z' \neq z_0$$

$$\mathbb{P}[B | C] = 0$$



Independence of Random Variables

13

Two r.v.s X, Y are said to be independent if, for all $x \in S_X$ and $y \in S_Y$, we have $\mathbb{P}[X = x, Y = y] = \mathbb{P}[X = x] \cdot \mathbb{P}[Y = y]$

We can show that if two r.v.s are independent, then this means that the value one r.v. takes does not influence the other r.v. to take some value more preferentially than others in any way

Proof:
$$\mathbb{P}[X = x \mid Y = y] = \frac{\mathbb{P}[X=x \mid Y=y]}{\mathbb{P}[Y=y]} = \frac{\mathbb{P}[X=x] \cdot \mathbb{P}[Y=y]}{\mathbb{P}[Y=y]} = \mathbb{P}[X = x]$$

Similarly, we also get $\mathbb{P}[Y = y \mid X = x] = \mathbb{P}[Y = y]$

Independence of r.v.s makes life extremely simple in ML algorithms but is very precious – not always available

Notation: If X, Y are independent, then we often write $X \perp\!\!\!\perp Y$



Independence of Random Variables

14

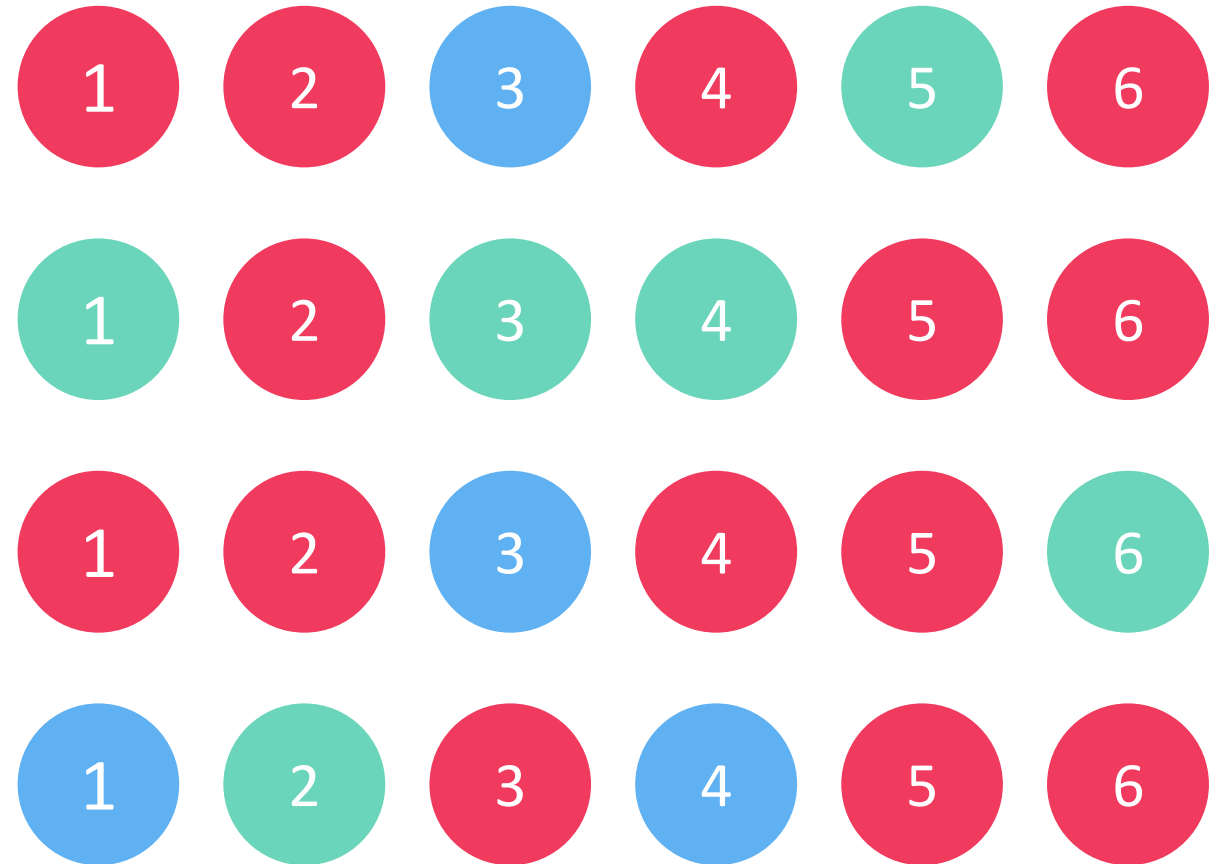
Two r.v.s X, Y are said to be independent if, for all $x \in S_X$ and $y \in S_Y$, we have $\mathbb{P}[X = x, Y = y] = \mathbb{P}[X = x] \cdot \mathbb{P}[Y = y]$

We have $\mathbb{P}_Y = \left[\frac{7}{12}, \frac{3}{12}, \frac{2}{12}\right]$

However, if we condition on $X = 1$, then we have

$\mathbb{P}_{Y|X=1} = \left[\frac{2}{4}, \frac{1}{4}, \frac{1}{4}\right]$ i.e. $X \not\perp Y$

For independence, we should have had $\mathbb{P}_{Y|X=x_0} = \mathbb{P}_Y$ for all $x_0 \in S_X$



Independence of Random Variables

15

Two r.v.s X, Y are said to be independent if, for all $x \in S_X$ and $y \in S_Y$, we have $\mathbb{P}[X = x, Y = y] = \mathbb{P}[X = x] \cdot \mathbb{P}[Y = y]$

We have $\mathbb{P}_Y = \left[\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right]$

We can verify that we have

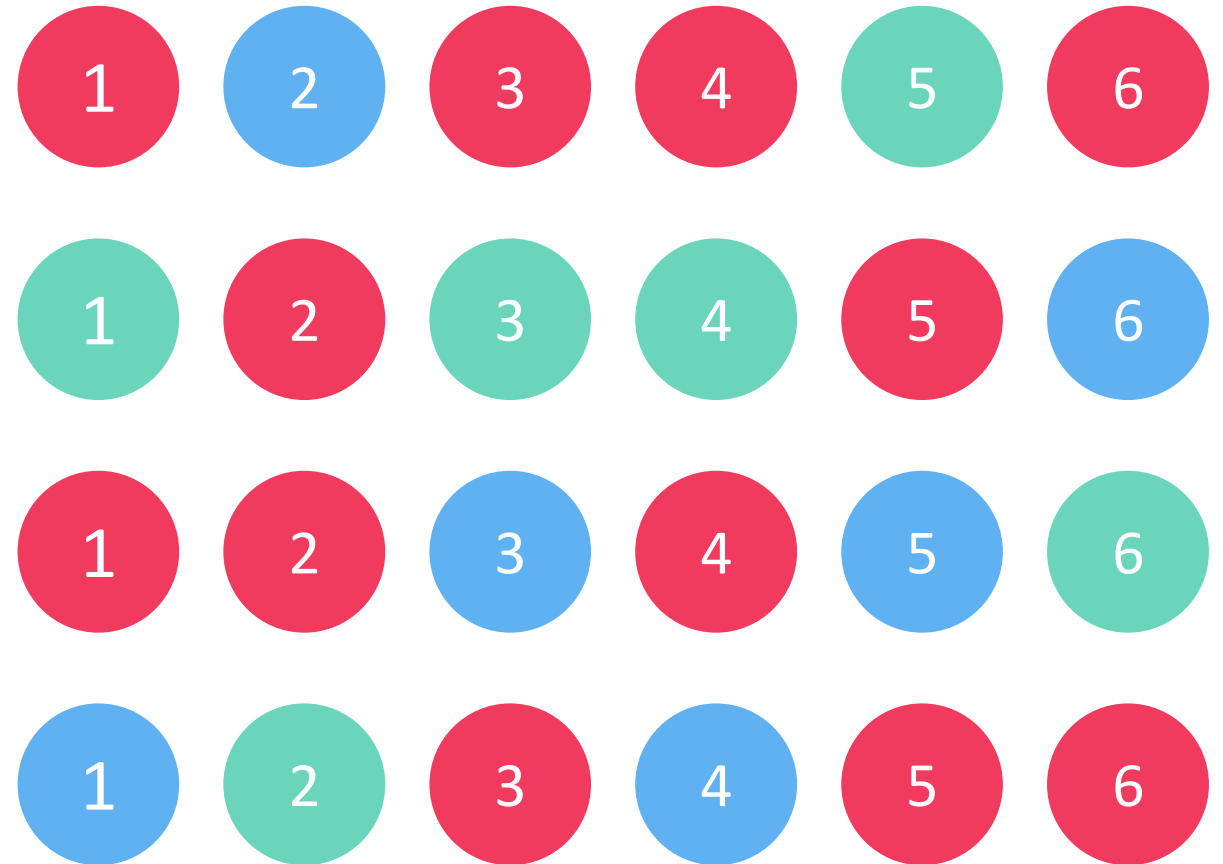
$\mathbb{P}_{Y|X=x_0} = \left[\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right] = \mathbb{P}_Y$ for

all $x_0 \in [6] = S_X$ i.e. $X \perp\!\!\!\perp Y$

Similarly, we can verify that

$\mathbb{P}_{X|Y=y_0} = \left[\frac{1}{6}, \frac{1}{6}, \dots, \frac{1}{6}\right] = \mathbb{P}_X$ for

all $y_0 \in [3] = S_Y$ i.e. $Y \perp\!\!\!\perp X$



Independence

$X \perp\!\!\!\perp Y$ means that X and Y will both take values according to their own (marginal) PMFs, happily unmindful of the values the other r.v. is taking!

Two r.v.s X, Y are said to be independent if

$$\text{we have } \mathbb{P}[X = x, Y = y] = \mathbb{P}[X = x] \cdot \mathbb{P}[Y = y]$$

Can you show that if $X \perp\!\!\!\perp Y$ (i.e. $\mathbb{P}_{Y|X=x_0} = \mathbb{P}_Y$ for all $x_0 \in S_X$) then we must always have $Y \perp\!\!\!\perp X$ (i.e. $\mathbb{P}_{X|Y=y_0} = \mathbb{P}_X$ for all $y_0 \in S_Y$) as well?

Hint: use the definition of independence

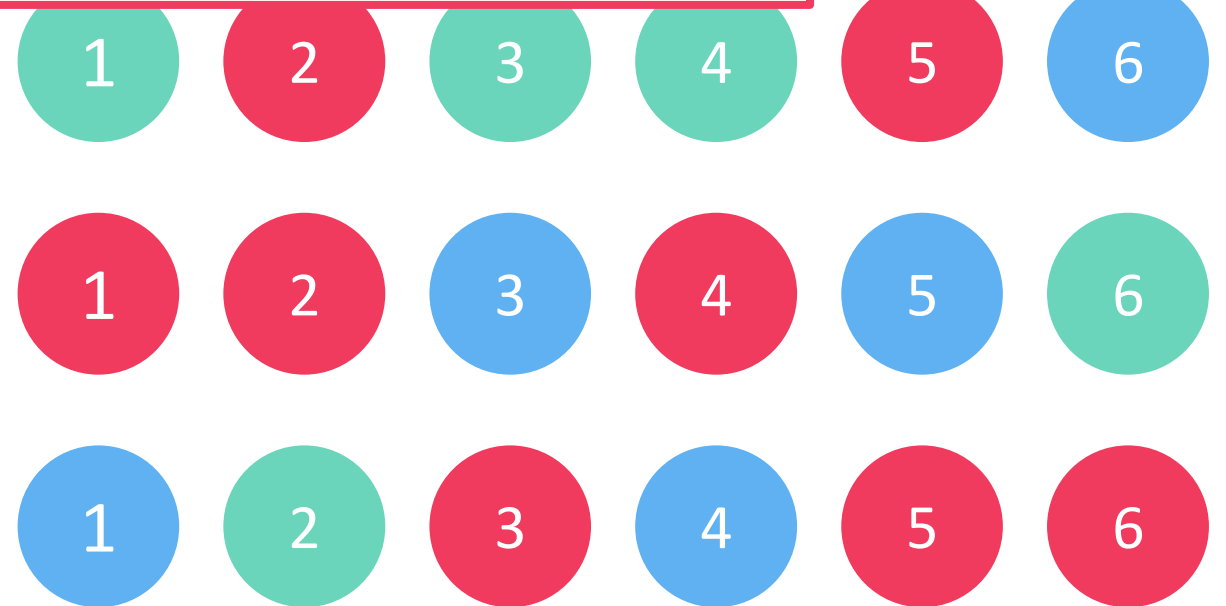
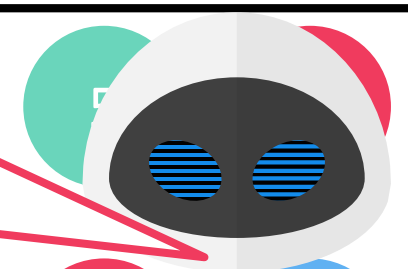
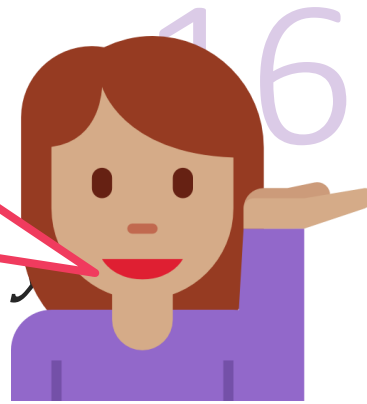
$$\mathbb{P}_{Y|X=x_0} = \left[\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right] = \mathbb{P}_Y \text{ for}$$

all $x_0 \in [6] = S_X$ i.e. $X \perp\!\!\!\perp Y$

Similarly, we can verify that

$$\mathbb{P}_{X|Y=y_0} = \left[\frac{1}{6}, \frac{1}{6}, \dots, \frac{1}{6}\right] = \mathbb{P}_X \text{ for}$$

all $y_0 \in [3] = S_Y$ i.e. $Y \perp\!\!\!\perp X$



Conditional Independence

17

If X, Y, Z are three r.v.s such that for all $x \in S_X, y \in S_Y, z \in S_Z$ we have

$$\mathbb{P}[X = x, Y = y \mid Z = z] = \mathbb{P}[X = x \mid Z = z] \cdot \mathbb{P}[Y = y \mid Z = z]$$

then we say that X and Y are *conditionally independent* given Z

Notation: $X \perp\!\!\!\perp Y \mid Z$

If X, Y are independent, then it is not necessary that they continue to be independent even if conditioned on a third random variable



Conditional Independence

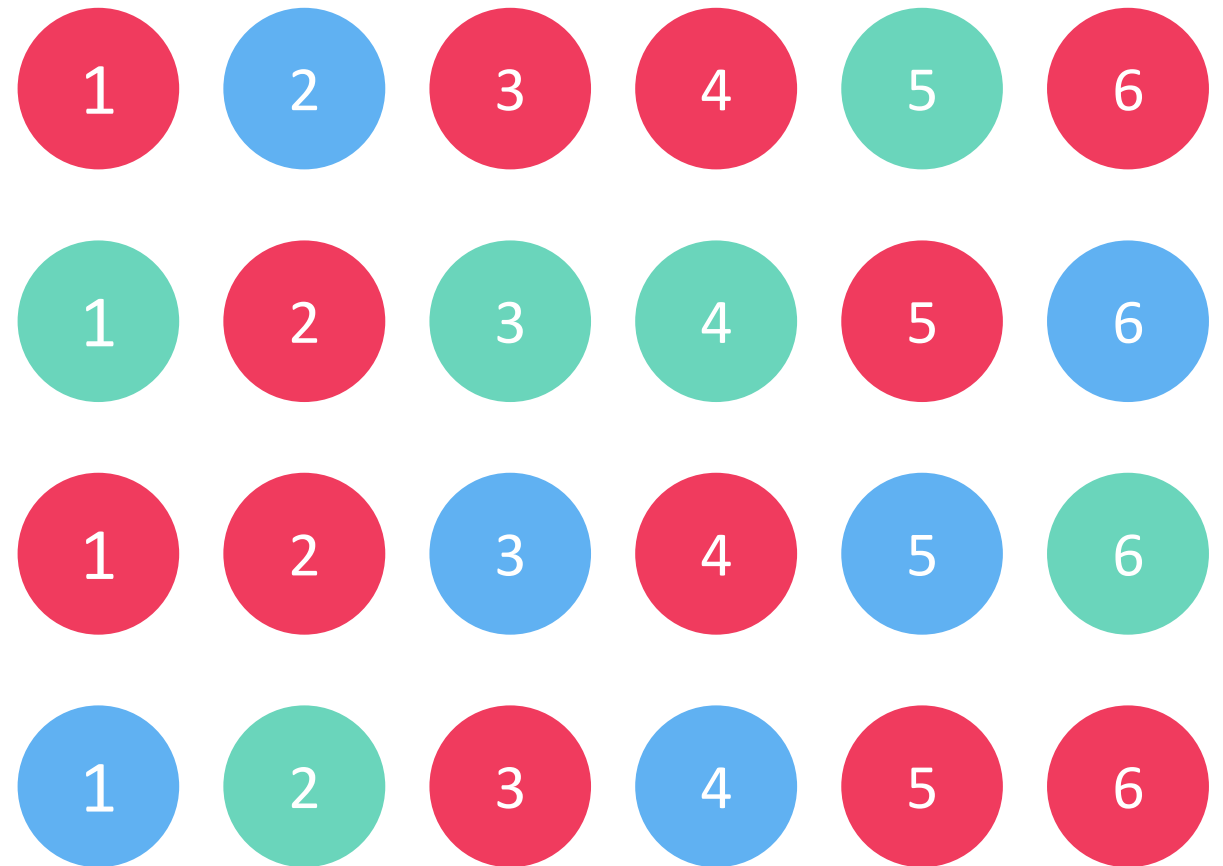
18

If X, Y, Z are three r.v.s such that for all $x \in S_X, y \in S_Y, z \in S_Z$ we have

$$\mathbb{P}[X = x, Y = y \mid Z = z] = \mathbb{P}[X = x \mid Z = z] \cdot \mathbb{P}[Y = y \mid Z = z]$$

This is the earlier example where we verified that $X \perp\!\!\!\perp Y$ and $Y \perp\!\!\!\perp X$. However it is easy to see that we do not have $X \perp\!\!\!\perp Y \mid Z$ since

$$\mathbb{P}[X = 2, Y = 2 \mid Z = 1] = 0 \neq \mathbb{P}[X = 2 \mid Z = 1] \cdot \mathbb{P}[Y = 2 \mid Z = 1]$$



Conditional Independence

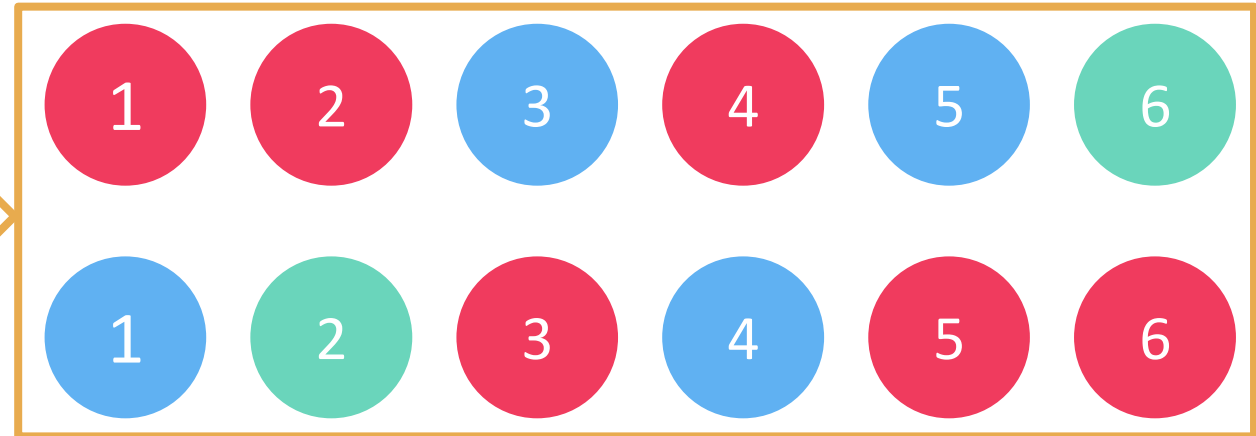
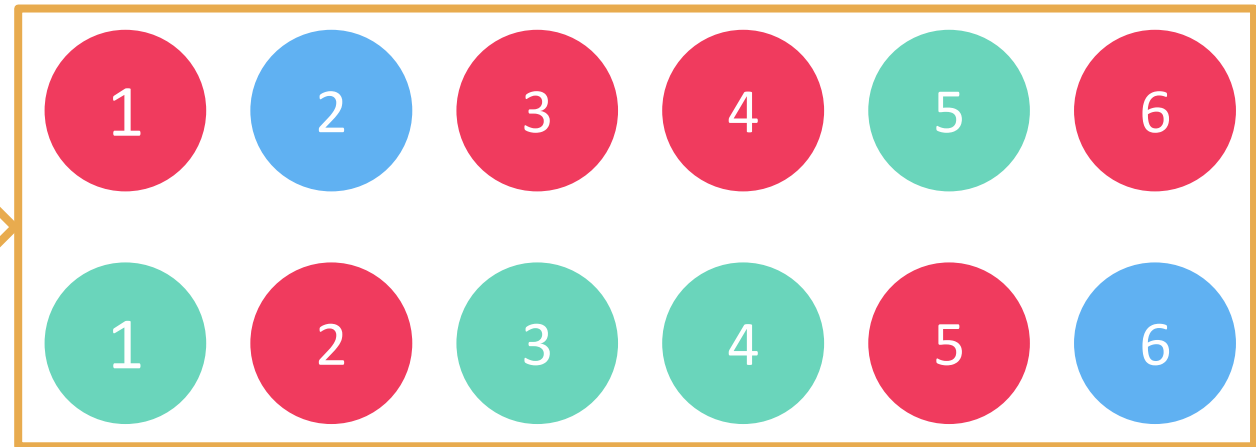
19

If X, Y, Z are three r.v.s such that for all $x \in S_X, y \in S_Y, z \in S_Z$ we have

$$\mathbb{P}[X = x, Y = y | Z = z] = \mathbb{P}[X = x | Z = z] \cdot \mathbb{P}[Y = y | Z = z]$$

This is the earlier example where we verified that $X \perp\!\!\!\perp Y$ and $Y \perp\!\!\!\perp X$. However it is easy to see that we do not have $X \perp\!\!\!\perp Y | Z$ since

$$\mathbb{P}[X = 2, Y = 2 | Z = 1] = 0 \neq \mathbb{P}[X = 2 | Z = 1] \cdot \mathbb{P}[Y = 2 | Z = 1]$$



Conditional Independence

20

If X, Y, Z are three r.v.s such that for all $x \in S_X, y \in S_Y, z \in S_Z$ we have

$$\mathbb{P}[X = x, Y = y \mid Z = z] = \mathbb{P}[X = x \mid Z = z] \cdot \mathbb{P}[Y = y \mid Z = z]$$

then we say that X and Y are *conditionally independent* given Z

Notation: $X \perp\!\!\!\perp Y \mid Z$

If X, Y are independent, then it is not necessary that they continue to be independent even if conditioned on a third random variable

Even if X, Y are not independent, it is still possible that there may exist a third random variable Z such that makes X, Y conditionally independent i.e. $X \perp\!\!\!\perp Y \mid Z$



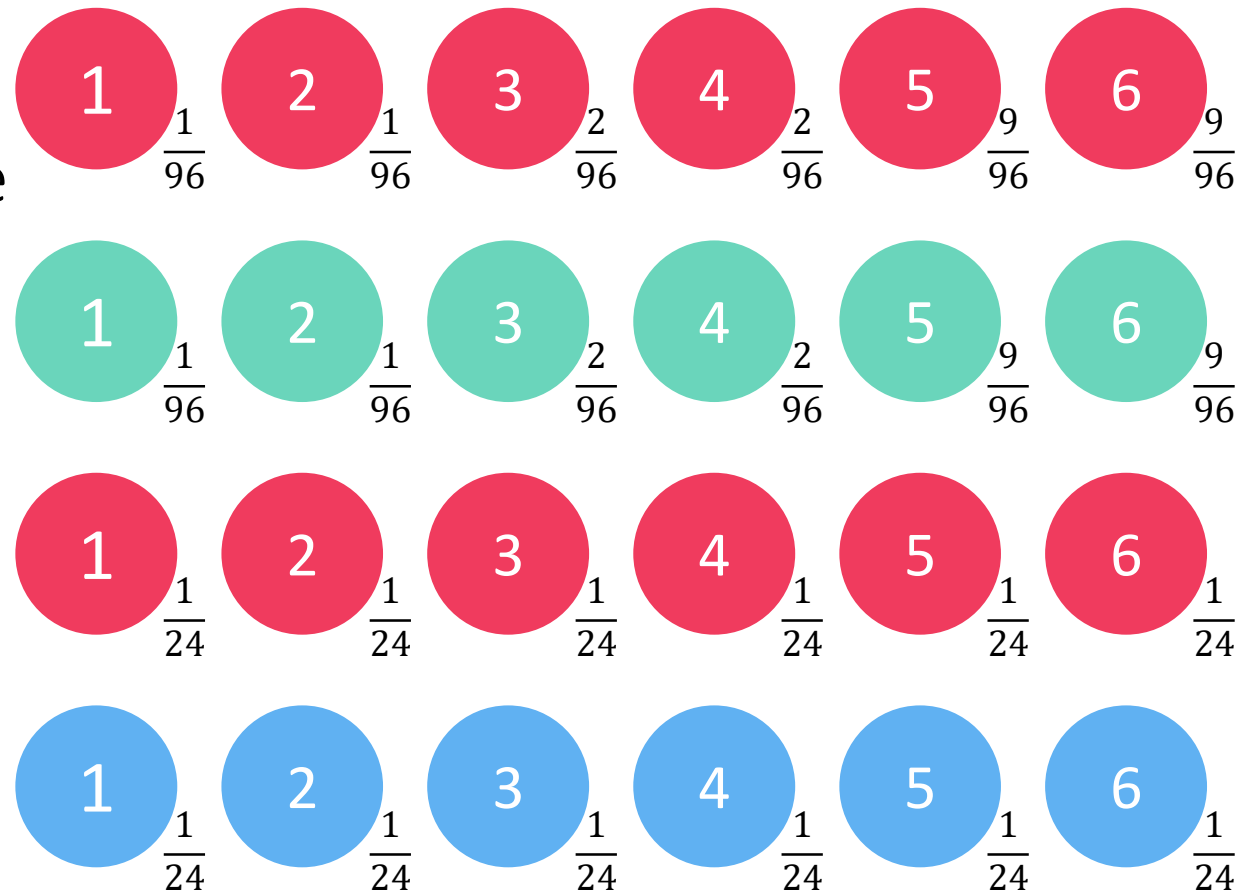
Conditional Independence

21

If X, Y, Z are three r.v.s such that for all $x \in S_X, y \in S_Y, z \in S_Z$ we have

$$\mathbb{P}[X = x, Y = y \mid Z = z] = \mathbb{P}[X = x \mid Z = z] \cdot \mathbb{P}[Y = y \mid Z = z]$$

The r.v.s number X and colour Y are not independent here. To see this, note that $\mathbb{P}[Y = 3 \mid X = 1] = 0.4 \neq 0.25 = \mathbb{P}[Y = 3]$. However, if we condition on a new random variable Z that distinguishes the first two rows from the last two rows, then X and Y become independent r.v.s i.e. $X \perp\!\!\!\perp Y$ but we do have $X \perp\!\!\!\perp Y \mid Z$



Conditional Independence

22

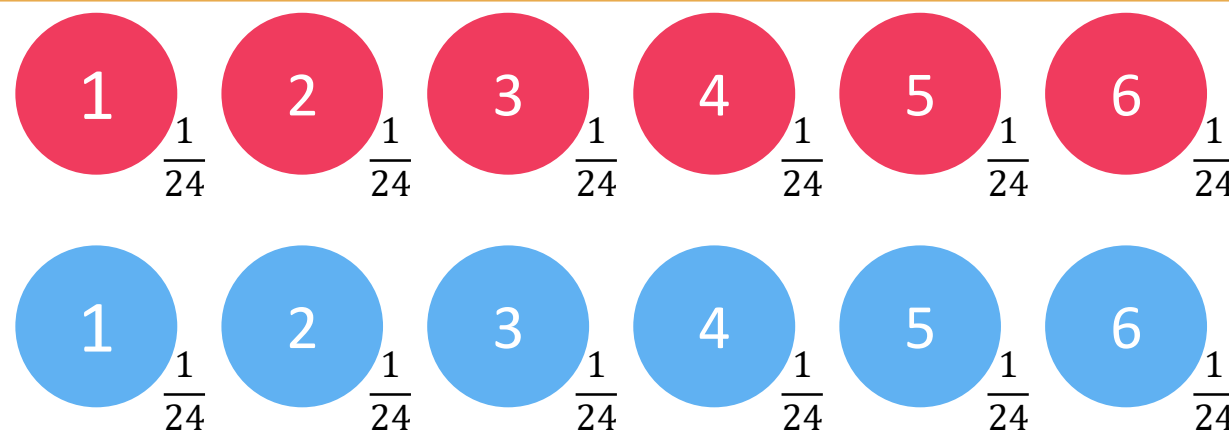
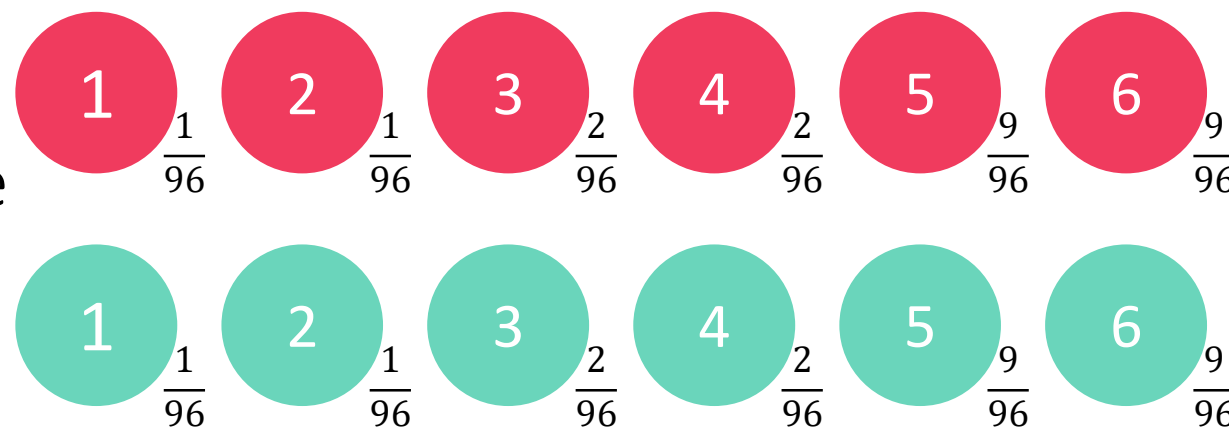
If X, Y, Z are three r.v.s such that for all $x \in S_X, y \in S_Y, z \in S_Z$ we have

$$\mathbb{P}[X = x, Y = y | Z = z] = \mathbb{P}[X = x | Z = z] \cdot \mathbb{P}[Y = y | Z = z]$$

The r.v.s number X and colour Y are not independent here. To note that $\mathbb{P}[Y = 3 | X = 1] = \frac{1}{6} \neq 0.25 = \mathbb{P}[Y = 3]$. However, if we condition on a new random variable Z that distinguishes the first two rows from the last two rows, they become independent r.v.s but we do have $X \perp\!\!\!\perp Y | Z$

$Z = 1$

$Z = 2$



Conditional Independence

23

Slightly more “practical” examples

If I throw a fair dice twice, the outcome on the first throw in no way influences the second outcome i.e. two outcomes are independent of each other and each takes values in $[6]$. However, if I additionally am told that the sum of the two numbers is 8, then the throws are no longer conditionally independent since for example we have

$$\mathbb{P}[X = 1, Y = 4 \mid Z = 8] = 0 \neq \mathbb{P}[X = 1 \mid Z = 8] \cdot \mathbb{P}[Y = 4 \mid Z = 8]$$

In the above, X is number on first throw, Y denotes second, Z is sum

Sample space is $[6] \times [6]$ i.e. all possible outcomes of two throws

Examples where non-independent r.v.s become conditionally independent are most commonly found in a branch of ML called graphical models.



Expectation of a Random Variable

24

The expectation of a random variable or its *expected value* is the mean or average value that random variable takes and is defined as

$$\mathbb{E}[X] = \sum_{x \in S_X} x \cdot \mathbb{P}[X = x]$$

Sometimes the notation used is just $\mathbb{E}X$ i.e. brackets are omitted

The name suggests that the r.v. is expected to take this value

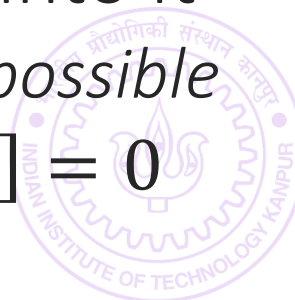
Some truth to this: if we sample from \mathbb{P}_X , “likely” to get a value “close” to $\mathbb{E}X$

What “close”, “likely” mean are topics in a learning theory course (e.g. CS777)

However, can be misleading – be careful not to read too much into it

$\mathbb{E}X$ need not be most likely value for X i.e. $\mathbb{E}X \neq \arg \max \mathbb{P}[X = x]$ possible

In fact, there are r.v. X which can never take this value i.e. $\mathbb{P}[X = \mathbb{E}X] = 0$



Rules of Expectation: Sum Rule

25

Linearity of Expectation: given two r.v. X, Y , no matter how they are defined, no matter whether independent or not, we always have

$$\mathbb{E}[X + Y] = \mathbb{E}X + \mathbb{E}Y$$

Proof: Let $Z \triangleq X + Y$ be a new r.v.. We have $\mathbb{E}[Z] = \sum_{z \in S_Z} z \cdot \mathbb{P}[Z = z]$. Now the only possible values for z are of the form $(x + y)$ where $x \in S_X, y \in S_Y$.

Thus, we have $\mathbb{E}Z = \sum_{x \in S_X} \sum_{y \in S_Y} (x + y) \cdot \mathbb{P}[X = x, Y = y]$. Note that even if multiple ways of getting a value z , all have been taken into account.

$$\begin{aligned} \sum_x \sum_y (x + y) \cdot \mathbb{P}[x, y] &= \sum_x \sum_y x \cdot \mathbb{P}[x, y] + \sum_x \sum_y y \cdot \mathbb{P}[x, y] \\ &= \sum_x x \sum_y \mathbb{P}[x, y] + \sum_y y \sum_x \mathbb{P}[x, y] = \sum_x x \mathbb{P}[x] + \sum_y y \mathbb{P}[y] = \mathbb{E}X + \mathbb{E}Y \end{aligned}$$

Note that the only result we used in our proof is the law of total probability in the second last step above which always holds no matter which r.v.s we have

Note: the same proof shows that $\mathbb{E}[X - Y] = \mathbb{E}X - \mathbb{E}Y$



Rules of Expectation: Scaling Rule

26

Given a r.v. X and a constant c , define a new r.v. $Y = c \cdot X$ i.e. on any outcome $\omega \in \Omega$, $Y(\omega) = c \cdot X(\omega)$, then $\mathbb{E}Y = c \cdot \mathbb{E}X$

Proof: any value y that Y takes is cx for some $x \in S_X$. Thus, we get
$$\mathbb{E}Y = \sum_{y \in S_Y} y \cdot \mathbb{P}[Y = y] = \sum_{x \in S_X} cx \cdot \mathbb{P}[X = x] = c \cdot \mathbb{E}X$$

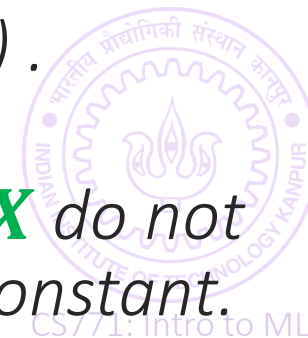
The expectation of a constant random variable is the constant itself

If we have a r.v. Z that always gives the same value c no matter what the outcome, then $\mathbb{P}[Z = c] = 1$ and $\mathbb{P}[Z = c'] = 0$ for all $c \neq c'$. Then $\mathbb{E}Z = c$

For any r.v. X , we always have $\mathbb{E}[X - \mathbb{E}X] = 0$

Proof: Create a dummy random variable Z that always takes the value $\mathbb{E}X$. Note that $\mathbb{E}X$ is a constant (does not depend on the outcome $\omega \in \Omega$). Linearity gives us $\mathbb{E}[X - Z] = \mathbb{E}X - \mathbb{E}Z = \mathbb{E}X - \mathbb{E}X = 0$

Note: notation is horrible here. In the expression $\mathbb{E}[\textcolor{red}{X} - \textcolor{green}{\mathbb{E}X}]$, $\textcolor{red}{X}$ and $\textcolor{green}{X}$ do not refer to two r.v.s or the same r.v. repeated. Instead, just read $\mathbb{E}X$ as constant.



Rules of Expectation: Scaling Rule

27

Given a r.v. X and a constant c , define a new r.v. $Y = c \cdot X$ i.e. on any outcome $\omega \in \Omega$, $Y(\omega) = c \cdot X(\omega)$, then $\mathbb{E}Y = c \cdot \mathbb{E}X$

Proof: any value y that Y takes is cx for some $x \in S_X$. Thus, we get

$$\mathbb{E}Y = \sum_{y \in S_Y} y \cdot \mathbb{P}[Y = y] = \sum_{x \in S_X} cx \cdot \mathbb{P}[X = x] = c \cdot \mathbb{E}X$$

The $\mathbb{E}X$ is a constant that does not depend on the outcome of any toss. For example, if we have a fair coin and create a r.v. X s.t. $X = 1$ for heads and $X = 0$ for tails, then $\mathbb{E}X = 0.5$ (since coin is fair) and clearly $\mathbb{E}X$ is a constant that does not depend on the outcome of any toss.



For any r.v. X , we always have $\mathbb{E}[X - \mathbb{E}X] = 0$

Proof: Create a dummy random variable Z that always takes the value $\mathbb{E}X$.

Note that $\mathbb{E}X$ is a constant (does not depend on the outcome $\omega \in \Omega$).

Linearity gives us $\mathbb{E}[X - Z] = \mathbb{E}X - \mathbb{E}Z = \mathbb{E}X - \mathbb{E}X = 0$

Note: notation is horrible here. In the expression $\mathbb{E}[\mathbf{X} - \mathbb{E}\mathbf{X}]$, \mathbf{X} and \mathbf{X} do not refer to two r.v.s or the same r.v. repeated. Instead, just read $\mathbb{E}X$ as constant.



Rules of Expectation

28

Law of the Unconscious Statistician (LOTUS)

Helps calculate expectations for complicated random variables easily

Suppose we have random variable X whose PMF we know \mathbb{P}_X

Suppose there is a weird function $g: S_X \rightarrow \mathbb{R}$ and we define a new random variable $Y \triangleq g(X)$. Can we calculate $\mathbb{E}Y$?

Calculating $\mathbb{E}Y$ directly would require us to first get hold of \mathbb{P}_Y – difficult!

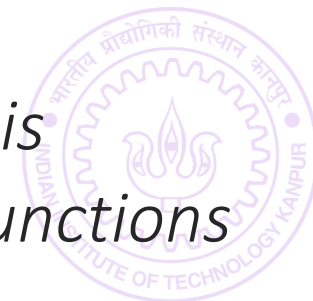
LOTUS gives us a way to use \mathbb{P}_X itself to calculate $\mathbb{E}Y$

$$\mathbb{E}Y = \mathbb{E}g(X) = \sum_{x \in S_X} g(x) \cdot \mathbb{P}[X = x]$$

Proof: *much the same way we proved linearity of expectation*

Works no matter what r.v. X we have, no matter how complicated g is

The function g does need to satisfy some very easy conditions – all functions we will look at in this course will satisfy these conditions



Rules of Expectation: Product Rule

29

If X, Y are two independent random variables, then we have stronger results on them $\mathbb{E}[X \cdot Y] = \mathbb{E}X \cdot \mathbb{E}Y$

Proof: Let $Z \triangleq XY$ be a new r.v.. We have $\mathbb{E}[Z] = \sum_{z \in S_Z} z \cdot \mathbb{P}[Z = z]$. Now the only possible values for z are of the form xy where $x \in S_X, y \in S_Y$.

Thus, we have $\mathbb{E}Z = \sum_{x \in S_X} \sum_{y \in S_Y} xy \cdot \mathbb{P}[X = x, Y = y]$. Note that even if multiple ways of getting a value z , all have been taken into account.

Using independence gives us $\mathbb{E}Z = \sum_{x \in S_X} \sum_{y \in S_Y} xy \cdot \mathbb{P}[X = x] \cdot \mathbb{P}[Y = y]$
 $= (\sum_x x \cdot \mathbb{P}[x]) \cdot (\sum_y y \cdot \mathbb{P}[y]) = \mathbb{E}X \cdot \mathbb{E}Y$

Warning: this result crucially uses independence: may fail if X, Y are not independent



Sample Mean

30

Suppose we have a r.v. X and we sample it again and again, say n times

E.g. we have a dice/coin and we throw/toss it again and again

Make sure that samples are independent of each other

For example in the coin case, do toss the coin fairly n times – do not just toss it once and then blindly repeat the value of the first toss n times

Using the values obtained in these repeated samples, say x_1, x_2, \dots, x_n , we can get a very good estimate $\mathbb{E}X$ if n is sufficiently large

Called sample mean, or sample expectation, or empirical mean

$$\hat{\mathbb{E}}X = \bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

Note: sample mean can give answers that need patient analysis



Sample

Suppos

Indeed! If we ask 1000 random Indians, how many children they have, the sample mean might come out to be 2.35. However, no Indian can have 2.35 children since number of children has to be an integer!

E.g. we have a dice/coin and we throw/toss it again and again

Make sure

For example
then blind

Using the values obtained in these repeated samples, say x_1, x_2, \dots, x_n , we can

get a

Called

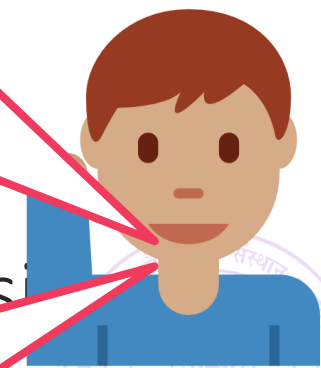
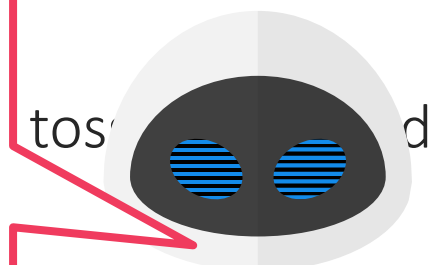
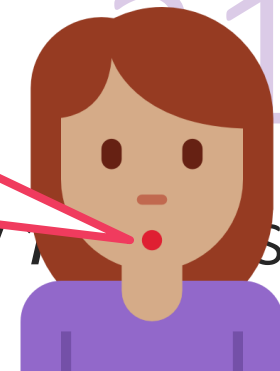
Interesting fact: the sample mean is the point which is the closest to all samples in terms of squared distance (**Proof:** use first order optimality)

$$\hat{\mathbb{E}}X = \arg \min_c \sum_{i=1}^n (x_i - c)^2$$

Note: s

Interesting fact: even the mean itself satisfies the nice property

$$\mathbb{E}X = \arg \min_c \mathbb{E}[(X - c)^2]$$



Mode of a Random Variable

32

The mode of a random variable is simply the value(s) that the r.v. takes with highest probability

Warning: a r.v. may have more than one mode value

$$\text{mode}(X) \triangleq \arg \max_{x \in S_X} \mathbb{P}[X = x]$$

Given a number of samples of X , can define *empirical mode* similarly

Simply the value that appears most frequently in the samples

$$\text{mode}(x_1, x_2, \dots, x_n) \triangleq \arg \max_{x \in S_X} \sum_{i=1}^n \mathbb{I}\{x_i = x\}$$

$$= \arg \max_{x \in \{x_1, \dots, x_n\}} \sum_{i=1}^n \mathbb{I}\{x_i = x\}$$

Note: mode of a random variable (or even samples) is always in S_X i.e. always a valid value that the r.v. can actually take (unlike expectation)



Mode of a Random Variable

33

The mode of a random variable is simply the value(s) that the r.v. takes with highest probability

Warning: a r.v. may have more than one mode value

$$\text{mode}(X) \triangleq \arg \max_{x \in S_X} \mathbb{P}[X = x]$$

Recall that $\mathbb{I}\{\text{blah}\} = 1$ if blah is true (or blah happens) else if blah does not happen or is false, $\mathbb{I}\{\text{blah}\} = 0$

The *empirical mode* similarly

$$\begin{aligned} \text{mode}(x_1, x_2, \dots, x_n) &\triangleq \arg \max_{x \in S_X} \sum_{i=1}^n \mathbb{I}\{x_i = x\} \\ &= \arg \max_{x \in \{x_1, \dots, x_n\}} \sum_{i=1}^n \mathbb{I}\{x_i = x\} \end{aligned}$$

Note: mode of a random variable (or even samples) is always in S_X i.e. always a valid value that the r.v. can actually take (unlike expectation)

Median of a Random Variable

34

The median of a random variable X is a value m that satisfies $\mathbb{P}[X \leq m] \geq 0.5$ as well as $\mathbb{P}[X \geq m] \geq 0.5$

The *empirical median* of a set of independent samples x_1, x_2, \dots, x_n of a random variable X is defined to be a value m such that as many samples are greater than or equal to m as are less than or equal to m

Often we talk about median income of a country – this is a value such that half the population earns at least that much value as income

To find the empirical median, first arrange samples in increasing order i.e.

$$x_1 \leq x_2 \leq \dots \leq x_n$$

If n is odd, then $m = x_{\frac{n+1}{2}}$. If n is even, then may be (infinitely) many

empirical medians but we often take $m = \frac{1}{2} (x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$

The empirical median gives a good estimate of median of the r.v. if n is large



Medi

Interesting Fact: The empirical median is the point which is the closest to all samples in terms of absolute distance (**Proof:** in notes)

$$\hat{\mathbb{E}}X = \arg \min_c \sum_{i=1}^n |x_i - c|$$

The me

$\mathbb{P}[X \leq m] \geq 0.5$ as well as $\mathbb{P}[X \geq m] \geq 0.5$

The *em*

of a ran

Interesting fact: even the median itself satisfies the nice property

$$\mathbb{E}X = \arg \min_c \mathbb{E}[|X - c|]$$

samples are greater than or equal to m as are less than or equal to m

Often we talk about median income of a country – this is a value such that half the population earns at least that much value as income

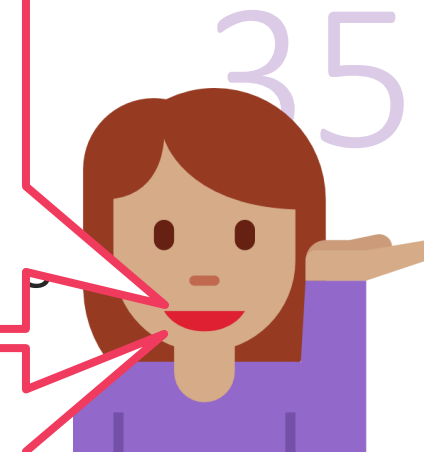
To find the empirical median, first arrange samples in increasing order i.e.

$$x_1 \leq x_2 \leq \dots \leq x_n$$

If n is odd, then $m = x_{\frac{n+1}{2}}$. If n is even, then may be (infinitely) many

empirical medians but we often take $m = \frac{1}{2} (x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$

The empirical median gives a good estimate of median of the r.v. if n is large



x_1, \dots, x_n
as many



Tells us how “spread out” are the values that an r.v. takes. Specifically, how far away from its expectation does the r.v. often take values

For a random variable X with expectation $\mu \triangleq \mathbb{E}X$, its variance, denoted as $\mathbb{V}[X]$ or $\text{Var}[X]$ or often just as σ^2 can be defined as

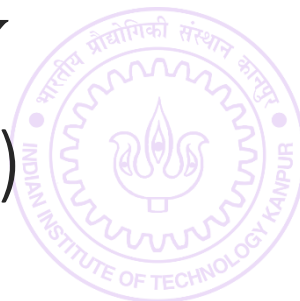
$$\sigma^2 = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[(X - \mathbb{E}X)^2] = \sum_{x \in S_X} (x - \mu)^2 \cdot \mathbb{P}[X = x]$$

Can be simplified to obtain another (equivalent) definition

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

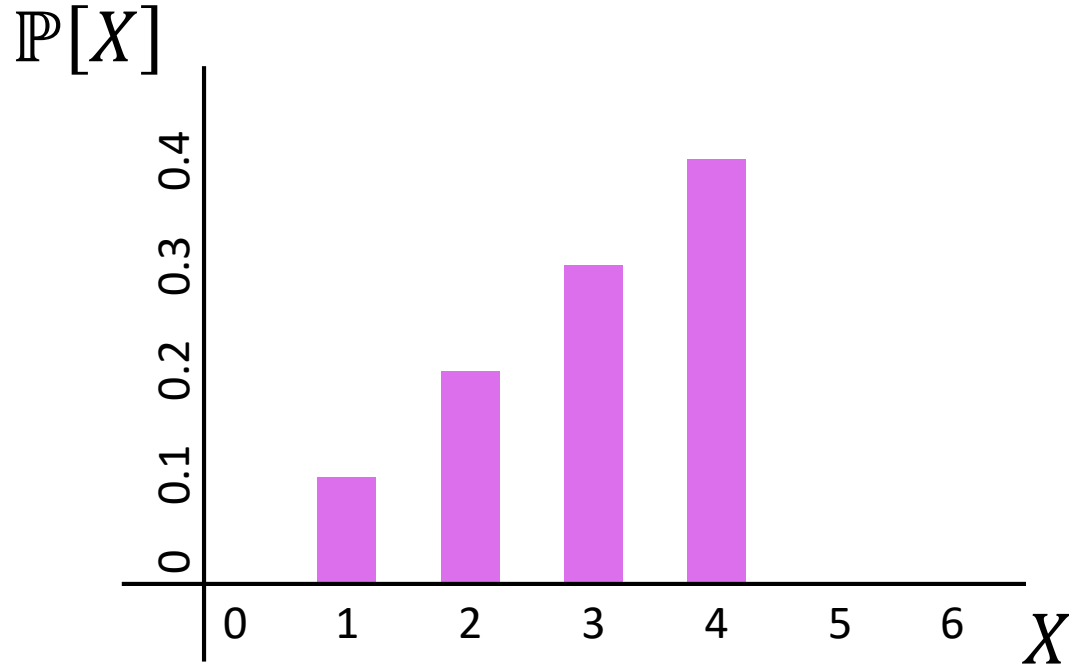
Notice: $(x - \mu)^2 \geq 0$ for all $x \in S_X$ which means $\mathbb{E}[(X - \mu)^2] \geq 0$ which means that $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$ for all r.v. X . Also $\mathbb{V}[X] \geq 0$ for all r.v. X

Standard deviation: the square root of the variance (denoted σ)

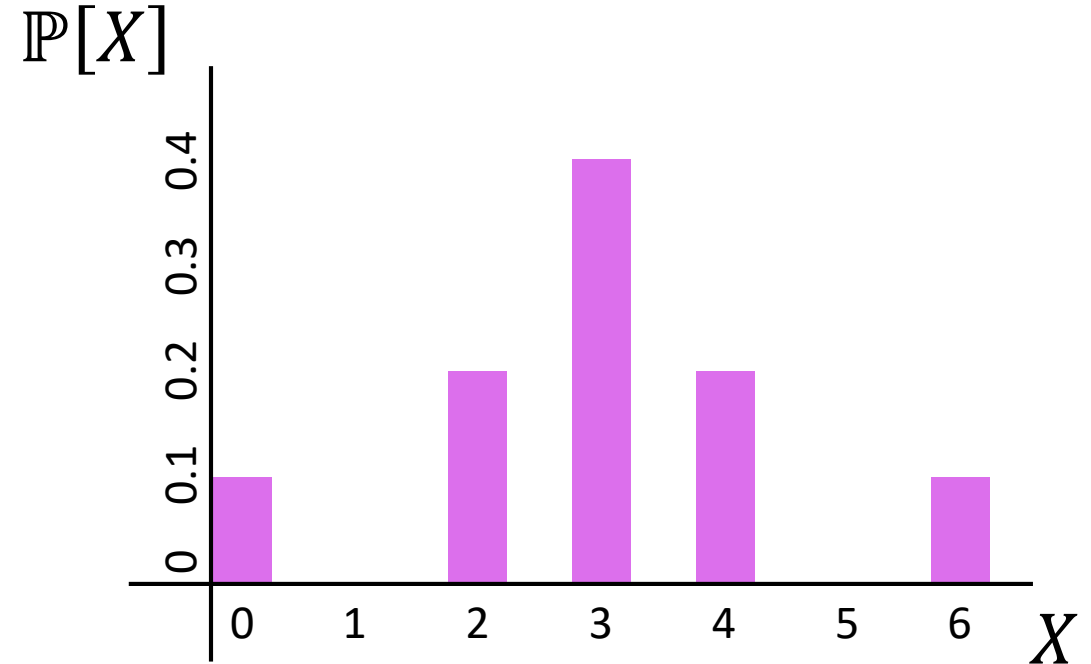


Example

37



$$\mathbb{E}X = 3, \text{med}(X) = 3, \\ \text{mode}(X) = 4, \mathbb{V}X = 1$$



$$\mathbb{E}X = 3, \text{med}(X) = 3, \\ \text{mode}(X) = 3, \mathbb{V}X = 2.2$$

This distribution has the same mean and median as the first one but is more “spread out” hence larger variance



Sample Variance

38

Given n independent samples x_1, \dots, x_n of a random variable X , the empirical variance can be calculated in two (equivalent) ways

First find the empirical mean $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$

Method 1: Calculate $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$

Method 2: First calculate $\hat{s} = \frac{1}{n} \sum_{i=1}^n x_i^2$ and then get $\hat{\sigma}^2 = \hat{s} - \hat{\mu}^2$

Both methods always give the same answer

Method 2 preferred when data not available all at once since it can be computed using running averages. Method 1 requires two passes over data

However, method 2 can be bad if \hat{s} and $\hat{\mu}^2$ are both very large and close

As before, if n is large, empirical variance is a good estimate of $V[X]$



Sample Variance

39

Given n independent samples x_1, \dots, x_n of a random variable X , the empirical variance can be calculated in two (equivalent) ways

First find the empirical mean $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$

Method 1: Calculate $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$

An effect called *catastrophic cancellation*. Basically, on computers, due to finite precision, for example, if we have $\hat{s} = 1000000000001$ and $\hat{\mu}^2 = 1000000000000$, then clearly $\hat{\sigma}^2 = 1$ but our computers may store $\hat{s} = 1000000000000$ to save space and ignore the error and cause us to get $\hat{\sigma}^2 = 0$

$$\hat{\sigma}^2 = \hat{s} - \hat{\mu}^2$$

since it can be two passes over data

However, method 2 can be bad if \hat{s} and $\hat{\mu}^2$ are both very large and close

As before, if n is large, empirical variance is a good estimate of $V[X]$



If we have two r.v.s X, Y then the covariance of these two r.v.s tell us how they behave in tandem

Example 1: let education level and income be two random variables defined on the sample space of all Indians – it is expected that if education of a person is higher than mean education level of all Indians, then their income should also be higher than mean income level of all Indians

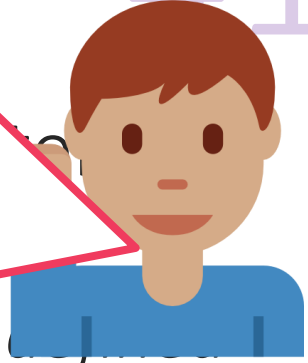
Example 2: let age and sleeping hours be two different r.v.s – it is expected that if age of a person is higher than mean age of all Indians, then the person would sleep fewer than the average number of hours (since children typically sleep more and old people tend to sleep less)

$$\begin{aligned}\text{Cov}(X, Y) &\triangleq \mathbb{E}[(X - \mathbb{E}X) \cdot (Y - \mathbb{E}Y)] = \mathbb{E}[XY] - \mathbb{E}X \cdot \mathbb{E}Y \\ &= \sum_{\omega \in \Omega} (X(\omega) - \mathbb{E}X) \cdot (Y(\omega) - \mathbb{E}Y) \cdot p_{\omega}\end{aligned}$$

Note that $\text{Cov}(X, X) = \mathbb{V}[X]$

Note that $\text{Cov}(X, Y)$ may be positive, negative or zero





We can estimate covariance using samples too. Suppose we are given values of X, Y on n outcomes (i.e. we sampled n outcomes $\omega_1, \dots, \omega_n$ and on each outcome ω_i , we return $(x_i, y_i) = (X(\omega_i), Y(\omega_i))$). Then sample covariance can be computed in two ways. First calculate empirical mean of X and Y

$$\hat{\mu}_X = \frac{1}{n} \sum_{i=1}^n x_i \text{ and } \hat{\mu}_Y = \frac{1}{n} \sum_{i=1}^n y_i$$

Method 1: Calculate $\widehat{\text{Cov}}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_X)(y_i - \hat{\mu}_Y)$

Method 2: First calculate $\hat{c} = \frac{1}{n} \sum_{i=1}^n x_i y_i$ and get $\widehat{\text{Cov}}(X, Y) = \hat{c} - \hat{\mu}_X \hat{\mu}_Y$

Just as before, both methods always give the same answer. Method 2 useful when data not available all at once but can be bad if \hat{c} and $\hat{\mu}_X \hat{\mu}_Y$ are both very large in magnitude but close together as well

person would sleep fewer than the average number of hours (since children typically sleep more and old people tend to sleep less)

$$\begin{aligned} \text{Cov}(X, Y) &\triangleq \mathbb{E}[(X - \mathbb{E}X) \cdot (Y - \mathbb{E}Y)] = \mathbb{E}[XY] - \mathbb{E}X \cdot \mathbb{E}Y \\ &= \sum_{\omega \in \Omega} (X(\omega) - \mathbb{E}X) \cdot (Y(\omega) - \mathbb{E}Y) \cdot p_{\omega} \end{aligned}$$

Note that $\text{Cov}(X, X) = \mathbb{V}[X]$

Note that $\text{Cov}(X, Y)$ may be positive, negative or zero



Rules of Variance

42

Suppose $b, c \in \mathbb{R}$ are any two constants and X, Y are any two r.v.s, then

Constant Rule: $\mathbb{V}[c] = 0$ i.e. if $Z \equiv c$ is a constant r.v. then $\mathbb{V}[Z] = 0$

Seems intuitive since a constant r.v. does not vary at all i.e. zero variance

Scaling Rule: $\mathbb{V}[c \cdot X] = c^2 \cdot \mathbb{V}[X]$

Shift Rule: $\mathbb{V}[X + c] = \mathbb{V}[X]$ i.e. if $W \triangleq X + c$ then $\mathbb{V}[W] = \mathbb{V}[X]$

Shifting a random variable does not change its “spread”

Sum Rule: $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y] + 2\text{Cov}[X, Y]$

Difference Rule: $\mathbb{V}[X - Y] = \mathbb{V}[X] + \mathbb{V}[Y] - 2\text{Cov}[X, Y]$



Rules of Variance

43

Suppose $b, c \in \mathbb{R}$ are any two constants and X, Y are any two r.v.s, then

Constant Rule: $\mathbb{V}[c] = 0$ i.e. if $Z \equiv c$ is a constant r.v. then $\mathbb{V}[Z] = 0$

Often used to deal with catastrophic cancellation by shifting the data to make it smaller in magnitude but leaving variance unchanged

Scaling

Shift Rule: $\mathbb{V}[X + c] = \mathbb{V}[X]$ i.e. if $W \triangleq X + c$ then $\mathbb{V}[W] = \mathbb{V}[X]$

Shifting a random variable does not change its “spread”

Sum Rule: $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y] + 2\text{Cov}[X, Y]$

Difference Rule: $\mathbb{V}[X - Y] = \mathbb{V}[X] + \mathbb{V}[Y] - 2\text{Cov}[X, Y]$



Rules of Covariance

44

Suppose $b, c \in \mathbb{R}$ are any two constants and X, Y are any two r.v.s, then

Constant Rule: $\text{Cov}[X, c] = 0$

Symmetry Rule: $\text{Cov}[X, Y] = \text{Cov}[Y, X]$

Scaling Rule: $\text{Cov}[bX, cY] = bc \cdot \text{Cov}[X, Y]$

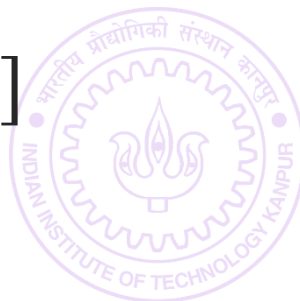
Shift Rule: $\text{Cov}[X + b, Y + c] = \text{Cov}[X, Y]$

If X, Y are independent then $\text{Cov}[X, Y] = 0$

Proof: $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}X \cdot \mathbb{E}Y = \mathbb{E}X \cdot \mathbb{E}Y - \mathbb{E}X \cdot \mathbb{E}Y = 0$

We applied the product rule for expectations above

Corollary: *If X, Y are independent r.v.s, then $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y]$*



R

In books/papers, you may come across a term called *correlation* which is a normalized version of covariance.

$$\rho_{X,Y} = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\mathbb{V}[X] \cdot \mathbb{V}[Y]}}$$

Su

Co

Sy

So

St

For any two r.v.s X, Y , we always have $\rho_{X,Y} \in [-1, 1]$. If $\rho_{X,Y} = 0$ then the two r.v.s are said to be *uncorrelated*. Note that if X, Y are uncorrelated, then also we have $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y]$. Warning: independent r.v.s are always uncorrelated but not all uncorrelated r.v.s need be independent.

Can estimate $\rho_{X,Y}$ using samples as well $\hat{\rho}_{X,Y} = \frac{\widehat{\text{Cov}}(X, Y)}{\sqrt{\hat{\sigma}_X^2 \hat{\sigma}_Y^2}}$

If X, Y are independent then $\text{Cov}[X, Y] = 0$

If $\rho_{X,Y} < 0$, this means that typically, whenever X takes larger values than its own mean, Y takes smaller values than its own mean and vice versa. If $\rho_{X,Y} > 0$, then this means that both r.v.s take values larger or smaller than their respective means together. $\rho_{X,Y} = 0$ means that typically, even if X takes a value larger than its mean, Y may take smaller or larger values than its own mean

