05.1 KHAN ACADEMY 05 : RANDOM VARIABLES
STATISTICS AND PROBABILITY

Random Variables

It is a random process where Outcomes → numbers.

Eg - $X = \begin{cases} 1 & \text{if heads} \\ 0 & \text{if tails} \end{cases}$ One advantage,

Y = sum of upward face after rolling 7 dice

Normally if we have to write, P (sum of upward face after rolling 7 dice > 30) = ?
In Random Variable, $P(Y > 30) = ?$

Discrete and Continuous Random Variables

Random Variables

Discrete
Distinct / separate value

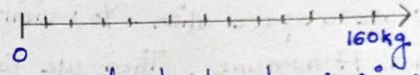
$X = \begin{cases} 1 & \text{Heads} \\ 0 & \text{Tails} \end{cases}$

X = Years that a random student was born in class
1994, 1995, 1996 ... etc

Mostly, Discrete random variables are countable.

Continuous
Any value in the interval

Y = Exact mass of a random human being



Y = No of ants born tomorrow in earth. with weight.

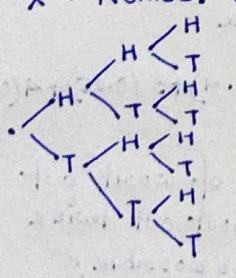
Y = Exact winning time men's 100m dash 2016 olympics.

Anyone can say it is discrete as it will have exact time i.e, 9.6 or 9.31.
But it can take any value 9.6 / 9.59 / 5.956789 any value on the scale and that is why continuous

But, if we say winning time men's 100m dash 2016 olympics (round to 2 decimal place)
Then it is discrete random variable as it will be 9.59

Constructing a probability distribution for random variable

X = Number of "heads" after 3 flip of a fair coin.



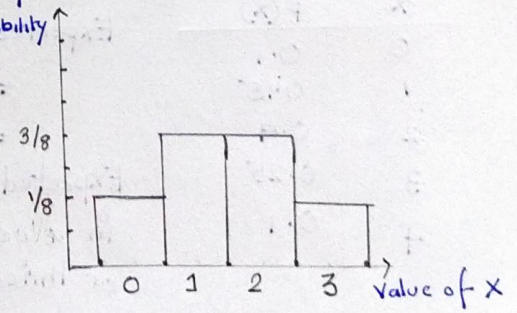
3) HHH
2) HHT
2) HTH
1) HTT
2) THH
1) THT
1) TTH
0) TTT

0) $P(X=0) = \frac{1}{8}$

1) $P(X=1) = \frac{3}{8}$

2) $P(X=2) = \frac{3}{8}$

3) $P(X=3) = \frac{1}{8}$



Frozen Yoghurt – How much line at frozen yoghurt at 4pm.

For next 50 days, we observed the line

Line size	Times Observed	Probability Estimate
0	24	$24/50 = 0.48 = 48\%$
1	18	$18/50 = 0.36 = 36\%$
2	8	$8/50 = 0.16 = 16\%$
	50	

So for next 500 days, how many number we will see 2 person line?

$500 \times (0.16) = 80$

So, 80 number of days approx (not exactly) we will see 2 person line.

Anthony DeNooN is analyzing his basketball statistics. The following table shows a probability model for the results from his next two free throws.

Outcome	Probability
Miss both Free throws	0.2
Make exactly one free throw	0.5
Make both free throws	0.1

Is this a valid model?

No because probability is not summing to 1.

You are a space alien. You visit planet Earth and abduct 97 chickens, 47 cows, and 77 humans. Then we randomly select one Earth creature from your sample to experiment on. Each creature has an equal probability of getting selected. Create a probability model to show how likely you are to select each type of Earth creature.

Type of Earth creature	Estimate	Probability
Chicken	$\frac{97}{97+47+77}$	$= \frac{97}{221}$
Cow	$\frac{47}{221}$	
Human	$\frac{77}{221}$	

So this is the probability model

Mean (expected value) of a discrete random variable

X = Number of workouts in a week.

X	P(X)
0	0.1
1	0.15
2	0.4
3	0.25
4	0.1

Expected value (αX) = (Mean) x = μ_x ← Population mean

$= (0.01)(0) + (1)(0.15) + (2)(0.4) + 3(0.25) + 4(0.1)$

$= 2.1$.

Expected value of X which is number of workouts in a week is 2.1. so we can workout out twice or thrice in a week but how 2.1 is possible? Sometime in a week I can workout 0 then next week 5. That's why after 10 weeks (suppose), 2.1 time avg weekly workout will be observed.

05.3 Variance $(X) = \text{Var}(X) = \frac{(x-\mu)^2}{N}$

$\text{Var}(X) = (0-2.1)^2 \times (0.1) + (1-2.1)^2 \times (0.15) + (2-2.1)^2 (0.4) + (3-2.1)^2 \times (0.25) + (4-2.1)^2 \times 0.1$

$= 1.19$

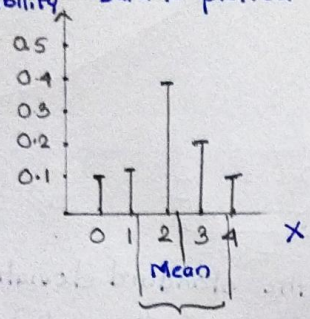
Standard Deviation $= \sigma_X = \sqrt{1.19} = 1.09$

Here we are not dividing by n in variance because.

In discrete ~~non~~ random variable, Variance $= (x-\mu)^2 \times (\text{probability})$

Probability is the proportion of data, which includes divide by n. That is why we do not divide by n.

Probability Data is plotted.



$\mu_X = 2.1$

$\text{Var}(X) = 1.19$.

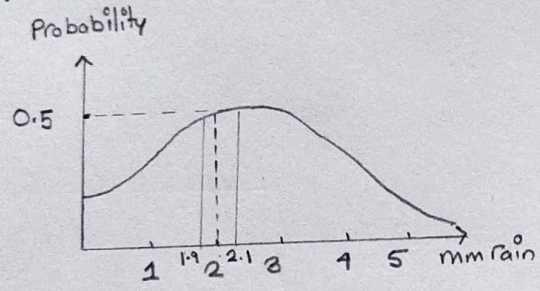
$\sigma(X) = 1.09$

So if we visually see also, it is making sense.

Continuous Random variable.

Probability Density Functions

Y = Exact amount of rain tomorrow



$P(Y=2) = 50\%$ not right answer

Because to get exact 2mm rain is impossible. It cannot be 1.99999/2.00001

Somewhere it will rain less or more.

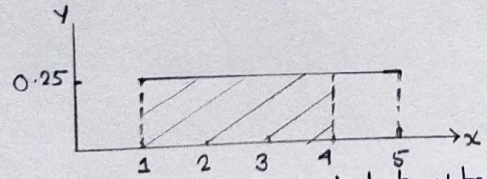
So we take some area ±.

$P(|Y-2|<1) = P(1.9 < Y < 2.1)$

= Area under Curve (1.9 and 2.1)

$= \int_{1.9}^{2.1} f(x)$ ∵ $f(x)$ is the area of whole curve

Consider the density curve below.

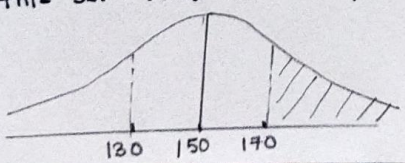


Here we have equal probability of density curve.

Probability that x is less than 4 $= (0.25) \times 3$

$= 0.75 = 75\%$

A set of middle school student heights are normally distributed with a mean of 150 cm and a standard deviation = 20cm. Let H be the height of randomly selected student from this set. Find and interpret $P(H > 170)$.



$P(H > 170) \Rightarrow$ ▨

$= 16\%$

1SD = 58%

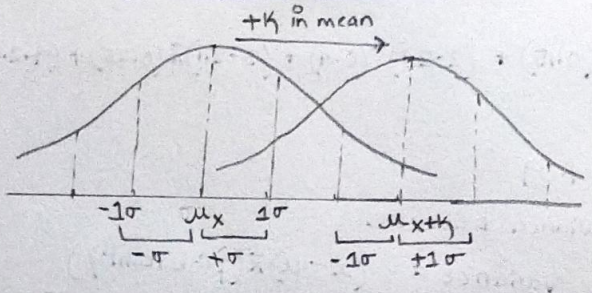
Remaining = 32% (both left & right side)

To get Right $= \frac{32}{2} = 16\%$

Impact of transforming (scaling and shifting) random variables

X = random variable,

$Y = X + K$ ← some constant (shift data)



this distribution will shift to right due to addition of K.

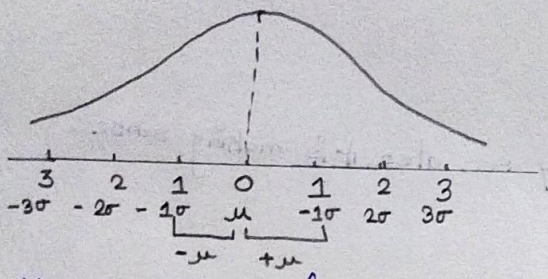
Older the mean was μ_x

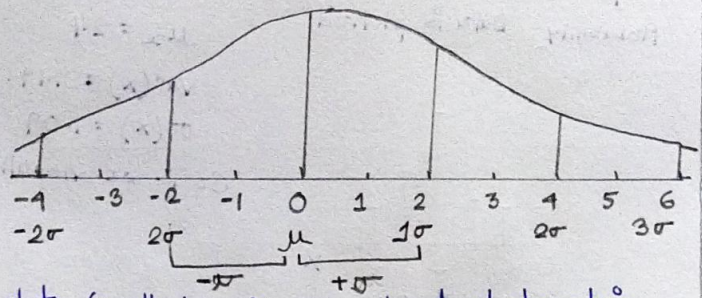
Now the new mean is μ_{x+K}

Even if $\mu_x \neq \mu_{x+K}$ (mean) because distance between them is K

but, $\sigma_K = \sigma_{\bar{x}+K}$ (standard deviation) standard deviation did not change

If we scale the data :; $Z = K X$ suppose 2.





We can observe, after scaling the data (multiplying) the standard deviation is changed from 1 unit to 2 unit and mean also change (not visually) but suppose 1,2,3 mean is 2. After scaling by 2, → 2,4,6 mean is 4.

Shifting, Mean is change. SD not change, Scaling, both mean & SD change.

05.5 Mean of sum and difference of random variables (Discrete example)

X = Number of dogs to see in a day. Suppose Expected (X) = Mean (X)
$E(X) = \mu_X = 3$.

Y = Number of cats to see in a day. $E(Y) = \mu_Y = 4$.

Averagely how many dogs and cats we can see in a day $= E(X+Y) = \mu_{X+Y}$
$= \mu_X + \mu_Y = 3+4 = 7$.

Averagely which we see most (Cats / Dogs) $= E(Y-X) = \mu_{X-Y} = \mu_Y - \mu_X = 4-3 = 1$.

We will see cat more than dog averagely

Variance of sum and difference of random variables (Continuous example)

X = Weight of cornflakes in box.
Suppose $E(X) = 16$ gram, weight
All the time weight of box will not be same.
$\sigma_X = 0.8$ gram, standard deviation
And weight range, $15 \le X \le 17$
(15: Minimum weight; 17: Maximum weight)

Y = Weight of cornflakes in a bowl.
Suppose $E(Y) = 4$ gram, weight
All the time weight of bowl will not be same.
$\sigma_Y = 0.6$ gram, standard deviation
Weight range, $3 \le Y \le 5$
(3: Minimum weight; 5: Maximum weight)

Total weight, $E(X+Y) = E(X) + E(Y) = 16 + 4 = 20$ grams.

Total Variance, $\text{Var}(X \pm Y) = \text{Var}(X) \pm \text{Var}(Y)$

$\text{Var}(X+Y) = 15+3 \le X+Y \le 17+5 = 18 \le X+Y \le 22$

$\text{Var}(X-Y) = 15-3 \le X-Y \le 17-5$ is wrong.
suppose $15-3 = 12$. but Y can go to range 5.
so correct will be $15-5 = 10$ (lowest range).
Othercase, $17-5 = 12$, but if $Y = 3$. then high $= 17-3 = 14$.
$\text{Var}(X-Y) = 15-5 \le X-Y \le 17-3 = 10 \le X-Y \le 14$.

So range of $\text{Var}(X \pm Y)$ is 4 and range of (X) and range of (Y) seprately is 2.
So range is increased. Even for standard deviation / variance will increase.

Calculate, standard deviation $(X+Y) = \sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2} = \sqrt{0.64+0.36} = \sqrt{1} = 1$

Whether we are adding or subtracting random variables, the variability will increase.
This is only true for Independent event where $\text{Var}(X \pm Y) = \text{Var}(X) \pm \text{Var}(Y)$

Suppose, x and y are dependent variables.
X = Number of hours, random person slept yesterday. So, $X+Y = 24$ hours
Y = Number of hours, same random person was awake yesterday $= X^{\text{complement}}$
So this is dependent event. Suppose $\text{Var}(X) = 4\text{hours}^2$ SD = 2hour.
$\text{Var}(Y) = 4\text{hour}^2$ SD = 2hour
$\text{Var}(X+Y) = 4+4 = 8\ \text{hour}^2$.
which is not possible.

So Independent events only can sum or subtract variance.

$X+Y = 24$ hour
but addition of dependent event don't sum up.

Variance of difference of random variable —

$X = E(X) = \mu_X$ (Mean), $Y = E(Y) = \mu_Y$ — (i)

$-Y = E(-Y) = -E(Y)$ — (ii)

X and Y both are independent variable.

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

$Var(Y) = E((Y-\mu_Y)^2) = \sigma_Y^2$

Suppose $Z = X+Y$

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

$\mu_Z = \mu_X + \mu_Y$

$Var(Z) = Var(X) + Var(Y)$

$\sigma_Z^2 = \sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$

$A = X-Y$

$E(A) = E(X-Y) = E(X)-E(Y)$

$\mu_A = \mu_X - \mu_Y$

$Var(A) = Var(X) - Var(Y)$

$\sigma_A^2 = \sigma_{X-Y}^2 = \sigma_{X+(-Y)}^2 = \sigma_X^2 + \sigma_{-Y}^2$ — (iii)

here $\sigma_{-Y}^2 = Var(-Y) = E((-Y-E(-Y)^2))$

from (ii), $E(-Y) = -E(Y)$.

$= E((-Y+E(Y)^2))$

Extract ~~Multiply by c~~ $= (-1)^2$ so it is not change

Eqn is same, symbol change $= E((-1)^2 E((Y-E(Y)^2))$

$= (1)E(Y-E(Y)^2)$

$= E(Y-E(Y)^2) = \sigma_Y^2$.

So $\sigma_{-Y}^2 = \sigma_Y^2$.

Therfore from (eqn (iii)) $\rightarrow \sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2$

REVISION OF COMBINING RANDOM VARIBALES

		Mean	Variance
Adding :	$T = X+Y$	$\mu_T = \mu_X + \mu_Y$	$\sigma_T^2 = \sigma_X^2 + \sigma_Y^2$
Subtracting :	$D = X-Y$	$\mu_D = \mu_X - \mu_Y$	$\sigma_D^2 = \sigma_X^2 + \sigma_Y^2$ [not subtract because we proved earlier]

- Make sure, variables are independent.
- Even we subtract two random variables, we still add their variances. If we subtract then it will increase overall variability in outcomes.
- To find standard deviation of combined distribution by taking square root of combined variance.