DATA DISTRIBUTION

Data Distribution

The distribution of a statistical data set (or a population) is a listing or function showing all the possible values (or intervals) of the data and how often they occur.

A. Probability Distribution: It is a listing of the probabilities of all the possible outcomes that could occur if the experiment was done.
It can be described as:
A diagram (Probability Tree)
A table

B. Frequency Distribution: It is a listing of observed /actual frequencies of all the outcomes of an experiment that actually occurred when experiment was done.

A. PROBABILITY DISTRIBUTION

- **Discrete Distribution:** Random Variable can take only limited number of values. Ex: No. of heads in two tosses.
- **Continuous Distribution:** Random Variable can take any value. Ex: Height of students in the class.

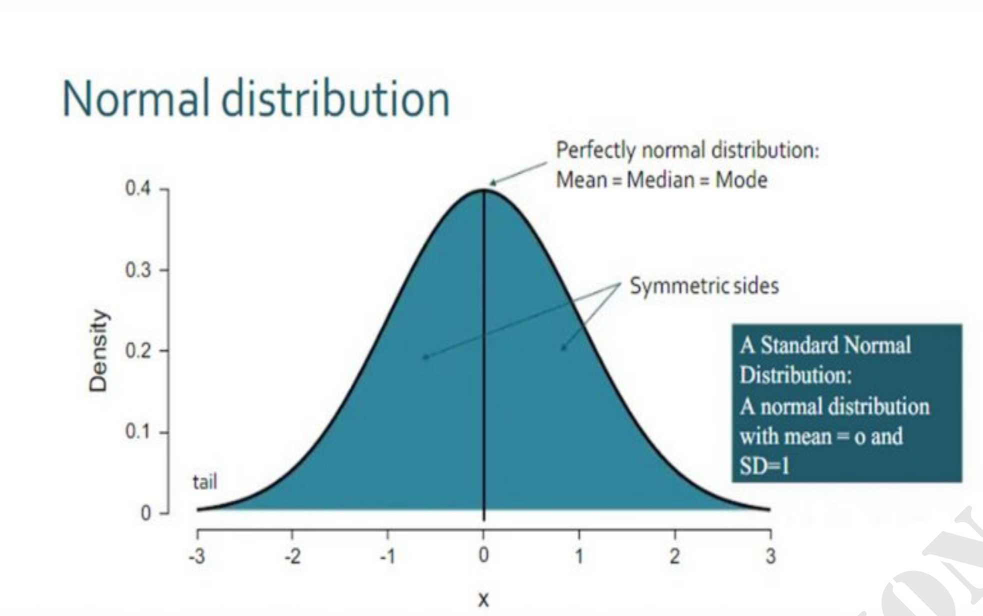
DATA DISTRIBUTION

❑ Normal Distribution

- Developed by eighteenth century mathematician — astronomer Karl Gauss, so also called Gaussian Distribution.
- It is a continuous PD i.e. random variable can take on any value within a given range. Ex: Height, Weight, Marks etc.
- Data is symmetrically distributed on both sides of mean and form a bell-shaped curve in frequency distribution plot
- Since it is symmetrical, its mean, median and mode all coincides i.e. all three are same.
- The tails are asymptotic to horizontal axis i.e. curve goes to infinity without touching horizontal axis.

Properties of Normal Distributions

- The mean. median. and Mode are equal.
- The normal Curve is and symmetric about the mean.
- Bell shaped curve with a single peak
- The total area under the curve is 1.
- Exactly half of the values are to the left of the center and the other half to the right.
- 68% of the observations fall **within 1 standard deviation of the mean**
- 95% of the observations fall within **2 standard deviations of the mean**
- **99.7% of the observations fall within 3 standard deviations of the mean**
- for a normal distribution, almost all values lie **within 3 standard deviations of the mean**

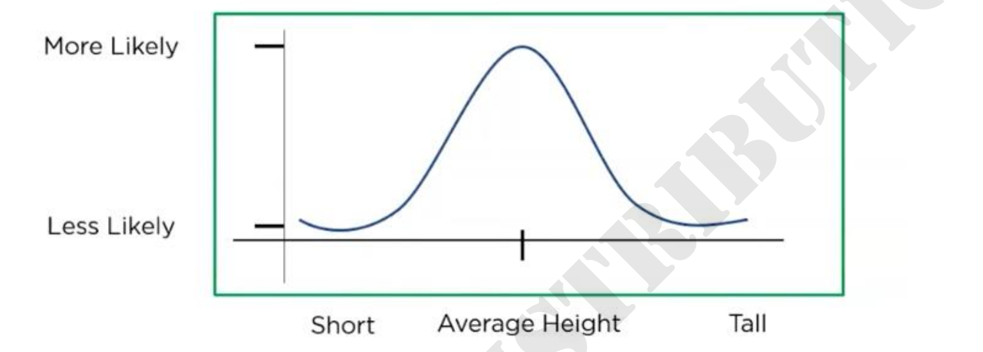


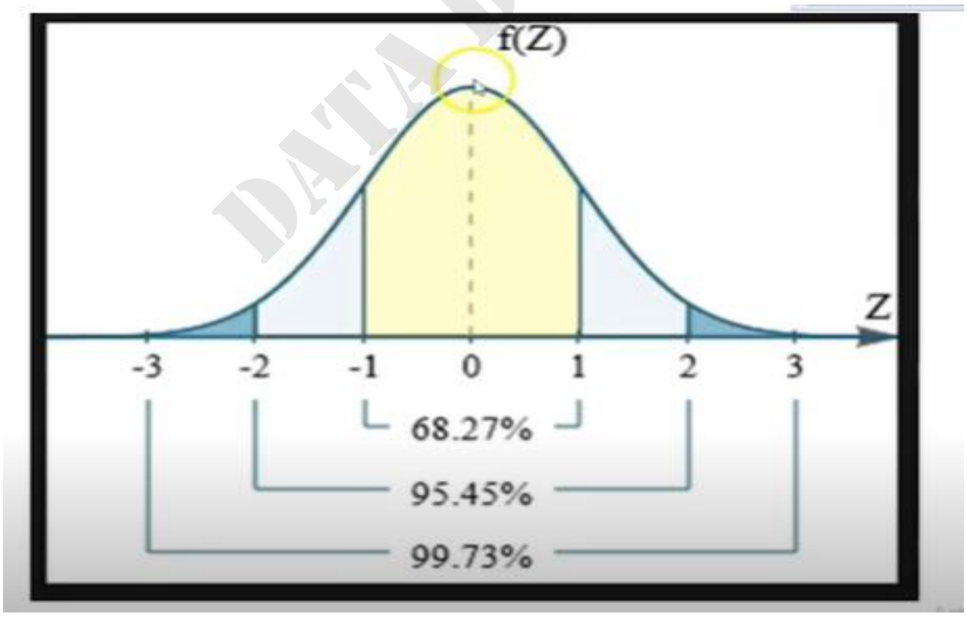
Normal distribution

Perfectly normal distribution:
Mean = Median = Mode

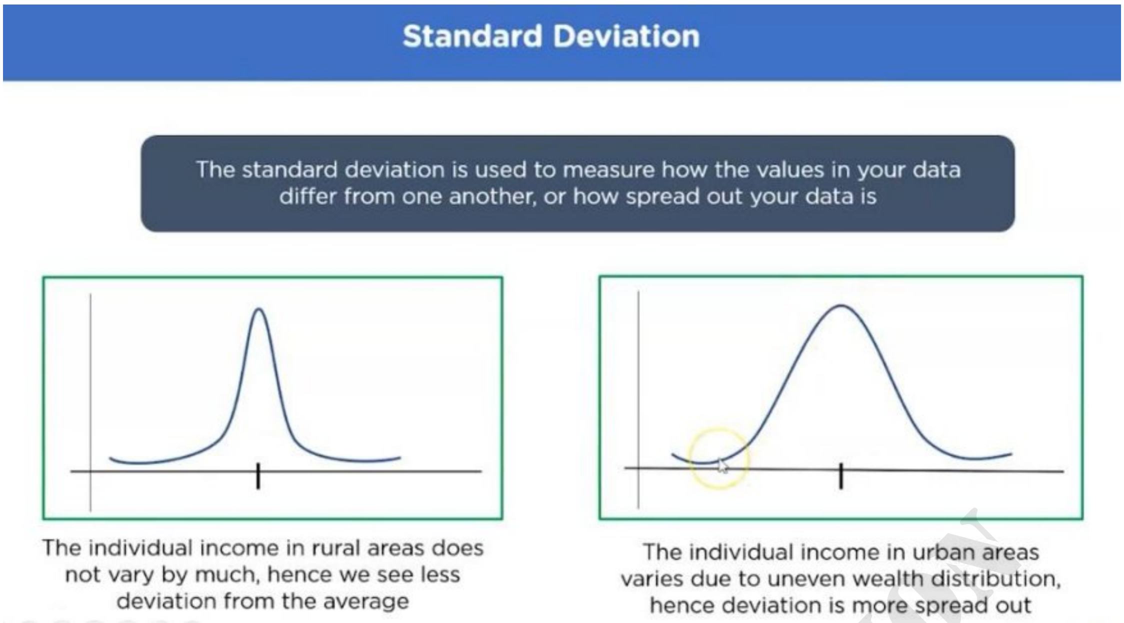
Symmetric sides

A Standard Normal Distribution:
A normal distribution with mean = o and SD=1





DATA DISTRIBUTION



Standard Deviation

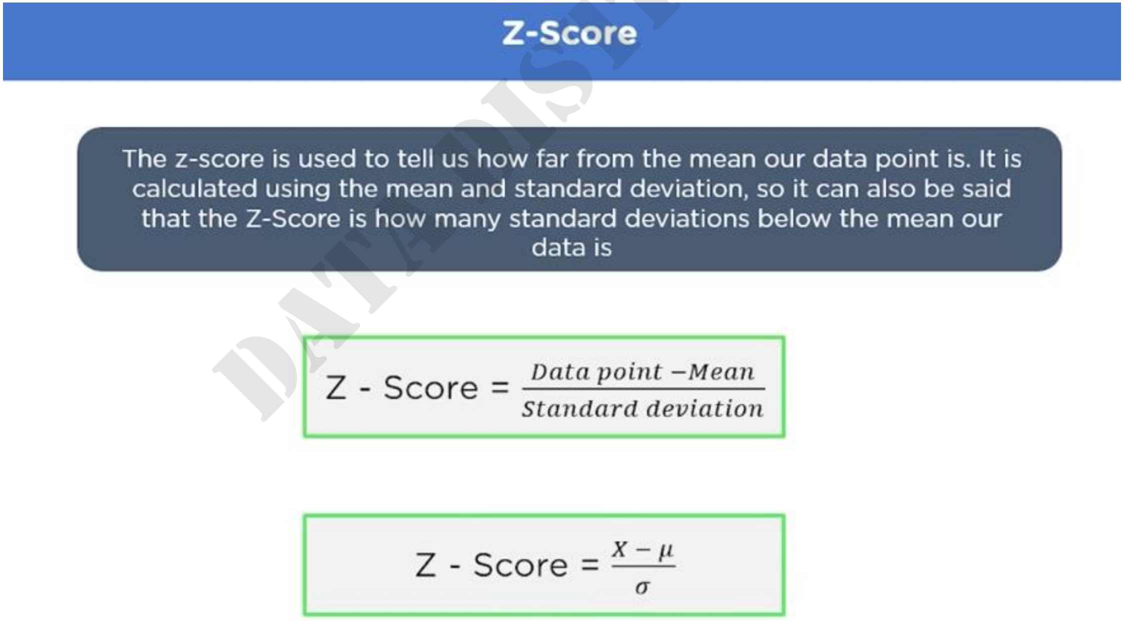
The standard deviation is used to measure how the values in your data differ from one another, or how spread out your data is

The individual income in rural areas does not vary by much, hence we see less deviation from the average

The individual income in urban areas varies due to uneven wealth distribution, hence deviation is more spread out

To draw a normal distribution, you need to know:

1. The average measurement. This tells you where the Center of the curve goes.
2. The standard deviation of the measurements, this tells you how wide the curve should be. And the width of the curve determines how tall it is. The wider the curve, the shorter. The narrower the curve. the taller.
3. Denser in the centre, less dense at the sides (tails),
4. It has two parameters: the mean and the standard deviation



Z-Score

The z-score is used to tell us how far from the mean our data point is. It is calculated using the mean and standard deviation, so it can also be said that the Z-Score is how many standard deviations below the mean our data is

$$\text{Z - Score} = \frac{\textit{Data point} - \textit{Mean}}{\textit{Standard deviation}}$$

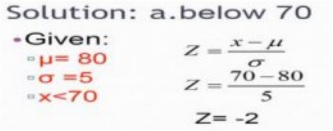
$$\text{Z - Score} = \frac{X - \mu}{\sigma}$$

DATA DISTRIBUTION

Example

A College conducts placement examination to all final year students. The examination scores of the 1000 examinees were approximately normally distributed with **mean score of 80** and **standard deviation of 5.** What is the probability that randomly chosen student got a score

1. below 70?
2. above 82?
3. Between 75 and 90?



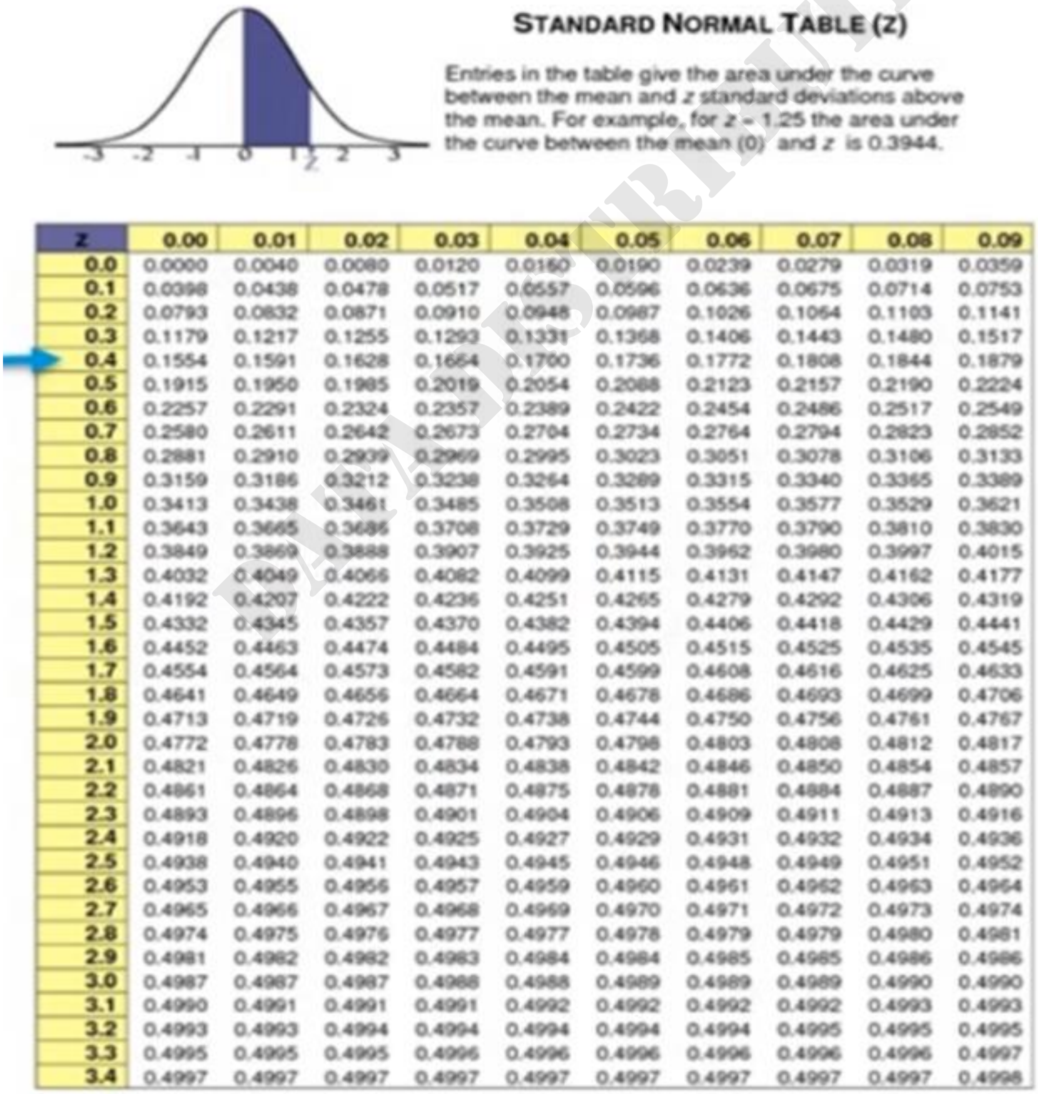
Solution: a.below 70

- Given:
  - μ= 80
  - σ =5
  - x<70

$$Z = \frac{x-\mu}{\sigma}$$

$$Z = \frac{70-80}{5}$$

Z= -2

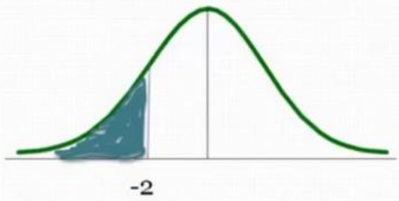


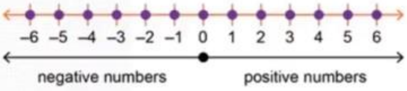
STANDARD NORMAL TABLE (Z)

Entries in the table give the area under the curve between the mean and z standard deviations above the mean. For example, for z = 1.25 the area under the curve between the mean (0) and z is 0.3944.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

DATA DISTRIBUTION

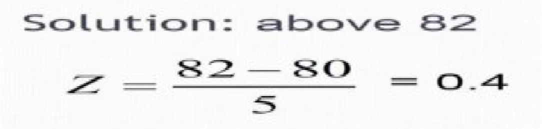




- P [x<70] = P[z<-2] = 0.5 – P[0<z<2]

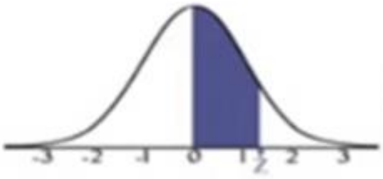
  =0.5 – 0.4772

  =0.0228 or 2.28%



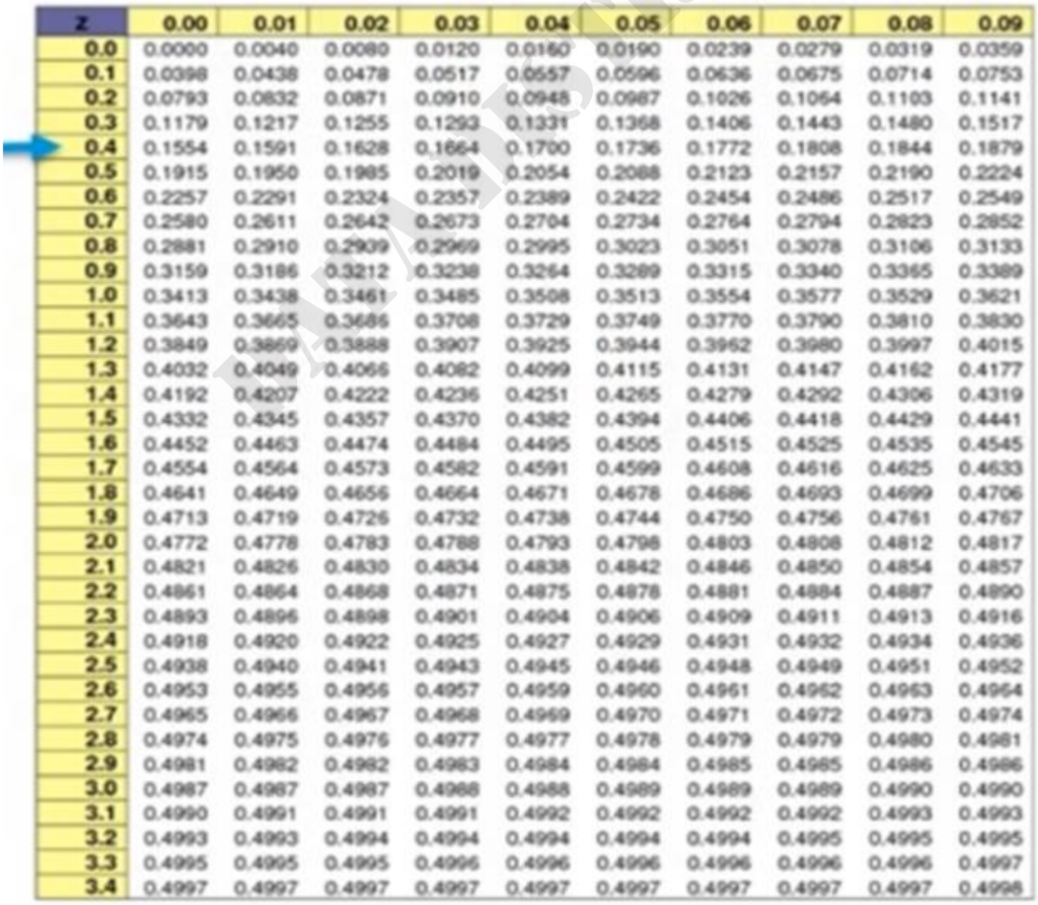
Solution: above 82

$$Z = \frac{82 - 80}{5} = 0.4$$



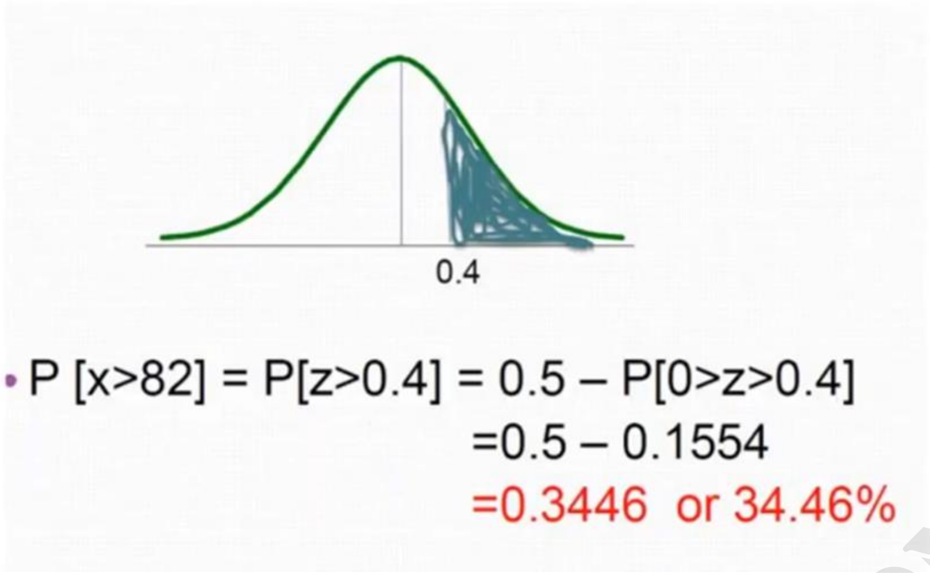
STANDARD NORMAL TABLE (Z)

Entries in the table give the area under the curve between the mean and z standard deviations above the mean. For example, for z = 1.25 the area under the curve between the mean (0) and z is 0.3944.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

DATA A DISTRIBUTION



- $P\ [x>82] = P[z>0.4] = 0.5 - P[0>z>0.4]$
  $=0.5 - 0.1554$
  $=0.3446$ or 34.46%

DATA DISTRIBUTION

Non-Gaussian (Non-Normal) distribution

- If the data is skewed on one side, then the distribution is non-normal.
- It may be

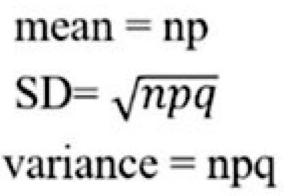
1. Binominal distribution or
2. Poisson distribution

❑ Binominal Distribution

- The binomial distribution is a common discrete distribution used in statistics, as opposed to a continuous distribution, such as the normal distribution.
- It was discovered by mathematician James bernouli
- It is the discrete Probability Distribution.
- In binominal distribution, The random experiment has only two possible outcomes 'success' and 'failure'. i.e. event can have only one of two possible outcomes such as yes/no, positive/negative, survival/death, and smokers/non-smokers.

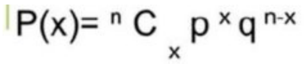
Success and Failure

- Consider an event associated to a random experiment. When a random experiment repeated a number of times, the event may or may not occur in each of those experiments.
- The occurrence of event may be named success and non-occurrence a 'failure'.
- Eg: in tossing a coin, there are 2 events, 'Head' and 'Tail'. One of them is a success and other is a failure.
- Consider an experiment which gives two possible outcomes success failure.
- This experiment is repeated n independent times.
- Let P be the probability of success and q be the probability of failure.
- The probability of success (or failure) lies between o and 1.
- Let us assume that x outcomes are success and the remaining n-x outcomes are failures.

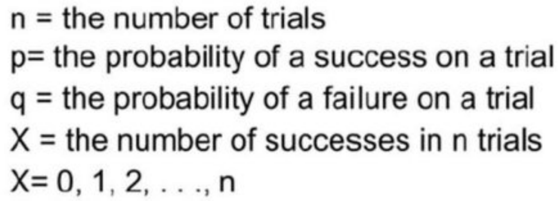


$$\text{mean} = np$$
$$SD = \sqrt{npq}$$
$$\text{variance} = npq$$

- The probability of this event is **P^xq^{n-x}** out of n trials.
- The x success can happen in any one of **nC_x** different ways.
- Therefore, the probability of getting x success in n trials is

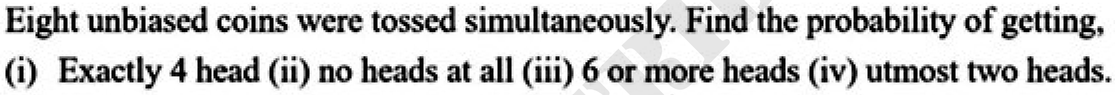


$$P(x) = {}^nC_x\, p^x q^{n-x}$$

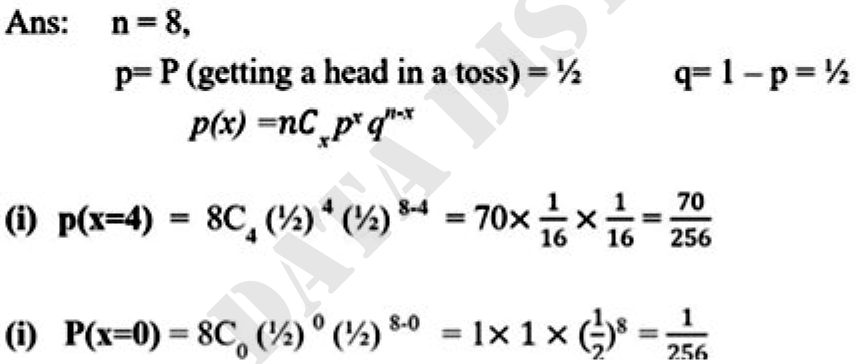


n = the number of trials
p= the probability of a success on a trial
q = the probability of a failure on a trial
X = the number of successes in n trials
X= 0, 1, 2, . . ., n

Qustion…



Eight unbiased coins were tossed simultaneously. Find the probability of getting,
(i) Exactly 4 head (ii) no heads at all (iii) 6 or more heads (iv) utmost two heads.



Ans: $n = 8$,
$p = P\text{ (getting a head in a toss)} = ½$ $\quad q = 1 - p = ½$
$$p(x) = nC_x p^x q^{n-x}$$

(i) $p(x=4) = 8C_4\,(½)^4\,(½)^{8-4} = 70 \times \frac{1}{16} \times \frac{1}{16} = \frac{70}{256}$

(i) $P(x=0) = 8C_0\,(½)^0\,(½)^{8-0} = 1 \times 1 \times (\frac{1}{2})^8 = \frac{1}{256}$

DATA DISTRIBUTION

❑ Poisson Distribution

- Named after French mathematician Simeon Denis Poisson.
- It gives us the probability of a given number of events happening in a fixed interval of time.
- The Poisson distribution is a discrete function, meaning that the event can only be measured as occurring or not as occurring, meaning the variable can only be measured in whole numbers.
- A poisson distribution is a measure of how many limes an event is likely to occur within "X" period of time.

What /s Poisson Distribution?

Poisson distribution is a limiting form of the binomial distribution in which n, the number of trials, becomes very large & p, the probability of success of the event is very very small.

Why we need Poisson Distribution

Poisson distribution used in cases where the chance of any individual event being a success is very small. The distribution is used to describe the behaviour of rare events. Examples;

- The number of defective screws per box of 5000 screws.
- The number of printing mistakes in each page of the first pro book.
- The number of air accidents in India in one year.
- Occurrence of number of scratches on a sheet of glass

Use of Poisson Distribution in Pharmaceutical

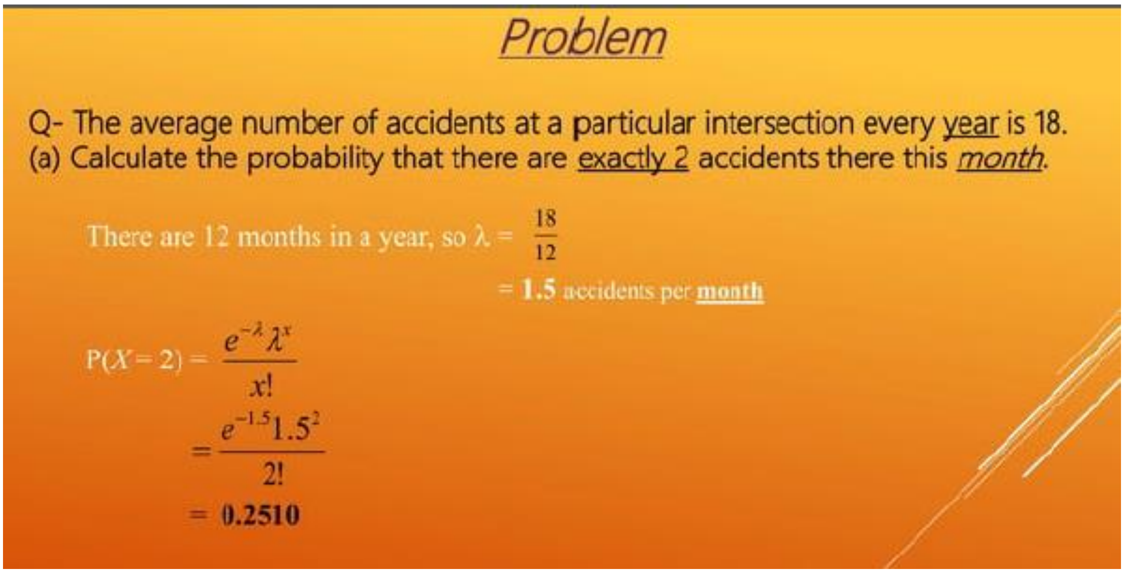
- Control limits for numbers of tablets rejected from an online Metal Detector during tablet compression cycle.
- Microbial counts in raw materials, products, and water for pharmaceutical use.
- Control limits for numbers of containers rejected from visual inspection of sterile production batches.
- Alert limits for microbial levels in cleanroom environment.
- Release limits for microbial counts in nonsterile products.

Condition Under Which Poisson Distribution is Used

- The random variable X should be discrete.
- A dichotomy exists i.e. happening of the event must be of two
- alternatives such as success & failure.
- Applicable is those cases where the number of trials n is very large and the probability of success p is very small but the mean np =is finite.
- Statistical independence is assumed.

Characteristics of Poisson Distribution

- Poisson Distribution is a discrete distribution.
- It depends mainly on the value of the mean m.
- This distribution is positively skewed to the left. With the increase in the value of the mean m, the distribution shift to the right and the skewness diminished
- if n is large & p is small, this distribution gives a close approximation to binomial distribution. Since the arithmetic mean of Poisson is same as that Binomial.
- Poisson distribution has only one parameter i.e. m, the arithmetic mean. Thus the entire distribution can be determined once the arithmetic mean is known
-



Problem

Q- The average number of accidents at a particular intersection every year is 18.
(a) Calculate the probability that there are exactly 2 accidents there this *month*.

There are 12 months in a year, so $\lambda = \frac{18}{12}$

$= \mathbf{1.5}$ accidents per **month**

$$P(X=2) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$= \frac{e^{-1.5}1.5^{2}}{2!}$$

$$= \mathbf{0.2510}$$

DATA DISTRIBUTION

B. FREQUENCY DISTRIBUTIONS

* Frequency :- It is the number of occarance of the any value in a Data Or

It is the no of times that any particular value comes in a data.

Eg :- Ravi takes a tablet thrice in a day, so '3' is the frequency of that data.

* Frequency Distribution

It is a tabular or graphical representation of data that displayed the no. of observation with in a given intervals.

In Simple words, It is the distribution of frequencies with respect to their given interval or particular categories

Eg :- Akash, Ravi, Pradeep drink a tea respectively

ⒶⓇⓅⒶⓇⒶⓅⒶⓇⓇⒶ

In this Raw data, Akash drinks the tea 5 times.
Ravi " 4 times
Pradeep " " 2 times.

The data can be shown as.

Sl. No	content	frequency
1	Akash	5
2	Ravi	4
3	Pradeep	2

* Types of frequency Distribution [F.D]

1) Discrete [un grouped] F.D

2) Continous [grouped] F.D

3) cumulative frequency distributions

DATA DISTRIBUTION

1) Discrete frequency Distributions

- Here Datas are presented in the form of ungrouped
- class intervals are not given.

eg :-

S.NO	Variables	frequency
1	Akash	5
2	Ravi	4
3	Pradeep	2

* examples of data commonly organized in this manner include gender, ethnicity, marital status, diagnostic category of study subject.. etc.

2) Continous frequency Distribution (Grouped)

- Here class intervals are used as variables.
- in this, Variable will be continous such as age, salary, etc are examined.

Eg :- Results of students got marks

Sl NO	Marks	No. of students
1	0-20	8
2	20-40	14
3	40-60	21

Components

i) class :- Groups according to size of data [(0-20), (20-30) etc.]

ii) classlimit :- smallest & largest possible measument in each class.

eg : 0-10, here

lower limit = 0

upper limit = 10

DATA DISTRIBUTION

iii) class interval = upper limit – lower limit

iv) class mark = middle value of class.

= $\frac{1}{2}$ (Lower limit + upper limit)

eg, in class '10-20',

classmark = $\frac{10+20}{2} = 15$

Types of continous F.D

A) Exclusive series

B) inclusive series.

Exclusive

class interval does not include upper class limit.

eg:-

marks	no. of students
0 – 20	6
20 – 40	11
40 – 60	32

inclusive

class interval that include upper class limit.

eg:-

marks	no. of students
0 – 19	6
20 – 39	11
40 – 60	32

3) cumulative F.D

- It is a type of distribution in which frequencies of variables are summed as one moves from the top of table to the bottom.

- Thus bottom categories would have a cumulative frequencies equivalent to the total size.

DATA DISTRIBUTION

eg:-

Sl. NO	~~Variables~~ (Name)	Frequency	cumuladive Frequency	cumolative %
1	Akash	10	10	25%
2.	Ravi	5	15	37.5%
3.	pradeep	8	23	57.5%
4	Neha	6	29	72.5%
5	Ganesh	11	$\boxed{40}$	100%.

Total no. od data.

DATA DISTRIBUTION