

We at The Data Monk hold the vision to make sure everyone in the IT industry has an equal stand to work in an open domain such as analytics. Analytics is one domain where there is no formal under-graduation degree and which is achievable to anyone and everyone in the World.

We are a team of 30+ mentors who have worked in various product-based companies in India and abroad, and we have come up with this idea to provide study materials directed to help you crack any analytics interview.

Every one of us has been interviewing for at least the last 6 to 8 years for different positions like Data Scientist, Data Analysts, Business Analysts, Product Analysts, Data Engineers, and other senior roles. We understand the gap between having good knowledge and converting an interview to a top product-based company.

Rest assured that if you follow our different mediums like our blog cum questions-answer portal www.TheDataMonk.com , our youtube channel - [The Data Monk](#), and our e-books, then you will have a very strong candidature in whichever interview you participate in.

There are many blogs that provide free study materials or questions on different analytical tools and technologies, but we concentrate mostly on the questions which are asked in an interview. We have a set of 100+ books which are available both on Amazon and on [The Data Monk e-shop page](#)

We would recommend you to explore our website, youtube channel, and e-books to understand the type of questions covered in our articles. We went for the question-answer approach both on our website as well as our e-books just because we feel that the best way to go from beginner to advance level is by practicing a lot of questions on the topic.

We have launched a series of 50 e-books on our website on all the popular as well as niche topics. Our range of material ranges from SQL, Python, and Machine Learning algorithms to ANN, CNN, PCA, etc.

We are constantly working on our product and will keep on updating it. It is very necessary to go through all the questions present in this book.

Give a rating to the book on Amazon, do provide your feedback and if you want to help us grow then please subscribe to our Youtube channel.

STATISTICS

Q1) What values can a Discrete Random Variable hold?

- a) Only Whole Numbers
- b) Only Floating Numbers
- c) Both Whole Numbers and Floating Numbers
- d) None

Ans) Option 'A'

EXPLANATION: A Discrete Random Variable can hold only whole numbers like 2,3,7,34,0 etc.

Q2) What values can a Continuous Random Variable hold?

- a) Only Whole Numbers
- b) Only Floating Numbers
- c) Both Whole Numbers and Floating Numbers
- d) None

Ans) Option 'C'

EXPLANATION: A Continuous Random Variable can hold both floating and whole numbers like 1.2,4,4.7,9.78,34 etc.

Q3) What are measures of Dispersion?

Ans) Measures of Dispersion talk about the distribution and uniformity of data.

Q4) Choose the odd one out of the following:

Range, Interquartile Range, Mode, Variance, Mean Deviation, Standard Deviation

Ans) 'Mode'

EXPLANATION: Range, Interquartile Range, Variance, Standard Deviation and Mean Deviation are the most commonly used Measures of Dispersion while, Mode is a measure of Central Tendency.

Q5) What is mean Deviation?

Ans) Mean deviation measures the average of the deviation of the sample observations from the sample mean.

Q6) Calculate the Mean Deviation of the following observations:

S.no.	Values (X)
1.	20
2.	32
3.	18
4.	10
5.	40

Ans) Follow the given series of steps in order to calculate the mean deviation of the given dataset,

Step 1) Calculate the **mean** of the values,

$$\text{So, mean} = \frac{20+32+18+10+40}{5} = \frac{120}{5} = \mathbf{24}$$

Step 2) Calculate the deviation of each observation from the mean,

Values	Deviation from Mean (mean = 24)
20	4
32	8
18	6
10	14
40	16

****Take the absolute value of the deviation****

Step 3) Calculate the **mean of the deviation** calculated for each sample,

$$\text{Mean deviation} = \frac{4+8+6+14+16}{5} = \frac{48}{5} = \mathbf{9.6}$$

So, this is how one can calculate the mean deviation, which in this case comes out to be **9.6**

Q7) What is the relation between Variance and Standard Deviation?

- a) Variance = (Standard Deviation)²
- b) Variance = (Standard Deviation)^{1/2}

- c) Variance = (Standard Deviation)³
 d) Variance = (Standard Deviation)^{1/3}

Ans) Option 'A'

EXPLANATION: Variance is the squared value of Standard Deviation.

Q8) What is Standard Deviation?

Ans) Standard Deviation is the root of mean of squares of deviations from the mean, i.e.,

$$\text{(Standard Deviation)} \quad \sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2},$$

Q9) Which of the following statements is/are true in reference to Standard Deviation and Variance:

- a) High value implies values are close to the mean.
 b) High value implies values are spread out.
 c) Low value implies values are close to the mean.
 d) Low value implies values are spread out.

Ans) Statement 'B' and 'C' are true.

Q10) Calculate the Variance and Standard Deviation of the given,

S.no.	Values (X)
1	20
2	32
3	18
4	10
5	40

Ans) Follow the given series of steps in order to calculate the mean deviation of the given dataset,

Step 1) Calculate the mean of the values,

$$\text{So, mean} = \frac{20+32+18+10+40}{5} = \frac{120}{5} = \mathbf{24}$$

Step 2) Calculate the deviation of each observation from the mean and square them,

Values	Deviation (X(i)-mean(X))	Square of deviation
20	4	16
32	8	64
18	6	36
10	14	196
40	16	256

Step 3) Calculate the mean of squares of deviation to get the VARIANCE,

$$\text{VARIANCE} = \frac{16+64+36+196+256}{5} = 113.6$$

Step 4) Calculate Standard Deviation, i.e., square root of variance,

$$\text{So, Standard Deviation} = \sqrt{113.6} = 10.65$$

So, this is how we can calculate the variance and standard deviation of the given dataset.

Q11) Compare the closeness of the values in the given two datasets on the basis of Standard Deviation and Variance.

Values of Dataset 1: {2, 8, -5, 7, -12}

Values of Dataset 2: {15, 21, -30, 11, -17}

Ans) For the given set of steps to find out the dataset with less distribution,

Step 1) Calculate mean of both the samples,

$$\text{Mean of Dataset 1} = \frac{2+8+(-5)+7+(-12)}{5} = 0$$

$$\text{Mean of Dataset 2} = \frac{15+21+(-30)+11+(-17)}{5} = 0$$

Step 2) Calculate the Variance and Standard Deviation of both the dataset,

Dataset 1:

- Variance = $(2^2+8^2+(-5)^2+7^2+(-12)^2)/5 = 57.2$
- Standard Deviation = $\sqrt{57.2} = 7.563$

Dataset 2:

- Variance = $(15^2 + 21^2 + (-30)^2 + 11^2 + (-17)^2) / 5 = 395.2$
- Standard Deviation = $\sqrt{395.2} = 19.88$

Step 3) Compare the values of standard Deviation of both the datasets and come up with the required results,

In this case we observe that the values of **Variance and Standard Deviation of Dataset 2 is much greater than that of Dataset 1**. So, we can say that values of Dataset 1 are relatively much closer than the values of Dataset 2.

Q12) What is the range of a sample?

Ans) Range is the Difference between the highest and lowest value of the sample.

Calc 4) Calculate the range of the given sample,

S.no.	Values (X)
1	20
2	32
3	18
4	10
5	40

Ans) Range of a sample = Highest Value – Lowest Value

- Highest Value in this sample = 40
- Lowest Value in this sample = 10

Range = 40 – 10 = 30.

Q13) What is Quartile Distribution?

Ans) The Dataset is divided into four equal parts known as **Lower (First) Quartile, Median (Second Quartile), Higher (Third Quartile)**.

Q14) Match the following:

Column A contains the three Quartiles while Column B contains the position of the Quartiles and Column C contains the percentile values that correspond a particular quantile, match the quartiles with their respective matches of other columns in an ordered dataset. (n=number of observations in the sample)

Column A	Column B	Column C
1) Q_1	a) $(\frac{3(n+1)}{4})^{\text{th}}$ term	I. 50%
2) Q_2	b) $(\frac{n+1}{4})^{\text{th}}$ term	II. 75%
3) Q_3	c) $(\frac{n+1}{2})^{\text{th}}$ term	III. 25%

Ans) 1) – b) – III, 2) – c) – I, 3) – a) – II.

EXPLANATION: Subsequent positions assigned to each of the quartile will show the position of the percentage of Values below that value in an ordered dataset.

Q15) Calculate the Quartile Values of the given set of values,

12, 34, -5, -12, 10, 2, -1, 15, 8

Ans) Follow the given series of steps in order to calculate the Quartile Values,

Step 1) Arrange the sample in ascending order, so here it goes,

-12, -5, -1, 2, 8, 10, 12, 15, 34

Step 2) Calculate the position of each Quartile,

- $Q_1 = (9+1)/4 = 10/4 = 2.5^{\text{th}}$ term
(This implies the mean of 2nd and 3rd term)
- $Q_2 = (9+1)/2 = 10/2 = 5^{\text{th}}$ term
- $Q_3 = (3(9+1))/4 = 30/4 = 7.5^{\text{th}}$ term
(This implies the mean of 7th and 8th term)

Step 3) Assign the values, so,

- $Q_1 = (-5-1)/2 = -3$
- $Q_2 = 10$
- $Q_3 = (12+15)/2 = 27/2 = 13.5$

Q16) What is inter quartile range (IQR),

- a) Median
- b) Median – Lower Quartile
- c) Median – Upper Quartile
- d) Upper Quartile – Lower Quartile

Ans) Option ‘D’

EXPLANATION: inter quartile range is the difference between the values of Upper Quartile and lower Quartile.

Q17) Calculate the inter quartile range (IQR) for the set of values given in Calc5.

Ans) From the above Calculation,

- Upper Quartile = 13.5
- Lower Quartile = -3

So, IQR = 13.5 – (-3) = 13.5+3 = 16.5

Q18) Match the measures of dispersion with their respective formulas,

Measures of dispersion	Formulas
1. Mean Deviation	a) $Q_3 - Q_1$
2. Variance	b) $(\sum_{i=1}^n (x(i) - \text{mean}(x)))/n$
3. Standard Deviation	c) $\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$
4. Range	d) $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$,
5. Interquartile Range	e) $\text{Max}(x_i) - \text{Min}(x_i)$

Ans) 1- b, 2-c, 3-d, 4-e, 5-a

Q19) Which of the following statements are true with respect to PDF (Probability Density Function) and PMF (Probability Mass Function)?

- a) PDF is for continuous random variable while PMF is for discrete random variable.
- b) PDF is for discrete random variable while PMF is for continuous random variable.
- c) Both PMF and PDF can be used for discrete random variable.
- d) Both PMF and PDF can be used for continuous random variable.

Ans) Option 'A'

EXPLANATION: PDF takes into account a range of values and gives the probability accordingly, while **PMF** takes into account individual discrete values and gives probability of particularly those discrete input values.

Q20) What is the formula for calculating probability over a given range of data (a, b) if the probability distribution function is f(x),

- a) $P(a < x < b) = P(b) - P(a)$
- b) $P(a < x < b) = \int_a^b f(x) \cdot dx$
- c) $P(a < x < b) = \frac{df(x)}{dx}$
- d) None of the above

Ans) Option 'B'

EXPLANATION: The probability between a given range can be calculated by calculating the area under the curve of the probability distribution function that surely can be achieved by integrating the function within the given range.

Q21) What is the sum of probabilities over the whole range of dataset within a probability distribution function?

- a) 0
- b) 0.5
- c) 1
- d) Cannot be determined

Ans) Option 'C'

EXPLANATION: The **sum of all the probabilities** lying with the range of PDF must **equal to 1**.

Q22) How will you calculate the probability of a value 'x' in a discrete variable dataset?

- a) $P(x) = P(X=x)$
- b) $P(x) = \int f(x)$
- c) $P(x) = \frac{df(x)}{dx}$
- d) None of the above

Ans) Option 'A'

EXPLANATION: To find the probability in a discrete variable dataset we should apply PMF and option 'A' shows the formula of PMF, i.e., the probability at one specific value (x in this case).

Q23) Choose the odd one out of the following:

Binomial Distribution, Poisson Distribution, Exponential distribution, Geometric distribution, Negative Binomial distribution, Hypergeometric distribution

Ans) 'Exponential distribution'

EXPLANATION: Exponential distribution is an example of **continuous distribution** where **PDF** can be applied while others are **discrete distribution** where **PMF** is applied.

Q24) Choose the odd one out of the following:

Exponential distribution, Gamma distribution, Normal distribution, Lognormal distribution, Poisson distribution.

Ans) 'Poisson distribution'

EXPLANATION: All others are continuous distributions where PDF can be applied while Poisson distribution is discrete distribution.

Q25) 'Probability distribution function can attain negative values' State whether the statement is TRUE or FALSE.

Ans) FALSE.

EXPLANATION: PDFs can only attain positive value or zero (*probability at a discrete point in PDF or continuous variable dataset is always zero*)

Q26) What kind of curve does a Gaussian Distribution follow,

- a) Parabolic
- b) Exponential
- c) Logarithmic
- d) Bell Curve

Ans) Option 'D'

EXPLANATION: A gaussian distribution follows the *bell curve*.

Q27) On what parameters is a gaussian distribution based on,

- a) Median and Variance
- b) Mean and Median
- c) Mean and Variance
- d) Mean and Standard Deviation

Ans) Option 'D'

EXPLANATION: The probability distribution of a Gaussian Distribution, takes into account the mean and standard deviation

Q28) What is the probability distribution function of a gaussian distribution?

Ans)

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

Following is the formula of probability distribution formula for gaussian distribution, here,

'x' denotes the variable,

Sigma denotes the standard deviation,

Meu denotes the mean.

Q29) 'There are no empirical percentage of values within different regions of the bell curve under gaussian distribution' State whether the following statement is true or false.

Ans) FALSE.

EXPLANATION: There indeed are standard percentage of values lying in different regions.

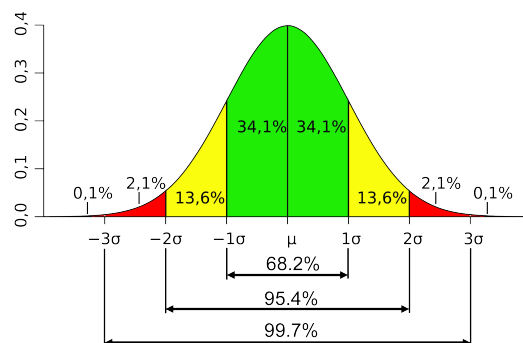
Q30) Match the columns.

Columns A contains specific ranges within the gaussian distribution and Column B contains the empirical percentage of values lying within that range.

Column A	Column B
a) Within one standard deviation of mean	1. 99.7%
b) Within one standard deviation of mean	2. 68%
c) Within one standard deviation of mean	3. 95%

Ans) a-2, b-3, c-1

EXPLANATION: The diagram mentioned below explains the values distribution in a gaussian bell curve.



Q31) Calculate the gaussian distribution for x = 5, mean = 7 and standard deviation = 2.

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

Ans) Putting the values in the formula

we get,

$$f(5, 7, 2) = 0.099 * 0.606 = \mathbf{0.06}$$

This tells us that for a hypothetical function that follows gaussian distribution with mean = 7 and standard deviation = 2, the probability at x = 5 will turn out to be 0.06

Q32) State whether the following statement is true or false,

'We can't estimate the probability distribution of a continuous random variable not following Gaussian Distribution through an empirical formula'

Ans) TRUE.

EXPLANATION: We probably cannot estimate the probability distribution of a random variable not following gaussian distribution by an empirical formula precise enough to estimate the probability distribution for every standard deviation. However, we can surely set a lower limit of probability distribution for random variable not following gaussian distribution with the help of ***Chebyshev's inequality***.

Q33) Which of the following is Chebyshev's inequality?

- a) $P(\text{mean} - k(\text{SD}) < x < \text{mean} + k(\text{SD})) > 1 - 1/k^2$
- b) $P(\text{mean} - k(\text{SD}) < x < \text{mean} + k(\text{SD})) > 1 - 1/k^3$
- c) $P(\text{mean} - k(\text{SD}) < x < \text{mean} + k(\text{SD})) > 1 - 1/k$
- d) None of the above

Here, k is the number of standard deviations,

Ans) Option 'A'

EXPLANATION: Option 'A' correctly states **Chebyshev's inequality**, For example,

$$\text{For } k=1, P(\text{mean} - 1(\text{SD}) < x < \text{mean} + 1(\text{SD})) > 1 - 1/1^2$$

$P(\text{mean} - 1(\text{SD}) < x < \text{mean} + 1(\text{SD})) > 0.$

For $k=2$, $P(\text{mean} - 2(\text{SD}) < x < \text{mean} + 2(\text{SD})) > 1 - 1/2^2$

$P(\text{mean} - 2(\text{SD}) < x < \text{mean} + 2(\text{SD})) > 3/4$ and so on.

Q34) What is the condition for a random variable say $X(x_1, x_2, x_3 \dots x_n)$ to belong to a Log Normal Distribution?

1. $\ln(x_i)$ should be valid.
2. $\ln(x_i)$ should follow a gaussian or normal distribution.

Options:

- a) Only statement 1 is required
- b) Only statement 2 is required
- c) Both the conditions are required
- d) None of the conditions are required

Ans) Option 'C'

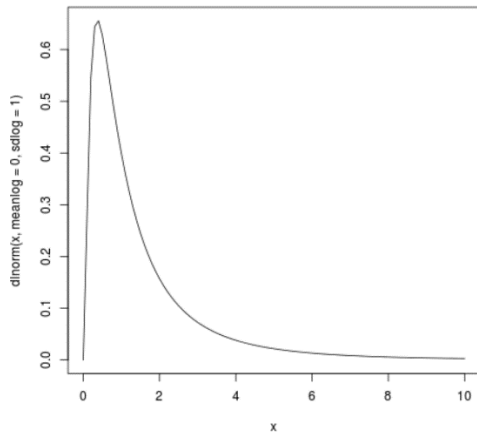
EXPLANATION: For a sample X , the log of its components should be valid and follow gaussian distribution in order to belong to a Log Normal Distribution.

Q35) What is the difference between the curve of gaussian distribution and log normal distribution?

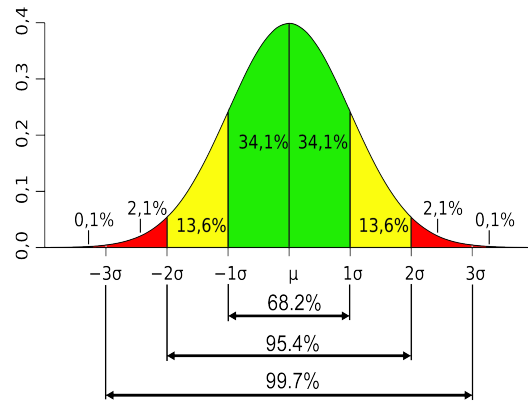
- a) Gaussian distribution is exactly symmetric (bell curve) while log normal distribution curve is skewed left.
- b) Gaussian distribution is exactly symmetric (bell curve) while log normal distribution curve is skewed right.
- c) Log normal distribution is exactly symmetric (bell curve) while gaussian distribution curve is skewed left.
- d) Log normal distribution is exactly symmetric while gaussian distribution curve is skewed left.

Ans) Option 'B'

EXPLANATION: Below mentioned are the gaussian distribution and log normal distribution curves which pretty well validate option 'B'.



Log Normal Distribution Curve



Gaussian Distribution Curve

Q36) When does a standard normal distribution occur?

- a) Mean = 0 and Standard deviation = 1.
- b) Mean = 1 and Standard deviation = 0.5
- c) Mean = 0.5 and Standard deviation = 1
- d) None of the above.

Ans) Option 'A'

Q37) What is Standard scaler and how is the score of standard scaler calculated?

Ans) Standard scaling is one of the techniques used for scaling the data to get a uniform distribution or to say that bringing all the whole data on a similar level. The score of standard scaler is calculated as,

$$\text{Score} = \frac{x(i) - \text{mean}(x)}{\text{standard deviation}(x)}$$

Q38) What is the standard scaler score referred to as?

- a) P-value
- b) Median
- c) Z-score
- d) None

Ans) Option 'C'

Q39) What is cumulative frequency distribution?

Ans) Cumulative frequency distribution finds out the number of observations that lie above a specific observation. The graph starts with the least value in the observation set and then moves on with appending the values of other observations till it covers all the observations.

Q40) How to calculate the probability within a range say (a, b) within a cumulative distribution function?

- a) $P(a < x < b) = P(b) - P(a)$
- b) $P(a < x < b) = \int_a^b f(x) \cdot dx$
- c) $P(a < x < b) = \frac{df(x)}{dx}$
- d) None of the above

Ans) Option 'A'

EXPLANATION: Cumulative distribution graph adds up the probability values rather than showing discrete values, so to find the values within a range one might want to subtract the lower probability from the higher probability.

Q41) What is log Normalization?

Ans) If a dataset follows log normal distribution,

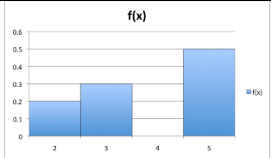
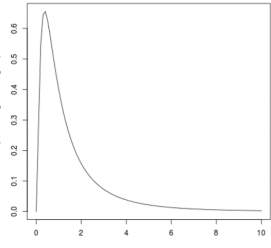
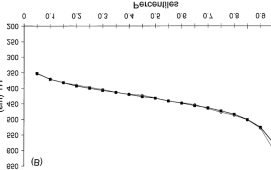
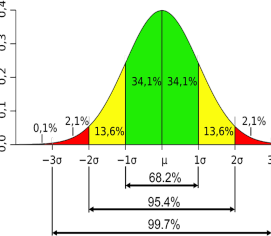
- The log of its individual values is taken which follows a ***gaussian distribution***.
- The gaussian distribution is converted to ***standard normal distribution*** with the help of ***z-score***.
- This converted dataset can further be used for various operations and comparisons if required.

Q42) Match the columns:

Column A contains various data distribution function or techniques.

Column B contains a formula or mathematical statement related to any of the data distribution functions or techniques mentioned in Column A.

Column C contains a graph related to any of the functions or techniques mentioned in Column A.

Column A	Column B	Column C
1. Gaussian Distribution	a) $P(x) = P(X=x)$	i. 
2. Cumulative Frequency Distribution	b) $f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	ii. 
3. Log Normal Distribution	c) $P(a < x < b) = P(b) - P(a)$	iii. 
4. Discrete Variable Distribution	d) $f_X(x) = \frac{1}{\sigma x \sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$	iv. 

Ans) 1-b-iv, 2-c-iii, 3-d-ii, 4-a-i

EXPLANATION: Above shown are various important ways of data distribution along with their formulas and respective graphs.

Q43) What is Central limit theorem?

Ans) According to central limit theorem, if we have a *dataset* (say X) with real values of *mean and standard deviation (SD)*, and create various sample out of it (say ' n ' number of samples) then the means of the samples will follow *normal distribution (bell curve)* and **mean of the means of all the samples will turns out to be near about the mean of original dataset with the standard deviation of the ratio of the standard deviation of original dataset (SD) and the root of number of samples (n).**

X (mean(X), SD) be our original dataset,

$x_1 x_2 x_3 \dots x_n$ be the samples created out of X ,

mean(x_1), mean(x_2), mean(x_3) ... mean(x_n) are the mean of all the samples, then these will follow normal distribution with $\frac{\sum \text{mean}(x(i))}{n} = \text{mean}(X)$ with **standard deviation** = $\frac{SD}{n}$.

Q44) What is the effect of sample size on Central limit theorem?

- a) Sample size has no effect on central limit theorem.
- b) Larger the sample size, more deviation between the mean of original dataset and the mean of means of samples.
- c) Larger the sample size, lesser is the deviation between the mean of original dataset and the mean of means of the sample
- d) None

Ans) Option 'C'

EXPLANATION: A larger sample size will ensure more efficient consideration of data and more suitable result in accordance with the central limit theorem.

Q45) What assumptions should be taken care of while applying Central Limit Theorem?

- a) Sample should draw values randomly from the dataset
- b) Each sample should be independent of each other
- c) Sample should be of sufficient large size
- d) All of the above

Ans) Option 'D'

EXPLANATION: Above mentioned points should be taken care of for efficient implementation of the *central limit theorem*.

Q46) What is Binomial Distribution?

Ans) Binomial distribution is a discrete probability distribution. (Mean it gives only two values, 1 or 0, true or false, success or failure)

Q47) What is the probability of a discrete dichotomous distribution with 'n' experiments, 'p' being the probability of 'True' and 'q' or '1-p' being the probability of 'false'? (Binomial Probability Distribution)

- a) $P(x:n,p) = {}^nC_x p^x (q)^{n-x}$
- b) $P(x:n,p) = {}^nP_x p^x (q)^{n-x}$
- c) $P(x:n,p) = {}^nC_x q^x (p)^{n-x}$
- d) None

Ans) Option 'A'

Q48) What are the mean and variance of a binomial distribution (notations same as above question)?

- a) np and npq respectively
- b) npq and np respectively
- c) nq and npq respectively
- d) npq and nq respectively

Ans) Option 'A'

Q49) Which of the following statements are true with respect to Poisson Distribution?

- a) Special Case of Binomial Distribution
- b) 'n' tends to infinity
- c) 'p' tends to zero
- d) np = 1 is finite

Ans) Option 'A' 'B' 'C' 'D'

EXPLANATION: All the above-mentioned statements are true with respect of *Poisson Distribution*.

Q50) Which of the following statements seem true with respect to Covariance?

- a) It defines the level of dependence of two variables
- b) Its value ranges from negative infinity to infinity

- c) Its value changes under different scales of measurement of variable
- d) All of the above

Ans) Option 'D'

EXPLANATION: Covariance is an extended form of variance for comparing the dependence of two variables.

Q51) Which of the following options holds true regarding the relationship of variance and covariance?

- a) $\text{Cov}(x, y) = \text{Variance}(x)$
- b) $\text{Cov}(x, y) = \text{Variance}(y)$
- c) $\text{Cov}(x, x) = \text{Variance}(x)$
- d) $\text{Cov}(y, y) = \text{Variance}(x)$

Ans) Option 'C'

EXPLANATION: $\text{Cov}(X, Y) = \frac{\sum (X_i - \bar{X})(Y_j - \bar{Y})}{n}$ and $\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ are the formulas of **covariance** and **variance** respectively and it is quite clear that replacing 'y' in covariance from 'x' will lead to the same formula as that of Variance, i.e., stated by option 'C'

Q52) Which of the following statements are true regarding the relation between covariance and correlation for two random variables A and B?

- a) There is no relation between them
- b) $\text{Covariance} = \text{Correlation} * \text{Standard deviation}(A) * \text{Standard deviation}(B)$
- c) $\text{Covariance} = \text{Correlation} * \text{Variance}(A) * \text{Variance}(B)$
- d) $\text{Correlation} = \text{Covariance} * \text{Standard deviation}(A) * \text{Standard deviation}(B)$

Ans) Option 'B'

EXPLANATION: People often interpret correlation and covariance as the same quantity, however there are various points of difference and the relation between them is as shown in option 'B'

Q53) Which of the following statements seem true with respect Correlation?

- a) It states the relation between two variables.
- b) Its value ranges between +1 and -1.
- c) Its values don't change on change the scale of variable in the dataset
- d) All of the above

Ans) Option 'D'

EXPLANATION: Correlation talks about the way variables change with respect to each other and options 'A' 'B' and 'C' pretty much talk about its specifics.

Q54) What is the common limitations of Covariance and Correlation?

- a) Lengthy and outdated process for comparison of variables.
- b) Applicable only for linear relationship.
- c) Both A and B
- d) None of the above.

Ans) Option 'B'

EXPLANATION: Both the concepts apply for linear relationship between variables, any other parabolic or logarithmic relation may seem to provide absurd results. ***Correlation usually is used more as compared to covariance due to its scalability factor.***

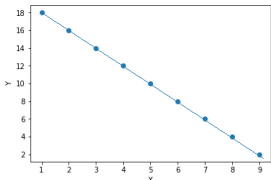
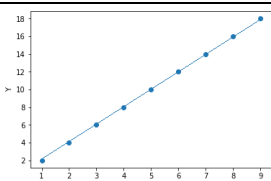
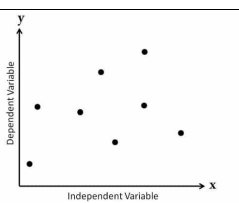
Q55) What of the following statements are true with respect to positive and negative Covariance for two random variables X and Y?

- a) If X and Y follow a direct proportionality (if x increases y increases; x decreases Y decreases) it is a case of positive Covariance.
- b) If X and Y follow an inverse proportionality (if x increases y decreases; x decreases y increases) it is a case of negative Covariance.
- c) If X and Y follow a direct proportionality (if x increases y increases; x decreases Y decreases) it is a case of negative Covariance.
- d) If X and Y follow an inverse proportionality (if x increases y decreases; x decreases y increases) it is a case of positive Covariance.

Ans) Option 'A' and 'B' are true.

Q56) Match the columns:

Column A contains the value of Correlation coefficients (r). Column B contains the implication of any of the values in Column A. Column C contains a graph showing the relationship between the two variables on the basis of their Correlation Values.

Column A	Column B	Column C
1. 0	a) High positive correlation (Value of one variable strictly increases with another variable)	i. 
2. 1	b) High negative correlation (Value of one variable strictly decreases with another variable)	ii. 
3. -1	c) No Correlation	iii. 

Ans) 1-c-iii, 2-a-ii, 3-b-i

Q57) State three major types of Correlation coefficient.

Ans) Here are the following majorly used correlation coefficients:

- *Pearson's correlation coefficient (r)*
- *Spearman rank correlation coefficient*
- *Kendall rank correlation coefficient*

Q58) Which of the following should be assumed for Pearson correlation?

- Linearity of the two variables
- Homoscedasticity (data equally distributed)
- Both variables should be normally distributed

d) All of the above

Ans) Option 'D'

Q59) When is the ideal situation for spearman rank correlation?

- a) When the data is mostly ordinal
- b) When the data contains equal ordinal and nominal points
- c) When the data is least ordinal
- d) It has nothing to do with ordinal or nominal data.

Ans) Option 'C'

EXPLANATION: Spearman rank correlation has nothing to do with the distribution of data (unlike Pearson that requires normally distributed data) but *spearman rank proves to be the most effective correlation analysis when the variables are least ordinal.*

Q60) What is true about concordant and discordant pairs?

- a) If $x_2 - x_1$ is same sign as that of $y_2 - y_1$ then it is a concordant pair.
- b) If $x_2 - x_1$ is same sign as that of $y_2 - y_1$ then it is a discordant pair.
- c) If $x_2 - x_1$ has a different sign as that of $y_2 - y_1$ then it is a concordant pair.
- d) If $x_2 - x_1$ has a different sign as that of $y_2 - y_1$ then it is a discordant pair.

Ans) Option 'A' and 'D' are true.

Q61) Which of the following is/are a parametric test?

- a) Pearson
- b) Spearman
- c) Kendall
- d) None

Ans) Option 'A'

EXPLANATION: Parametric tests are those tests which are based on initial assumptions while non parametric are those which are not based on any initial assumptions.

Q62) What is the difference between Pearson Correlation and Point-Biserial Correlation?

- a) Pearson is common for categorical variables while Point-Biserial takes into account numerical variables
- b) They have different formulas
- c) Pearson has one dichotomous while Point-Biserial has two numerical variables
- d) Pearson takes into account numerical variables while Point-Biserial has one dichotomous variable.

Ans) Option 'D'

EXPLANATION: Both the correlation coefficients have the same formula but Point-Biserial takes into account a dichotomous variable.

Q63) Match the Correlation coefficients with their respective formulas.

Coefficient	Formula
1. Pearson	a) $\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$
2. Spearman Rank	b) $r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$
3. Kendall Rank	c) $\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$

Ans) 1-b, 2-c, 3-a

EXPLANATION: Following are the formulas of respective correlation coefficients. The variables denote the following,

- n_c – Number of concordant pairs
- n_d – Number of discordant pairs
- n – sample size

- d_i – difference between rank of corresponding variables
- x_i – Value of x (for i^{th} observation)
- y_i – Value of y (for i^{th} observation)

Q64) What is skewness?

Ans) Skewness is the deviation of random variables from normal distribution, i.e., the graph of the variables is **not a symmetric bell curve**, while it is **distorted** a bit towards the left or right depending on the values.

Q65) What are the types of skewness?

Ans) Skewness is of two types:

1. *Positive Skewness*
2. *Negative Skewness*

Q66) Which of the following are referred to as right skewed and left skewed respectively?

- a) Negative Skewness and Positive Skewness respectively
- b) Positive Skewness and Negative Skewness respectively
- c) No Skewness and Presence of skewness respectively
- d) Presence of skewness and no skewness respectively

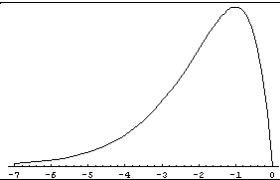
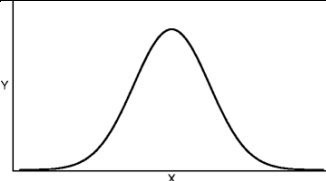
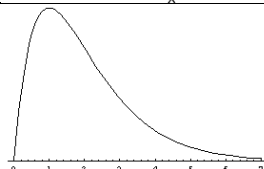
Ans) Option 'B'

EXPLANATION: Positive skewness is often referred to as Right skewed (will be more understandable when referred to the graph of positive skewness) and Negative skewness is often referred to as left skewed.

Q67) Match the columns.

Column A contains various types of skewness and Column B corresponds to graphs related to the options in Column A.

Column A	Column B
----------	----------

1. Positive (Right) Skewness	a) 
2. Negative (Left) Skewness	b) 
3. No Skewness	c) 

Ans) 1-c, 2-a, 3-b

EXPLANATION) Column B options 'b' seems like a normal distribution and hence has no skewness, option 'a' has a longer left leg, hence is negatively skewed, while option 'c' has longer right leg and is thus positively skewed.

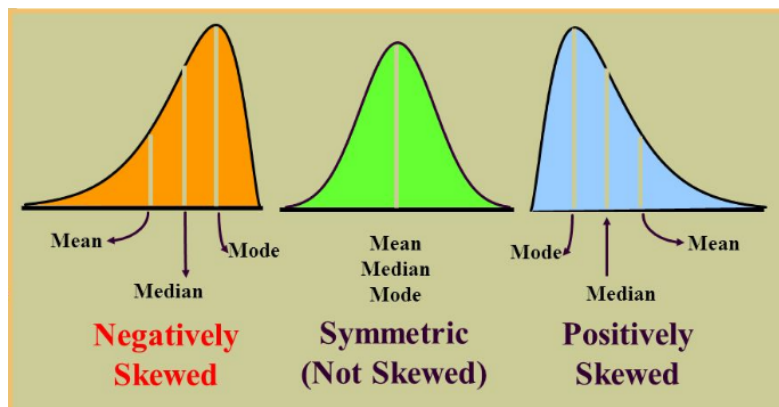
Q68) Match the columns.

Column A contains different types of skewness while column B contains the relation of mean, median and mode in different conditions as mentioned in Column A.

Column A	Column B
1. Positive Skewness	a) Mean = Median = Mode
2. Negative Skewness	b) Mean > Median > Mode
3. No Skewness	c) Mode > Median > Mean

Ans) 1-b, 2-c, 3-a.

EXPLANATION:



The relation between mean, median and mode seems pretty obvious and understandable from the diagram given above.

Q69) Which of the following is an example of positive skewed dataset?

- a) Income Distribution
- b) Life span of human beings
- c) Both
- d) None of the above

Ans) Option 'A'

EXPLANATION: There are very few people with huge amount of wealth or income (Big businessmen and industrialists) leading to a distribution with a *tail towards the right (Positively Skewed)*

Q70) Which of the following is an example of negative skewed dataset?

- a) Income Distribution
- b) Life span of human beings
- c) Both
- d) None of the above

Ans) Option 'B'

EXPLANATION: There are relatively very few number of people who die at an early age leading to a *tail at the start or the left side (Negatively Skewed)*.

Q71) Which of the following is the correct formula of Pearson's Mode skewness?

- a) $\frac{\text{Mean} - \text{Median} - \text{Mode}}{\text{Variance}}$
- b) $\frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$
- c) $\frac{\text{Mean} - \text{Mode}}{\text{Variance} * \text{Standard Deviation}}$

d) $\frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$

Ans) Option 'D'

Q72) Which of the following is the Pearson's Median Skewness?

a) $\frac{3\text{Mean} - \text{Median}}{\text{Standard Deviation}}$

b) $\frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$

c) $\frac{\text{Mean} - 3\text{Median}}{\text{Standard Deviation}}$

d) $\frac{\text{Mean} - \text{Median}}{3\text{Standard Deviation}}$

Ans) Option 'A'

Q73) How can we estimate skewness with the help of Quartiles?

- a) We can't determine skewness with the help of Quartiles
- b) Yule's Coefficient
- c) Pearson's Second Skewness Coefficient
- d) Inter quartile range

Ans) Option 'B'

EXPLANATION: Yule's Coefficient is used to determine the skewness with the help of Quartiles.

$$\text{Yule's Coefficient} = \frac{Q(3) - 2Q(2) + Q(1)}{Q(3) - Q(1)}$$

Q74) Which of the following are major data measurement or data collection scales in statistics?

- a) Ordinal Data
- b) Nominal Data
- c) Ratio Data

- d) Interval Data
- e) All of the above

Ans) Option 'E'

EXPLANATION: Data in statistics can be measured under the above mentioned four different scales.

Q75) What is Ordinal Data?

Ans) An ordered categorical dataset is called Ordinal Data.

Q76) What is Nominal Data?

Ans) An unordered categorical dataset (dataset however that can't be ordered at times) is called Nominal Data.

Q77) What is Interval Data?

Ans) Data which is measured at a particular interval is referred to as interval data.

Q78) What is Ratio Data?

Ans) Ratio Data is the data measurement techniques that is much like interval data. It however has the ability to set a value as '**True Zero**' which is not in interval data.

Q79) Which of the following Data measurement scales is used for categorical datasets?

- a) Ordinal
- b) Nominal
- c) Interval
- d) Ratio

Ans) Option 'A' and 'B'

Q80) Which of the following Data measurement scales is used for numerical datasets?

- a) Ordinal
- b) Nominal
- c) Interval
- d) Ratio

Ans) Option 'C' and 'D'

Q81) Match the Columns.

Column A contains a data measurement scales and Column B contains an example related to any of the scales.

Scales	Example Statement
1) Interval	a) Gender of a group of people
2) Ratio	b) Satisfaction scale for a hotel
3) Ordinal	c) Time
4) Nominal	d) Salary

Ans) 1-c, 2-d, 3-b, 4-a

EXPLANATION:

- The outcomes of Gender of a group of people will lead to a certain number of people being male and other female, it however has **nothing to do with the ordering** of male and female and as it is **categorical**, it came under **Nominal Data Measurement scale**.
- Satisfaction reviews for a hotel will be like 'Highly Satisfied' 'Neither Satisfied or Unsatisfied' and 'Not Satisfied at all', here **ordering** the variables becomes quite essential and it is a **categorical** dataset.
- Time for any event is measured in **intervals** and it deals with **numerical values**.
- Salary is represented in intervals, and for an unemployed person can be set to zero, **Ratio data** should be the ideal choice over here.

Q82) What is the difference between Ratio and Interval Data Measurement scales?

Ans) Interval doesn't have an entity 'True Zero' while Ratio does have a 'True Zero'. Also, Ratio allows multiplication and division of variables that Interval doesn't.

Q83) What is the difference between Ordinal and Nominal Data measurement scales?

Ans) Variables are **ordered** in Ordinal data while they are **unordered** in Nominal data.

Q84) What is the advantage of Nominal Data over Ordinal Data?

Ans) Respondents have the freedom to express themselves the way they want to in Nominal Data, while they are restricted to a set of options in Ordinal data. *Example, a hotel review asking for a feedback text is more efficient than providing a satisfaction scale to the customers to rate from.*

Q85) What is the advantage of Ordinal Data over Nominal Data?

Ans) Ordinal Data makes it easier for the researcher to deal with the data without wasting time. Giving freedom to express in Nominal Data comes with a disadvantage of a lot of **irrelevant data that needs to be filtered**. However, this is not required in Ordinal Data. Example, a hotel review asking for a feedback text can precisely understand the customers' concerns but it will surely take a lot of time and will come with a significant amount of irrelevant data.

Q86) Which of the following measures of central tendency can be used for data analysis in both Nominal and Ordinal data?

- a) Mean
- b) Median
- c) Mode
- d) None

Ans) Option 'C'

EXPLANATION: Mode is a measure of central tendency that has nothing to do with the ordering of data and can be applied to categorical variables.

Q87) Which of the following measures of Central Tendency can be used only for Ordinal Data but not Nominal Data?

- a) Mean
- b) Median
- c) Mode
- d) None

Ans) Option 'B'

EXPLANATION: Median requires ordering of dataset and can be applied to categorical variables.

Q88) What upper hand does Numerical Data measurement scales (Interval and Ratio) possess over Categorical ones (Nominal and Ordinal)?

Ans) Data Collection and Analysis in numerical terms has always been better than categorical terms. One can **apply calculation** on numerical data to obtain a more detailed description that can't be done in case of categorical data.

However, with the ongoing advances in Natural Language Processing, one day this might not be a let down point for categorical data

Q89) What is scaling of data?

Ans) Scaling of data refers to the operation of bringing the whole dataset on a similar level or scale so that further analysis of the data can take place more efficiently.

Q90) Which of the following methods are used for scaling of data?

- a) Standardization
- b) Normalization
- c) Both
- d) None

Ans) Option 'C'

Q91) How the data is scaled from Normalization?

Ans) Normalization brings the whole dataset within 0 and 1 leading to a uniform scale for the whole dataset.

Q92) How is data scaled from Standardization?

Ans) Data is scaled on the basis of Standard Distribution in this case (*Mean = 0 and Standard deviation = 1*)

Q93) Choose the odd one out of the following,

KNN, K-mean clustering, K Nearest Neighbor, Linear Regression, XgBoost

Ans) XgBoost

EXPLANATION: All the algorithms except XgBoost require **Scaling**.

Q94) Choose the odd one out of the following,

AdaBoost, XgBoost, Random Forest, Decision Tress, Logistic Regression

Ans) Logistic Regression

EXPLANATION: Logistic Regression requires **Scaling**, others do not.

Q95) What is a Truncated Normal Distribution?

- a) Normal distribution in which the random variable is bounded from above or below or both
- b) Here negative elements are set to zero
- c) Few elements are outside a following range
- d) None

Ans) Option 'A'

Q96) What is rectified normal distribution?

- a) Normal distribution in which the random variable is bounded from above or below or both
- b) Here negative elements are set to zero
- c) Few elements are outside a following range
- d) None

Ans) Option 'B'

Q97) What is a Univariate Analysis?

Ans) In Univariate Analysis the data we are analyzing is only one variable. Example being Mean, Median, Mode, Standard Deviation etc.

Q98) What is Bivariate Analysis?

Ans) Analysis dealing with two variables is called Bivariate Analysis. It always has a 'X' and 'Y' or independent and dependent variables.

Q99) What is Multivariate Analysis?

Ans) It is a bit similar to Bivariate analysis but the variables taken into consideration is more than two.

Q100) State the data visualization techniques that can be used for Univariate, Bivariate and Multivariate Analysis.

- **Univariate** – Histogram, Box plot, Violin Plot
- **Bivariate** – Scatter Plot, Regression Plot
- **Multivariate** – We can't visualize data with more than two variables into consideration. Three variables will lead to a 3D distribution and further will go on with increasing number to variables to take into consideration.

