Statistics:

Total Marks: 40

Each question 10 marks

**Question: 1**

**What is the meaning of six sigma in statistics? Give proper example**

The term "Six Sigma" refers to a statistical measure of how far a process deviates from perfection. A process that operates at six sigma has a failure rate of only 0.00034%, which means it produces virtually no defects.

**The Six Sigma Tools**

1. Cause and Effect Analysis
2. Flow Chart
3. Pareto Chart
4. Histogram
5. Check Sheet
6. Scatter Plot
7. Control Chart

**Six Sigma Techniques**

1. Brainstorming

Brainstorming is the key process of any problem-solving method and is often utilized in the "improve" phase of the DMAIC methodology. It is a necessary process before anyone starts using any tools. Brainstorming involves bouncing ideas and generating creative ways to approach a problem through intensive freewheeling group discussions. A facilitator, who is typically the lead Black Belt or Green Belt, moderates the open session among a group of participants.

1. Root Cause Analysis/The 5 Whys

This technique helps to get to the root cause of the problems under consideration and is used in the "analyze" phase of the DMAIC cycle.

In the 5 Whys technique, the question "why" is asked, again and again, finally leading up to the core issue. Although "five" is a rule of thumb, the actual number of questions can be greater or fewer, whatever it takes to gain clarity.

1. Voice of the Customer

This is the process used to capture the "voice of the customer" or customer feedback by either internal or external means. The technique is aimed at giving the customer the best products and services. It captures the changing needs of the customer through direct and indirect methods. The voice of the customer technique is used in the "define' phase of the DMAIC method, usually to further define the problem to be addressed.

1. The 5S System

This technique has its roots in the Japanese principle of workplace energies. The 5S System is aimed at removing waste and eliminating bottlenecks from inefficient tools, equipment, or resources in the workplace. The five steps used are Seiri (Sort), Seiton (Set In Order), Seiso (Shine), Seiketsu (Standardize), and Shitsuke (Sustain).

1. Kaizen (Continuous Improvement)

The Kaizen technique is a powerful strategy that powers a continuous engine for business improvement. It is the practice continuously monitoring, identifying, and executing improvements. This is a particularly useful practice for the manufacturing sector. Collective and ongoing improvements ensure a reduction in waste, as well as immediate change whenever the smallest inefficiency is observed.

1. Benchmarking

Benchmarking is the technique that employs a set standard of measurement. It involves making comparisons with other businesses to gain an independent appraisal of the given situation. Benchmarking may involve comparing important processes or departments within a business (internal benchmarking), comparing similar work areas or functions with industry leaders (functional benchmarking), or comparing similar products and services with that of competitors (competitive benchmarking).

1. Poka-yoke (Mistake Proofing)

This technique's name comes from the Japanese phrase meaning "to avoid errors," and entails preventing the chance of mistakes from occurring. In the poka-yoke technique, employees spot and remove inefficiencies and human errors during the manufacturing process.

1. Value Stream Mapping

The value stream mapping technique charts the current flow of materials and information to design a future project. The objective is to remove waste and inefficiencies in the value stream and create leaner operations. It identifies seven different types of waste and three types of waste removal operations.

**Six Sigma Levels**

1. White Belt

This is the simplest stage, where:

Any newcomer can join.

People work with teams on problem-solving projects.

The participant is required to understand the basic Six Sigma concepts.

1. Yellow Belt

Here, the participant:

Takes part as a project team member.

Reviews process improvements.

Gains understanding of the various methodologies, and DMAIC.

1. Green level

This level of expertise requires the following criteria:

Minimum of three years of full-time employment.

Understand the tools and methodologies used for problem-solving.

Hands-on experience on projects involving some level of business transformation.

Guidance for Black Belt projects in data collection and analysis.

Lead Green Belt projects or teams.

1. Black Level

This level includes the following:

Minimum of three years of full-time employment

Work experience in a core knowledge area

Proof of completion of a minimum of two Six Sigma projects

Demonstration of expertise at applying multivariate metrics to diverse business change settings

Leading diverse teams in problem-solving projects.

Training and coaching project teams.

1. Master Black Belt

To reach this level, a candidate must:

Be in possession of a Black Belt certification

Have a minimum of five years of full-time employment, or Proof of completion of a minimum of 10 Six Sigma projects

A proven work portfolio, with individual specific requirements, as given here, for instance.

Have coached and trained Green Belts and Black Belts.

Develop key metrics and strategies.

**Six Sigma Career Choices and Salary Prospects-**

Six Sigma is widely adopted by many industries such as manufacturing, healthcare, finance, and retail, and offers a range of career opportunities with attractive salary prospects. Here are some career choices and salary prospects in Six Sigma:

Six Sigma Consultant: A Six Sigma consultant advises organizations on process improvements, identifies areas for cost savings, and develops strategies for implementation. The average salary for a Six Sigma consultant is around $96,000 per year.

Six Sigma Project Manager: A Six Sigma project manager oversees Six Sigma projects, manages project teams, and ensures successful implementation of process improvements. The average salary for a Six Sigma project manager is around $107,000 per year.

Six Sigma Black Belt: A Six Sigma Black Belt is responsible for leading Six Sigma projects, training team members, and ensuring sustained process improvements. The average salary is around $110,000 per year.

Six Sigma Master Black Belt: It is the highest level of Six Sigma certification and is responsible for leading organizational Six Sigma initiatives, coaching and mentoring Six Sigma Black Belts and Green Belts, and driving business transformation. The average salary for a Six Sigma Master Black Belt is around $140,000 per year.

Quality Manager: A Quality Manager ensures that an organization's products or services meet customer expectations, industry standards, and regulatory requirements. Six Sigma certification can be valuable for this role, and the average salary for a Quality Manager is around $91,000 per year.

Overall, Six Sigma offers various career opportunities with competitive salary prospects. Individuals with Six Sigma certification can expect higher salaries and better job prospects than those without certification.

**Question: 2**

**What type of data does not have a log-normal distribution or a Gaussian distribution? Give proper example**

1)Exponential distributions do not have a log-normal distribution or a Gaussian distribution. In fact, any type of data that is categorical will not have these distributions as well.

Example: Duration of a phone car, time until the next earthquake, etc.

2)Many random variables have distributions that are asymptotically Gaussian but may be significantly non-Gaussian for small numbers. For example the Poisson Distribution, which describes (among other things) the number of unlikely events occurring after providing a sufficient opportunity for a few events to occur. It is pretty non-Gaussian unless the mean number of events is very large. The mathematical form of the distribution is still Poisson, but a histogram of the number of events after many trials with a large average number of events eventually looks fairly Gaussian.

For me, the best examples come from my field of research (astrophysical data analysis). For example, something that comes up all the time is that we detect stars in astronomical images and solve for their celestial coordinates. My current project uses images about 1.5 degrees on a side and typically detects 60 to 80 thousand stars per image, with the number well modeled as a Poisson Distribution, assuming that the image is not of a star cluster surrounded by mostly empty space. That’s about 8 or 9 stars per square arcminute. If we cut out “postage stamps”from the image that are half an arcminute per side, then the mean number of detected stars in them is about 2. If we do that for (say) 1000 postage stamps and make a histogram of the number of detected stars in them, it will not look very Gaussian, but as we increase the size of the postage stamps, it becomes asymptotically Gaussian.

What generally never becomes Gaussian, however, is the Uniform Distribution. A histogram of the stars’ right ascensions or declinations (the azimuthal and elevation angles used in astronomy) looks a lot like a step function, i.e., flat within the image boundaries. The positions are not uniformly spaced, but they are distributed in the same way as a uniformly distributed random variable for any size postage stamp, including the entire image.

Another example is the location of the centers of raindrop ripples on a pond; they are not uniformly spaced in (say) the east-west direction, but they are uniformly distributed.

The simplest example is the distribution of numbers that show up on the top of a fair die after a large number of throws. Each number from 1 to 6 will occur with approximately equal frequency. Increasing the number of throws will not tend to produce a bell-shaped histogram, in fact the fractional occurrence will approach a constant 1/6 over the possible numbers.

3) Examples of non-Gaussian distributions include the exponential distribution, Poisson distribution, log-normal distribution, Weibull distribution, gamma distribution, and chi-square distribution. Each distribution has its own characteristics and applications in different fields. For instance, the exponential distribution is used to model time between events, the Poisson distribution describes the occurrence of rare events, and the log-normal distribution represents data with positive skewness, such as financial returns.

4) I’ll give a general example for time series data.

Here are sample time series for two measured variables: one is Gaussian-distributed (top), and the other is not (bottom).

On the right, I tallied the measurements in a histogram. This can help us check if a variable is Gaussian or not.

Non-Gaussian distributed time series data arise when the mean or noise statistics vary with time.

If the mean varies with time, the variable could be non-stationary / time-varying (its trend changes with time), auto- or cross-correlated (it changes depending on its previous value or the values of other variables), or its value is computed from the values of other Gaussian variables but in a nonlinear way.

**Question: 3**

What is the meaning of the five-number summary in Statistics? Give proper example

**Calculate the 5 Number Summary-**

The concept of a 5 number summary is a way to describe a distribution using 5 numbers. This includes minimum number, quartile-1, median/quartile-2, quartile-3, and maximum number. This concept of 5 number summary comes under the concept of Statistics which deals with the collection of data, analyzing it, interpreting, and presenting the data in an organized manner.

**5 Number Summary**

As told in the above paragraph, It gives a rough idea how the given dataset looks like by representing minimum value, maximum value, median, quartile values, etc. To understand better the 5 number summary concept look at the below pictorial representation of 5 number summary

**Calculating 5 number summary**

In order to find the 5 number summary, we need the data to be sorted. If not sort it first in ascending order and then find it.

Minimum Value: It is the smallest number in the given data, and the first number when it is sorted in ascending order.

Maximum Value: It is the largest number in the given data, and the last number when it is sorted in ascending order.

Median: Middle value between the minimum and maximum value. Below is the formula to find median,

Median = (n + 1)/2th term

Quartile 1: Middle/center value between the minimum and median value. We can simply identify the middle value between median and minimum value for a small dataset. If it is a big dataset with so many numbers then better to use a formula,

Quartile 1 = ((n + 1)/4)th term

Quartile 3: Middle/center value between median and maximum value.

Quartile 3 = (3(n + 1)/4)th term

**Question: 4**

What is correlation? Give an example with a dataset & graphical representation on jupyter Notebook

A correlation is the statistical summary of the relationship between two sets of variables.

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.

**Plotting Correlation Matrix using Python:**

**Correlation means an association, It is a measure of the extent to which two variables are related.**

1. Positive Correlation: When two variables increase together and decrease together. They are positively correlated. ‘1’ is a perfect positive correlation. For example – demand and profit are positively correlated the more the demand for the product, the more profit hence positive correlation.

2. Negative Correlation: When one variable increases and the other variable decreases together and vice-versa. They are negatively correlated. For example, If the distance between magnet increases their attraction decreases, and vice-versa. Hence, a negative correlation. ‘-1’ is no correlation

3. Zero Correlation( No Correlation): When two variables don’t seem to be linked at all. ‘0’ is a perfect negative correlation. For Example, the amount of tea you take and level of intelligence.

Note:Half Answer in Jupyter NoteBook **statAns4.ipynb**