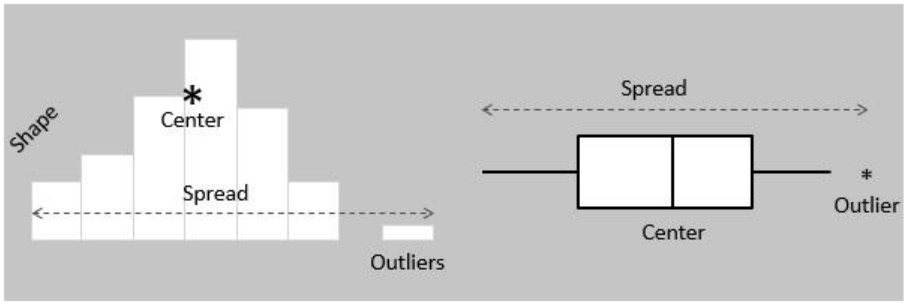
STATISTICS WORKSHEET

1. **D**
2. **A**
3. **A**
4. **C**
5. **C**
6. **A**
7. **C**
8. **B**
9. **B**

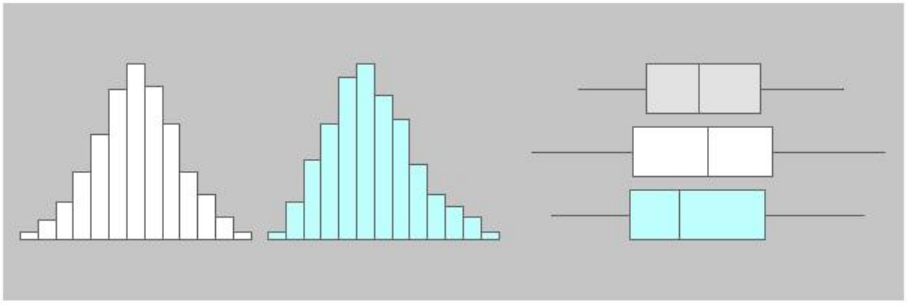
10. What is the difference between a box plot and histogram?

Answer: Histograms and box plots are graphical representations for the frequency of numeric data values. They aim to describe the data and explore the central tendency and variability before using advanced statistical analysis techniques.

Both histograms and box plots allow to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points.



Both histograms and box plots are used to explore and present the data in an easy and understandable manner. Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets. They are less detailed than histograms and take up less space.



11.How to select metrics?

Answer: - Performance metrics are defined as information and project-specific data used to characterize and assess an organization's quality, capabilities, and skills. Performance metrics are defined differently in every industry and can change based on a company's services and

products. Common performance metrics include sales, return on investment, customer satisfaction, industry and consumer reviews, and a company's reputation with its consumers.

One of the biggest problems associated with continuous improvement and problem solving is the selection of the most appropriate performance measures or quality performance metrics.

ORGANIZATIONAL AND ENTERPRISE-LEVEL PERFORMANCE MEASURES

The world of commerce and industry uses a variety of financial performance measures or performance metrics at the organizational and enterprise level. These include ratios such as return on investment (ROI) and return on net assets (RONA).

These ratios and other non-financial ratios such as market share and name recognition index, are dependent variables that numerically describe the level of success or failure of an organization for a specific period of time, such as one quarter of a fiscal year.

However, the ways in which organizations achieve these levels of success or failure is of greater importance (see Figure 1 below). Independent variables, such as customer satisfaction indices, defect rates, and supplier capability indices, provide this information. When these factors reflect well on an organization, their dependent variables are much more likely to reflect overall enterprise success.

These metrics can be treated as dependent variables with an entirely new set of independent variables such as conveyor speeds, temperature settings, spindle speeds, and work-in-process (WIP) levels. The independent variables are direct measures of the processes that make up the enterprise systems creating products and services that generate organizational income.

When determining what performance measures or performance metrics to use for a system, process, or step within a process, it may be helpful to first determine what is important to customers (either internal or external) and can be measured or counted.

If this cannot be answered directly, consider what is important to the customer that cannot be directly measured or counted but can be assessed indirectly, using one or more proxy measures.

A sequence of top-down performance measures or performance metrics can demonstrate this approach to continuous improvement.

Independent Variables	Dependent Variables
Customer satisfaction	Market share
On-time delivery Competitive price Consistent quality Uncompromising service Brand name recognition	Customer satisfaction
Robust design	Consistent quality
Statistical quality control Six Sigma and/or ppm Design of experiments Analysis of variance Statistical process control	Robust design
Process inputs (5Ms and an E)	Statistical process control
Conveyer speed Operating temperature Operator skill level Humidity level Measurement capability	Process inputs (5Ms and an E)

Figure 1: Relationships of independent and dependent variables as performance metrics

PERFORMANCE MEASUREMENTS & METRICS EXAMPLE

Baseball is a sport that involves players with assigned positions and roles who also work together on a greater team. Baseball is one way to illustrate how organizations can use a variety of performance measurements or performance metrics to assess the success or failure of individuals, as well as their teams.

- Pitchers are often evaluated by their earned run average (ERA), their total number of strikeouts, and the number of hits they yield. For example, ERA is the number of runs scored against a pitcher per nine innings pitched. An ERA of 2.05 indicates that for multiple games, a pitcher had 2.05 runs scored against him or her for each nine innings pitched.
- Position players are often evaluated on both offensive and defensive skills.
- Offensively, a position player is rated on his or her hitting skills, including batting average (BA), runs batted in (RBI), and on-base percentage (OBP). BA is the number of hits per number of times at bat. A BA of .400 is exceptional, but rarely achieved. A BA of .300 is good. A BA of .100 is typical of most pitchers, who aren't usually considered skilled hitters.
- Defensively, a position player is ranked according to fielding percentage (FP), or the ability to catch and throw a ball without committing an error. This includes catching a batted ball and securing an out, throwing the ball to the right base to secure a putout, or making an accurate throw to the correct base. FP is the number of fielding opportunities played without an error divided by the total number of opportunities.

12.How do you assess the statistical significance of an insight?

Answer: Steps in Testing for Statistical Significance

1) State the Research Hypothesis
2) State the Null Hypothesis
3) Select a probability of error level (alpha level)
4) Select and compute the test for statistical significance
5) Interpret the results

1) State the Research Hypothesis

A research hypothesis states the expected relationship between two variables. It may be stated in general terms, or it may include dimensions of direction and magnitude. For example,

General: The length of the job training program is related to the rate of job placement of trainees.
Direction: The longer the training program, the higher the rate of job placement of trainees.
Magnitude: Longer training programs will place twice as many trainees into jobs as shorter programs.
General: Graduate Assistant pay is influenced by gender.
Direction: Male graduate assistants are paid more than female graduate assistants.
Magnitude: Female graduate assistants are paid less than 75% of what male graduate assistants are paid.

2) State the Null Hypothesis

A null hypothesis usually states that there is no relationship between the two variables. For example,

There is no relationship between the length of the job training program and the rate of job placement of trainees.
Graduate assistant pay is not influenced by gender.

A null hypothesis may also state that the relationship proposed in the research hypothesis is not true. For example,

Longer training programs will place the same number or fewer trainees into jobs as shorter programs.
Female graduate assistants are paid at least 75% or more of what male graduate assistants are paid.

Researchers use a null hypothesis in research because it is easier to disprove a null hypothesis than it is to prove a research hypothesis. The null hypothesis is the researcher's "straw man." That is, it is easier to show that something is false once than to show that something is always true. It is easier to find disconfirming evidence against the null hypothesis than to find confirming evidence for the research hypothesis.

3) TYPE I AND TYPE II ERRORS

Even in the best research project, there is always a possibility (hopefully a small one) that the researcher will make a mistake regarding the relationship between the two variables. There are two possible mistakes or errors.

The first is called a Type I error. This occurs when the researcher assumes that a relationship exists when in fact the evidence is that it does not. In a Type I error, the researcher should accept the null hypothesis and reject the research hypothesis, but the opposite occurs. The probability of committing a Type I error is called alpha.

The second is called a Type II error. This occurs when the researcher assumes that a relationship does not exist when in fact the evidence is that it does. In a Type II error, the researcher should reject the null hypothesis and accept the research hypothesis, but the opposite occurs. The probability of committing a Type II error is called beta.

Generally, reducing the possibility of committing a Type I error increases the possibility of committing a Type II error and vice versa, reducing the possibility of committing a Type II error increases the possibility of committing a Type I error.

Researchers generally try to minimize Type I errors, because when a researcher assumes a relationship exists when one really does not, things may be worse off than before. In Type II errors, the researcher misses an opportunity to confirm that a relationship exists, but is no worse off than before.

In this example, which type of error would you prefer to commit?

> Research Hypothesis: El Nino has reduced crop yields in County X, making it eligible for government disaster relief.
> Null Hypothesis: El Nino has not reduced crop yields in County X, making it ineligible for government disaster relief.

If a Type I error is committed, then the County is assumed to be eligible for disaster relief, when it really is not (the null hypothesis should be accepted, but it is rejected). The government may be spending disaster relief funds when it should not, and taxes may be raised.

If a Type II error is committed, then the County is assumed to be ineligible for disaster relief, when it really is eligible (the null hypothesis should be accepted, but it is rejected). The government may not spend disaster relief funds when it should, and farmers may go into bankruptcy.

In this example, which type of error would you prefer to commit?

Research Hypothesis: The new drug is better at treating heart attacks than the old drug
Null Hypothesis: The new drug is no better at treating heart attacks than the old drug

If a Type I error is committed, then the new drug is assumed to be better when it really is not (the null hypothesis should be accepted, but it is rejected). People may be treated with the new drug, when they would have been better off with the old one.

If a Type II error is committed, then the new drug is assumed to be no better when it really is better (the null hypothesis should be rejected, but it is accepted). People may not be treated with the new drug, although they would be better off than with the old one.

INTERPRET THE RESULTS

If the computed value for Chi Square equals or exceeds the value indicated in the table for the given level of alpha and degrees of freedom, then the researcher can assume that the observed relationship between the two variables exists (at the specified level of probability of error, or alpha), and reject the null hypothesis. This gives support to the research hypothesis.

13. Give examples of data that doesn't have a Gaussian distribution, nor log-normal.
Answer:

- Allocation of wealth among individuals
- Values of oil reserves among oil fields (many small ones, a small number of large ones)

14. Give an example where the median is a better measure than the mean.
Answer: If the score of students in a class are 1,2,3,4,20
So if we calculate the mean = 51+2+3+4+20= 530 =6
Median =3
So, median is better or appropriate measure because 20 is much greater than other numbers and because of 20 the mean has come out to 6.
∴ Its better to take median than mean.

15. What is the Likelihood?
Answer: The term "probability" refers to the possibility of something happening. The term Likelihood refers to the process of determining the best data distribution given a specific situation in the data. When calculating the probability of a given outcome, you assume the model's parameters are reliable.