

B.C.A study

UNIT-1:Population, Sample and Data Condensation

Population

In statistical terms, population refers to the complete set of individuals or objects that have a certain characteristic or attribute of interest. When conducting research or making statistical inferences, it is often not possible or practical to study every member of the population. Instead, a sample of the population is selected and studied in order to make inferences about the larger group.

The population can be described using various measures such as mean, median, mode, variance, and standard deviation. These measures provide information about the central tendency, dispersion, and distribution of the population data.

It's important to note that when working with a sample rather than the complete population, the estimates obtained from the sample may not perfectly reflect the characteristics of the population. This is due to sampling error, which can be reduced by increasing the sample size or using a more representative sample.

In summary, in statistics, population refers to the complete group of individuals or objects of interest and the study of the population helps to describe and understand the characteristics of the group

Sample and Data Condensation

Sample and data condensation are two important concepts in statistics.

A sample is a subset of the population that is selected for study. Sampling is an important aspect of statistical analysis because it allows researchers to make inferences about a population based on a smaller, more manageable subset of data. There are various sampling methods, including random sampling, stratified sampling, and cluster sampling, each with its own strengths and weaknesses.

Data condensation refers to the process of summarizing and simplifying a large amount of data into a more manageable form. This can be done using various methods such as frequency tables, histograms, and box plots, which can provide a visual representation of the data and help to identify patterns and trends.

Another method of data condensation is descriptive statistics, which involves calculating measures such as mean, median, mode, variance, and standard deviation to summarize the data. Descriptive statistics can provide important insights into the distribution of the data and help to identify outliers and other features of interest.

In summary, sample and data condensation are important concepts in statistics that help researchers to work with smaller subsets of data and summarize large amounts of information in a more manageable form

Definition and scope of statistics

Statistics is a branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of data. It is a tool used to make sense of data and draw meaningful conclusions and inferences about a population based on a sample of data.

The scope of statistics is broad and includes a wide range of applications in fields such as biology, social sciences, engineering, economics, and many others. Some of the key areas in which statistics is used include:

1. Data collection and analysis: Statistics is used to design studies and experiments, collect data, and analyze it to draw conclusions and make inferences about a population.
2. Descriptive statistics: This involves summarizing and describing data using measures such as mean, median, mode, and standard deviation, as well as visualizing data using techniques such as histograms and box plots.
3. Inferential statistics: This involves using a sample of data to make inferences about a population. Inferential statistics involves the use of statistical models and hypothesis testing to determine the likelihood of a relationship between variables or to make predictions about future events.
4. Probability: Statistics also involves the study of probability, which is used to model and understand random events and make predictions about the likelihood of certain outcomes.
5. Survey design and analysis: Statistics is used to design and analyze surveys, which are used to gather information about a population, such as opinions or attitudes.

In summary, the scope of statistics is wide and involves the collection, analysis, interpretation, and presentation of data for a variety of purposes and applications.

concept of population and sample with Illustration

The concepts of population and sample are important in statistical analysis and research.

A population refers to the complete set of individuals or objects being studied. It is the entire group of interest and includes all relevant data. For example, the population of a country includes all individuals who live in that country.

A sample, on the other hand, is a subset of the population. It is a smaller group of individuals or objects selected from the population for study. Sampling is done because it is often not feasible or practical to study the entire population. For example, if a researcher wants to study the income level of a population, it may not be possible to gather data from every individual. In this case, a sample of the population can be selected and studied instead.

Here's a simple illustration to help understand the concept:

Imagine a bag containing 100 marbles, with 50 red and 50 blue marbles. The bag represents the population. If we take a sample of 10 marbles from the bag, this represents the sample. The sample may contain 6 red marbles and 4 blue marbles, which is just a smaller representation of the population and not a perfect representation of the population

Raw data, attributes and variables

Raw data refers to unprocessed data that has been collected from various sources, such as surveys, experiments, or databases. Raw data is usually in its original form and hasn't been manipulated or analyzed.

Attributes are characteristics or features of the data. For example, in a study of a population of individuals, the attributes might include age, gender, education level, and income.

Variables are attributes that can take on different values. For example, in a study of a population of individuals, the variable "age" can take on different values for each person in the population, such as 20, 25, 30, etc. In statistical analysis, variables are used to answer questions and make predictions.

In summary, raw data is the starting point for any analysis, while attributes and variables are used to describe and analyze the data

What is Frequency Distribution?

Frequency distribution is defined as the first method that is used to organize data in an effective way. Frequency distribution performs the systematic investigation of the raw data. The data is first arranged by frequency distribution and then set as frequency table.

Frequency distribution is defined as the systematic representation of different values of variables along with the corresponding frequencies; it is classified on the basis of class interval.

Class interval is defined as the size of each class into which a range of variables is divided and represented as histogram or bar graph.

Types of Class Intervals

Class intervals are divided into two different categories, exclusive and inclusive class intervals. Here is the example to both:

1. Exclusive Class Interval

The class interval where the upper limit of previous data entry is the same as the lower limit of next data entry is called an exclusive data interval. For consideration,

S. No	Marks	No. of students
1	0-20	8
2	20-40	7
3	40-60	3

2. Inclusive Class Interval

The class interval where the upper limit of previous data entry is the same as the lower limit of next data entry is called an exclusive data interval. For consideration,

S. No	Marks	Number of students
1	1-20	7
2	21-40	9
3	41-60	8

Also Read | Introduction to Bayesian Statistics

What is Discrete and Continuous Frequency Table Distribution?

Frequency distribution is further classified into two types based upon class interval. Named as discrete frequency table and continuous frequency table. Here are the examples:

1. Discrete Frequency Table

If the class interval of data is not given, it is termed as a discrete frequency distribution. For example,

S. no.	Number of items	Number of packets
1	1	23
2	2	12
3	3	34
4	4	20
5	5	72
	Total	163

2. Continuous Frequency Table

When the class intervals are available within the data, it is called a continuous frequency distribution. For consideration,

S. No	Marks	Number of students
1	0-10	5
2	20-30	7
3	30-40	12
4	40-50	32
5	50-60	4
	Total	60

Types of Frequency Distribution Methods

There are two types of frequency distribution methods:

1. Grouped frequency distribution.
2. Ungrouped frequency distribution.

1. Grouped Frequency Distribution

As the name suggests, grouped frequency distribution is well defined and distributed into groups. When the variables are continuous the data is gathered as grouped frequency distribution. Different measures are taken during data collection, such as age, salary, etc. The entire data is classified into class intervals. For consideration,

Family Income	Number of persons
Below-20,000	52
20,001-30,000	14
30,001-40,000	6
40,001-50,000	8

2. Ungrouped Frequency Distribution

As the name suggests, ungrouped frequency distribution doesn't consist of well-distributed class intervals. Ungrouped frequency distribution is applied on discrete data rather than continuous one. Examples of such data usually include data related to gender, marital status, medical data etc. For consideration,

Variable	Number of persons
GENDER	
Female	19
Male	22
MARITAL STATUS	
Single	32
Married	4
Divorced	4

Other Types of Frequency Distribution

1. Cumulative Frequency Distribution

Cumulative frequency distribution is also known as percentage frequency distribution. Percentage distribution reflects the percentage of samples whose scores fall in the specific group and number of scores.

This type of distribution is quite useful for comparison of data with the findings of other studies having different sample sizes. In this type of distribution, percentages and frequencies are summed up in a single table. For consideration,

Score	Frequency	Percentage	Cumulative frequency	Cumulative percentage
1	4	8	4	8
2	14	28	32	64
4	6	12	10	20
5	8	16	18	36
7	8	16	40	80
8	6	12	46	92

9	4	8	50	<div>...</div>	100
---	---	---	----	----------------	-----

A WordPress.com Website.

B.C.A study

Unit-2:Measure of central tendency

A measure of central tendency is a statistical calculation used to determine a single value that summarizes the central location of a set of data. The central location of the data provides a quick summary of the main characteristics of the data, which can be useful in making predictions or drawing conclusions about the data. There are three common measures of central tendency: mean, median, and mode.

1. Mean: The mean, or average, is calculated by adding up all the values in the data set and then dividing by the number of values. The mean provides a good representation of the central location of the data if the data is evenly distributed. However, if there are extreme values in the data set, the mean can be significantly impacted, making it a less reliable measure of central tendency.
2. Median: The median is the middle value of a data set when the values are ordered from smallest to largest. It provides a good representation of the central location of the data if there are extreme values in the data set, as it is not affected by outliers.
3. Mode: The mode is the value that occurs most frequently in a data set. The mode can provide a good representation of the central location of the data if the data is not evenly distributed, such as in the case of categorical data.

In conclusion, the appropriate measure of central tendency to use depends on the type and distribution of the data, as well as the goals of the analysis.

concept of central tendency

Central tendency refers to a single value or typical value that summarizes the central location of a set of data. The idea is to find a single value that best represents the “center” of the data and gives a quick summary of its main characteristics. There are several measures of central tendency, including mean, median, and mode. Each measure provides a different perspective on the central location of the data and may be more or less appropriate depending on the type and distribution of the data, as well as the goals of the analysis.

Mean, median, and mode are three common measures of central tendency. Mean (average) is calculated by summing up all the values in a data set and dividing by the number of values. Median is the middle value of a data set when the values are ordered from smallest to largest. Mode is the value that occurs most frequently in a data set.

In general, mean is a good representation of central tendency if the data is evenly distributed, but it can be significantly impacted by extreme values (outliers). Median provides a good representation of central tendency if there are outliers, as it is not affected by extreme values. Mode provides a good representation of central tendency if the data is not evenly distributed, such as in the case of categorical data.

Central tendency is a useful tool in data analysis, as it provides a simple and quick summary of a large and complex data set. It can also help identify patterns, make predictions, and draw conclusions about the data

quirements of good measures of central tendency

For a measure of central tendency to be considered “good”, it should satisfy the following requirements:

1. Uniqueness: There should be a single value that summarizes the central location of the data, not multiple values or ranges.
2. Representativeness: The measure should provide an accurate representation of the central location of the data, capturing its main characteristics.
3. Stability: The measure should not change significantly with small variations in the data set.
4. Insensitivity to extreme values: The measure should not be greatly affected by outliers or extreme values in the data set.

Example: Suppose we have the following data set: 1, 2, 3, 4, 1000

Mean: $(1 + 2 + 3 + 4 + 1000) / 5 = 200.8$

Median: 3 (when the values are ordered from smallest to largest)

Mode: None (no value occurs more than once)

In this example, the median is a better measure of central tendency as it is not greatly affected by the extreme value of 1000. The mean is significantly impacted by this value and is not representative of the central location of the data. The lack of a mode suggests that this is not a good measure of central tendency for this data set.

Arithmetic mean

The arithmetic mean, also known as the average, is calculated by summing up all the values in a data set and dividing by the number of values. Here's how to calculate the arithmetic mean with an example:

Example: Suppose we have the following data set of 5 values: 2, 4, 5, 8, 9

Step 1: Sum up all the values: $2 + 4 + 5 + 8 + 9 = 28$

Step 2: Divide the sum by the number of values (n): $28 \div 5 = 5.6$

Step 3: The result, 5.6, is the arithmetic mean of the data set.

So, the average of the data set is 5.6. This value provides a single representation of the central location of the data and can be used for making predictions or drawing conclusions about the data.

Median

Median is a measure of central tendency that represents the middle value of a set of data when the values are ordered from smallest to largest. The median provides a good representation of the central location of the data if there are extreme values or outliers, as it is not affected by these values.

Here's how to calculate the median with an example:

Example: Suppose we have the following data set of 7 values: 2, 4, 6, 8, 9, 10, 12

Step 1: Order the values from smallest to largest: 2, 4, 6, 8, 9, 10, 12

Step 2: If the number of values in the data set is odd, the median is the middle value. In this case, the median is 9.

Step 3: If the number of values in the data set is even, the median is the average of the two middle values. In this case, the median is $(8 + 9) / 2 = 8.5$.

So, the median of this data set is 8.5. This value provides a good representation of the central location of the data, as it is not affected by the presence of extreme values

Median Formula for Ungrouped Data

The following steps are helpful while applying the median formula for ungrouped data.

- Step 1: Arrange the data in ascending or descending order.
- Step 2: Secondly, count the total number of observations 'n'.
- Step 3: Check if the number of observations 'n' is even or odd.

Median Formula When n is Odd

The median formula of a given set of numbers, say having 'n' odd number of observations, can be expressed as:

Median = $[(n + 1)/2]^{\text{th}}$ term

$$\text{Median} = \left(\frac{n+1}{2} \right)^{\text{th}} \text{ observation}$$

Median Formula When n is Even

The median formula of a given set of numbers say having 'n' even number of observations, can be expressed as:

Median = $[(n/2)^{\text{th}} \text{ term} + ((n/2) + 1)^{\text{th}} \text{ term}]/2$

$$\text{Median} = \frac{\frac{n^{\text{th}}}{2} \text{ obs.} + \left(\frac{n}{2} + 1 \right)^{\text{th}} \text{ obs.}}{2}$$

Example: The age of the members of a weekend poker team has been listed below. Find the median of the above set.

{42, 40, 50, 60, 35, 58, 32}

Solution:

Step 1: Arrange the data items in ascending order.

Original set: {42, 40, 50, 60, 35, 58, 32}

Ordered Set: {32, 35, 40, 42, 50, 58, 60}

Step 2: Count the number of observations. If the number of observations is odd, then we will use the following formula: Median = $[(n + 1)/2]^{\text{th}}$ term

Step 3: Calculate the median using the formula.

$$\text{Median} = [(n + 1)/2]^{\text{th}} \text{ term}$$

$$= (7 + 1)/2^{\text{th}} \text{ term} = 4^{\text{th}} \text{ term} = 42$$

$$\text{Median} = 42$$

Median Formula for Grouped Data

When the data is continuous and in the form of a frequency distribution, the median is calculated through the following sequence of steps.

- Step 1: Find the total number of observations(n).
- Step 2: Define the class size(h), and divide the data into different classes.
- Step 3: Calculate the cumulative frequency of each class.
- Step 4: Identify the class in which the median falls. (Median Class is the class where $n/2$ lies.)

- Step 5: Find the lower limit of the median class(l), and the cumulative frequency of the class preceding the median class (c).

Now, use the following formula to find the median value.

Median for Grouped Data

$$\text{Median} = l + \left[\frac{\frac{n}{2} - c}{f} \right] \times h$$

Application of Median Formula

Let us use the above steps in the following practical illustration to understand the application of the median formula.

Illustration: There are 5 top management employees in an organization. The salaries given to the employees are \$5,000, \$6,000, \$4,000, \$8,000, and \$7,500. Using the median formula calculates the median salary.

Solution: We will follow the given steps to find the median salary.

- Step 1: Sorting the given data in increasing order, \$4,000, \$5,000, \$6,000, \$7,500, and \$8,000.
- Step 2: Total number of observations = 5
- Step 3: The given number of observations is odd.
- Step 4: Using median formula for odd observation, Median = $[(n + 1)/2]^{\text{th}}$ term
- Median = $[(5+1)/2]^{\text{th}}$ term. = $6/3 = 3^{\text{rd}}$ term. The third term is \$6,000.

The median salary is \$6,000.

How to Find Median?

We use a median formula to find the median value of given data. For a set of **ungrouped data**, we can follow the below-given steps to find the median value.

- Step 1: Sort the given data in increasing order.
- Step 2: Count the number of observations.
- Step 3: If the number of observations is odd use median formula: Median = $[(n + 1)/2]^{\text{th}}$ term
- Step 4: If the number of observations is even use median formula: Median = $[(n/2)^{\text{th}}$ term + $(n/2 + 1)^{\text{th}}$ term]/2

Example: The height (in centimeters) of the members of a school football team have been listed below.

{142, 140, 130, 150, 160, 135, 158, 132}

Find the median of the above set.

Solution:

Step 1:

Arrange the data items in ascending order.

Original set: {142, 140, 130, 150, 160, 135, 158, 132}

Ordered Set: {130, 132, 135, 140, 142, 150, 158, 160}

Step 2:

Count the number of observations.

Number of observations, $n = 8$

If number of observations is even, then we will use the following formula:

$$\text{Median} = [(n/2)^{\text{th}} \text{ term} + ((n/2) + 1)^{\text{th}} \text{ term}] / 2$$

Step 3:

Calculate the median using the formula.

$$\text{Median} = [(n/2)^{\text{th}} \text{ term} + ((n/2) + 1)^{\text{th}} \text{ term}] / 2$$

$$\text{Median} = [(8/2)^{\text{th}} \text{ term} + ((8/2) + 1)^{\text{th}} \text{ term}] / 2$$

$$= (4^{\text{th}} \text{ term} + 5^{\text{th}} \text{ term})/2$$

$$= (140 + 142)/2$$

$$= 141$$

For a set of **grouped data**, we can follow the following steps to find the median:

When the data is continuous and in the form of a frequency distribution, the median is calculated through the following sequence of steps.

- Step 1: Find the total number of observations(n).
- Step 2: Define the class size(h), and divide the data into different classes.
- Step 3: Calculate the cumulative frequency of each class.
- Step 4: Identify the class in which the median falls. (Median Class is the class where $n/2$ lies.)
- Step 5: Find the lower limit of the median class(l), and the cumulative frequency(c).
- Step 6: Apply the formula for median for grouped data: $\text{Median} = l + \frac{(n/2 - cf)}{f} \times h$

mode

The mode is the value that occurs most frequently in a data set. It provides a good representation of central tendency for data sets that are not evenly distributed, such as categorical data.

Here's how to calculate the mode with an example:

Example: Suppose we have the following data set of 6 values: 2, 4, 4, 6, 8, 8

Step 1: Count the frequency of each value in the data set:

Value: 2 Frequency: 1 Value: 4 Frequency: 2 Value: 6 Frequency: 1 Value: 8 Frequency: 2

Step 2: Find the value(s) with the highest frequency. In this case, both 4 and 8 have a frequency of 2.

Step 3: Both 4 and 8 are the modes of the data set, as they both occur with the same highest frequency.

So, the modes of this data set are 4 and 8. These values provide a good representation of central tendency for this data set, as they capture the most common values in the data. Note that it is possible for a data set to have more than one mode, or no mode at all.

The mode for ungrouped data

is the value that appears most frequently in the data set.

Example: Consider the following set of numbers: {1, 2, 2, 3, 4, 4, 4, 5}

The mode of this data set is 4, as it appears three times, which is more than any other value.

Formula

Statisticians use the mode formula in **statistics** (<https://www.wallstreetmojo.com/statistics/>) to know the highest frequency in a group of data or distribution. They take the most repeated data as the Mode of distribution. It is one of the three important measures related to the central tendency besides mean and median.

Here is the formula that statisticians and analysts use in the calculation of a data set in statistics:

$$\text{Mode}_g = L + h \frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_2)}$$

Where the modal class = the one with the highest frequency data interval;

- L = lower limit of the said modal class
- h = size of the class interval
- f_m = modal class frequency
- f_1 = frequency of the class which precedes the modal class; and
- f_2 = frequency of the class which succeeds the modal class

Formula To Find The Mode Of Ungrouped Data

For doing so, one has to first arrange the data in ascending or descending manner in terms of their values. After the arranging, one must mark the data values, which are repeated more often. Amongst all the frequent data values, the one having the highest frequency of occurring in the data set is the modal value or the most common value for the set.

Formula To Find The Mode Of Grouped Data

To find the value of grouped data, one has to identify the class interval with the most frequency, known as the modal class. After doing so, one calculates the class size by subtracting the lower limit from the upper limit. Finally, statisticians use the the most common value formula to calculate the Mode for the grouped data after putting all the values in it, as shown below:

$$\text{Mode} = L + h \frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_2)}$$

Calculation Example

Here is a mode calculation example to understand the concept and its usage.

Class Interval	0–10	10–20	20–30	30–40	40–50
Frequency	8	5	10	4	7

Modal class = 20-30 as it has the data with the highest frequency

Size of the class interval, $h = 10$

The lower limit of the above modal class, $L = 20$

Frequency of the modal class, $f_m = 10$

Frequency of the class which precedes the modal class, $f_1 = 5$; and

Frequency of the class which succeeds the modal class $f_2 = 4$

Therefore, putting all the values in the formula of mode, $M =$

$$L + h \frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_2)}$$

One can get Mode = $20 + 10 \frac{(10 - 5)}{(10 - 5) + (10 - 4)}$

$$= 20 + 10 \cdot \frac{5}{11}$$

$$\text{i.e., } (220 + 50) / 11$$

$$= 270 / 11$$

or, **Mode = 24.54 for the above data set.**

Example

As discussed below, the best way to understand the basics of this concept is through a mode example.

Let us assume an inventory manager has to know which stock is mostly purchased by the customers and must be replenished accordingly. Therefore, the inventory manager prepares a list of the items in the warehouse with the respective product type and purchase code as below:

PURCHASE CODE	PRODUCT TYPE
A	SUNGLASSES
B	KIDS GARMENTS
C	LAPTOPS
D	MOBILE
E	SUNGLASSES

PURCHASE CODE	PRODUCT TYPE
F	MOBILE
G	KIDS GARMENTS
H	LAPTOPS
I	MOBILE
J	SUNGLASSES
K	KIDS GARMENTS
L	MOBILE
M	LAPTOPS
N	MOBILE
O	KIDS GARMENTS

For the most common value calculation, one should categorize the above data in the frequency of buying by the customers as below:

TYPE	FREQUENCY
SUNGLASSES	3
KIDS GARMENTS	4
LAPTOPS	3
MOBILE	5

As a result of the table above, one finds that mobile is bought more frequently than other items in the warehouse inventory. Therefore, **mobile is the mode of the data set of our example**. Thus, the mode in statistics helps with inventory management.

Harmonic Mean, Geometric mean for grouped and ungrouped data.

The Harmonic Mean and the Geometric Mean are two types of average measures used in statistics.

For ungrouped data, the Harmonic Mean is calculated as the reciprocal of the arithmetic mean of the reciprocals of the individual values, and is used when finding the average rate, such as speed. The Geometric Mean, on the other hand, is calculated as the n th root of the product of n values, and is used when finding the average growth rate.

For grouped data, the Harmonic Mean is calculated by dividing the total number of observations by the sum of the reciprocals of the class frequencies, and the Geometric Mean is calculated by finding the n th root of the product of the class frequencies.

In general, the Harmonic Mean is a better measure of central tendency for data sets with extremely large or small values, while the Geometric Mean is a better measure for data sets with values close to each other

Harmonic Mean (H.M.)

Harmonic Mean is defined as the reciprocal of the arithmetic mean of reciprocals of the observations.

(a) H.M. for Ungrouped data

Let x_1, x_2, \dots, x_n be the n observations then the harmonic mean is defined as

$$\text{H. M.} = \frac{n}{\sum_{i=1}^n \left(\frac{1}{x_i} \right)}$$

Example 5.11

A man travels from Jaipur to Agra by a car and takes 4 hours to cover the whole distance. In the first hour he travels at a speed of 50 km/hr, in the second hour his speed is 64 km/hr, in third hour his speed is 80 km/hr and in the fourth hour he travels at the speed of 55 km/hr. Find the average speed of the motorist.

Solution:

x	50	65	80	55	Total
$1/x$	0.0200	0.0154	0.0125	0.0182	0.0661

$$\begin{aligned} \text{H. M.} &= \frac{n}{\sum \left(\frac{1}{x_i} \right)} \\ &= \frac{4}{0.0661} = 60.5 \text{ km/hr} \end{aligned}$$

Average speed of the motorist is 60.5km/hr

(b) H.M. for Discrete Grouped data:

For a frequency distribution

$$H. M. = \frac{N}{\sum_{i=1}^n f_i \left(\frac{1}{x_i} \right)}$$

Example 5.12

The following data is obtained from the survey. Compute H.M

Speed of the car	130	135	140	145	150
No of cars	3	4	8	9	2

Solution:

x_i	f_i	$\frac{f_i}{x_i}$
130	3	0.0231
135	4	0.0091
140	8	0.0571
145	9	0.0621
150	2	0.0133
Total	N = 26	0.1648

$$H. M. = \frac{N}{\sum_{i=1}^n f_i \left(\frac{1}{x_i} \right)}$$

$$= \frac{26}{0.1648}$$

$$H.M = 157.77$$

(c) H.M. for Continuous data:

$$\text{The Harmonic mean H.M.} = \frac{N}{\sum_{i=1}^n f_i \left(\frac{1}{x_i} \right)}$$

Where x_i is the mid-point of the class interval

Geometric Mean

A geometric mean is a mean or average which shows the central tendency of a set of numbers by using the product of their values. For a set of n observations, a geometric mean is the n th root of their product. The geometric mean G.M., for a set of numbers x_1, x_2, \dots, x_n is given as

$$\text{G.M.} = (x_1 \cdot x_2 \dots x_n)^{1/n}$$

$$\text{or, G. M.} = \left(\prod_{i=1}^n x_i \right)^{1/n} = \sqrt[n]{x_1 \cdot x_2 \dots x_n}.$$

The geometric mean of two numbers, say x , and y is the square root of their product $x \times y$. For three numbers, it will be the cube root of their products i.e., $(x \cdot y \cdot z)^{1/3}$.



[A WordPress.com Website.](https://bcastudyguide.com/unit-2measure-of-central-tendency/)

B.C.A study

Unit-3: Measure of dispersion

Dispersion is a statistical concept that refers to the extent to which data points in a set are spread out from each other. There are several measures of dispersion, including:

1. Range: It's the difference between the largest and the smallest value in a dataset.
2. Interquartile Range (IQR): It's the difference between the third quartile and the first quartile, which represents the range of the middle 50% of the data.
3. Variance: It's a measure of the spread of a set of data around its mean. Variance is the average of the squared differences between each data point and the mean.
4. Standard Deviation: It's the square root of the variance and provides a measure of how far each data point is from the mean.
5. Mean Absolute Deviation (MAD): It's the average of the absolute differences between each data point and the mean.
6. Coefficient of Variation (CV): It's the ratio of the standard deviation to the mean, expressed as a percentage, and provides a measure of relative dispersion.

These measures of dispersion help us to understand how much the data is spread out and how the data points are distributed around the central value

Concept of dispersion

Dispersion, also known as variability or scatter, is a statistical concept that measures how spread out the values in a set of data are. It provides information about the distribution of the data, such as how much the data points vary from the center of the distribution and how much they vary from each other.

In other words, dispersion reflects the degree of variation or spread in the data. A set of data with high dispersion means that the data points are widely spread out, while a set of data with low dispersion means that the data points are clustered closely together.



Di
sc
ov
er
H
ali
b
un
c

There are several measures of dispersion, including range, variance, standard deviation, interquartile range, mean absolute deviation, and coefficient of variation, which are used to describe the spread of a set of data. These measures help us to understand the shape of the distribution and the degree of variation in the data, and provide important information for making decisions and predictions

Absolute and relative measure of dispersion

Dispersion measures can be classified into two categories: absolute and relative measures.

1. **Absolute Measures of Dispersion:** These measures describe the spread of the data in absolute terms, such as the difference between the largest and smallest values, or the average difference between each data point and the mean. Examples of absolute measures of dispersion are:
 - **Range:** The difference between the largest and smallest values in a set of data. For example, if the largest value is 8 and the smallest value is 2, then the range is $8 - 2 = 6$.
 - **Mean Absolute Deviation (MAD):** The average of the absolute differences between each data point and the mean. For example, if the data set is $[1, 2, 3, 4, 5]$ and the mean is 3, the MAD would be $(|1-3| + |2-3| + |3-3| + |4-3| + |5-3|)/5 = (2 + 1 + 0 + 1 + 2)/5 = 1.2$
2. **Relative Measures of Dispersion:** These measures describe the spread of the data relative to the mean or some other central value, such as the standard deviation, which is expressed as a proportion of the mean. Examples of relative measures of dispersion are:
 - **Variance:** The average of the squared differences between each data point and the mean. For example, if the data set is $[1, 2, 3, 4, 5]$ and the mean is 3, the variance would be $((1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2)/5 = (4 + 1 + 0 + 1 + 4)/5 = 2.4$.
 - **Standard Deviation:** The square root of the variance. For example, if the variance is 2.4, the standard deviation would be $\sqrt{2.4} = 1.55$.
 - **Coefficient of Variation (CV):** The ratio of the standard deviation to the mean, expressed as a percentage. For example, if the mean is 100 and the standard deviation is 10, the CV would be $(10/100)*100 = 10\%$.

Relative measures of dispersion are particularly useful when comparing data sets with different units or scales, as they provide a normalized measure of spread that is independent of the size of the data

range variance, Standard deviation, Coefficient of variation

Here are examples to illustrate the concepts of range, variance, standard deviation, and coefficient of variation:

1. Range: The range is the difference between the largest and smallest values in a set of data. For example, consider the following set of numbers: [1, 2, 3, 4, 5]. The largest value is 5 and the smallest value is 1, so the range is $5 - 1 = 4$.
2. Variance: Variance is a measure of the spread of a set of data around its mean. It is the average of the squared differences between each data point and the mean. For example, consider the data set [1, 2, 3, 4, 5] with a mean of 3. The variance would be calculated as follows:

$$((1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2)/5 = (4 + 1 + 0 + 1 + 4)/5 = 2.4.$$

3. Standard Deviation: Standard deviation is the square root of the variance and provides a measure of how far each data point is from the mean. For the data set [1, 2, 3, 4, 5] with a variance of 2.4, the standard deviation would be $\sqrt{2.4} = 1.55$.
4. Coefficient of Variation (CV): CV is the ratio of the standard deviation to the mean, expressed as a percentage. It provides a measure of relative dispersion, allowing for comparisons between data sets with different units or scales. For example, consider a data set with a mean of 100 and a standard deviation of 10. The CV would be $(10/100)*100 = 10\%$

[A WordPress.com Website.](#)

B.C.A study

Unit-5: Probability

sample space

A sample space is a collection of all possible outcomes of a random experiment. It is the set of all possible results of a random process, or a set of possible values of a random variable. The sample space provides a framework for understanding probability and statistical analysis, as it represents all the possible outcomes of an event. The elements of a sample space are known as sample points or outcomes. For example, if you roll a dice, the sample space would be the set {1, 2, 3, 4, 5, 6}.

Events and Probability

In probability theory, an event is a set of outcomes of a random experiment. It is a collection of one or more possible outcomes from a sample space. An event can be either simple, consisting of a single outcome, or it can be complex, consisting of multiple outcomes. The probability of an event is a measure of the likelihood that the event will occur, expressed as a number between 0 and 1, where 0 represents that the event is impossible and 1 represents that the event is certain to occur.

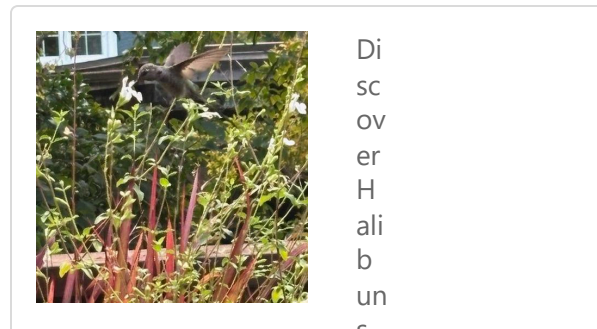
The probability of an event can be calculated using the following formula:

$$P(A) = \text{Number of favorable outcomes} / \text{Total number of possible outcomes}$$

where A is the event of interest, and the numerator and denominator are taken from the sample space.

There are two types of events: mutually exclusive and non-mutually exclusive events. Mutually exclusive events are events that cannot occur at the same time, and their sample spaces do not overlap. For example, the event “rolling a 4 on a die” and the event “rolling a 5 on a die” are mutually exclusive events because they cannot occur simultaneously. On the other hand, non-mutually exclusive events are

events that can occur simultaneously. For example, the event “rolling an even number on a die” and the event “rolling a number greater than 4 on a die” are non-mutually exclusive events because they can occur at the same time.



In addition, the concept of conditional probability is used to describe the probability of an event given that another event has occurred. The formula for conditional probability is:

$$P(A|B) = P(A \text{ and } B) / P(B)$$

where A and B are two events and P(A and B) is the probability of both events occurring simultaneously.

Probability theory is an important tool in many fields, including statistics, economics, finance, and engineering, and it plays a crucial role in understanding and modeling random processes

Experiments and random experiments

An experiment is an activity that is performed with the intention of observing the result. A random experiment is a type of experiment in which the outcome is not determined by any predetermined pattern or rule, but instead is influenced by chance. The outcome of a random experiment is uncertain, and can be any one of a set of possible outcomes.

Examples of random experiments include flipping a coin, rolling a dice, drawing a card from a deck, and measuring the height of a randomly selected person. In each of these examples, there is a set of possible outcomes and the actual outcome is not determined in advance, but instead is determined by chance.

The results of a random experiment are modeled using a sample space, which is the set of all possible outcomes of the experiment. Each outcome in the sample space is assigned a probability, which represents the likelihood that the outcome will occur. The sum of the probabilities of all outcomes in the sample space is equal to 1, which means that the sum of the probabilities of all possible outcomes of a random experiment is 1.

The study of random experiments and their outcomes is known as probability theory, which is an important branch of mathematics that is widely used in many fields, including statistics, finance, and engineering. Probability theory provides a framework for understanding and modeling random processes and for making predictions about the outcomes of random experiments

Ideas of deterministic and non-deterministic experiments

A deterministic experiment is an experiment in which the outcome is completely determined by the initial conditions and the underlying rules governing the experiment. In a deterministic experiment, the same initial conditions and rules will always produce the same outcome. For example, calculating the product of 2 multiplied by 3 is a deterministic experiment because the outcome (6) is determined by the initial conditions (the numbers 2 and 3) and the rule for multiplication.

In contrast, a non-deterministic experiment is an experiment in which the outcome is not determined by the initial conditions and rules. The outcome of a non-deterministic experiment is influenced by chance or randomness, and the same initial conditions and rules can produce different outcomes. For example, flipping a coin is a non-deterministic experiment because the outcome (heads or tails) is not determined by the initial conditions (the coin) and the rule for flipping, but instead is influenced by chance.

Definition of sample space

A sample space is the set of all possible outcomes of a random experiment. The sample space provides a framework for understanding and modeling probability and random processes, as it represents all the possible outcomes of an event. The sample space is used to assign probabilities to the individual outcomes, which represent the likelihood that each outcome will occur.

discrete sample space

A discrete sample space is a sample space in which the possible outcomes are finite and can be listed or enumerated. The sample space is called “discrete” because the outcomes are separated or distinct from each other. For example, the sample space for rolling a six-sided die is $\{1, 2, 3, 4, 5, 6\}$, which is a discrete sample space because the possible outcomes are finite and can be listed.

In contrast, a continuous sample space is a sample space in which the possible outcomes are not finite and cannot be listed. Instead, the outcomes form a continuous range of values. For example, the sample space for measuring the height of a randomly selected person is a continuous sample space because the possible heights form a continuous range of values.

events

An event is a set of outcomes of a random experiment. It represents a collection of one or more possible outcomes from the sample space. In probability theory, events are used to model and analyze the outcomes of random experiments.

There are several types of events, including:

1. **Simple events:** A simple event is an event that consists of a single outcome from the sample space. For example, rolling a 6 on a six-sided die is a simple event.
2. **Compound events:** A compound event is an event that consists of multiple outcomes from the sample space. For example, rolling an even number on a six-sided die is a compound event.
3. **Union of events:** The union of two or more events is the event that consists of all outcomes that belong to at least one of the events. The symbol for the union of events A and B is $A \cup B$. For example, if A represents the event “rolling a 4 on a die” and B represents the event “rolling a 5 on a die”, then $A \cup B$ represents the event “rolling a 4 or 5 on a die”.
4. **Intersection of events:** The intersection of two or more events is the event that consists of all outcomes that belong to all of the events. The symbol for the intersection of events A and B is $A \cap B$. For example, if A represents the event “rolling an even number on a die” and B represents the event “rolling a number greater than 4 on a die”, then $A \cap B$ represents the event “rolling a 6 on a die”.
5. **Mutually exclusive events:** Mutually exclusive events are events that cannot occur at the same time. In other words, they have no common outcomes in their sample spaces. For example, the events “rolling a 4 on a die” and “rolling a 5 on a die” are mutually exclusive because they cannot occur at the same time.
6. **Complementary event:** The complementary event of an event A is the event that consists of all outcomes from the sample space that do not belong to event A. The symbol for the complement of event A is \bar{A} . For example, if A represents the event “rolling an even number on a die”, then \bar{A} represents the event “rolling an odd number on a die”.
7. **Exhaustive events:** Exhaustive events are events that together make up the entire sample space. In other words, they are events that cover all possible outcomes of a random experiment. For example, the events “rolling an even number on a die” and “rolling an odd number on a die” are exhaustive because they together make up the entire sample space of rolling a die.

Classical definition of probability

The classical definition of probability is a method of defining the probability of an event based on the ratio of the number of favorable outcomes to the number of possible outcomes in the sample space. The classical definition of probability is expressed as follows:

Probability of an event A = Number of favorable outcomes of A / Total number of possible outcomes in the sample space

In other words, the classical definition of probability states that the probability of an event is proportional to the number of favorable outcomes of that event divided by the total number of possible outcomes in the sample space. The classical definition of probability assumes that all outcomes in the sample space are equally likely, so the probability of an event can be calculated by counting the number of favorable outcomes and dividing by the total number of possible outcomes.

For example, if you have a fair six-sided die, the probability of rolling a 4 is $1/6$ because there is only 1 favorable outcome (rolling a 4) out of 6 possible outcomes (rolling a 1, 2, 3, 4, 5, or 6)

Definition of conditional probability Definition of independence of two events

The addition theorem of probability states that the probability of the union of two or more events is equal to the sum of the probabilities of the individual events minus the sum of the probabilities of the intersections of the events. The theorem can be expressed as follows:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

where A and B are two events and $A \cup B$ is their union.

For three events A, B, and C, the addition theorem of probability can be expressed as follows:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

This theorem provides a useful way to calculate the probability of a compound event in terms of the probabilities of individual events and their intersections

Conditional probability is the probability of an event occurring given that another event has already occurred. It is used to quantify the relationship between two events. The definition of conditional probability is expressed as follows:

$$P(A | B) = P(A \cap B) / P(B)$$

where A and B are two events and $P(A | B)$ is the probability of event A occurring given that event B has already occurred. The symbol “|” is read as “given” or “conditional on”.

The conditional probability of event A given event B is only defined when the probability of event B is greater than zero. This is because event B must have already occurred for the conditional probability to be meaningful.

Independence of two events means that the occurrence of one event does not affect the probability of the other event occurring. In other words, the events are not dependent on each other. Two events A and B are independent if and only if:

$$P(A \mid B) = P(A)$$


In other words, the probability of event A occurring given that event B has already occurred is equal to the probability of event A occurring regardless of the occurrence of event B.

A simple numerical problem involving conditional probability and independence can be stated as follows:

Problem: You have a box containing 3 red balls and 7 blue balls. You choose a ball randomly from the box and then replace it. Then you choose a second ball randomly from the box. What is the probability that the first ball is red and the second ball is blue?

Solution: Let event A be the event that the first ball is red and event B be the event that the second ball is blue. Since the balls are replaced after each draw, the events are independent, so we have:

$$P(A \cap B) = P(A) * P(B \mid A) = (3/10) * (7/10) = 21/100$$

So the probability that the first ball is red and the  second ball is blue is 21/100

[A WordPress.com Website.](#)

B.C.A study

Unit-6:Statistical Quality Control

Introduction

Statistical Quality Control (SQC) is a method used to monitor and control the quality of a product or service by using statistical techniques and methods. It is a systematic approach to ensuring that the products manufactured meet the specified quality standards and requirements. SQC helps organizations to identify and control the sources of variability in the production process, which can lead to defects or nonconformance.

SQC involves collecting data, analyzing the data using statistical tools, and making decisions based on the results of the analysis. The data collected can come from a variety of sources, including production processes, inspection results, and customer feedback. The statistical techniques used in SQC can range from simple statistical measures, such as mean and standard deviation, to more complex statistical models, such as control charts and Design of Experiments (DOE).

The goal of SQC is to continuously improve the quality of products or services by reducing variability and improving processes. This can lead to increased customer satisfaction, lower costs due to reduced waste and rework, and improved competitiveness in the market.

Overall, SQC is a powerful tool for organizations that want to improve their quality and competitiveness by using data and statistical analysis to make informed decisions about their processes and products

control limits

Control limits, in the context of statistical quality control, are lines or boundaries that are plotted on a control chart to distinguish between normal and abnormal behavior of a process. The control limits are calculated from the data collected from the process and are used to determine if the process is in statistical control or not.

There are typically two types of control limits used in SQC: upper control limit (UCL) and lower control limit (LCL). The UCL is the upper boundary or limit beyond which any data point is considered to be an out-of-control point, indicating that the process has deviated from its normal behavior. The LCL is the lower boundary or limit below which any data point is considered to be an out-of-control point, indicating that the process has deviated from its normal behavior.

Control limits are important in SQC because they help to determine if a process is operating consistently and within the expected limits. If a data point falls outside of the control limits, it can be an indicator of a problem with the process and can trigger an investigation to identify and correct the root cause.

In summary, control limits provide a statistical framework for detecting and correcting variations in a process, and are essential for continuous improvement and maintaining quality control in an organization

specification limits

Specification limits, in the context of statistical quality control, are the predetermined bounds that define the acceptable range for a product characteristic or process output. The specification limits define the criteria for conformance or non-conformance of a product to the established standards or customer requirements.

For example, in the manufacturing of a component, the specification limits may specify the acceptable range of dimensions, weight, strength, or other characteristics of the finished product. If a product falls outside of the specification limits, it is considered to be non-conforming and may be rejected or reworked.

Specification limits are established based on customer requirements, industry standards, and the manufacturer's own goals for quality and performance. They serve as the target for the process and are used as a basis for setting control limits in statistical quality control.

In summary, specification limits are an essential component of quality control in an organization, as they provide a clear definition of the acceptable quality criteria for products or services, and serve as a benchmark for continuous improvement and process control

tolerance limits

Tolerance limits, in the context of statistical quality control, are the acceptable bounds for deviation from the target or specification limits for a product characteristic or process output. Tolerance limits define the range within which a product or process can vary while still meeting the customer requirements and quality standards.

For example, in the manufacturing of a component, the tolerance limits may specify the acceptable range of dimensions, weight, strength, or other characteristics of the finished product, which may vary slightly from the target or specification limits. If a product falls within the tolerance limits, it is considered to be

conforming, even if it is not exactly the same as the target or specification limits.

Tolerance limits are established based on the customer requirements, industry standards, and the manufacturer's own goals for quality and performance. They serve as a flexible range for the process and help to account for normal variations in the production process.

In summary, tolerance limits are an important component of quality control in an organization, as they provide a level of flexibility for the production process and help to ensure that products or services meet the customer requirements and quality standards, even if they are not exactly the same as the target or specification limits

process and product control

Process and product control are two key concepts in statistical quality control (SQC) that are used to monitor and improve the quality of a product or service.

Process control refers to the techniques and methods used to monitor and control the production process to ensure that it is operating consistently and within the expected limits. The goal of process control is to detect and correct variations in the process, so that the process remains in statistical control and produces products or services that meet the established quality criteria.

Product control, on the other hand, refers to the techniques and methods used to monitor and control the quality of the finished product or service. The goal of product control is to detect and correct defects or nonconformities in the product, so that the product meets the established quality standards and customer requirements.


In SQC, both process control and product control are important for ensuring that the final product or service meets the required quality criteria. Process control helps to maintain consistency in the production process, while product control helps to detect and correct defects or nonconformities in the finished product


Control charts for X and R

What is it?

An X-bar and R (range) chart is a pair of control charts used with processes that have a subgroup size of two or more. The standard chart for variables data, X-bar and R charts help determine if a process is stable and predictable. The X-bar chart shows how the mean or average changes over time and the R chart shows how the range of the subgroups changes over time. It is also used to monitor the effects of process improvement theories. As the standard, the X-bar and R chart will work in place of the X-bar and s or median and R chart. To create an X-bar and R chart using software, download a copy of SQCpack.

What does it look like?

The X-bar chart, on top, shows the mean or average of each subgroup. It is used to analyze central location. The range chart, on the bottom, shows how the data is spread . It is used to study system variability.

g-chart

(<https://www.pqsystems.com/quality-solutions/statistical-process-control/SQCpack/samples/x-bar-and-range.png>)

When is it used?

You can use X-bar and R charts for any process with a subgroup size greater than one. Typically, it is used when the subgroup size falls between two and ten, and X-bar and s charts are used with subgroups of eleven or more.

Use X-bar and R charts when you can answer yes to these questions:

1. Do you need to assess system stability?
2. Is the data in variables form?
3. Is the data collected in subgroups larger than one but less than eleven?
4. Is the time order of subgroups preserved?

Getting the most

Collect as many subgroups as possible before calculating control limits. With smaller amounts of data, the X-bar and R chart may not represent variability of the entire system. The more subgroups you use in control limit calculations, the more reliable the analysis. Typically, twenty to twenty-five subgroups will be used in control limit calculations.

X-bar and R charts have several applications. When you begin improving a system, use them to **assess the system's stability**.

After the stability has been assessed, **determine if you need to stratify the data.** You may find entirely different results between shifts, among workers, among different machines, among lots of materials, etc. To see if variability on the X-bar and R chart is caused by these factors, collect and enter data in a way that lets you stratify by time, location, symptom, operator, and lots.

You can also use X-bar and R charts to **analyze the results of process improvements.** Here you would consider how the process is running and compare it to how it ran in the past. Do process changes produce the desired improvement?

Finally, use X-bar and R charts for **standardization.** This means you should continue collecting and analyzing data throughout the process operation. If you made changes to the system and stopped collecting data, you would have only perception and opinion to tell you whether the changes actually improved the system. Without a control chart, there is no way to know if the process has changed or to identify sources of process variability.

Control charts for number of defective {n-p chart}

NP CONTROL CHARTS

An np control chart is used to look at variation in yes/no type attributes data. There are only two possible outcomes: either the item is defective or it is not defective. The np control chart is used to determine if the number of defective items in a group of items is consistent over time. The subgroup size (the number of item in the group) must be the same for each sample.

A product or service is defective if it fails, in some respect, to conform to specifications or a standard. For example, customers like invoices to be correct. If you charge them too much, you will definitely hear about it and it will take longer to get paid. If you charge them too little, you may never hear about it. As an organization, it is important that your invoices be correct. Suppose you have decided that an invoice is defective if it has the wrong item or wrong price on it. You could then take a random sample of invoices (e.g., 100 per week) and check each invoice to see if it is defective. You could then use an np control chart to monitor the process.

You use an np control chart when you have yes/no type data. This type of chart involves counts. You are counting items. To use an np control chart, the counts must also satisfy the following two conditions:

1. You are counting n items. A count is the number of items in those n items that fail to conform to specification.
2. Suppose p is the probability that an item will fail to conform to the specification. The value of p must be the same for each of the n items in a single sample.

If these two conditions are met, the binomial distribution can be used to estimate the distribution of the counts and the np control chart can be used. The control limits equations for the np control chart are based on the assumption that you have a binomial distribution. Be careful here because condition 2 does not always hold. For example, some people use the p control chart to monitor on-time delivery on a monthly basis. A p control chart is the same as the np control chart, but the subgroup size does not have to be constant. You can't use the p control chart unless the probability of each shipment during the month being on time is the same for all the shipments. Big customers often get priority on their orders, so the probability of their orders being on time is different from that of other customers and you can't use the p control chart. If the conditions are not met, consider using an individuals control chart.

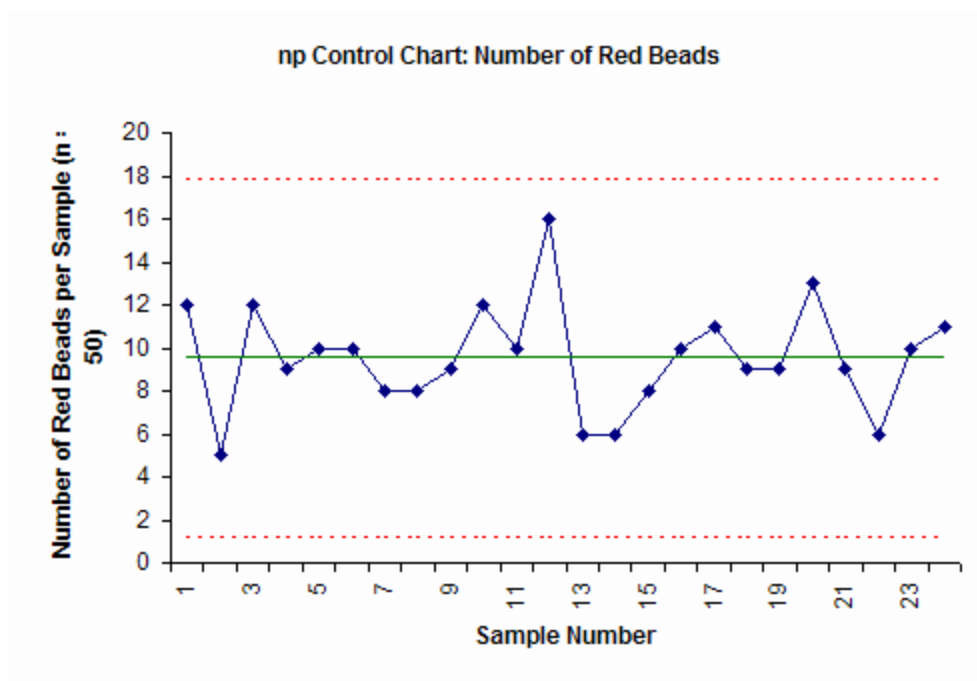
NP CONTROL CHART EXAMPLE: RED BEADS

The red bead experiment described in last month's newsletter is an example of yes/no data that can be tracked using an np control chart. In this experiment, each worker is given a sampling device that can sample 50 beads from a bowl containing white and red beads. The objective is to get all white beads. In this case, a bead is "in-spec" if it is white. It is "out of spec" if it is red. So, we have yes/no data – only two possible outcomes. In addition, the subgroup size is the same each time, so we can use an np control chart.

Data from one red bead experiment are shown below. The numbers represent the number of red beads each person received in each sample of 50 beads.

WORKER	DAY 1	DAY 2	DAY 3	DAY 4
Tom	12	8	6	9
David	5	8	6	13
Paul	12	9	8	9
Sally	9	12	10	6
Fred	10	10	11	10
Sue	10	16	9	11

The np control chart from this data is shown below.



The np control chart plots the number of defects (red beads) in each subgroup (sample number) of 50. The center line is the average. The upper dotted line is the upper control. The lower dotted line is the lower control limit. As long as all the points are inside the control limits and there are no patterns to the points, the process is in statistical control. We know what it will produce in the future. While we don't know the exact number of red beads a person will draw the next time, we know it will be between about 2 and 17 (the control limits) and average about 10.

STEPS IN CONSTRUCTING AN NP CONTROL CHART

The steps in constructing the np chart are given below. The data from above is used to demonstrate the calculations.

1. Gather the data.

- ” a. Select the subgroup size (n). Attributes data often require large subgroup sizes (50 – 200). The subgroup size should be large enough to have several defective items. **The subgroup size must be constant.**

In the red bead example, the subgroup size is 50.

b. Select the frequency with which the data will be collected. Data should be collected in the order in which it is generated.

c. Select the number of subgroups (k) to be collected before control limits are calculated. You can start a control chart with as few as five to six points but you should recalculate the average and control limits until you have about 20 subgroups.

d. Inspect each item in the subgroup and record the item as either defective or non-defective. If an item has several defects, it is still counted as one defective item.

e. Determine np for each subgroup.

np = number of defective items found

f. Record the data.

2. Plot the data

- ” a. Select the scales for the control chart.

b. Plot the values of np for each subgroup on the control chart.

c. Connect consecutive points with straight lines.

3. Calculate the process average and control limits.

- ” a. Calculate the process average number defective:

$$\bar{np} = \frac{\sum np}{k} = \frac{np_1 + np_2 + \dots + np_k}{k} = \frac{229}{24} = 9.54$$

” where np_1, np_2 , etc. are the number of defective items in subgroups 1, 2, etc. and k is the number of subgroups.

In the red bead example, each of the six workers had 4 samples. So, $k = 24$. The total number of red beads (summing all the data in the table above) is 229. Thus, the average number of defective items (red beads) in each sample is 9.54.

b. Draw the process average number defective on the control chart as a solid line and label.

c. Calculate the control limits for the np chart. The upper control limit is given by UCL_{np} . The lower control limit is given by LCL_{np} .

$$UCL_{np} = \bar{np} + 3\sqrt{\bar{np}(1 - (\bar{np}/n))} = 9.54 + 3\sqrt{9.54(1 - (9.54/50))} = 17.87$$

$$LCL_{np} = \bar{np} - 3\sqrt{\bar{np}(1 - (\bar{np}/n))} = 9.54 - 3\sqrt{9.54(1 - (9.54/50))} = 1.20$$

The control limits for the red bead data are calculated by substituting the value of 9.54 for the average number defective and the value of 50 for the subgroup size in the equations above. This gives an upper control limit of 17.87 and a lower control limit of 1.20.

d. Draw the control limits on the control chart as dashed lines and label.

4. Interpret the chart for statistical control.

” a. The following tests for statistical control are valid.

- Points beyond the control limits
- Length of runs test
- Number of runs test

The red bead control chart (shown above) is in statistical control. All the points are between the control limits and there are no patterns. Remember that the upper control limit represents the largest number we would expect if only common cause of variation is present. The lower control limit represents the smallest number we would expect. As long as the process stays the same, we can predict what will happen in the future. Each person will have between 2 and 17 red beads in the sample. This won't change until we fundamentally change the process.

c-chart

What is it?

A c-chart is an attributes control chart used with data collected in subgroups that are the same size. C-charts show how the process, measured by the number of nonconformities per item or group of items, changes over time. Nonconformities are defects or occurrences found in the sampled subgroup. They can be described as any characteristic that is present but should not be, or any characteristic that is not present but should be. For example a scratch, dent, bubble, blemish, missing button, and a tear would all be nonconformities. C-charts are used to determine if the process is stable and predictable, as well as to monitor the effects of process improvement theories. C-charts can be created using software products like SQCpack.

What does it look like?

The c-chart shows the number of nonconformities in subgroups of equal size.

 c control chart

(<https://www.pqsystems.com/quality-solutions/statistical-process-control/SQCpack/samples/c-chart.png>)



[A WordPress.com Website.](#)