Basics of Mechanical Engineering-3

Prof. J. Ramkumar

Prof. Amandeep Singh Oberoi

Department of Mechanical Engineering

Indian Institute of Technology, Kanpur

Week 12

Lecture 50: Basics of Engineering Statistics

Welcome back to the course Basics of Mechanical Engineering 3. This is the third course in the series of Basics of Mechanical Engineering 1, 2, and 3, which we have been presenting to you over the last three sessions. Basics of Mechanical Engineering 1 covered the basics of engineering mechanics, strength of materials, theory of machines, and some parts of machine design, which are suitable for first-year and second-year B.Tech engineering and diploma students.

The second part of the course series discussed basic manufacturing topics, such as casting, machining, welding, and other manufacturing processes. In this part, we will cover thermodynamics and fluid mechanics. In all three courses, we have included tutorial sessions to help you understand real-life problems, such as how to handle data related to manufacturing, design, boiler efficiency, and similar topics. We have also conducted virtual laboratory demonstrations to show you how to operate experimental setups in a virtual environment from home using your laptop, including how to click, switch on/off, take readings, and perform calculations.

This is the 12th week of the third part, which is the final week of the entire three-course series. In this session, I will discuss a topic that is not directly part of manufacturing but is essential for understanding and working in manufacturing. That topic is engineering statistics. In the numerical examples I have discussed, I primarily used averages. I also used the median, which is the middle point.

These are all part of statistics, which you have been studying since school. Now, I will provide more insight into the real need for statistics in engineering. I will start with measures of central tendency, such as mean, mode, median, and mid-range, and expand on these concepts beyond what you already know. Then, I will discuss the types of

statistics, including descriptive and inferential statistics. We will explore these categories in detail.

Then we'll talk about distributions. When we talk about taking the data or collecting the data, when the data is collected, first the data is cleaned for any missing values, for any things which are not required or which are outliers. All those things are taken away. Then, we try to work on the data to plot something to understand the behavior, like I told you to plot the characteristic curve between the efficiency and the input pressure in the virtual laboratory demonstrations. I will also try to show you those in the Excel sheet this week.

Then, I will talk about hypothesis testing, which is inferential statistics. Whenever we design any experimental theory, when I say experimental theory, theory is based upon the experiments, number of experiments are run. When all the time these experiments are giving exactly similar results. When I say exactly similar, I am not talking about same. Exactly similar because there could be 0.00001% error. I am talking about five zeros and then one. For example, errors are always there.

It could be at six sigma level, nine sigma level or maybe three sigma level. What is sigma here? That is the standard deviation that I will talk to you about this week. Let me start with the very first lecture of week 12, which talks about the basics of engineering statistics.

Contents

I will just introduce what statistics is and the importance of it. Use of statistics in mechanical engineering: types of statistics, application areas, and limitations. Some terms in statistics are already familiar, while others require more understanding of the kinds of statistics. Stages in statistical investigation: data collection and data presentation. What is statistics?

Introduction

- Statistics is a branch of mathematics, yet a science, which concerns with collecting, organizing, analyzing and interpreting numerical data.
- It is recognized as a distinct scientific discipline due to its broad applications across numerous fields, including science, economics, healthcare and social sciences.
- It helps make sense of complex data through quantitative models.
- It plays a critical role in decision-making in fields like weather forecasting, stock market analysis, insurance and data science.

https://learnwithexamples.org/learn-statistics-for-beginners/ 3

Statistics is a branch of mathematics, yet a science, which concerns collecting, organizing, analyzing, and interpreting numerical data. You see the steps. What is collecting? Then organizing, then analyzing, then interpreting. There are five steps anyway. I'll talk about that because it's not only interpreting; we also need to present the data. I will talk about that when I go through the slides further. It is recognized as a distinct scientific discipline due to its broad applications across numerous fields, including science, economics, healthcare, social science, and so on.

It helps make sense of complex data through quantitative models. It plays a critical role in decision-making in fields like weather forecasting, stock market analysis, insurance, and data science. Talk about the very general data presentation, graphical presentation. I will talk about line diagrams, bar graphs, and pie charts. Nowadays, you are getting different kinds of charts. In the stock market, they show you tree charts.

They show you a fan chart. Different kinds of charts are there which give you deeper information for specifically understanding a particular kind of chart. There are courses available. People are producing their own personal courses where, for one hour, they will teach you how to read this chart and how to present your data in this specific kind of chart.

Importance of Statistics

- Statistics and statistical analysis play a vital role in many fields such as business, engineering, economics and healthcare, while monitoring / observing from one's own office.

- It helps to:
  - Gather, classify and simplify data
  - Provide concrete information about problems and issues
  - Facilitate reliable and objective decision-making
  - Present facts and interpretations in a precise and definite form

Now, why is statistics important? Statistics and statistical analysis play a vital role in many fields such as business, engineering, economics, and healthcare. While monitoring and observing from one's own office, it helps to gather, classify, and simplify data. It provides concrete information about problems and issues. It facilitates reliable and objective decision-making. It presents facts and interpretations in a precise and definite form. So, statistics is the study of data. When the domain of study is the properties of data, it comes under descriptive statistics. When it is about concluding from the data, it comes under inferential statistics. This I will talk about.

Statistics in Mechanical Engineering

- Engineering statistics combines engineering and statistics using scientific methods for analyzing data.
- Engineering statistics involves data concerning manufacturing processes such as:
  - ✓ component dimensions,
  - ✓ tolerances,
  - ✓ type of material, and
  - ✓ fabrication process control.
- Mechanical engineers use the principles of calculus, statistics and other advanced subjects in math for analysis, design and troubleshooting in their work.
- Using statistical methods is essential for mechanical engineers to design, validate hypothesise, maintain quality control and create innovative solutions.

Engineering statistics combines engineering and statistics using scientific methods for analyzing data. So we can compare, predict, and formalize a specific method on how to make a specific comparison between two kinds of setups, maybe mechanical setups, or two kinds of boilers.

What is the efficiency of boiler 1, the efficiency of boiler 2, the number of boilers to be tested, the number of nozzles to be tested and put on the boiler, to understand the quality of the nozzle to be put on the boiler. Now, all these things can be done using statistics, but we need to have data for that. Engineering statistics involve data concerning manufacturing processes such as component dimensions.

If I'm talking about a nozzle, it must fit precisely in a specific slot in the boiler. That nozzle must have a tolerance equivalent to the boiler's specific hole or thread. We discussed tolerances in the first part of the Basics of Mechanical Engineering course series. Tolerances could be clearance tolerance, or in this case, I am talking about a fit. It could be a clearance fit.

It could be an interference fit. It could be a transition fit. For the specific kind of fit, if we suppose we have 100 nozzles, we have to select 5 out of them. We might have to test all the nozzles, maybe. This is my sensor survey.

Out of all the nozzles, we will see the dimensions of the nozzle and the dimension of the hole where it is to be fit. Then we will measure. Then these component dimensions are the part of the data collection. Then come tolerances. An interference fit is a very tight fit.

For example, putting a pen in a pocket is a very loose fit. Putting the cap of the pen on is a slight interference fit. That is why it makes a click sound. Then, a transition fit is between the interference fit and the clearance fit. Type of material.

That is very important as well. If you are talking about boilers, we will talk about the nozzles made out of material that can bear the high temperature of the steam. Generally, mild steel nozzles could not be used directly there. SS, copper; these kinds of nozzles would be used there. Fabrication process control. When I say fabrication process control, if we say first is the kind of fabrication process that we need to select, whether it is manufacturing through 3D printing or it is manufacturing through regular machining. Regular machining and this kind of the processes would be CNC machining (Computer Numerical Control) machining.

In those, we could see whether it is a one component that it is required is manufactured in two parts that is part one or part two then welded together or single part is manufactured in just one go. Fabrication process control that is also part of the data that is concerning manufacturing processes.

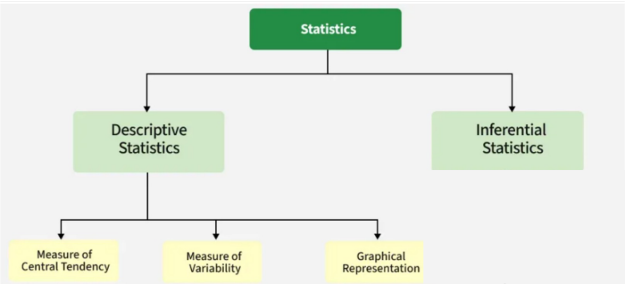
Mechanical engineers use principles of calculus, statistics and other advanced subjects for analysis, design and troubleshooting in their work. Using statistical methods is essential for mechanical engineers to design, validate, hypothesize, maintain quality control, and create innovative solutions. What are the types of statistics?

Types of Statistics

- There are two main branches of statistics:

1) Descriptive statistics

1) Inferential statistics



There are multiple types of statistics. I will talk about descriptive and inferential statistics. Other than this, we have predictive statistics. What is predictive? We predict the future. We predict the demand. We predict the weather.

That is predictive statistics. Prescriptive statistics. After descriptive and inferential, prescriptive prescribing something. Giving a decision, giving a detailed information or this should be the call that you should take on a specific kind of a selection of a boiler, selection of the machinery that you will overcome on, selection of the material that you have. That is prescription based upon the statistical background. That is prescription based upon an inference.

What is it inference, that is inferential statistics. This we will keep talking about in this week throughout descriptive statistics have measures of central tendency, that is mean, mode, median. Measures of variability, and graphical presentation. That is, we describe the data only. We collect the data. We describe what is the data kind of. That is descriptive statistics.

Types of Statistics

1) **Descriptive Statistics:**

- The initial phase of statistical analysis, focusing on data processing tasks such as collection, organization, presentation and analysis.
- Highlights key features of the data sample without making inferences or conclusions beyond the observed dataset (sample).
- Descriptive statistics describes the nature or characteristics of the observed data quantitatively without making conclusions or generalization.

To talk about little detail and some examples in this, descriptive statistics, the initial phase of statistical analysis is Focusing on data processing tasks such as collection, organization, presentation, and error analysis.cIt highlights key features of a data sample without making inferences. This is very important without making inferences because if inferences are made, it becomes inferential statistics. And no conclusions beyond the observed dataset or sample are taken. Descriptive statistics describes the nature and characteristics of the observed data quantitatively without making conclusions or generalizations. When we say quantitatively, we are talking about using a sample.

We talk about what is sampling inspection and what is census inspection. We call it inspection or survey. Sampling is when we take a small part of the overall lot. Census is when the whole part is done. We will talk about that and why sampling is important. That part will also be covered.

Descriptive Statistics

The following are some examples of descriptive statistics:

1. Machining: The average surface roughness of 50 turned steel samples was measured to be 2.5 µm Ra.

1. Compressors: During last week's test, the discharge pressure of a centrifugal compressor ranged from 6 bar to 9 bar.

1. Boilers: The efficiency of a boiler operated at IIT Kanpur's power plant last month averaged 82%.

Some examples of descriptive statistics in machining. See the statement. The average surface roughness of 50 steel samples was measured to be 2.5 micrometers Ra. That is, surface roughness is the parameter.

I will call it; this is the response parameter that has been taken from a specific kind of observation. What is the number of observations? 50 turn-steel samples. And what is the final average value? That is 2.5 micrometer Ra.

We are just reporting it. We are just describing the data. Nothing is an inference out of it. Compressors, during last week's test, the discharge pressure of a centrifugal compressor ranged from 6 bar to 9 bar. This is a complete statement about the last week's test.

It tells about the time period when the test was taken. That is, we are describing the data only. The parameter here is discharge pressure. Of what? Which equipment?

Centrifugal compressor. This ranges from 6 bar to 9 bar. Boilers. The efficiency of a boiler operated at IIT Kanpur's power plant last month averaged 82%. There are boilers here.

Just beside IIT Kanpur, we have a small area called Panki where the power plants are located. The boiler efficiency of Panki, which operated for IIT Kanpur, monthly averaged at 82%. This is the boiler efficiency. This is a descriptive statement.

Descriptive Statistics

The following are some examples of descriptive statistics:

4. IC Engines: In a performance test, the brake thermal efficiency of a 4-cylinder diesel engine was found to be 28% at full load.

4. Heat Exchangers: The log mean temperature difference (LMTD) recorded across a shell-and-tube heat exchanger was 45°C during steady-state operation.

4. Nozzles: The velocity of steam measured at the nozzle exit during an experiment varied between 420 m/s and 460 m/s.

In IC engines, during a performance test, the brake thermal efficiency of a 4-cylinder diesel engine was found to be 28% at full load. You see the statement, the parameter, brake thermal efficiency. When I say parameter, this is a response parameter.

The equipment under study, then we talk about the value, what is the value that came out, and what is the condition? This is the environmental condition. Using the word environmental here does not mean we are talking about the overall environment. We are talking about the condition in which this is run.

In full load, it is 28%. In half load, it could be 46% or something. In heat exchangers, the log mean temperature difference (LMTD) recorded across a shell-and-tube heat exchanger was 45 degrees during steady-state operation. This is the environmental condition. 45 degrees centigrade is the value.

Shell and tube. The heat exchanger is the system or equipment under study. What is being recorded or what is being studied. The parameter is LMTD. Nozzles.

The velocity of steam measured at the nozzle exit during an experiment varied between 420 meters per second and 160 meters per second. This is the range. I'm giving you certain terms here. For example, I'm talking about the response parameter. I'm talking about the range.

I'm talking about the system under study. I'm talking about the condition. These are statistical terms, which I'll talk about in the coming slides as well.

Descriptive Statistics

Some Descriptive Statistics measures commonly used are measures of:

1. Central tendency, and

1. Measures of variability

- **Measures of central tendency** include Mean, Median and Mode.
- **Measures of variability** include Standard Deviation (or variance), Minimum and Maximum values of variables, Kurtosis and Skewness.

Now, in descriptive statistics, there are two major measures of kind. One is central tendency. Another is the measure of variability. Central tendency is mean, median, and mode. This means if the data is there, where is the center of the data? Mean or average is the center. Median is between the counts.

That is the center. Mode is the maximum value. It is also the center. This is a measure of central tendency. Other than that, mid-range is also a center.

Then, across the center, we have variability. That is, we talk about standard deviation or variance, minimum, maximum values of variables, quartiles, and skewness. From the center, what is the spread? The mean could be whatever we have taken. Suppose the mean was 2.5 micrometers for the roughness value.

2.5 micrometer roughness. The standard deviation could be this broad spread or it could be a very small spread. That means it could vary from 2 micrometers to 3 micrometers. The average is 2.5 micrometers. It could vary from 0.5 micrometers to 4.5 micrometers.

But the average is 2.5 micrometers. So, what is the spread? This is variance or standard deviation; that is how it is varying. Minimum value and maximum value—that is also a range. Maximum value minus minimum value is the range.

4.5 micrometers minus 0.5 micrometers is 4, which is a bigger range. Then, 3 micrometers minus 2 micrometers is 1 micrometer, which is a smaller range. So, this is a measure of variability. This also we will talk about in detail. Then comes the other type, which is inferential statistics.

Inferential Statistics

2) Inferential statistics:

- It involves making generalizations and drawing conclusions (inferences) about specific characteristics of a population based on information obtained from samples.
- It includes methods like hypothesis testing, confidence intervals, regression analysis, analysis of variance (ANOVA), and chi-square tests.
- These techniques allow researchers to analyze a sample's data and make predictions, generalizations, or conclusions about a larger population.

It involves making generalizations and drawing conclusions or inferences about specific characteristics of a population based on information obtained from samples. It includes methods like hypothesis testing, confidence intervals, regression analysis, analysis of variance, chi-square tests, etc. I'll cover some of these methods here this week. These techniques allow researchers to analyze a sample's data and make predictions, generalizations, or conclusions about a larger population.

Inferential Statistics

The following are some examples of Inferential Statistics:

1. Machining: Based on tool wear data from 30 cutting tests, it is estimated that the average tool life in turning AISI 1040 steel is 45 minutes under the same cutting conditions.

1. Compressors: From experiments on five test units, it is predicted that the average isentropic efficiency of this type of compressor will be around 78% for all manufactured units.

1. Boilers: Using heat loss data from 10 trial runs, it is forecast that the boiler efficiency will drop by 5% after one year of continuous operation.

Let us see examples in the same fields that I talked about: machining, compressors, boilers, IC engines, etc. In machining, an inferential statement you see, based on tool wear data from 30 cutting tests, is that the average tool life in turning AISI 1040 steel is 45 minutes under the same cutting conditions. This is an inferential statement.

In compressors, for experiments on 5 test units, it is predicted that the average isentropic efficiency of this type of compressor will be around 78% for all manufacturing units. Let us see the statement. It talks about something taken from a small number, only 5 test units, and then they say for all manufacturing units.

That is a small number tested. Only 5 units are tested. And we say throughout the manufacturing unit, this is replicable. This is known as replicability. It is known as the interchangeability of the results as well.

Interchangeability, replicability—these are terms when a small sample gives similar results for a specific confidence. We call it the confidence interval, which we will talk about. They say 78% for all manufacturing units. This is the isentropic efficiency that would be there. So, this is known as sampling.

That also we will talk about. Boilers using heat loss from 10 trial runs—it is forecast that the boiler efficiency will drop by 5% after one year of continuous operation.

Inferential Statistics

The following are some examples of Inferential Statistics:

4. IC Engines: From a set of endurance tests, it is inferred that diesel engines using biodiesel blends can reduce CO emissions by about 20% compared to conventional diesel across similar models.

4. Heat Exchangers: Analyzing test results from three prototypes, it is estimated that the new design of compact heat exchangers will achieve 15% higher effectiveness in industrial use.

4. Nozzles: From CFD simulations validated by experiments, it is projected that the supersonic nozzle design will deliver an exit velocity of ~1200 m/s when scaled up to full size.

IC engines. From a set of endurance tests, it is inferred that diesel engines using biodiesel blends can reduce carbon monoxide emissions by about 20% compared to conventional diesel across similar models. They say biodiesel blends reduce carbon dioxide emissions. Heat exchangers. Analyzing test results from three prototypes. It is estimated that the new design of a compact heat exchanger will achieve 15% higher effectiveness in industrial use. This is an inference. It's a kind of deterministic statement.

Nozzles. From CFD simulations, that is, computational fluid dynamics simulations validated by experiments, it is predicted that the supersonic nozzle design will deliver an exit velocity of around 1200 meters per second when scaled up to full size. So, these are inferential statements.

These have a lot of data, I would say analysis or analytics as well behind them. Hypothesis testing has been conducted. Regression analysis might have been performed. Analysis of variance might have been performed. So, based on percentage tests like I talked about one of the tests in the previous slide, chi-square test or maybe T-test, all these inferences have come.

Descriptive vs Inferential Statistics

	Descriptive Statistics	Inferential Statistics
1	It gives information about raw data which describes the data in some manner.	It makes inference about population using data drawn from the population.
2	It helps in organizing, analyzing and to present data in a meaningful manner.	It allows us to compare data, make hypothesis and predictions.
3	It is used to describe a situation.	It is used to explain the chance of occurrence of an event.
4	It explain already known data and limited to a sample or population having small size.	It attempts to reach the conclusion about the population.
5	It can be achieved with the help of charts, graphs, tables etc.	It can be achieved by probability.

Now, let me have a quick glance over the overall differences between descriptive and inferential statistics. Descriptive statistics give information about raw data, which describes the data in some manner. Inferential statistics make inferences about the population using data drawn from the population.

Descriptive statistics help in organizing, analyzing, and presenting the data in a meaningful manner. On the other hand, inferential statistics help us to compare the data, make hypotheses, and predict based upon the data.

Descriptive statistics are used to describe a situation only. However, inferential statistics help us to explain the chance of occurrence of an event. Descriptive statistics explain already known data. They are limited to a sample or a population having a small size. However, the aim of inferential statistics is to reach conclusions about the population.

The work of descriptive statistics can be achieved with the help of charts, graphs, and tables, which are made only to describe what kind of data it is. In inferential statistics, probability is always used.

Application Areas

- **Engineering:**
  Enhancing product designs, testing product performance, determining reliability and maintainability and developing safer flight control systems for airports, planning for technology.

- **Business :**
  Estimating the volume of retail sales, designing efficient inventory management systems, producing auditing and accounting procedures, improving industrial working conditions and assessing the market potential for new products.

Certain applications of engineering statistics. In engineering, it helps in enhancing product designs, testing product performance, determining reliability and maintainability, developing safer flight control systems for airports, etc., and planning technology. In business, it assists in estimating the volume of retail sales, designing efficient inventory management systems, producing auditing and accounting procedures, improving industrial working conditions, and assessing the market potential for new products.

Application Areas

- **Quality Control:**
  In this, techniques are developed for evaluating quality through proper sampling, in-process control, consumer surveys, and experimental design in product development.
  Recognizing its significance, many large organizations have established their own Statistical Quality Control Departments.

- **Economics:**
  In economics, key indicators such as trade volume, labor force size, and living standards are measured. This field also involves analyzing consumer behavior, computing national income accounts, and formulating economic laws, with regression analysis being extensively used.

In quality control, statistical techniques are developed for evaluating quality through proper sampling, in-process control, consumer surveys, and experimental design in product development.

Recognizing its significance, many large organizations have established their own SQC department, statistical quality control department, or what you call a quality department. There are large organizations such as BSI and TUV, which help us in developing our own QMS (Quality Management Systems), generally known by the term ISO systems. For example, IIT Kanpur has a METEC facility that is ISO 13485 certified. So whatever is being done in this facility, the material can be tracked, the people working on it can be tracked, and the processes can be tracked. Prototype development is happening in this facility.

So here, the Quality Management System is taken from BSI. It is British Standards that developed these standards for medical devices, and statistical quality control, statistical data analysis, and statistical data presentation are used to maintain the quality within specific limits. That is, quality conformance is being assured there.

In economics, key indicators such as trade volume, labor force size, and living standards are measured. This field also involves analyzing consumer behavior, computing national income accounts, and formulating economic laws, with regression analyses being extensively used.

Application Areas

- **Health and Medicine:**
  Statistics play a crucial role in developing and testing new drugs, enhancing medical care, and diagnosing and treating diseases, with inferential statistics being particularly significant.

- **Biology:**
  Statistics help explore species interactions with their environment, create theoretical models of the nervous system and study genetic evolution.

- **Psychology:**
  Statistics are used to measure learning ability, intelligence and personality traits, as well as to develop psychological scales and understand abnormal behavior.

Health and medicine statistics play a crucial role in developing and testing new drugs, enhancing medical care, diagnosing and treating diseases, with influential statistics being particularly significant. In biology, statistics help to explore species interactions with their environment, create theoretical models of the nervous system, and study genetic evolution; in psychology, statistics are used to measure learning ability, intelligence, and personality traits, as well as to develop psychological skills and understand abnormal behavior. With so much of the use of statistics, there is always a limitation or demerits associated with it.

Limitations of Statistics

1. Statistics focuses and provides results based on data that can be measured and quantified, rather than individual cases.

1. It does not address qualitative aspects, it concentrates solely on numerical data.

1. Statistical results can not become universal truth, hence can be misused.

1. There could be cheating through scales using statistics.

What are the limitations of statistics? Statistics focuses on and provides results based on data that can be measured and quantified rather than individual cases. If the data is not a right fit for the present scenario, the current scenario, whatever statistical results are there based on the data, that is, the past data, would not fit right to the current scenario. It does not address qualitative aspects. That is, it concentrates solely on numerical data, though there are qualitative aspects that could be also addressed. So, quantified or statistics could be used in them. For example, yes or no.

If multiple times you do yes and no, you purchase something on Amazon, you purchase a nozzle for your maybe laser cutting. For the specific nozzle, you will have a rating. Rating is suppose 4.2. From there, where this 4.2 number has come? When they say 4.2

rating is there from the 5000 users. Different users are given different rating that is a qualitative rating. Qualitative data, out of the 5 they have given whatever star, 2, 3, 4, 5, 5, 5, 4, 2, 3. These are all average and average is 4.2. This is how the qualitative data, they have given only quality. Out of the 5 stars, this is the star I give, then converting it to quantitative is 4.2 is a quantity.

Now a number that has come, so that is one, but most of the quantitative part does not allow the qualitative aspects of the problem to be addressed; statistical results cannot become universal truth, hence can be misused, there could be cheating through skills using statistics. Now once we have known the limitations of statistics.

Statistical Terms

- The study of engineering statistics involves research activities. For this, the following main terms are used:

- **Population:** A large group of data or a large number of measurements is called population. A population can be finite or infinite. e.g. number of students in a class, number of engines produced in an industry over a span of time etc.
- **Sample:** A subset of data taken from some large population or process is a sample. It is a portion of the population selected for detailed analysis.
- **Random sample:** If each item in the population has an equal opportunity of being selected, it is called a random sample. This definition is applicable for both infinite and finite population.
- **Data:** Certainly known facts from which conclusions may be drawn.

Let me now talk about some of the terms which I use in statistics. I've been using certain terms here. I have been talking about the population and the sample. I'm talking about qualitative and quantitative data. What are these terms? And I will keep using these terms throughout this week. What are these terms that we talk about? Population. This is a large group of data, a large number of measurements.

That is known as population. A population can be finite or infinite. For example, the number of students in a class, or the number of engines produced in an industry over a span of time. These are finite populations. Then comes sample.

A subset of data taken from some large population or process is a sample. It is a portion of the population selected for detailed analysis. As I said, for example, in the manufacturing of a nozzle itself, there were 100 nozzles. We tested only 5 out of them. That is 5% of the overall population.

The sample size is there. The total population size is 100. That is also called the lot size. And we have 5 as the sample size. Random sample: if each item in the population has an equal opportunity of being selected, it is called random sampling.

This definition is applicable to both infinite and finite populations. Data, commonly known facts from which conclusions can be drawn, is data.

Statistical Terms

- **Statistical data:** Raw material for a statistical investigation which are obtained whenever measurements or observations are made.

It is of two types:

i. **Quantitative data:** data of a certain group of individuals which is expressed numerically.

ii. **Qualitative data:** data of a certain group of individuals that is not expressed numerically.

On basis of source, data is said to be **Primary data and Secondary data.**

Statistical data. Raw material for a statistical investigation, which is obtained whenever measurements or observations are made. This is statistical data. It is very important to understand here. I am looking at the last line here. Primary data and secondary data. Data can be taken from both sources. If data is collected right now through experimentation, that is primary data.

If data is taken from the recorded portion that is already present in books, journals, or specific websites. For example, ProWise is a website that provides data about the economy and financial growth. Economic Input-Output Life Cycle Analysis (EIOLCA) is a website that provides data about the overall carbon footprint produced for specific types

of manufacturing. For example, in the automobile or pharmaceutical industries, what is the carbon footprint? This is all secondary data taken from verified sources, which are also somewhat authentic. Primary data is when I collect my own data, performing all measurements on the specific five nozzles I have selected. These five samples are measured using a vernier caliper, micrometer, non-contact method, etc. That is primary data being collected.

That is data collection. We will discuss the steps of statistical analysis. Data collection is the first step. Now, there are two types of data. Quantitative data: data from a certain group of individuals expressed numerically.

Qualitative data: data from a certain group of individuals not expressed numerically. Everything we are discussing now is quantitative data. The height of the specific sample size you have taken. The size or diameter of the nozzle. The weight of the nozzle you have selected.

These are all quantitative numbers. Qualitative data. Nozzle color: acceptable, not acceptable—qualitative data. Nozzle received, not received—qualitative data. Whether the color is black, white, or which box you put it in—qualitative data. Nozzle: correct, not correct, whatever we talked about—acceptable, not acceptable. This is all qualitative data. When I talk about the rating between 1 to 5, that is also qualitative data. Rating 1, 2, 3, 4, 5—this could be quantified. Quantitative data could also be qualified. Both ways, the data could have a kind of change of scales.

Statistical Terms

Variable:

- A variable is a **factor or characteristic** that can take on **different possible values or outcomes**.
- A variable can be qualitative or quantitative (numeric).
- Example: Income, height, weight, sex, age, etc. of a certain group of individuals are examples of variables.

Survey (Experiment):

- Survey are of two types :

1. **Census Survey:** A way of obtaining data referring the entire population including a total coverage of the population.
2. **Sample Survey:** A way of obtaining data referring to a portion of the entire population consisting only a partial coverage of the population.

Variable. A variable is a factor or characteristic that can take on different possible values or outcomes. We talked about the efficiency of the boiler—that would be this. The size of the specific component—that is, 2.5 micrometer roughness. The variable here is roughness. I talked about response parameters. Whatever response parameters I talked about in the examples given for descriptive and inferential statistics—those were all variables. Variables can be qualitative or quantitative—that is, numeric, or we also call it metric.

When we talk about quantitative and qualitative, it is non-metric. Examples include income, height, weight, sex, and age. These are all kinds of variables for a certain group of individuals, and they are examples of variables. Survey, or you can call it inspection. These are of two types: census survey and sample survey.

A census survey is a way of obtaining data referring to the entire population, including total coverage of the population. A sample survey is a way of obtaining data referring to a portion of the entire population, consisting of only partial coverage of the population. A census, as you know, is also known as 100% inspection.

Each and every component, each and every part that you are making, would be tested for a specific parameter, such as the diameter of the nozzle or the internal diameter of the nozzle that goes into the hole—a specific diameter. If you survey or inspect all 100 nozzles you have produced—the total lot size—and you test all 100, that is a census survey.

If you select a specific sample of 5 or 10, or whatever number you decide, that is a sample survey. A sample survey has its advantages. The time spent on this overall inspection is less. In a census survey, sometimes non-destructive testing is used. For example, if you need to see the internal properties of the nozzle, you might have to cut the nozzle and examine its internal properties using multiple techniques.

That is non-destructive testing. If you do it for all of them, you would have nothing left. All 100 nozzles would be destroyed. In a sample survey, you destroy only five nozzles to test the behavior of the internal material characteristics of the remaining 95 pieces. So, that is a sample survey.

Both have their own pros and cons. Generally, we do sample surveys to save resources—time, money, manpower, and other resources. Then, That is why we have inferential

statistics based on hypothesis testing. This will be covered in the third lecture of this week.

Steps/stages in Statistical Investigation

1. **Collection of data**
1. **Organization of Data**
3. **Presentation of data**
4. **Analyzing of data**
4. **Interpretation of data:**

Stages or steps in statistical investigation. First is we collect the data. Data collection could be primary, could be secondary. Primary data is anyway one of the ways which specific parameter you are going to work upon, which range you are going to work upon, what is your overall domain of the work that is primary data selection criteria. Secondary data is also very important to be thought of. which kind of the data source you are trying to refer to and about that data source there will be very large amount of data.

Nowadays we talk about big data analytics, we talk about a lot of other techniques which are there, for example, hadoop; multiple techniques are there. Large data is there each second; weather reporting is happening at each area in kanpur. Each second, what is the weather reporting that is happening? This generates a large amount of data throughout india, throughout world. Out of this overall data, what is the kind of secondary data that is under your study you need to select?

That is the right collection of the data. Then organizing the data. Organizing the data in a way so that you can plan your analysis accordingly. For example, if I selected 100 nozzles, specific data, 100 pieces, organizing them. I will take maybe inspections in two

stages. In the first stage, I will take three samples. In the second stage, I will take two samples. First, I will take two or three samples to be tested; if those are correct, then it is okay. If those are not correct, if one of them is not good, then we will take another two samples. This is the way we organize the data.

Then presentation of the data. This is very important. We will talk about graphical and tabular forms of data presentation. Whatever tables you see, when you have a population, the tables are drawn with fields, rows, and columns; that is a tabular presentation of the data. Graphical presentation is when you try to draw between two columns or between certain rows, etc., when you try to present the data in a graphical form; that is the graphical presentation of the data. Analysis of data, then interpretation of the data—this is all we are talking about in statistical investigation.

Steps/stages in Statistical Investigation

1. **Collection of data:**
   - The process of gathering information about the variable of interest.
   - Data serves as the input for statistical investigations.
   - Data can be obtained from primary sources (direct collection) or secondary sources (existing data).

2. **Organization of Data:** Organization of data includes three major steps.
   i. **Editing:** Checking and removing inconsistencies and irrelevant data.
   ii. **Classification**: Grouping the collected and edited data.
   iii. **Tabulation:** Arranging the classified data into tables.

Collection of the data. This is the process of gathering information about the variable of interest. Data serves as the input for statistical investigations. Data can be obtained from primary sources, that is direct collection, or secondary sources, that is existing data. Organization of data. Organization of data includes three major steps. Editing the data, that is checking or removing inconsistencies or irrelevant data, whether there is a missing data points, whether there is an extra data points which are not required. Classification. Grouping the collected and edited data.

Classification of data: When I am talking about the organization of data, this could also be in the form of an 80-20 rule, which is a Pareto rule. For 80% of the sample I am talking about—80% of the sample that is there—I will try to do all the testing there. Whatever testing is done in the 80%, whatever results come, I will try to validate those results in the next 20%. It is the 80-20 rule. Within the sample, if I am talking about 80-20, regarding the nozzle example, the total lot size was 100%.

The sample size is 5. 80% is 4 pieces, which are being tested. Whatever results come, the remaining 1 piece is 20%. We see whether the results from that 80% of the sample size—4 pieces—are valid for the other 20%, which is the 1 nozzle that is there. Then, tabulation of data is also organizing and arranging the classified data into tables.

Steps/stages in Statistical Investigation

3. **Presentation of data:**
   - Displaying data through charts, pictures, diagrams, histograms and graphs etc.
   - It aims to make the data easily understandable and visually appealing.

4. **Analyzing of data:**
   - Involves calculating statistical measures such as averages and measures of dispersion.
   - Includes hypothesis testing, regression analysis, and other mathematical operations.
   - Advanced analysis may require knowledge of higher-level mathematics

Presentation of data: displaying data through charts, pictures, diagrams, histograms, and graphs, etc. It aims to make data easily understandable and visually appealing. Analysis of data: this involves calculating statistical measures such as averages and measures of dispersion. This includes hypothesis testing, regression analysis, and other mathematical operations. Advanced analysis may require knowledge of higher-level mathematics.

Steps/stages in Statistical Investigation

5. **Interpretation of data:**

   - Interpreting statistical results to draw valid conclusions and inferences.
   - This stage requires significant skill and accuracy.
   - Correct interpretation leads to valid conclusions and informed decisions.
   - Incorrect interpretation can result in misleading conclusions, undermining the study's objectives.

Interpretation of data. Interpreting statistical results to draw valid conclusions and inferences. This stage requires significant skill and accuracy. Correct interpretation leads to valid conclusions and informed decisions. Incorrect interpretation can result in misleading conclusions and undermine study objectives. I'll talk about Type 1 and Type 2 errors. That is, incorrect interpretation could also result in something unfavorable.

Data Collection

Classification of data based on source:

1. **Primary:** data collected for the purpose of specific study. It can be obtained by:
   - Direct personal observation
   - Direct or indirect oral interviews
   - Administrating questionnaires
2. **Secondary:** refers to data collected earlier for some purpose other than the analysis currently being undertaken. It can be obtained from:
   - External Secondary data Sources (govt. and non govt. publications, gazettes)
   - Internal Secondary data Sources: the data generated within the organization in the process of routine business activities (e.g. production records, inventories).

   Qualitative data are non-numeric.

   Quantitative data are numeric.

Just to recall the concept of primary and secondary data, with this, I will close this part or the first part of this week. Classification of data based on source: primary data, that is, data collected for the purpose of a specific study. It can be obtained by direct personal observation, direct or indirect oral interviews. When I am talking about data for administrative purposes, that is, interviews or, I will say, experiments. Administering questionnaires. Secondary data refers to data collected earlier for some purpose other than the current analysis being undertaken.

It can be obtained from external secondary data sources, that is government and non-government publication, gadgets, etc., or internal secondary data sources, that is data generated within organization in the process of routine business activities, for example, production records, inventory, etc.

For example, if I need to do something about my nozzle manufacturing itself. In the nozzle manufacturing, I could see some other student at IIT Kanpur itself has done some experimentation on nozzles. What were the results that he has taken on the specific kind of the machine where he has done the testing? That is, within the organization I am talking about.

Right? Then, external data sources. There are handbooks on the nozzles for boilers. There are handbooks on the Nozzles itself, the kind of nozzle being used. What are the handbooks?

Those are kind of the gadgets or the handbooks which are there. Manufacturing handbooks. Those could be referred to those are external secondary data sources then data could be qualitative that is non-numeric data could be quantitative that is numeric. I am stressing this a lot because this. I will keep using in the forthcoming lectures. With this I'm closing this lecture.

I'll come to the next second lecture where I'll talk about the descriptive statistics process in detail, then I will talk about inferential statistics in the third part. And further, then I will talk about the hypothesis testing and design of experiments. Then I will close this course.

Thank you.