Statistics is the science of data. It's about collecting, analyzing, interpreting, and presenting data to make informed decisions. Think of it as a toolbox with many different tools (mathematical formulas and methods) that can be used to understand and make sense of the world around us.

There are two main types of statistics:

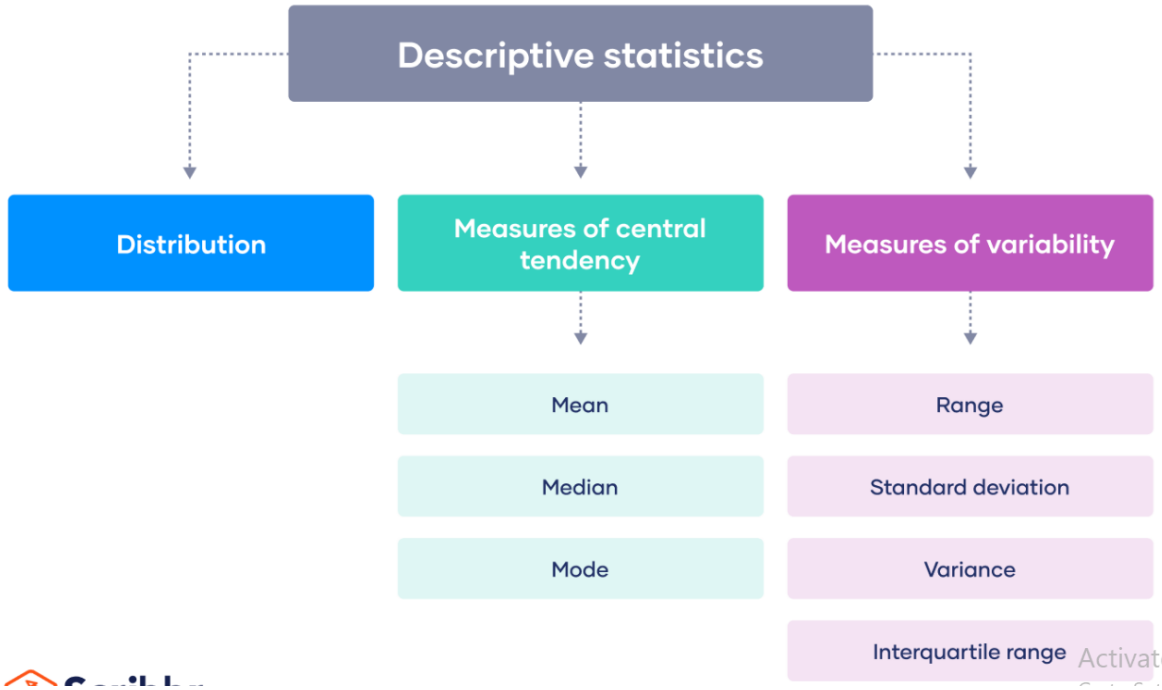
Descriptive statistics:

This involves summarizing a set of data using numbers like the mean, median, mode, standard deviation, and variance. It also includes presenting data in visual forms like graphs and charts. Descriptive statistics help us to get a basic understanding of the data, such as its central tendency, spread, and distribution.

Types of descriptive statistics

There are 3 main types of descriptive statistics:

- The distribution concerns the frequency of each value.
- The central tendency concerns the averages of the values.
- The variability or dispersion concerns how spread out the values are.
-



Example:

If we want to study the popularity of different leisure activities by gender. Let's distribute a survey and ask participants how many times they did each of the following in the past year:

- Go to a library
- Watch a movie at a theater
- Visit a national park

The data set is the collection of responses to the survey. Now we can use descriptive statistics to find out the overall frequency of each activity (distribution), the averages for each activity (central tendency), and the spread of responses for each activity (variability).

Inferential statistics

This involves using sample data to draw conclusions about a larger population. This is often done through hypothesis testing, where we make a guess about the population and then use statistical methods to see if our guess is correct. Inferential statistics helps us to make predictions about things we haven't actually measured, and it's essential for scientific research and decision-making in many fields.

Inferential statistics unlock the door to understanding populations beyond the data we directly observe. It's all about using hypothesis testing, a powerful tool that lets us draw conclusions about entire groups based on just a sample. Think of it like peering into a telescope to study distant galaxies - with a little statistical magic, we can see the bigger picture from a tiny glimpse.

The Steps of Hypothesis Testing:

1. **Defining the Rivals:**
    - Null Hypothesis (H_0): This is the "status quo" assumption, stating there's no significant difference between groups or variables. It's the starting point we challenge.
    - Alternative Hypothesis (H_1): This is the exciting underdog, proposing an actual difference exists. It's what we hope to uncover!
2. **Setting the Bar:**
    - Significance Level (α): This is the threshold for rejecting the null hypothesis. Imagine it as a tightrope - if the evidence against H_0 crosses this line, it falls! We typically set α at 5%, meaning a 5% chance of rejecting H_0 even when it's true (false positive).

3. **Drawing the Line in the Sand:**

    - Rejection Region: This zone on the statistical spectrum marks where the evidence is strong enough to ditch H_0. If our sample data lands here, it's like finding a smoking gun - H_0 is toast!

4. **The Showdown:**

    - Comparing Samples: We analyze the sample data using statistical tests like t-tests and ANOVAs, comparing means, medians, or rankings across groups.

5. **The Verdict:**

    - Rejecting H_0: If the sample data falls within the rejection region, we've got significant evidence to say "bye-bye, null hypothesis!" There's likely a real difference between the groups or variables.
    - Failing to Reject H_0: If the data stays outside the rejection zone, we can't definitively say there's a difference. But remember, this doesn't mean there isn't one - it just means our evidence wasn't strong enough to overturn the status quo.

Different tests are like specialized tools in your statistical toolbox. Here are some common ones:

- Parametric Tests (e.g., t-test, ANOVA): These assume the data follows a specific distribution (like normal) and are more powerful if that assumption holds.
- Non-Parametric Tests (e.g., Spearman's rank, Wilcoxon): These work with any data, regardless of distribution, but may be less powerful.

Beyond Means and Differences:

Inferential statistics go beyond just comparing groups. We can also use it to:

- Correlation Tests (e.g., Pearson's r, Spearman's r): These assess the strength and direction of relationships between variables.
- Regression Tests: These explore how changes in one variable (predictor) influence another (outcome).

Hypothesis testing is a powerful tool, but it's not a magic wand. Always consider the context, limitations, and potential for error when interpreting results. There’s a whole world of statistical tests out there, each with its own strengths and weaknesses. Choose the right one for your question and data!

In addition to these two main types, there are many other branches of statistics, such as:

- Probability: This is the study of chance and uncertainty. It's used in many areas of statistics, such as hypothesis testing and regression analysis.
- Bayesian statistics: This is a type of statistics that uses probability to update our beliefs about something as we learn more evidence. It's becoming increasingly popular in many fields, such as machine learning and artificial intelligence.

Population:

The term population refers to a collection of people or objects that share common observable characteristics. For example, a population could be all of the people who live in your city, all of the students enrolled in a particular university, or all of the people who are afflicted by a certain disease (e.g., all women diagnosed with breast cancer during the last five years). Generally, researchers are interested in particular characteristics of a population, not the characteristics that define the population but rather such attributes as height, weight, gender, age, heart rate, and systolic or diastolic blood pressure.

- Defined by a specific characteristic or condition of interest.
- Can be finite (e.g., all patients undergoing a surgical procedure) or infinite (e.g., all living organisms on Earth).
- Serves as the ultimate reference point for your research findings.

Sample:

In making inferences about populations, we use samples. A subset of the population selected for study. Ideally, it should be representative of the whole population. It's often more feasible, cost-effective, and faster to study a sample than the entire population. Samples can be biased, meaning they don't accurately reflect the whole population. This can lead to misleading conclusions.

- A subset of the population chosen for study.
- Ideally, a representative microcosm reflecting the characteristics of the entire population.
- Enables efficient data collection and analysis, paving the path to understanding the broader population.

Parameters and statistics are both terms used in statistics to describe data, but they differ in their scope:

Parameter:

- A fixed characteristic of an entire population.
- Represents the true value for a specific aspect of the population, like the mean, median, or standard deviation.
- Usually unknown, as measuring every individual in a population is often impractical or impossible.
- Examples: The true mean income of the US population, the actual proportion of defective products in a production line.

Statistic:

- An estimate of a parameter based on a sample of the population.
- Calculated from the data of the sample (e.g., sample mean, sample median, sample standard deviation).
- Used to infer the value of the corresponding parameter for the entire population.
- Examples: The average income of a random sample of 1000 US residents, the percentage of defective products found in a quality control test of 50 items.

Key differences:

Feature	Parameter	Statistic
Scope	Entire population	Sample of the population
Value	Fixed and unknown	Estimated from sample data
Purpose	Describes the true characteristic	Provides an estimate of the population characteristic
Example	Population mean	Sample mean

Relationship between parameters and statistics:

- **Sampling error:** Statistics are estimates, and they will always have some degree of error compared to the true parameter value. This error is called sampling error, and it decreases as the sample size increases.
- **Confidence intervals:** We can calculate confidence intervals around a statistic to express the range of values within which the true parameter is likely to fall. Wider confidence intervals indicate higher sampling error and less certainty about the estimate.
- **Statistical tests:** We can use statistical tests to assess the likelihood that a given sample statistic reflects a real difference in the population parameters. For example, comparing the average income of two different groups to see if they truly differ.

Examples:

1a. Think of a large garden with all the plants as the population. You want to know the average plant height in this garden (parameter). But measuring every plant is too tiring! So, you randomly pick a few plants and measure their heights (statistic). Based on these few plants, you get an estimate of the average height for the entire garden.

1b. Health researchers might estimate the prevalence of a disease in a community based on a random sample of residents (statistic) to understand the true burden of the disease in the population (parameter).

1c. Political pollsters might calculate the confidence interval around their estimated election outcome based on their sample size to express the uncertainty in their forecast.

1d. A manufacturing company might monitor the sample percentage of defective products from their production line (statistic) to infer the overall quality of their production (parameter).

Variables

We will look at a simple example first, then refer back to it to introduce some key concepts.

Example 1:- A group of 2266 Canadian newborns are classified by birthweight and whether the mother had to lift as part of her job. (yes or no). Interest is in whether babies born to mothers who had to lift at work are likely to be smaller. The data from the study might be held in a computer file looking something like this:

Table 1

Subject ID Number	Birthweight	lifting at work (0-no, 1=yes)
1	3300	0
2	4100	0
3	3970	1
4	3840	0
.....		
.....		
2265	4000	0
2266	3300	0

The purpose of this lecture is to introduce concepts and methods to help you produce and interpret summaries of these and similar data to answer questions like that posed above.

Subject and Variables

In example 1, and in most health studies, information (data) was collected on the characteristics of persons included in the study (study subjects). In example 1, these characteristics are birthweight and mother lifting at work. Other examples of characteristics are sex, environmental exposures, treatments, blood pressure, and experience of disease. We call the characteristics variables, because the value a variable takes (for example birthweight in grams, whether lifted at work) varies from person to person.

In Table 1, each row represents a subject, and each column a variable. This is the way data are usually kept by computers.

In this example, and often, study subjects are persons. However sometimes rather than measuring characteristics (variables) for persons we measure them for other “units”, for example households, towns, hospitals, or mosquitoes. In statistical terminology, we still call these units subjects. A value of each variable is needed for each subject.

Explanatory and response variables

A variable will usually be measured for one of two purposes: -

• It is an outcome of interest. In example 1 above, birthweight is the outcome. Outcome variables are also called response variables (dependent variables).

• It is a factor that influences (or might influence) the outcome. In the example, lifting is such a variable. These are often called explanatory variables (independent variables).

Types of Variables

Variables can also be classified by the types of values they can take. The main ones are:

Qualitative

- **Binary (dichotomous) variables**, where the values are two different categories; for example, was the subject dead or alive at the end of the study, whether or not a person is vaccinated, or being of low birthweight or not.
- **Categorical variables**, taking as values several different categories that are distinct from each other; for example, marital status (never married, married, widowed, divorced), or blood group.
- **Ordered categorical variables**, for which the different categories are ordered on some scale; for example, severity of disease (mild, moderate, severe) or a disability score.

Quantitative (numerical) variables, where some quantity is measured on a well-defined scale with units; for example, weight, blood pressure, number of episodes of asthma in a fixed period.

- **Continuous variables,** these take on an infinite range of values within a specific interval. For example, height can be 1.6 meters, 1.7 meters, 1.71 meters, and so on.

- **Discrete variables,** these take on a finite set of distinct values, often with no meaningful order between them. For instance, the number of children in a family could be 0, 1, 2, or 3.

Example:

A randomised controlled trial for a new drug for the treatment of hypertension. The response (outcome) is blood pressure or change in blood pressure. The principal explanatory variable is whether a subject is assigned to the new drug or control.

Exercise:

Question 1 What would be the response and explanatory variables in the following studies?

1a A questionnaire survey of whooping cough vaccination status in a sample of boys and girls, aiming to identify if socio-economic or ethnic status determined whether children were vaccinated.

1b A study of the occurrence of whooping cough in children, aiming to identify how effective vaccination was at preventing whooping cough.

Question 2 Write down the types of variables you identified in question 1.