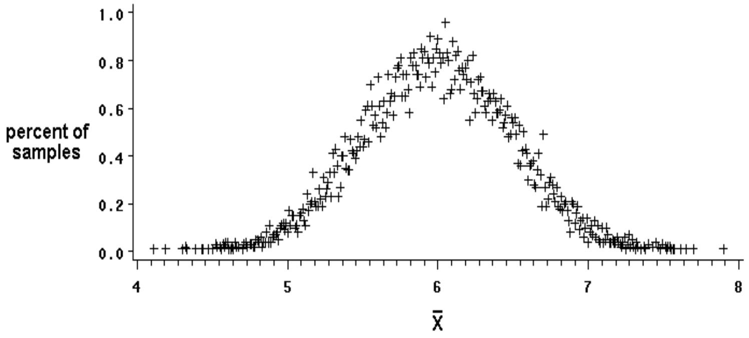
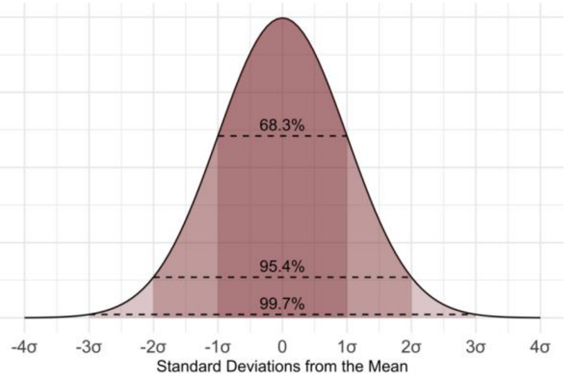
Linear Algebra and Probability





Prepared by: Sudip Pokhrel

Contents

1. **Introduction** [2 Hrs]
    1.1 Meaning and Definition of Statistics
    1.2 Application of Statistics in IT
    1.3 Variable and its types
    1.4 Limitation

Introduction to Statistics

Statistics:

- The word statistics has two meanings:
- In the most common usage – ***statistics*** refers to numerical facts
- The number that represents –

  a) annul income

  b) age

  c) the percentage of students who scored grade A

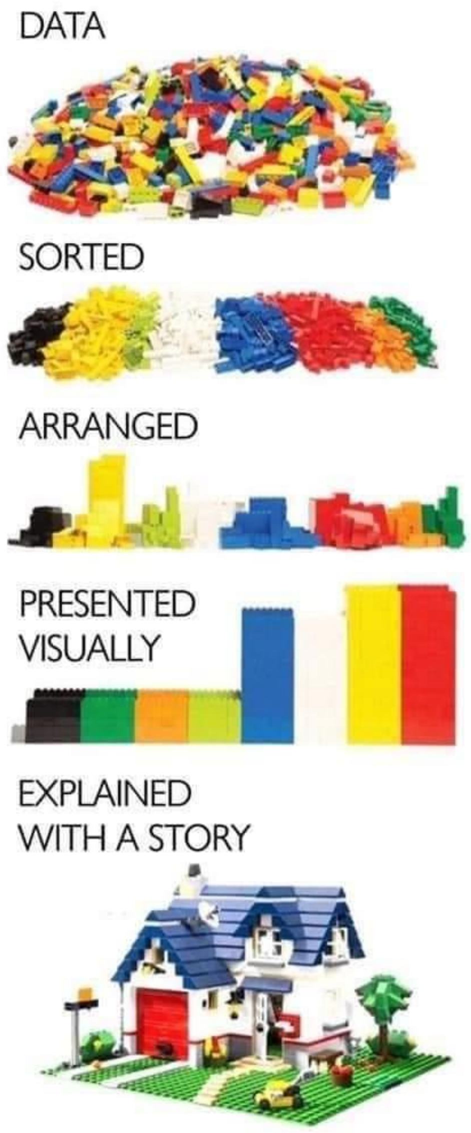
  d) the starting salary of a typical college graduate
- What will be other examples of ***statistics***? ……………..

The following examples present some statistics:

- Approximately 30% of Google's employees were female in July 2014 (*USA TODAY*, July 24, 2014).
- In 2013, author James Patterson earned $90 million from the sale of his books (*Forbes*, September 29, 2014).
- As per the CBS report, the hotel and restaurant, manufacturing and transportation sectors of Nepal will witness negative growth of 16.3 percent, 1.1 percent and 2.3 percent, respectively, in the current fiscal year (*The Himalayan Times,* April 30, 2020).

- The second meaning of ***statistics*** refers to the field or discipline of study.
- ***Statistics*** is the science of collecting, analyzing, presenting, and interpreting data, as well as making decisions based on such analyses.
- A comprehensive definition given by **Croxton and Cowden** is:

“Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data”



- **“Statistics is the science which deals with the collection, classification, and interpretation of numerical facts.”**
  — *Boddington*
- **“Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.”**
  — *U.S. National Research Council*
- **“Without data, you're just another person with an opinion.”**
  — *W. Edwards Deming*
- **“Statistics is the grammar of science.”**
  — *Karl Pearson*
- **“Statistics is the art of learning from data.”**
  — *David Hand*

- Statistical methods help us make ***scientific and intelligent decisions***.
- Decisions made by using statistical methods are called ***educated guesses.***
- Decisions made without using statistical (or scientific) methods are called ***pure guesses*** and, hence, may prove to be unreliable.
- For example: …….

Characteristics of Statistics

1. Deals with Numerical Data

- Statistics focuses primarily on **quantitative data** (numbers).
- Qualitative facts (like gender, caste, or satisfaction) must be **converted into numerical form** (e.g., coding) to be analyzed statistically.

2. Based on Aggregates, Not Individuals

- Statistics considers **groups or sets of data**, not isolated observations.
- Example: The average salary of a company is meaningful, but the salary of one employee is not statistically significant on its own.

3. Collected for a Purpose

- Data must be collected with a **specific objective** in mind.
- Purpose-driven data collection ensures the relevance and applicability of the results.

4. Subject to Variation

- Statistical data can vary due to **randomness**, **sampling error**, or **natural diversity**.
- Understanding variability is key to making reliable inferences.

5. Affected by Multiple Factors

- Statistical outcomes are often influenced by several **independent variables**.
- For example, crop yield depends on rainfall, fertilizer, soil type, etc.

6. Facilitates Comparison

- Statistics enables **comparison across time, groups, or regions** by using measures like averages, ratios, percentages, and growth rates.

7. Can Be Misleading If Misused

- While statistics aims to present truth, it can be **distorted** if the data is incorrect, biased, or presented unethically.

8. Scientific and Systematic

- Statistical processes follow a **structured approach** — data collection, classification, tabulation, analysis, and interpretation.

Scope/Importance of Statistics

The scope of statistics refers to the wide range of disciplines and real-world areas where statistical methods are applied.
Statistics is not confined to mathematics alone; it plays a crucial role in various social, economic, scientific, industrial, and technological fields.

Accounting: Generally, the number of individual accounts receivable is large and time-consuming to check their validity. Based on sample data, auditors make conclusions as to whether the accounts receivable amount shown on the client's balance is acceptable or not.

Daily Life: We use statistics to interpret weather forecasts, sports results, opinion polls, and personal finance data. Enables rational thinking and data-informed decisions.

Finance: Financial analysis uses a variety of statistical information and methods to guide its investment recommendations.

Economics: Economists use a variety of statistical information and methods in making forecasting, planning and formulating economic policies, price index numbers, unemployment rates, manufacturing capacity utilization, human development indicator indices, and quality control charts etc.

Research and Scientific Advancement: Statistics is the backbone of scientific research. It enables researchers to collect data systematically, test hypotheses, and validate findings. Example: Clinical trials in medicine use statistical analysis to determine drug effectiveness.

Business and Management: Used for market analysis, sales forecasting, quality control, and decision-making. Helps businesses understand customer preferences, optimize operations, and reduce risks.

Education: Educational institutions use statistics to assess student performance, curriculum effectiveness, and enrollment trends. Researchers use statistical tools for educational assessments and surveys.

Social Sciences: Statistics helps analyze social issues like poverty, literacy, health, and migration. Supports the formulation of evidence-based social policies.

Information Technology: Statistics plays a vital role in the IT sector, supporting data-driven decision-making, automation, system optimization, and innovation. The key areas include:

1. Data Analytics & Business Intelligence

- Statistical methods are used to analyze **structured and unstructured data**.
- Helps identify **patterns**, **trends**, and **customer behavior** in massive datasets.
- Tools: R, Python, SQL, Power BI.

2. Machine Learning & Artificial Intelligence

- Core algorithms in **supervised**, **unsupervised**, and **reinforcement learning** rely heavily on statistical theory.
- Example: Regression, classification, clustering, and probabilistic models.

3. Software Quality Assurance (SQA)

- Statistics is used for **defect analysis**, **reliability estimation**, and **testing effectiveness**.
- Techniques like **control charts**, **regression**, and **sampling** improve software quality.

4. Cybersecurity & Anomaly Detection

- Statistical models help detect **unusual behavior**, fraud, and security breaches.
- Involves **time series analysis**, **Bayesian models**, and **probabilistic intrusion detection**.

5. Database Management & Optimization

- Statistics aid in **query optimization**, **indexing**, and **data mining**.
- DBMS systems (like Oracle and SQL Server) use statistics to improve performance.

6. Cloud Computing & Resource Allocation

- Statistical modeling ensures **load balancing**, **resource prediction**, and **failure estimation** in cloud environments.

7. User Experience (UX) & A/B Testing

- Uses **hypothesis testing**, **confidence intervals**, and **probability** to evaluate design choices, layouts, and functionality.

8. Network Performance & Simulation

- Queuing theory and stochastic processes model **traffic flow**, **latency**, and **packet loss** in IT networks.

Limitations of Statistics

While statistics is a powerful tool for data analysis and decision-making, it has several **inherent limitations**. Awareness of these limitations helps in applying statistical methods **appropriately and ethically**.

1. Statistics Cannot Study Qualitative Phenomena Directly

- Non-measurable attributes like honesty, intelligence, leadership, or love **cannot be directly analyzed** using statistics.
- These need to be quantified or **subjectively coded**, which may reduce accuracy.

2. Statistics Deals with Aggregates, Not Individuals

- Statistics is designed to work with **groups or large datasets**.
- It **cannot provide meaningful insights for individual cases**.
- Example: Average income data says nothing about one person’s situation.

3. Results Can Be Misleading

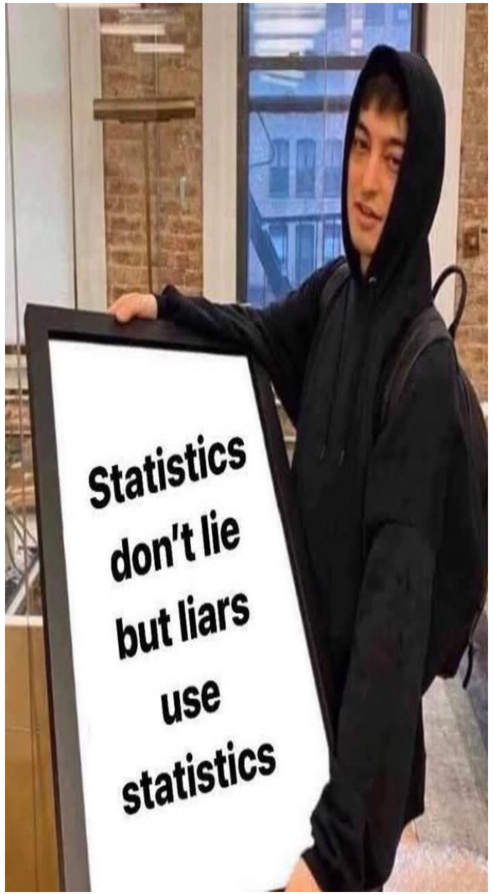
- Incorrect data collection, inappropriate methods, or biased interpretation can produce **misleading conclusions**.
- Example: A company claims that the **average salary** of its employees is ₹80,000 per month to attract top talent. However, the **majority of employees** earn between ₹30,000 and ₹50,000, while a few top executives earn over ₹5 lakhs per month.
- The **mean (average)** salary is inflated by a few high earners.

4. Requires Skilled Interpretation

- Statistical results are often **open to interpretation**, and conclusions require **expert judgment**.
- Inexperienced users may misapply techniques or misread findings.

5. Does Not Establish Causation

- Statistics show **correlation**, not **cause-and-effect**.
- Example: A correlation between exercise and academic performance doesn't prove that exercise causes better grades.



6. Sensitive to Errors

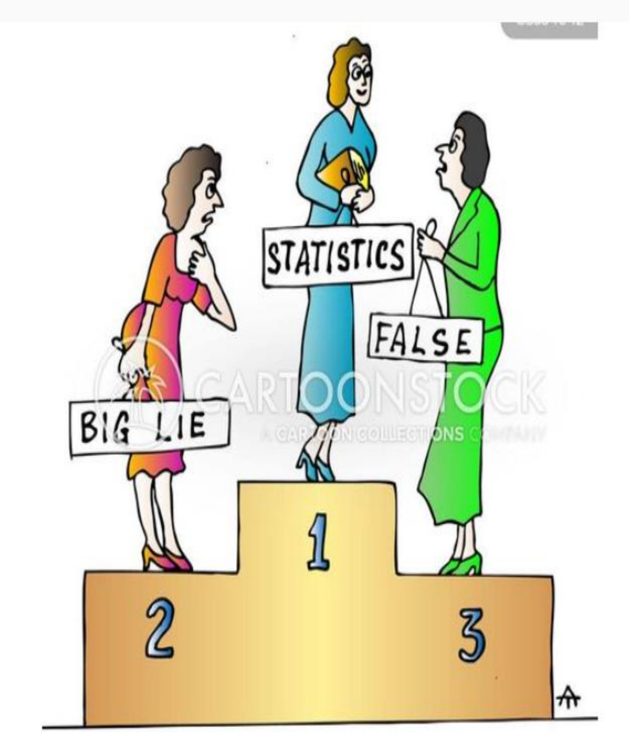
- Small **errors in data** collection, classification, or computation can lead to **incorrect results**.
- Statistical tools assume data quality and completeness.

7. Can Be Misused Deliberately

- Statistics can be **manipulated** to support specific agendas or misleading narratives.
- “Lies, damned lies, and statistics” highlights this risk.

8. Depends on Assumptions

- Many statistical methods rely on assumptions like **normality**, **independence**, or **linearity**.
- If these assumptions are violated, results may be **invalid or misleading**.



Distrust of Statistics

Despite its usefulness, **statistics is often met with skepticism or distrust**. This is primarily because it can be **manipulated**, **misunderstood**, or **misused**, either intentionally or unintentionally.

1. Can Be Misused

- Statistics may be **deliberately used to mislead** others. Selective use of data, biased sampling, or intentional omission of context can make statistics support almost any argument.
- **Example:** Showing a percentage increase without mentioning the actual base values.

2. Difficult to Understand

- Many people lack the statistical literacy needed to interpret results correctly. Terms like “confidence interval” or “standard error” may **confuse readers**, leading to mistrust.

3. Different Sources, Different Results

- On the same issue, different reports may present **contradictory statistics**, which makes people **doubt their reliability**.
- **Example:** Two surveys may show different unemployment rates due to differences in definitions or methodology.

4. Bias in Data Collection

- If the data is collected from a **non-representative sample** or in a biased manner, the result will be flawed, even if calculations are correct.
- **Example:** A satisfaction survey done only among loyal customers will not reflect true public opinion.

Branches of Statistics

- **Descriptive statistics** are methods for organizing and summarizing data.
- For example, tables or graphs are used to organize data, and descriptive values such as the average score are used to summarize data.
- A descriptive value for a population is called a **parameter** and a descriptive value for a sample is called a **statistic**.

Parameter vs Statistic

Aspect	Parameter	Statistic
Definition	A **parameter** is a **numerical value** that describes a **characteristic of a population**.	A **statistic** is a **numerical value** that describes a **characteristic of a sample**.
Data Source	Derived from the **entire population**.	Derived from a **subset (sample)** of the population.
Nature	**Fixed and constant** (if the population doesn't change).	**Variable**, because it changes with different samples.
Usage	Represents the **true value** in the population (often unknown).	Used to **estimate or infer** the population parameter.
Symbol Notation	Common symbols: **μ** (mean), **σ** (standard deviation), **P** (proportion)	Common symbols: **$\bar{x}$** (mean), **s** (standard deviation), **$\hat{p}$** (proportion)
Example	Average age of **all employees** in a company (μ).	Average age of **a sample of employees** from that company ($\bar{x}$).
Application	Used in **theoretical models** and **population-level decision-making**.	Used in **sample analysis**, **surveys**, and **research studies**.

- **Inferential statistics** are methods for using sample data to make general conclusions (inferences) about populations.
- Because a sample is typically only a part of the whole population, sample data provide only limited information about the population. As a result, sample statistics are generally imperfect representatives of the corresponding population parameters.

Aspect	Descriptive Statistics	Inferential Statistics
1. Definition	Descriptive statistics deals with **summarizing**, **organizing**, and **presenting** data in a meaningful way.	Inferential statistics involves **drawing conclusions**, **making predictions**, or **generalizing** about a population based on a sample.
2. Purpose	To **describe the characteristics** of a given dataset without making generalizations beyond it.	To **infer**, **estimate**, or **test hypotheses** about a larger population from which the sample is drawn.
3. Data Type Used	Uses **entire data** from a population or a sample; focuses on **what is observed**.	Uses **sample data** to make inferences about a **larger population** that may not be fully observed.
4. Techniques Involved	- Measures of central tendency (mean, median, mode) - Measures of dispersion (range, variance, standard deviation) - Graphs and tables (bar charts, pie charts, histograms)	- Hypothesis testing (e.g., t-test, chi-square) - Confidence intervals - Regression analysis - ANOVA, sampling distributions
5. Output	Provides **facts, summaries**, and **visualizations** about the dataset (e.g., average, spread, shape).	Produces **conclusions**, **probability-based estimates**, and **decision rules** about populations.
6. Application Area	Useful in **reporting**, **data visualization**, **exploratory data analysis**, and **business summaries**.	Common in **scientific research**, **market studies**, **clinical trials**, and **policy decision-making**.
7. Example	- Finding the **average test score** of students in a class. - Creating a **bar graph** of employee ages.	- Predicting the **election outcome** based on polling data. - Testing if a new medicine is more effective than the existing one.

Things to remember....

- A descriptive study may be performed either on a **sample or on a population**. Only when an inference is made about the population, based on information obtained from the sample, does the study become inferential.
- Descriptive statistics and inferential statistics are interrelated. You must almost always use techniques of descriptive statistics to organize and summarize the information obtained from a sample before carrying out an inferential analysis.

Basic Terms

Population or target population: The collection of all elements/members whose characteristics are being studied.

For example:……………….

Sample: A portion/fraction of the population of interest.

For example: ……………

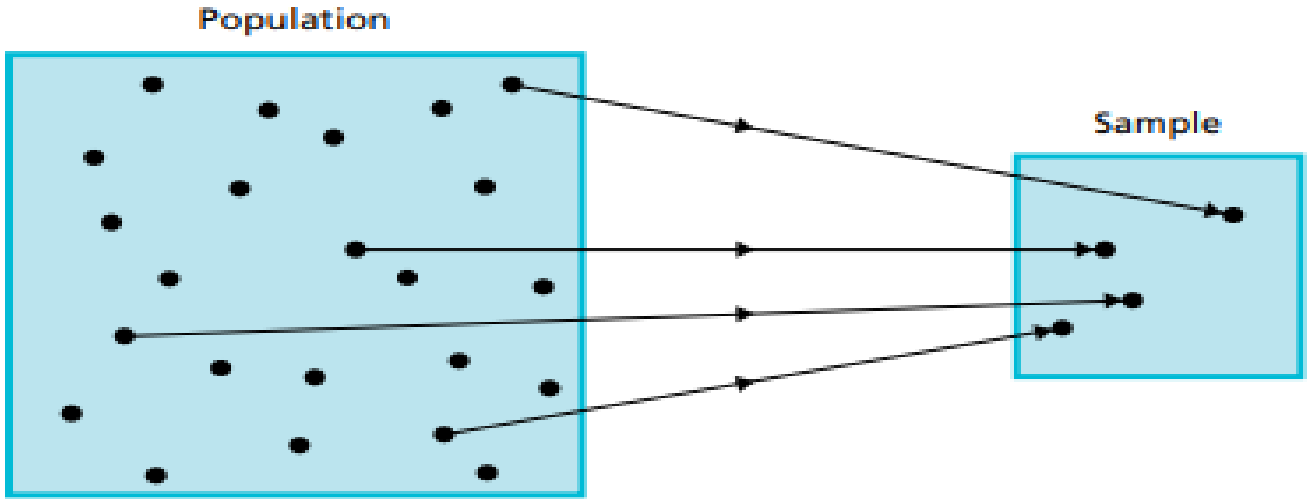
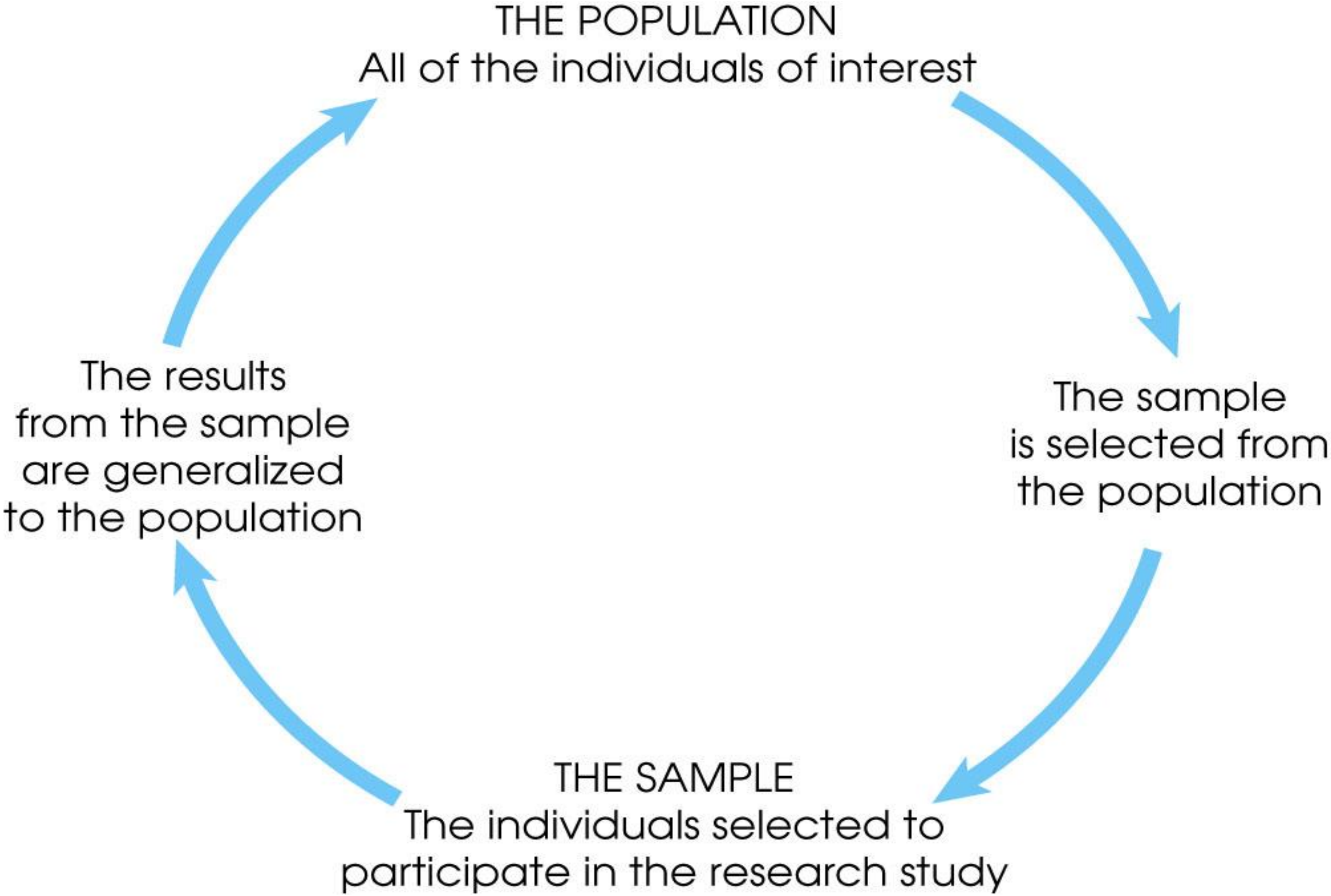


Fig1. the relation between population and sample

Goal of Sample:

Usually populations are so large that a researcher cannot examine the entire group. Therefore, a **sample** is selected to represent the population in a research study. The goal is to use the results obtained from the sample to help answer questions about the population.



THE POPULATION
All of the individuals of interest

The sample
is selected from
the population

THE SAMPLE
The individuals selected to
participate in the research study

The results
from the sample
are generalized
to the population

Basic terms continued…..

Survey:

A survey is a research method used for collecting data from *a predefined group* of respondents to gain information and insights into various topics of interest.

Census:

procedure of systematically calculating, acquiring and recording information about the members of a given population.

Sample Survey:

procedure of systematically calculating, acquiring and recording information from only a portion of a population of interest.

- **Variable**

- A variable is a characteristic under study that assumes different values for different elements.

- A variable is often denoted by letters x, y, or z

- The value of a variable for an element is called an ***observation or measurement***.

- **Data**

- A collection of information/observations

- The goal of statistics is to help researchers organize and interpret the data.

Types of Variables

- Some variables (such as the height of person, price of groceries) can be measured numerically, whereas others (such as occupation, income sources) cannot.
- Variables are classified into two types:

  a) Quantitative Variable

  b) Qualitative Variable

i) Quantitative Variable

- A variable that can be measured **numerically** is called a quantitative variable.
- The data collected on a quantitative variable are called ***quantitative data***.
- Example: Number of workers: 23, 24, 25, 15, 19, 18
- Other examples:

- Annual Gross sale
- No. of accidents
- Weight of a laptop
- Temperature
- No. of gadgets owned

- As you can see from the above examples that certain quantitative variable can assume may **be countable** or **noncountable**
- Quantitative variables may be classified into two categories
    a) **Discrete Variable**
    b) **Continuous Variable**

A) Discrete Variable

- Variable whose values are countable.
- In other words, a discrete variable can assume only certain values with no intermediate values.
- For example:

- No. of accidents
- The no. of daily admissions in a general hospitals
- The no. of people visit bank in on any day
- The no. of books in a library

B) Continuous Variable

- A variable that can assume any numerical value over a certain interval or intervals is called a continuous variable.
- Example:

- Price of book: USD105.6
- Annual salary
- Body temperature
- Expenditure on food on any day
- The time it takes to complete a certain task

ii) Qualitative or Categorical Variable

- A variable that cannot assume a numerical value but can be classified into two or more nonnumeric categories is called a qualitative or categorical variable.
- The data collected on such a variable are called qualitative data.
- Examples:

- Gender of a person
- A person’s blood type
- Occupation
- Modes of transportation

Figure summarizes the different types of variables.

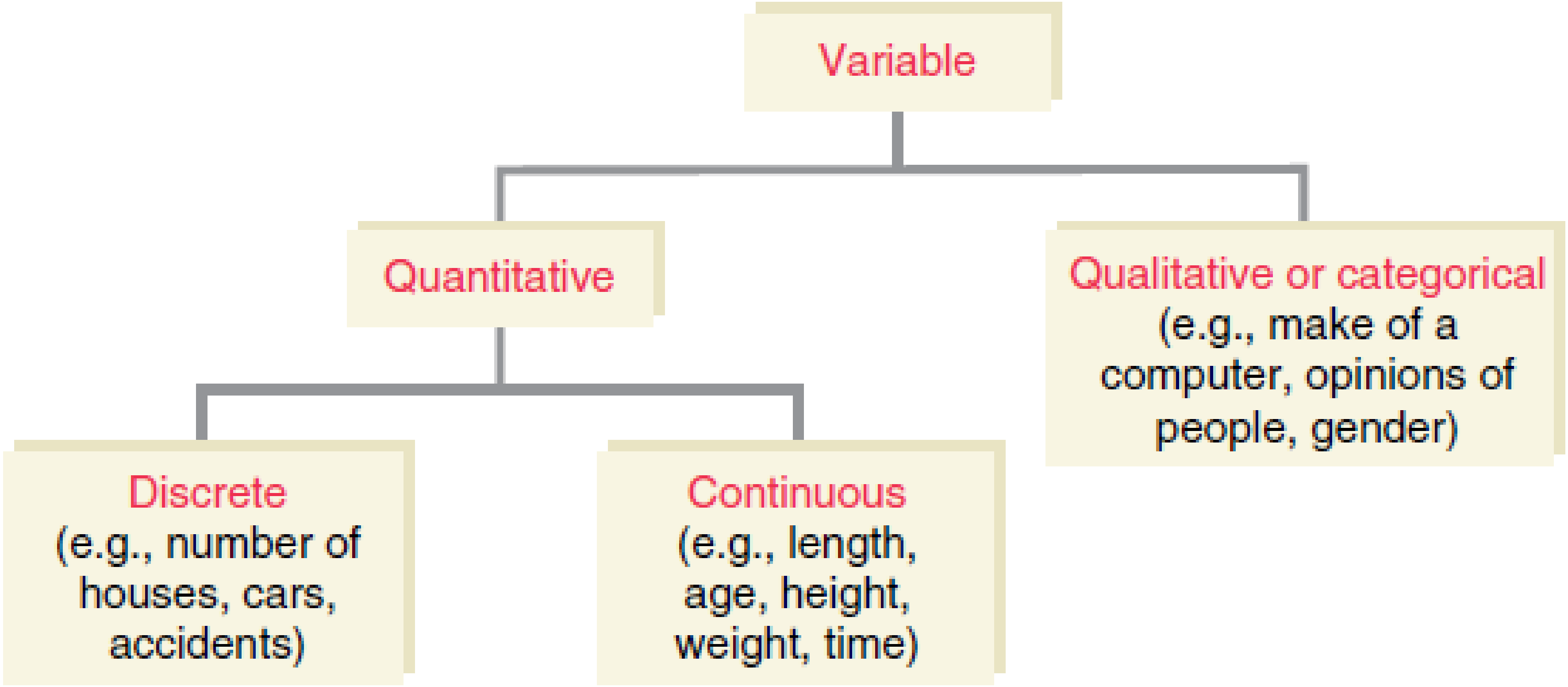

Variable
Quantitative
Qualitative or categorical
(e.g., make of a computer, opinions of people, gender)
Discrete
(e.g., number of houses, cars, accidents)
Continuous
(e.g., length, age, height, weight, time)

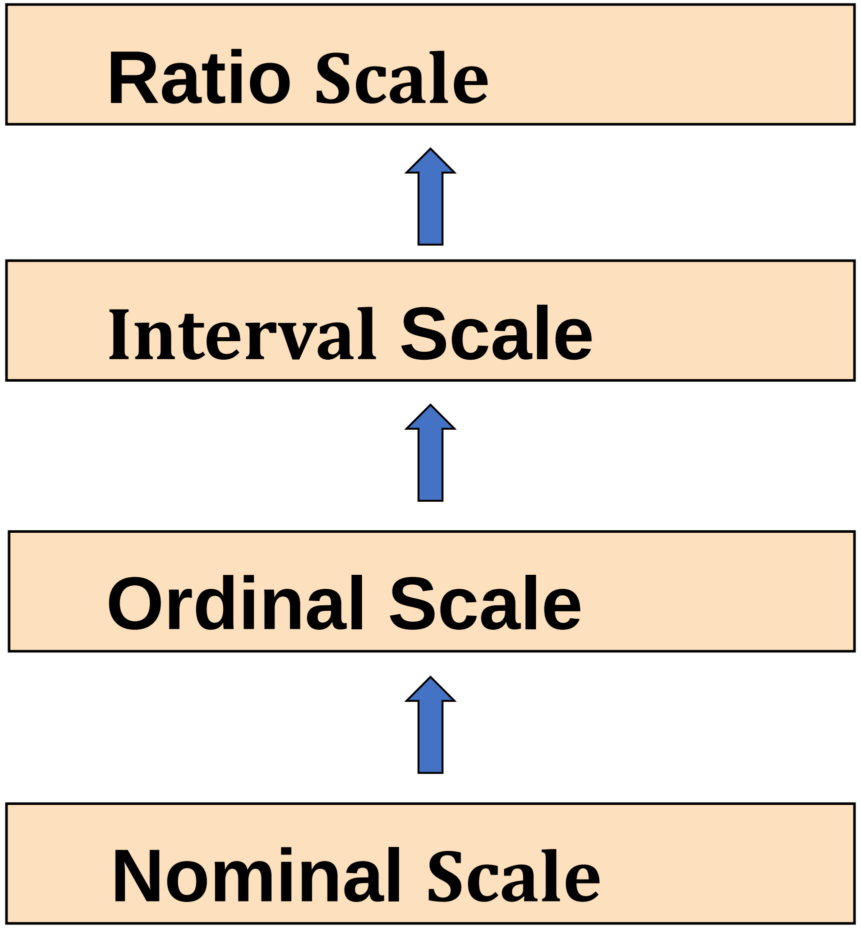
Figure Types of variables.

Measuring Variables

- When carrying out any kind of data collection or analysis, it’s essential to understand the nature of the data you’re dealing with
- Within your dataset, you’ll have different variables—and these variables can be recorded to varying degrees of precision. This is what’s known as the **level of measurement.**

- When gathering data, you collect different types of information, depending on what you hope to investigate or find out.
- For example, if you wanted to analyze the spending habits of people living in Kathmandu, you might send out a survey to 500 people asking questions about their income, their exact location, their age, and how much they spend on various products and services. These are your variables: data that can be measured and recorded, and whose values will differ from one individual to the next.
- Level of measurement is important as it determines the type of statistical analysis you can carry out. As a result, it affects both the nature and the depth of insights you're able to glean from your data.
- Certain statistical tests can only be performed where more precise levels of measurement have been used, so it's essential to plan in advance how you'll gather and measure your data.

Four Types of Measurement Scales



Differences between measurements, true zero exists

Ratio Scale

Highest Level

(Strongest forms of measurement)

Differences between measurements but no true zero

Interval Scale

Higher Levels

Ordered Categories (rankings, order, or scaling)

Ordinal Scale

Categories (no ordering or direction)

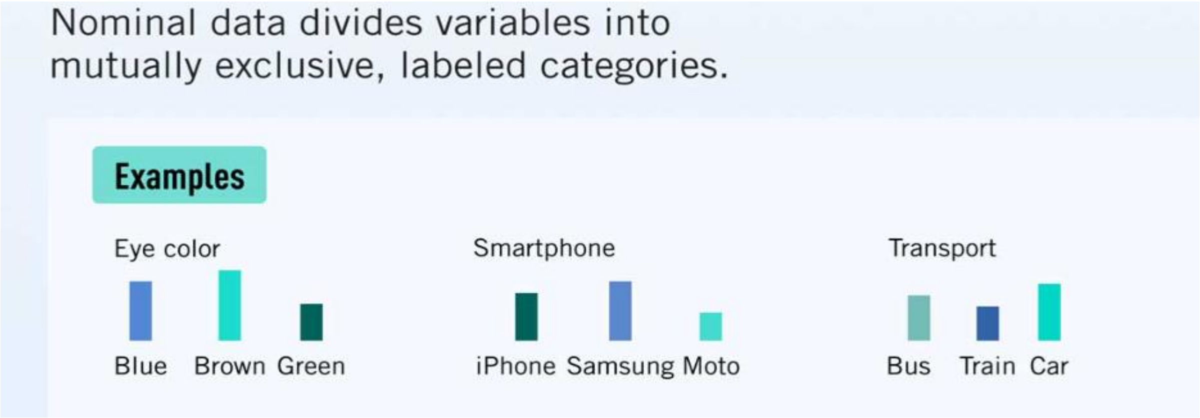
Nominal Scale

Lowest Level

(Weakest form of measurement)

Nominal Scale:

- The nominal level is the first level of measurement and the simplest. It classifies and labels variables qualitatively. In other words, it divides them into named groups without any quantitative meaning.
- It's important to note that, even where numbers are used to label different categories, these numbers don't have any numerical value.
- The **nominal scale** simply categorizes variables according to qualitative labels (or names). These labels and groupings don't have any order or hierarchy to them, nor do they convey any numerical value.



Nominal data divides variables into mutually exclusive, labeled categories.

Some examples of nominal data include:

- Eye color (e.g. blue, brown, green)
- Nationality (e.g. German, Cameroonian, Lebanese)
- Personality type (e.g. introvert, extrovert, ambivert)
- Employment status (e.g. unemployed, part-time, retired)
- Political party voted for in the last election (e.g. party X, party Y, party Z)
- Type of smartphone owned (e.g. iPhone, Samsung, Google Pixel)

How to analyze nominal data?

- **Descriptive statistics for nominal data**

Descriptive statistics describe or summarize the characteristics of the dataset. Two useful descriptive statistics for nominal data are:

- Frequency distribution

- Mode

- **Statistical tests for analyzing nominal data**

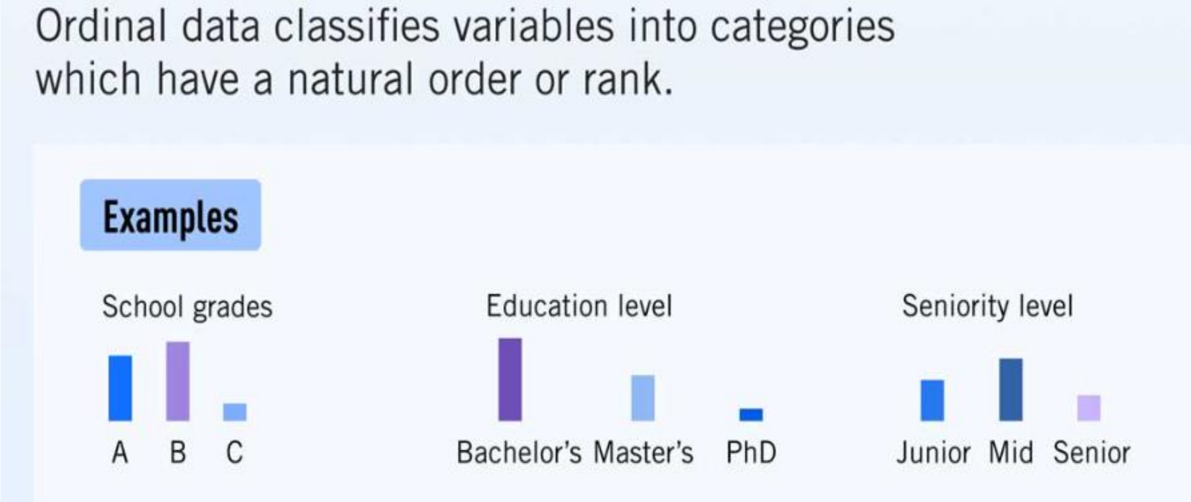
can analyze nominal data using certain non-parametric statistical tests, namely:

- Chi-square test of goodness of fit

- Chi-square test of independence

Ordinal Scale

- The ordinal scale also categorizes variables into labeled groups, and these categories have an **order or hierarchy** to them.
- For example, you could measure the variable “income” on an ordinal scale as follows: low income, medium income, and high income.
- Another example could be level of education, classified as follows: high school, master’s degree, doctorate. These are still qualitative labels (as with the nominal scale), but they follow a hierarchical order.



Ordinal data classifies variables into categories which have a natural order or rank.

Some examples of ordinal data include:

- Academic grades (A, B, C, and so on)
- Happiness on a scale of 1-10 (this is what's known as a **Likert scale**)
- Satisfaction (extremely satisfied, quite satisfied, slightly dissatisfied, extremely dissatisfied)
- Income (high, medium, or low). **Note that income is not an ordinal variable by default**; it depends on how you choose to measure it.
- Seniority level at work (junior, mid-level, senior)

How to analyze ordinal data?

- **Descriptive statistics for ordinal data**

The following descriptive statistics can be used to summarize ordinal data:

- Frequency distribution
- The mode and/or the median
- The range

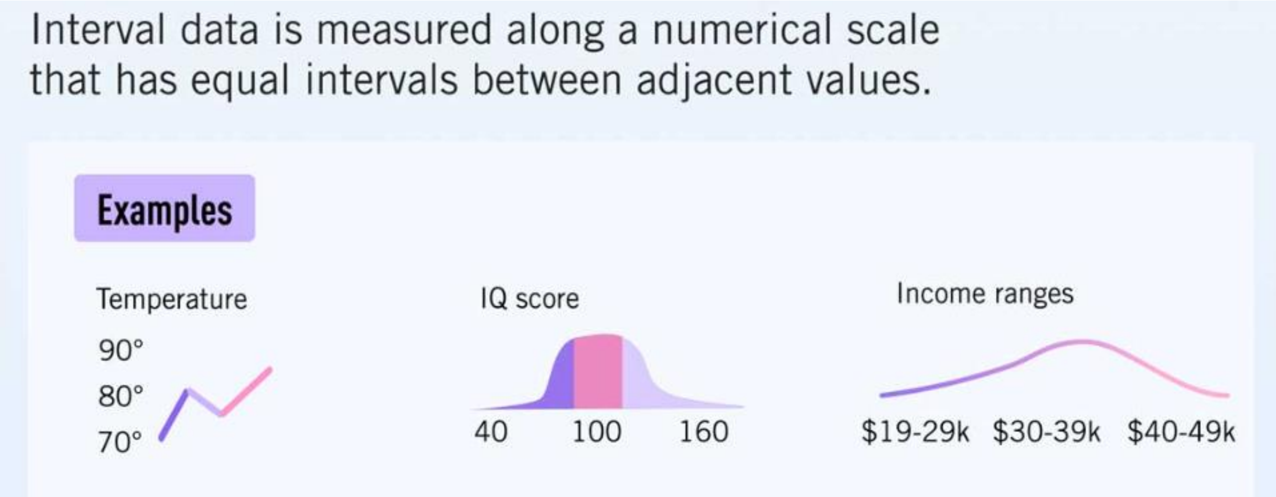
- **Statistical tests for analyzing ordinal data**

Can analyze ordinal data using certain non-parametric statistical tests, namely:

- Mood’s median test
- Mann-Whitney U-test
- Wilcoxon matched-pairs signed-rank test
- Kruskal-Wallis H test
- Spearman’s rho (rank correlation coefficient)

Interval Scale

- Interval scale is a quantitative measurement scale where there is order, the difference between the two variables is meaningful and equal, and the presence of zero is arbitrary.
- e.g. temperature measurements in Fahrenheit and Celsius, or the pH scale. Interval data always lack what's non as a 'true zero.'
- In short, this means that interval data can contain negative values and that a measurement of 'zero' can represent a quantifiable measure of something.



Interval data is measured along a numerical scale that has equal intervals between adjacent values.

Some examples of interval data

- Temperature in Fahrenheit or Celsius (-20, -10, 0, +10, +20, etc.)
- Times of the day (1pm, 2pm, 3pm, 4pm, etc.)
- Income level on a continuous scale ($10K, $20K, $30K, $40K, and so on)
- IQ scores (100, 110, 120, 130, 140, etc.)
- pH (pH of 2, pH of 4, pH of 6, pH of 8, pH of 10, etc.)
- SAT scores (900, 950, 1000, 1050, 1100 etc.)
- Credit ratings (20, 40, 60, 80, 100)
- Dates (1740, 1840, 1940, 2040, 2140, etc.)

How to analyze interval data?

- **Descriptive statistics for interval data**

- Frequency distribution
- Central tendency: Mode, median, and mean
- Variability: Range, standard deviation, and variance

- **Statistical tests for analyzing interval data**

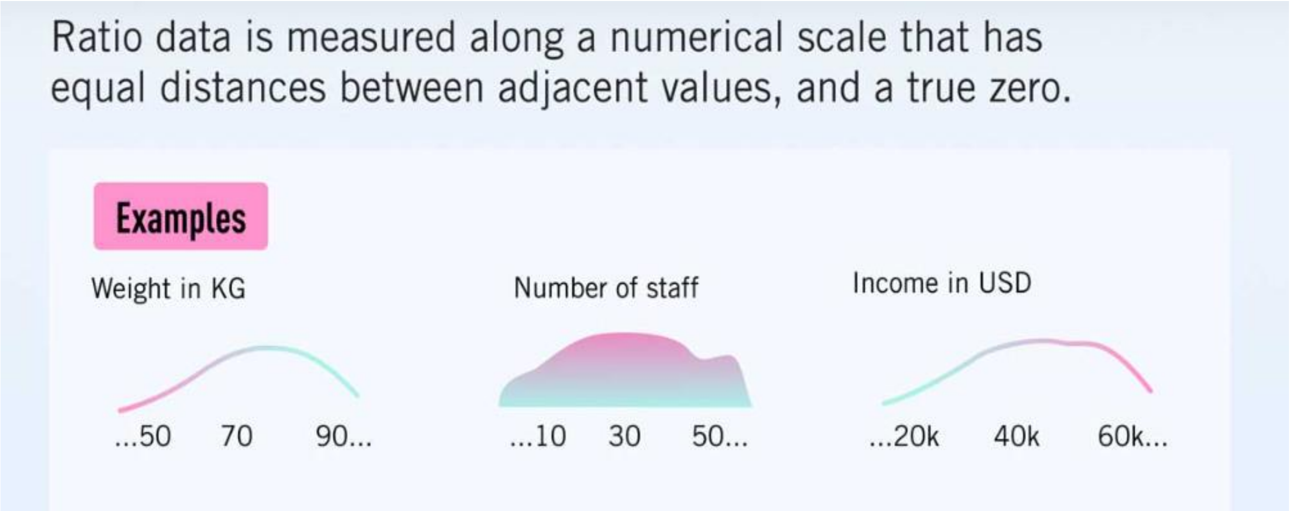
To analyze quantitative datasets, it is best to use parametric tests rather than non-parametric tests (more commonly used for qualitative data, i.e. nominal and ordinal data).

To highlight, some parametric tests you can use to explore interval data are:

- T-test
- Analysis of variance (ANOVA)
- Pearson correlation coefficient
- Linear regression

Ratio Scale

- The ratio scale of measurement is similar to the interval scale in that it also represents the quantity and has equality of units.
- However, this scale also has an absolute zero (no numbers exist below zero).
- Very often, physical measures will represent ratio data (for example, height and weight).
- A good example of ratio data is weight in kilograms. If something weighs zero kilograms, it truly weighs nothing—compared to temperature (interval data), where a value of zero degrees doesn’t mean there is “no temperature,” it simply means it’s extremely cold!



Ratio data is measured along a numerical scale that has equal distances between adjacent values, and a true zero.

Examples of ratio data

Ratio variables can be discrete (i.e. expressed in finite, countable units) or continuous (potentially taking on infinite values). Here are some examples of ratio data:

- Weight in grams (continuous)
- Number of employees at a company (discrete)
- Speed in miles per hour (continuous)
- Length in centimeters (continuous)
- Age in years (continuous)
- Income in dollars (continuous)
- Sales made in one month (discrete)

How to analyze Ratio data?

- **Descriptive statistics for ratio data**

- Frequency distribution
- Central tendency: Mode, median, and mean
- Variability: Range, standard deviation, variance, and coefficient of variation

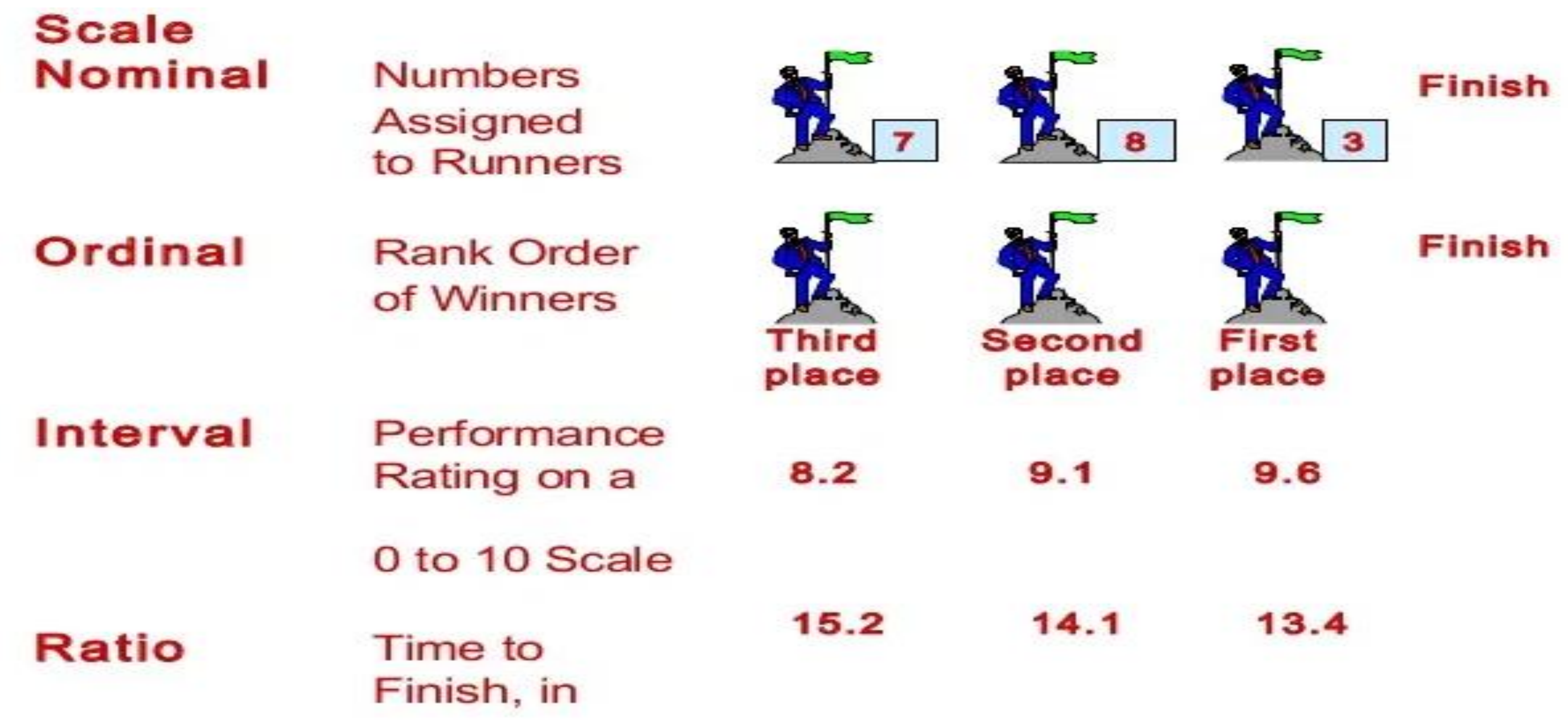
- **Statistical tests for analyzing ratio data**

To analyze quantitative datasets, it is best to use parametric tests rather than non-parametric tests (more commonly used for qualitative data, i.e. nominal and ordinal data).

To highlight, some parametric tests you can use to explore interval data are:

- T-test
- Analysis of variance (ANOVA)
- Pearson correlation coefficient
- Linear regression

Example: Scale of measurement



Scale					
Nominal	Numbers Assigned to Runners	7	8	3	Finish
Ordinal	Rank Order of Winners	Third place	Second place	First place	Finish
Interval	Performance Rating on a 0 to 10 Scale	8.2	9.1	9.6	
Ratio	Time to Finish, in	15.2	14.1	13.4	

Thank You