



STATISTICS FOR DATA SCIENCE

A Complete Guide for Aspiring Data Scientists



Introduction to Statistics

What is Statistics?

Definition: Statistics is the science of collecting, organizing, summarizing, analyzing, and interpreting data to draw meaningful insights and conclusions. It's like becoming a detective with numbers—you gather clues (data), organize them, and reveal the hidden stories they tell.

Simple Example: Imagine you're a teacher. Statistics helps you understand not just individual students' scores, but the overall class performance—who's doing well, who needs help, and what's the average level.

Why Statistics is Important in Data Science

Importance & Use:

Statistics is the foundation of data science. Without statistics, you're just looking at random numbers. With statistics, you can:

- **Find patterns** in large amounts of data
- **Make predictions** about future trends
- **Make confident decisions** based on evidence, not guesses
- **Validate your findings** to ensure they're not just random luck
- **Build reliable machine learning models** that actually work

What it tells about data:

Statistics reveals whether your data makes sense, if there are patterns, if something is unusual, and what the data really represents. It separates the **signal** (real patterns) from the **noise** (random variations).

How it helps in decision-making:

A business might see that their average sales went up 10%, but statistics shows you whether that's a real improvement or just random chance. Only then can you confidently say "let's expand this strategy."



Section 1: Types of Data

Data comes in different forms. Understanding these forms is crucial because different data requires different statistical techniques.

1 Qualitative Data (Categorical Data)

Definition: Qualitative data describes qualities or categories—things you can't measure with numbers directly. These are words, labels, or categories.

What it tells about data: It tells you *what kind* of things you have, not how many or how much.

Nominal Data

Definition: Nominal data are categories with no particular order or ranking. Each category is equally important, and you can't rank them.

Examples:

- Gender (Male, Female, Other)
- Colors (Red, Blue, Green)
- Brand names (Apple, Samsung, Nokia)
- Country names (India, USA, Canada)

Why is this important & what is the use?

Nominal data helps you categorize things for organization and comparison. You can count how many people are in each category.

Real-World Application: In a marketing survey, nominal data helps identify customer demographics for targeted advertising campaigns.

What it tells about data:

It shows diversity and distribution. If you have 100 customers from 5 countries, nominal data tells you how many are from each country.

How it helps in decision-making:

A company might discover that 60% of their customers are from India and 40% from the USA. This informs where to focus marketing efforts.

Ordinal Data

Definition: Ordinal data are categories with a specific order or ranking. The order matters, but the distance between categories isn't equal.

Examples:

- Customer satisfaction (Poor, Fair, Good, Excellent)
- Education level (High School, Bachelor's, Master's, PhD)
- Survey ratings (1-star, 2-star, 3-star, 4-star, 5-star)
- Movie ratings (Flop, Average, Hit, Blockbuster)

Why is this important & what is the use?

Ordinal data lets you rank things and identify preferences or progression. It's more informative than nominal data.

Real-World Application: Customer feedback surveys using Likert scales (Strongly Disagree to Strongly Agree) help businesses measure customer satisfaction levels.

What it tells about data:

It shows hierarchy and preference. You know that "Excellent" is better than "Good," but you can't say "Excellent is twice as good as Fair."

How it helps in decision-making:

If most customers rate your product as "Good" but very few say "Excellent," you know where to improve. This guides product development decisions.

2 Quantitative Data (Numerical Data)

Definition: Quantitative data consists of actual numbers that you can measure and perform mathematical operations on.

Discrete Data

Definition: Discrete data consists of countable numbers. You can only have whole numbers; you can't have fractions.

Examples:

- Number of students in a class (20, 35, 50—not 20.5)
- Number of cars in a parking lot (15, 16, 17—not 15.7)
- Number of website visits per day (100, 250, 500)
- Number of customer complaints per week (5, 8, 12)

Why is this important & what is the use?

Discrete data is used when you're counting things that can't be divided.

Real-World Application: E-commerce websites track discrete data like number of orders, cart abandonment rates, and unique visitors to measure business performance.

What it tells about data:

It shows exact quantities. You know exactly how many items there are.

How it helps in decision-making:

If you know exactly how many products are defective in a batch of 1000, you can calculate the defect rate and decide if the batch meets quality standards.

Continuous Data

Definition: Continuous data consists of numbers that can have decimal values.

You can have infinite possible values within a range.

Examples:

- Height of students (165.3 cm, 170.8 cm, 175.2 cm)
- Weight (65.5 kg, 72.3 kg)
- Temperature (23.5°C, 28.1°C)
- Time taken to complete a task (45.2 minutes, 52.8 minutes)
- Price of products (Rs. 299.99, Rs. 1,599.50)

Why is this important & what is the use?

Continuous data is used for measurements where precision matters and you can have any value within a range.

Real-World Application: Financial institutions use continuous data like stock prices, interest rates, and currency exchange rates for trading algorithms and risk assessment.

What it tells about data:

It shows precise measurements. You can describe data very accurately.

How it helps in decision-making:

A manufacturer knowing the exact weight distribution of products can identify if there are quality issues and adjust machines for better precision.

Data Types Summary

Type	Subtype	Key Feature	Real-World Example
Qualitative	Nominal	Categories without order	Gender, Colors, Countries
Qualitative	Ordinal	Categories with order	Ratings, Education Levels
Quantitative	Discrete	Countable whole numbers	Students, Cars, Visits
Quantitative	Continuous	Measurable decimal values	Height, Weight, Temperature

Key Takeaways:

- **Qualitative Data describes qualities (words, categories)**
- **Quantitative Data involves numbers and measurements**
- **Nominal = categories without order**
- **Ordinal = categories with order**
- **Discrete = countable whole numbers**
- **Continuous = measurable decimal values**

Section 1 of Statistics Classes - Data Types

Created for Educational Purposes - Statistics for Data Science

Data Science with Vamsi

© 2025 Data Science Education Guide