

Statistics Study Notes

What is Statistics?

Statistics is the science of collecting, organizing, analyzing, and interpreting data to make informed decisions. It helps us understand patterns, trends, and relationships in data, and draw meaningful conclusions about the world around us.

Two Main Branches of Statistics

Descriptive Statistics

Purpose: Summarizes and describes data

What it does: Organizes, displays, and summarizes data in a meaningful way

Tools: Mean, median, mode, graphs, charts, standard deviation

Inferential Statistics

Purpose: Makes predictions and generalizations

What it does: Uses sample data to draw conclusions about a larger population

Tools: Hypothesis testing, confidence intervals, regression analysis

Example: Average score of students in your class is 85

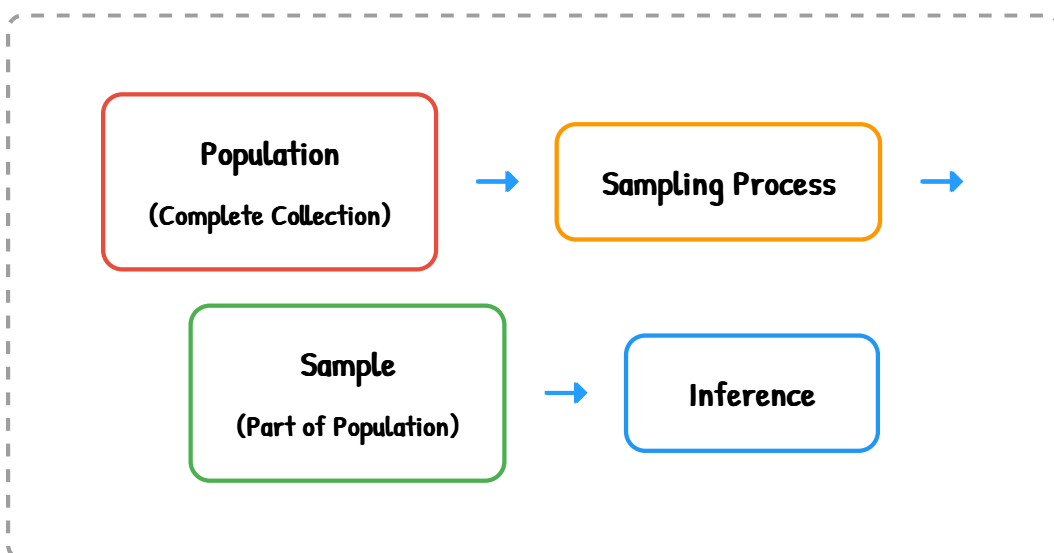
Example: Based on 100 students, we predict all students in the school average 85

Real World Example

- **Descriptive:** A restaurant tracks that it serves an average of 150 customers per day, with peak hours between 7-9 PM
- **Inferential:** Based on surveying 200 customers, the restaurant predicts that 70% of all their customers prefer vegetarian options

Basic Statistical Terms

Understanding the relationship between population, sample, parameters, and statistics is fundamental to statistical analysis. We rarely can study an entire population, so we use samples to make inferences.



Key Definitions:

- **Population** : Complete collection of all items/individuals to be studied
- **Sample** : A subset or part of the population selected for study
- **Parameter** : A numerical characteristic of a population (μ , σ , N)
- **Statistic** : A numerical characteristic of a sample (\bar{x} , s , n)
- **Inference** : Process of drawing conclusions about population from sample data

Example 1: Netflix Viewership Study

- **Population**: All 230 million Netflix subscribers worldwide
- **Sample**: 10,000 randomly selected subscribers
- **Parameter**: Average watch time of ALL subscribers (unknown)
- **Statistic**: Average watch time of 10,000 sampled subscribers (3.2 hours/day)
- **Inference**: We estimate all Netflix subscribers watch approximately 3.2 hours per day

Example 2: University Exam Performance

- **Population**: All 15,000 students at the university
- **Sample**: 500 students randomly surveyed
- **Parameter**: True average study hours of all 15,000 students
- **Statistic**: Average study hours of 500 surveyed students (4.5 hours/day)
- **Inference**: We estimate all students study about 4.5 hours daily

Statistical Notations

Different symbols are used to distinguish between population parameters and sample statistics. This is crucial because we use different formulas for each.

Measure	Population Parameter	Sample Statistic
Mean	μ (mu)	\bar{x} (x-bar)
Standard Deviation	σ (sigma)	s
Variance	σ^2	s^2
Proportion (having attribute)	P	p
Proportion (not having)	$Q = 1 - P$	$q = 1 - p$
Correlation Coefficient	ρ (rho)	r
Number of Elements	N	n

Descriptive Statistics Overview

Descriptive statistics helps us understand and communicate what the data looks like. It provides two main types of information: where the data centers (central tendency) and how spread out it is (variability).

Descriptive Statistics

Central Tendency

Mean
Mode
Median
Percentile
Quartile

Variability

Range
Interquartile Range
Standard Deviation

Measures of Central Tendency

Central tendency tells us the "typical" or "central" value in a dataset. It answers the question: "What is a representative value for this data?" Different measures are useful in different situations.

Why Do We Use Central Tendency?

- To summarize large datasets with a single representative value
- To compare different groups or datasets

- To identify the "typical" value in a distribution
- To make data easier to understand and communicate

1. Mean (Average)

- Sum of all values divided by count
- Most commonly used measure of center
- Affected by extreme values (outliers)
- Best used when data is symmetrical without outliers

$$\text{Mean} = \text{Sum of all values} / \text{Number of values}$$

Simple Example:

Test scores: 10, 11, 14, 9, 6

$$\text{Mean} = (10 + 11 + 14 + 9 + 6) / 5 = 50 / 5 = 10$$

Real Example 1: E-commerce Sales

Daily sales for a week: \$2000, \$2500, \$3000, \$2200, \$2800, \$3500, \$2600

Mean = \$2,657 per day

Use: The store can plan inventory based on average daily sales

Real Example 2: Student Study Hours

Weekly study hours: 15, 20, 18, 22, 19, 17, 21

Mean = 18.86 hours per week

Use: Students can compare their study time to the class average

2. Mode

- Most frequently occurring value in the dataset
- Can have multiple modes (bimodal, multimodal) or no mode
- Useful for categorical data
- Not affected by extreme values

Simple Example:

Data: 10, 11, 14, 9, 6, 10

Mode = **10** (appears twice, more than any other value)

Real Example 1: Shoe Store Inventory

Shoe sizes sold: 7, 8, 9, 9, 9, 10, 10, 11, 9, 8

Mode = Size 9 (sold most frequently)

Use: Store should stock more size 9 shoes

Real Example 2: Restaurant Orders

Most ordered dishes: Pizza, Burger, Pizza, Pasta, Pizza, Burger, Pizza

Mode = Pizza

Use: Restaurant should ensure Pizza ingredients are always in stock

3. Median

- Middle value when data is arranged in order
- If even count: average of two middle values
- Not affected by extreme values (robust measure)
- Better than mean when data has outliers

Example 1 (Odd count):

Data: 10, 11, 14, 9, 6

Ordered: 6, 9, **10**, 11, 14

Median = **10** (middle value)

Example 2 (Even count):

Data: 10, 11, 14, 9, 6, 11

Ordered: 6, 9, **10**, **11**, 11, 14

Median = $(10 + 11) / 2 = 10.5$

Real Example 1: House Prices

House prices in neighborhood: \$200K, \$220K, \$250K, \$280K, \$2M

Mean = \$590K (misleading due to \$2M outlier)

Median = \$250K (better representation of typical house price)

Use: Median gives a better sense of what most people pay

Real Example 2: Employee Salaries

Salaries: \$40K, \$45K, \$50K, \$55K, \$60K, \$500K (CEO)

Mean = \$125K (skewed by CEO salary)

Median = \$52.5K (typical employee salary)

Use: Median better represents what most employees earn

4. Percentile & Quartile

- **Median** : Divides data into 2 equal parts (50th percentile)
- **Percentile** : Divides data into 100 equal parts
- **Quartile** : Divides data into 4 equal parts (25%, 50%, 75%)
- Helps understand position of a value relative to the entire dataset

Steps to Calculate Percentile:

1. Arrange data in ascending order
2. Calculate location: $i = P \times (n) / 100$
3. If i is whole number → Average of i -th and $(i+1)$ -th values
4. If i is not whole → Round up to next whole number position

Example:

Data: 6, 9, 10, 11, 11, 14

Q1 (25th percentile) = 9 → 25% of values are ≤ 9

Q2 (50th percentile / Median) = 10.5 → 50% of values are ≤ 10.5

Q3 (75th percentile) = 11 → 75% of values are ≤ 11

Real Example 1: Exam Scores (Detailed Calculation)

Scenario: 15 students took a test. Find Q1 (25th percentile), Q2 (50th percentile), Q3 (75th percentile)

Raw scores: 45, 52, 67, 71, 75, 78, 80, 82, 85, 87, 90, 92, 95, 98, 100

n (sample size) = 15

Finding Q1 (First Quartile – 25th Percentile):

Step 1: Calculate position $\rightarrow i = (P/100) \times n = (25/100) \times 15 =$

3.75

Step 2: Since 3.75 is NOT a whole number, round UP to **4**

Step 3: The 4th value in ordered list = **71**

Q1 = 71 marks

Meaning: 25% of students scored 71 or less

Finding Q2 (Second Quartile – 50th Percentile / Median):

Step 1: Calculate position $\rightarrow i = (50/100) \times 15 = 7.5$

Step 2: Since 7.5 is NOT a whole number, round UP to **8**

Step 3: The 8th value in ordered list = **82**

Q2 = 82 marks (Median)

Meaning: 50% of students scored 82 or less

Finding Q3 (Third Quartile – 75th Percentile):

Step 1: Calculate position $\rightarrow i = (75/100) \times 15 = 11.25$

Step 2: Since 11.25 is NOT a whole number, round UP to **12**

Step 3: The 12th value in ordered list = **92**

Q3 = 92 marks

Meaning: 75% of students scored 92 or less

Visual representation:

45, 52, 67, **71 (Q1)**, 75, 78, 80, **82 (Q2)**, 85, 87, 90, **92 (Q3)**, 95, 98, 100

Real Use: If you scored 85, you're between Q2 and Q3, performing

better than at least 50% but not in the top 25%

Real Example 2: App Loading Time (Detailed Calculation)

Scenario: A developer measures app loading time for 12 users

Loading times (seconds): 1.2, 1.5, 1.8, 2.0, 2.1, 2.3, 2.5, 2.8, 3.0, 3.5, 4.0, 5.2

n (sample size) = 12

Finding Q1:

$$i = (25/100) \times 12 = 3 \text{ (whole number!)}$$

When i is a whole number: Average of 3rd and 4th values

3rd value = 1.8 , 4th value = 2.0

$$Q1 = (1.8 + 2.0) / 2 = \mathbf{1.9 \text{ seconds}}$$

Meaning: 25% of users experience load time of 1.9s or less

Finding Q2 (Median):

$$i = (50/100) \times 12 = 6 \text{ (whole number!)}$$

Average of 6th and 7th values

6th value = 2.3 , 7th value = 2.5

$$Q2 = (2.3 + 2.5) / 2 = \mathbf{2.4 \text{ seconds}}$$

Meaning: Typical user waits about 2.4 seconds

Finding Q3:

$$i = (75/100) \times 12 = 9 \text{ (whole number!)}$$

Average of 9th and 10th values

9th value = 3.0 , 10th value = 3.5

$$Q3 = (3.0 + 3.5) / 2 = \mathbf{3.25 \text{ seconds}}$$

Meaning: 75% of users wait 3.25s or less

Symbol Summary:

P = Percentile value (25, 50, 75)

n = Number of data points

i = Position/Location in ordered list

Q1, Q2, Q3 = First, Second, Third Quartiles

Real Use: Developer can claim "75% of users experience load time under 3.25 seconds"

Measures of Variability (Spread)

Variability measures tell us how spread out or scattered the data is.

Two datasets can have the same mean but very different spreads.

Understanding variability is crucial for making informed decisions.

Why Do We Use Variability Measures?

- To understand consistency and reliability of data
- To identify outliers and unusual values
- To assess risk and predictability
- To compare the consistency of different groups

Why Spread Matters: Two Delivery Services

Service A: Delivery times: 28, 29, 30, 31, 32 minutes (Mean = 30)

Service B: Delivery times: 10, 20, 30, 40, 50 minutes (Mean = 30)

Both have same mean, but Service A is more consistent and reliable!

Use: Variability helps choose the more predictable service

1. Range

- Difference between highest and lowest values
- Simple measure but heavily affected by extreme values (outliers)
- Gives quick sense of total spread
- One extreme value can make range misleading

$$\text{Range} = \text{Maximum Value } (x_{\max}) - \text{Minimum Value } (x_{\min})$$

Simple Example (Without Outlier):

Data: 6, 9, 10, 11, 11, 14

$$x_{\min} = 6, x_{\max} = 14$$

$$\text{Range} = 14 - 6 = 8$$

Problem: How Outliers Affect Range

Case 1: Employee Salaries (Without Outlier)

Salaries in thousands: 40, 45, 50, 55, 60

$$\text{Min} = \$40\text{K}, \text{Max} = \$60\text{K}$$

$$\text{Range} = 60 - 40 = \$20\text{K}$$

Interpretation: Reasonable salary spread of \$20K

Case 2: Same Company but Including CEO (With Outlier)

Salaries in thousands: 40, 45, 50, 55, 60, 500 (CEO)

$$\text{Min} = \$40\text{K}, \text{ Max} = \$500\text{K (outlier)}$$

$$\text{Range} = 500 - 40 = \$460\text{K}$$

Problem: Range jumped from \$20K to \$460K because of ONE outlier!

Misleading: Makes it seem like huge salary differences exist among regular employees

Solution: Use IQR (Interquartile Range)

For the same data with CEO:

Ordered: 40, 45, 50, 55, 60, 500

$$Q1 = 45\text{K}, Q3 = 60\text{K}$$

$$\text{IQR} = 60 - 45 = \$15\text{K}$$

IQR focuses on middle 50% and ignores the CEO outlier!

This \$15K better represents typical salary variation

Another Example: Delivery Times

Normal Days: 25, 28, 30, 32, 29, 27, 31 minutes

$$\text{Range} = 32 - 25 = 7 \text{ minutes (good consistency)}$$

One Day with Traffic Jam: 25, 28, 30, 32, 29, 27, 120 minutes (outlier)

$$\text{Range} = 120 - 25 = 95 \text{ minutes (looks terrible!)}$$

Problem: One bad day makes the service look unreliable

Using IQR:

$Q1 = 27$ minutes, $Q3 = 32$ minutes

$IQR = 5$ minutes

IQR shows service is actually consistent (5-min variation) for typical deliveries

Key Lesson: When to Use Range vs IQR

- **Use Range when:** You want to know absolute limits (min to max)
- **Use IQR when:** Data has outliers and you want typical variation
- **Symbol meanings:** x_{\min} = minimum value, x_{\max} = maximum value, $Q1$ = first quartile, $Q3$ = third quartile

2. Interquartile Range (IQR)

- Range of middle 50% of data
- Less affected by outliers than regular range
- Shows where most typical values lie
- Used to identify outliers: values beyond $Q1 - 1.5 \times IQR$ or $Q3 + 1.5 \times IQR$

$$IQR = Q3 - Q1$$

Simple Example:

Data: 6, 9, 10, 11, 11, 14

$Q1 = 9$, $Q3 = 11$

$IQR = 11 - 9 = 2$

This means the middle 50% of data spans only 2 units

Box-and-Whisker Plot



Real Example 1: Customer Wait Times

Wait times at a bank (minutes): 2, 3, 5, 6, 7, 8, 9, 10, 12, 25

Q1 = 5 minutes, Q3 = 10 minutes

IQR = 5 minutes (middle 50% wait between 5–10 minutes)

The 25-minute wait is an outlier

Use: Bank can promise "most customers served within 10 minutes"

Real Example 2: Website Response Time

Response times (ms): 100, 120, 150, 180, 200, 220, 250, 280, 300, 1500

Q1 = 150ms, Q3 = 280ms

IQR = 130ms

The 1500ms is an outlier (possible server issue)

Use: Focus on improving the typical range, investigate outliers separately

3. Standard Deviation (σ or s) & Variance (σ^2 or s^2)

→ Most important measure of spread in statistics

→ Shows average distance of data points from the mean (\bar{x} or μ)

- **Variance (σ^2 or s^2)** = Average of squared deviations from mean
- **Standard Deviation (σ or s)** = Square root of variance (in original units)
- Small SD = data points close to mean (consistent/predictable)
- Large SD = data points far from mean (variable/unpredictable)

Sample Standard Deviation (s):

$$s = \sqrt{[\Sigma (x - \bar{x})^2 / (n - 1)]}$$

Where: x = each value, \bar{x} = sample mean, n = sample size

Σ (sigma) = sum of all values

We divide by $(n-1)$ for sample to get better estimate

Population Standard Deviation (σ):

$$\sigma = \sqrt{[\Sigma (x - \mu)^2 / N]}$$

Where: μ (mu) = population mean, N = population size

We divide by N for entire population

Step-by-Step Calculation with Symbols:

x (value)	$x - \bar{x}$ (deviation)	$(x - \bar{x})^2$ (squared dev.)
100	$100 - 100 = 0$	$0^2 = 0$
101	$101 - 100 = 1$	$1^2 = 1$
99	$99 - 100 = -1$	$(-1)^2 = 1$
102	$102 - 100 = 2$	$2^2 = 4$

$$98 \quad 98 - 100 = -2 \quad (-2)^2 = 4$$

$$100 \quad 100 - 100 = 0 \quad 0^2 = 0$$

$$\begin{aligned} \bar{x} &= 100 \\ \Sigma(x - \bar{x}) &= 0 \quad \Sigma(x - \bar{x})^2 = 10 \end{aligned}$$

Step 1: Calculate mean $\rightarrow \bar{x} = (100+101+99+102+98+100)/6 = 100$

Step 2: Find deviation $(x - \bar{x})$ for each value

Step 3: Square each deviation to make all positive

Step 4: Sum all squared deviations $\rightarrow \Sigma(x - \bar{x})^2 = 10$

Step 5: Calculate variance $\rightarrow s^2 = 10 / (6-1) = 10 / 5 = 2$

Step 6: Take square root $\rightarrow s = \sqrt{2} = 1.414$

Results:

Variance (s^2) = 2

Standard Deviation (s) = 1.414

Meaning: On average, values differ from mean by about 1.4 units

Real Example 1: Daily Coffee Shop Sales (Easy to Understand)

Coffee Shop A:

Daily cups sold: 98, 100, 99, 101, 102, 100

Mean (\bar{x}) = 100 cups/day

Variance (s^2) = 2

Standard Deviation (s) = 1.4 cups

What this means in simple words:

Each day, sales are typically within 1.4 cups of the average (100 cups).

So most days: $100 - 1.4 = 98.6$ to $100 + 1.4 = 101.4$ cups

Interpretation: Very predictable! Owner can confidently order supplies

Business decision: Stock around 100–102 cups worth of ingredients daily

Coffee Shop B:

Daily cups sold: 70, 90, 100, 110, 130, 100

Mean (\bar{x}) = 100 cups/day (same as Shop A!)

Variance (s^2) = 400

Standard Deviation (s) = 20 cups

What this means in simple words:

Each day, sales typically vary by 20 cups from average!

Range: $100 - 20 = 80$ to $100 + 20 = 120$ cups

Interpretation: Very unpredictable! Sales swing wildly

Business decision: Need flexible inventory, might waste ingredients or run out

KEY INSIGHT:

Both shops sell 100 cups on average, but Shop A is reliable (SD=1.4) while Shop B is risky (SD=20)

Low SD = Consistent = Easy to plan

High SD = Variable = Hard to predict

Real Example 2: Student Test Preparation Hours

Student Group A (Disciplined):

Study hours per day: 4, 4.5, 4, 4.2, 4.3, 4

Mean (\bar{x}) = 4.17 hours

Standard Deviation (s) = 0.19 hours (about 11 minutes)

Simple meaning: Students study consistently around 4 hours daily

Daily variation is only ± 11 minutes from average

Tells us: Very disciplined study routine

Likely result: Predictable, steady improvement

Student Group B (Irregular):

Study hours per day: 1, 6, 2, 7, 3, 5

Mean (\bar{x}) = 4 hours (almost same as Group A!)

Standard Deviation (s) = 2.28 hours

Simple meaning: Study time swings wildly from 1 to 7 hours

Daily variation is ± 2.3 hours from average

Tells us: No consistent routine, cramming some days

Likely result: Less effective learning, higher stress

Real Example 3: Internet Speed at Home

Provider A (Fiber):

Speed tests (Mbps): 98, 100, 99, 101, 100, 102

Mean (μ) = 100 Mbps

Standard Deviation (σ) = 1.4 Mbps

Variance (σ^2) = 2

What SD tells you: Speed is super stable, only varies by ± 1.4 Mbps

What Variance adds: The squared value (2) is still small, confirming low variation

Real meaning: Your Zoom calls never freeze, gaming is smooth

Reliability: You always get what you pay for

Provider B (Cable):

Speed tests (Mbps): 60, 80, 100, 120, 90, 150

Mean (μ) = 100 Mbps (same advertised speed!)

Standard Deviation (σ) = 30 Mbps

Variance (σ^2) = 900

What SD tells you: Speed varies wildly by ± 30 Mbps

What Variance adds: The large squared value (900) emphasizes huge variability

Real meaning: Sometimes fast (150), sometimes slow (60)

Reliability: Frustrating during peak hours, unreliable for work-from-home

Why we need BOTH Variance and Standard Deviation:

Variance (σ^2): Used in advanced calculations and statistical tests

Standard Deviation (σ): Easier to interpret because it's in same units as data

For internet speed: SD of 30 Mbps makes sense to us

Variance of 900 Mbps² is hard to visualize, but useful for math

Understanding Standard Deviation in Daily Life

- **Low SD:** Things are consistent, predictable, reliable (like your favorite restaurant)
- **High SD:** Things vary a lot, unpredictable, risky (like weather or stock prices)
- **Variance (s^2 or σ^2):** Mathematical version, useful for calculations
- **Standard Deviation (s or σ):** Practical version, tells you "typical difference from average"

- **Rule of thumb:** About 68% of data falls within 1 SD of the mean
- **Use case:** Compare consistency between two options with same average

Quick Summary

When to Use Each Measure:

Measure	Best Used When	Avoid When
Mean	Data is symmetrical, no outliers	Data has extreme values
Median	Data has outliers or is skewed	Need to consider all values
Mode	Categorical data, finding most common	Continuous data with no repeats
Range	Quick sense of total spread	Presence of outliers
IQR	Data has outliers, need robust measure	Want to include all data points
Standard Deviation	Need precise measure of spread	Severe outliers present

Key Insights:

- **Central Tendency:** Tells us the "typical" value (where data centers)
- **Variability:** Tells us about consistency and spread (how data differs)
- **Population vs Sample:** Always use correct symbols (μ, σ vs \bar{x}, s)
- **Outliers:** Use median and IQR when outliers present
- **Interpretation:** Low SD = consistent/predictable; High SD = variable/risky
- **Context Matters:** Choose measures based on data characteristics and question