

Indian Institute of Technology Madras
ZANZIBAR CAMPUS

Z3003 · DATA MINING: SCALING TO LARGE DATA

Central Tendency & Dispersion

Lecture 3 · Friday 18 September 2026 · …and why exact statistics don't always scale

Dr. Innocent Nyalala | BS Semester 5 (DSAI) | IIT Madras Zanzibar

Session Outline

IIT Madras Zanzibar

1. (0–4 min) Recap: which statistic each attribute type allows
2. (4–16 min) Mean, median, mode, midrange — by hand on 9 farms
3. (16–30 min) Range, quartiles, IQR, boxplot, variance and standard deviation
4. (30–36 min) The 500 farms: skew and the five-number summary
5. (36–44 min) At scale: one-pass mean and variance; approximate median
6. (44–50 min) Exercise and summary

Learning Outcomes

IIT Madras Zanzibar

By the end of this session, you will be able to:

Compute	mean, weighted mean, median, mode and midrange by hand
Compute	quartiles, IQR, outlier fences, variance and standard deviation
Choose	mean or median from the shape (skew) of the data
Read	a five-number summary and a boxplot
Explain	why the mean scales to a billion rows and the exact median does not

Central Tendency & Dispersion

IIT Madras Zanzibar

→ **Recap: Type Decides the Statistic**

Central Tendency: Mean, Median, Mode

Dispersion: Range, Quartiles, IQR, Variance

Our 500 Farms, Summarised

Statistics at Scale: One Pass, Little Memory

In-Class Exercise

Recap: The Type Decides the Statistic

IIT Madras Zanzibar

Attribute type	Example	Allowed centre	Allowed spread
Nominal	crop	mode (Clove)	count of categories
Ordinal	soil: Poor < Fair < Good	mode, median	range of ranks, quartiles
Numeric	farm size, yield, rainfall	mode, median, mean	range, IQR, variance, std. dev.

Today: numeric attributes. Centre answers 'what is typical?'; spread answers 'how different are the farms from each other?'

Central Tendency & Dispersion

IIT Madras Zanzibar

Recap: Type Decides the Statistic

→ **Central Tendency: Mean, Median, Mode**

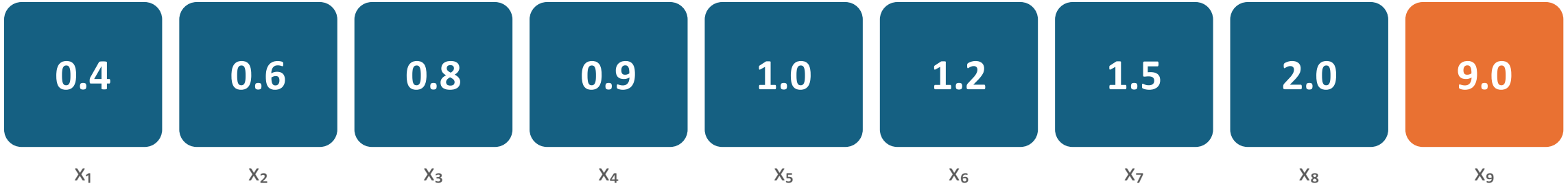
Dispersion: Range, Quartiles, IQR, Variance

Our 500 Farms, Summarised

Statistics at Scale: One Pass, Little Memory

In-Class Exercise

Our Tiny Example: 9 Farms in One Village



Symbol	Read as	Here
n	number of values	9
x_i	the i-th value	$x_1 = 0.4, \ldots, x_9 = 9.0$
Σ	'sum of' — add them all	$\Sigma x_i = 17.4$

Watch the 9.0 ha estate: it is real, not an error — and it will pull some statistics far more than others.

The Mean: Share Everything Equally

$$\bar{x} = (\Sigma\, x_i) \div n = (x_1 + x_2 + \ldots + x_n) \div n$$

$\bar{x}$ ('x-bar') = sample mean; μ ('mu') = population mean

1 **Add**
0.4 + 0.6 + 0.8 + 0.9 + 1.0 + 1.2 + 1.5 + 2.0 + 9.0 = 17.4

2 **Divide by n = 9**
17.4 ÷ 9 = 1.93 ha

3 **Without the estate**
8.4 ÷ 8 = 1.05 ha — one farm moved the mean by 0.88 ha

Analogy: pool all the land and share it equally — each farmer gets 1.93 ha, yet 8 of 9 farms are smaller.

Weighted Mean: When Values Count Differently

IIT Madras Zanzibar

$$\bar{x} = (\Sigma\, w_i\, x_i) \div (\Sigma\, w_i)$$

w_i = weight of value i (how much it counts)

Batch	kg (w_i)	price \$/kg (x_i)	$w_i \times x_i$
1	100	8	800
2	300	7	2,100
3	600	6	3,600
Total	**1,000**		**6,500**

1 **Plain mean (wrong here)**
(8 + 7 + 6) ÷ 3 = \$7.00

2 **Weighted mean**
6,500 ÷ 1,000 = \$6.50 per kg

Also weighted: your course grade — Quiz 1 20%, mini-project 20%, Part 2 60%.

The Median: The Middle Farm

0.4	0.6	0.8	0.9	**1.0**	1.2	1.5	2.0	9.0

position 5 = median

odd n: value at position $(n + 1) \div 2$

n = 9 → position 5 → 1.0 ha

even n: average the two middle values

n = 8 (no estate) → (0.9 + 1.0) ÷ 2 = 0.95 ha

	with estate	without estate	moved by
mean	1.93	1.05	0.88 ha
median	1.00	0.95	0.05 ha

Mode and Midrange

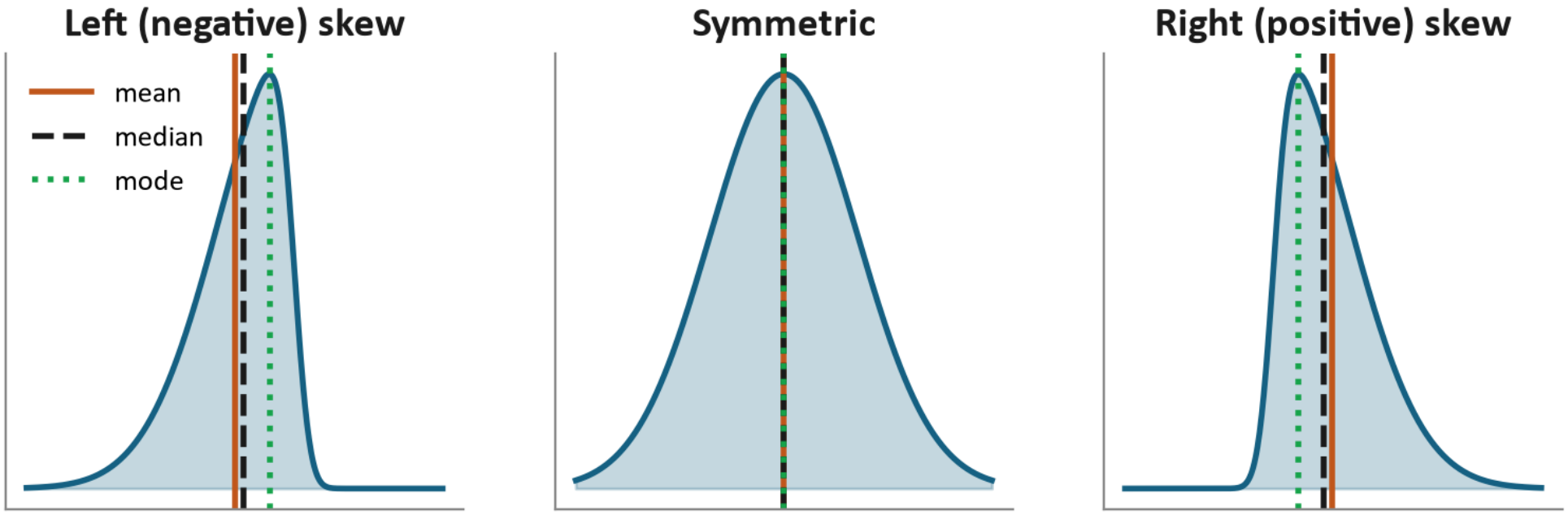
IIT Madras Zanzibar

1 **Mode = most frequent value**
9 farms: every value once → no mode. 500 farms: modal crop = Clove (142 farms)

2 **Midrange = (min + max) ÷ 2**
(0.4 + 9.0) ÷ 2 = 4.7 ha — bigger than 8 of the 9 farms

Statistic	9 farms	Uses	Robust to the estate?
mean	1.93	every value	no
median	1.00	the order	yes
mode	none	frequencies	yes
midrange	4.70	only min and max	worst

Skew: Where Mean, Median and Mode Sit



Rule of thumb: mean ≫ median → right skew → report the median (incomes, farm sizes, balances). Mean ≈ median → symmetric → either works.

Central Tendency & Dispersion

IIT Madras Zanzibar

Recap: Type Decides the Statistic

Central Tendency: Mean, Median, Mode

→ **Dispersion: Range, Quartiles, IQR, Variance**

Our 500 Farms, Summarised

Statistics at Scale: One Pass, Little Memory

In-Class Exercise

Why Spread Matters: Same Centre, Different Farms

IIT Madras Zanzibar

Village	Farm sizes (ha)	Mean	Spread
A	2, 2, 2, 2, 2, 2	2.0	none — all equal
B	0.5, 0.5, 0.5, 3.5, 3.5, 3.5	2.0	large — two very different groups

1. **Range**
   max − min

2. **Quartiles & IQR**
   spread of the middle half

3. **Variance**
   average squared distance from the mean

4. **Std. deviation**
   √variance, in original units

Range and Quartiles

IIT Madras Zanzibar

0.4	0.6	0.8	0.9	1.0	1.2	1.5	2.0	9.0

lower half → Q1 = (0.6 + 0.8) ÷ 2 = 0.7

upper half → Q3 = (1.5 + 2.0) ÷ 2 = 1.75

1. **Range**
   9.0 − 0.4 = 8.6 ha (set by the estate alone)

2. **Q2 = median**
   1.0 ha

3. **IQR = Q3 − Q1**
   1.75 − 0.7 = 1.05 ha: the middle half of farms

Note: software interpolates quartiles slightly differently; on big data the methods agree. Use the halves method by hand.

Outliers and the Five-Number Summary

IIT Madras Zanzibar

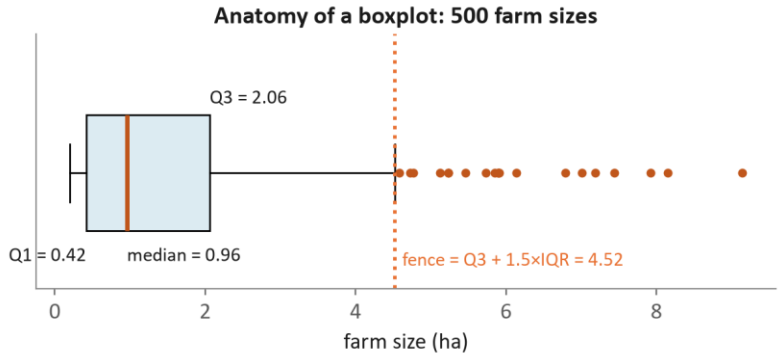
$$\text{upper fence} = Q3 + 1.5 \times \text{IQR}$$

1.75 + 1.5 × 1.05 = 3.325 ha → 9.0 is an outlier

$$\text{lower fence} = Q1 - 1.5 \times \text{IQR}$$

0.7 − 1.575 < 0 → no small outliers

	min	Q1	median	Q3	max
9 farms	0.4	0.7	1.0	1.75	9.0
500 farms	0.20	0.42	0.96	2.06	9.14



500 farms: fence = 4.52 ha

20 farms flagged — real large estates, not errors

Variance and Standard Deviation, Step by Step

IIT Madras Zanzibar

$$s^2 = \Sigma\,(x_i - \bar{x})^2 \div (n - 1) \qquad s = \sqrt{s^2}$$

s^2 = sample variance; s = standard deviation (same units as the data)

x_i	0.4	0.6	0.8	0.9	1.0	1.2	1.5	2.0	9.0
$x_i - \bar{x}$	-1.53	-1.33	-1.13	-1.03	-0.93	-0.73	-0.43	+0.07	+7.07
$(x_i - \bar{x})^2$	2.35	1.78	1.28	1.07	0.87	0.54	0.19	0.00	49.94

1 **Sum of squares**
58.02 (the estate alone: 49.94)

2 **Divide by n − 1 = 8**
$s^2 = 7.25$ ha²

3 **Square root**
$s = 2.69$ ha (without the estate: 0.51)

Why n − 1? deviations are measured from $\bar{x}$, which is itself fitted to the sample; dividing by n would slightly underestimate the true spread.

Central Tendency & Dispersion

IIT Madras Zanzibar

Recap: Type Decides the Statistic

Central Tendency: Mean, Median, Mode

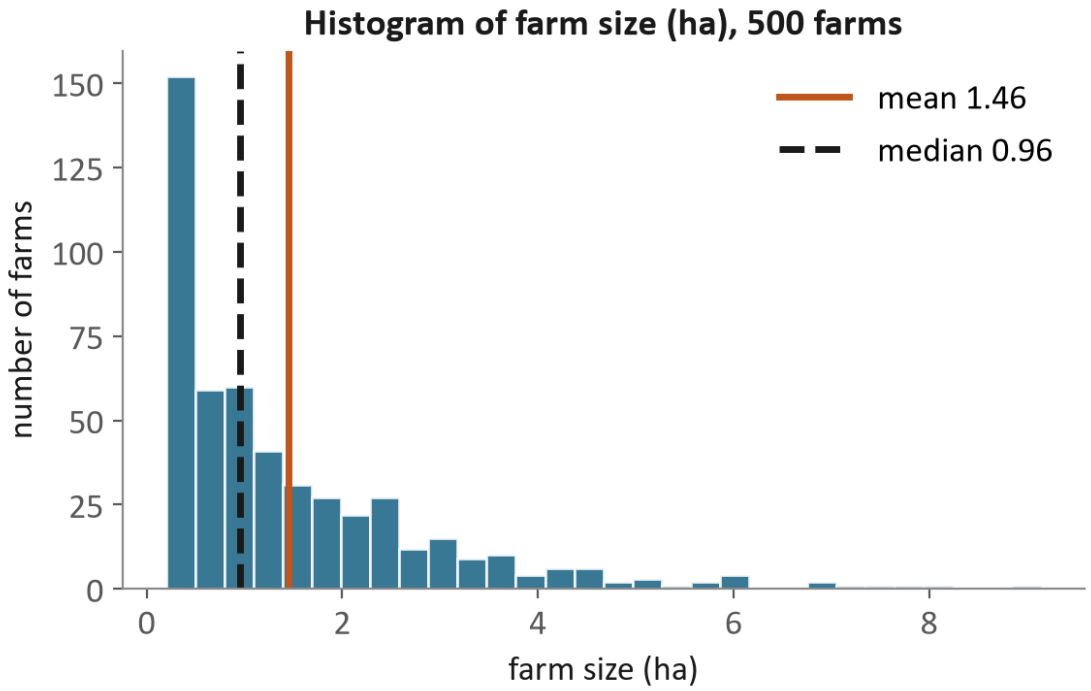
Dispersion: Range, Quartiles, IQR, Variance

→ **Our 500 Farms, Summarised**

Statistics at Scale: One Pass, Little Memory

In-Class Exercise

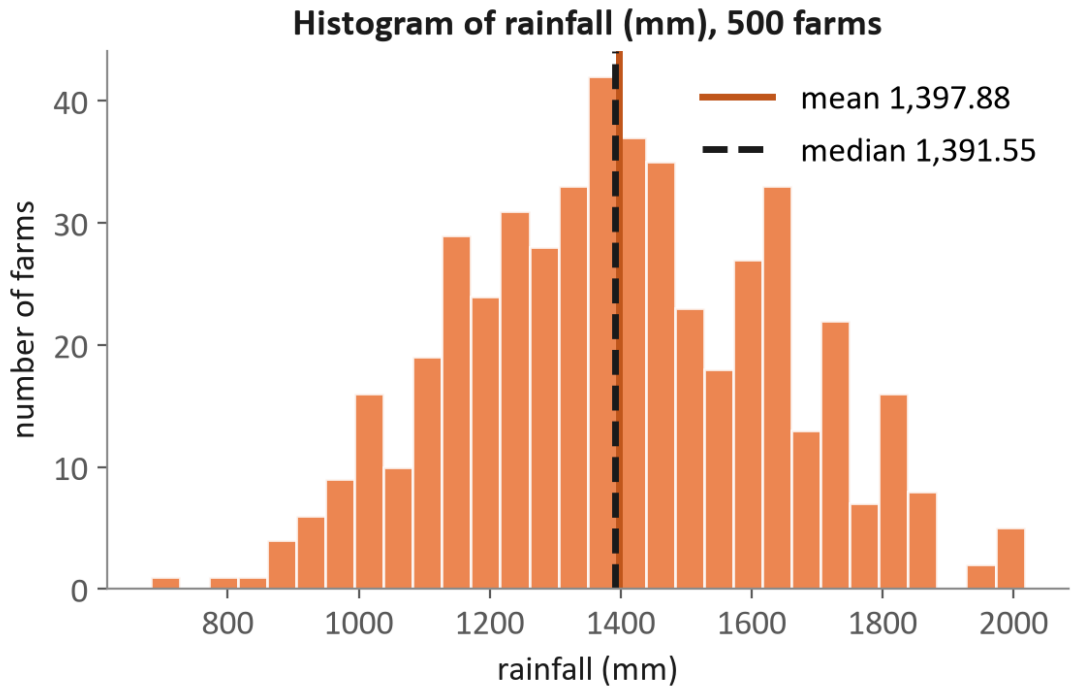
Farm Size, All 500 Farms: Right-Skewed



statistic	farm size
mean	1.46 ha
median	0.96 ha
std. dev. s	1.45 ha
IQR	1.64 ha
skewness	1.93 (right)

Decision: planning for a 'typical' farm of 1.46 ha overstates most farms. The median, 0.96 ha, describes them better.

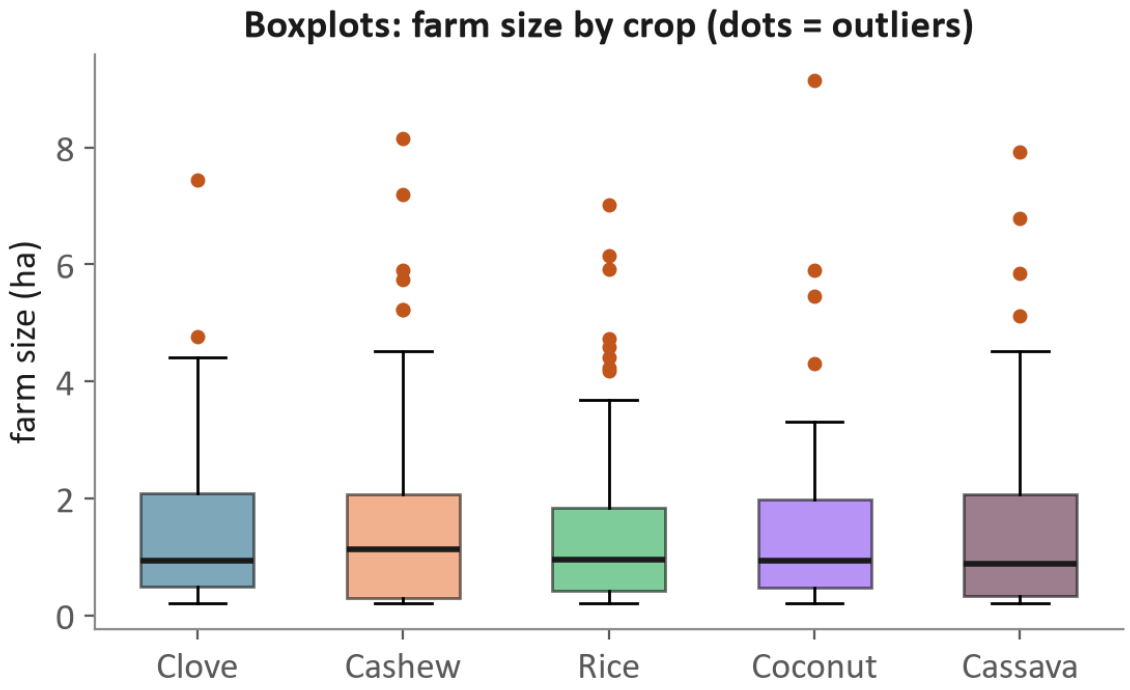
Rainfall, All 500 Farms: Symmetric



statistic	rainfall
mean	1398 mm
median	1392 mm
std. dev. s	243 mm
IQR	361 mm
skewness	0.05 (symmetric)

Rule: for bell-shaped data, about 68% of values lie within one std. dev. of the mean: 1155–1641 mm.

Comparing Groups with Boxplots



1. **Line**
   median

2. **Box**
   Q1 to Q3 — the middle half

3. **Whiskers**
   furthest point inside the fences

4. **Dots**
   outliers beyond 1.5 × IQR

Reading: median sizes are close across crops (Cashew 1.12, Cassava 0.87, Clove 0.92, Coconut 0.93, Rice 0.95 ha); every crop has a few large farms.

Central Tendency & Dispersion

IIT Madras Zanzibar

Recap: Type Decides the Statistic

Central Tendency: Mean, Median, Mode

Dispersion: Range, Quartiles, IQR, Variance

Our 500 Farms, Summarised

→ **Statistics at Scale: One Pass, Little Memory**

In-Class Exercise

The Mean in One Pass: Keep Two Numbers

IIT Madras Zanzibar

Each new value x	Update
count	$n \leftarrow n + 1$
sum	$\Sigma x \leftarrow \Sigma x + x$
sum of squares	$\Sigma x^2 \leftarrow \Sigma x^2 + x^2$

$$\text{mean} = \Sigma x \div n$$

$$\text{variance} = \Sigma x^2 \div n - \text{mean}^2$$

population form (divide by n)

500 farms	n	Σx	Σx²	mean	variance	σ
farm size	500	727.77	2,102.95	1.4555	4.2059 − 2.1186 = 2.0873	1.445 ha

O(1) memory: three numbers whatever the size — 500 farms or 6 billion payments. Sums from different machines can simply be added.

The Exact Median Does Not Scale

IIT Madras Zanzibar

	Mean	Exact median
Memory	2–3 numbers (O(1))	every value (O(n))
Work	one pass, n steps	sort: about n log n steps
Split across machines	add the sums and counts	median of medians ≠ true median
Week 4 name	algebraic (from distributive sum, count)	holistic

Example: machine 1 holds {1, 2, 3} (median 2); machine 2 holds {4, 5, 6, 7, 8} (median 6). Median of medians = 4, but the true median of all eight values is 4.5.

Approximate Median from a Histogram

IIT Madras Zanzibar

$$\text{median} \approx L_1 + [\,(N/2 - (\Sigma\,\text{freq})_l) \div \text{freq_median}\,] \times \text{width}$$

L_1 = lower edge of the median bin; $(\Sigma\,\text{freq})_l$ = count below it; freq_median = count inside it

bin (ha)	farms	cumulative
0.0–0.5	152	152
0.5–1.0	109	261 ≥ 250
1.0–1.5	67	328
…	…	…

1 **Middle position**
N/2 = 500/2 = 250 → bin 0.5–1.0

2 **Plug in**
0.5 + (250 − 152) ÷ 109 × 0.5

3 **Estimate vs exact**
0.95 ha vs 0.96 ha — 20 counts instead of 500 values

Scales: memory = number of bins; bin counts from many machines simply add. Other tools: sampling, sketches (t-digest).

Central Tendency & Dispersion

IIT Madras Zanzibar

Recap: Type Decides the Statistic

Central Tendency: Mean, Median, Mode

Dispersion: Range, Quartiles, IQR, Variance

Our 500 Farms, Summarised

Statistics at Scale: One Pass, Little Memory

→ **In-Class Exercise**

In-Class Exercise (5 min, pairs)

IIT Madras Zanzibar

IN-CLASS TASK

Seven mobile-money balances in a village (thousand TZS), sorted: 2, 3, 3, 4, 5, 6, 40

1. Compute the mean, median, mode and midrange.
2. Compute Q1, Q3, the IQR and the upper fence. Is 40 an outlier?
3. Compute the variance s^2 and standard deviation s (divide by $n - 1$).
4. Which centre would you report for 'a typical user', and why?

Solution

IIT Madras Zanzibar

1 **Centre**
mean = 63 ÷ 7 = 9.0 · median = 4 · mode = 3 · midrange = (2 + 40) ÷ 2 = 21

2 **Quartiles**
lower half 2, 3, 3 → Q1 = 3 · upper half 5, 6, 40 → Q3 = 6 · IQR = 3 · fence = 6 + 4.5 = 10.5 → 40 is an outlier

3 **Spread**
squared deviations 49, 36, 36, 25, 16, 9, 961 → sum 1,132 → s^2 = 1,132 ÷ 6 = 188.7 → s = 13.7

4 **Typical user**
median 4 — the mean 9.0 is larger than 6 of the 7 balances (right skew)

Summary & Exam Checkpoints

IIT Madras Zanzibar

Centre

mean (uses all values, not robust) · median (robust) · mode (any type) · midrange (least robust)

Spread

range · quartiles and IQR (robust) · variance and std. dev. (n − 1)

Shape

mean ≫ median → right skew → report the median; outliers beyond Q3 + 1.5 × IQR

At scale

mean, variance: one pass, O(1) memory · median: holistic → approximate with a histogram

What's Next

IIT Madras Zanzibar

Wed **Lecture 4 (23 Sep): Covariance, Correlation & Graphics**
how two attributes move together; histograms, boxplots, scatter and Q-Q plots

Thu **Lecture 5 (24 Sep): Similarity & Distance**
Euclidean, Manhattan, Jaccard, cosine; mixed attribute types

Tut **Tutorial 2**
correlation, distance and similarity on the 500 farms

Practise: compute the five-number summary of yield in pandas: df.yield_tonnes.describe().

References & Further Reading

- Han, J., Pei, J. & Tong, H. (2023). Data Mining: Concepts and Techniques, 4th ed., Section 2.2.1. Morgan Kaufmann.
- Leskovec, J., Rajaraman, A. & Ullman, J. (2020). Mining of Massive Datasets, 3rd ed., Ch. 4 (streams). mmds.org.
- Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.
- Welford, B. P. (1962). Note on a Method for Calculating Corrected Sums of Squares. Technometrics 4(3).
- Dunning, T. & Ertl, O. (2019). Computing Extremely Accurate Quantiles Using t-Digests. arXiv:1902.04023.