Stat.BD

APPLIED BIOSTATISTICS & RESEARCH METHODOLOGY • SESSION 4

Describing Numerical Data

Where the middle is, how spread out it is, and which pair of numbers to print

Muhtasim Munif Fahim, MSc
Mentor, Stat.BD
Research Associate, Data Science Research Lab
Department of Statistics & Data Science, University of Rajshahi

Session 4 of 30

Two good papers, two opposite habits

The Sylhet survey
cross-sectional, $n = 1{,}810$

Age	48 (41 to 59)
Body mass index	20.3 (18.1 to 22.9)
Waist circumference	77.0 (69.7 to 84.8)
Years of schooling	2 (1 to 5)

Median (interquartile range) for every numerical row. Not one mean anywhere in the paper.

COBRA-BPS
cluster randomised trial, $n = 2{,}645$

Age	58.8 ± 11.5
Systolic BP, intervention	146.7 ± 22.4
Systolic BP, control	144.7 ± 21.0
Diastolic BP, intervention	89.1 ± 14.7

Mean $\pm$ standard deviation for every numerical row. Not one median.

Two careful papers, two opposite conventions, and neither is an oversight. The choice was made by the shape of the columns, and working out how they made it is what this session is for.

Khanam R, et al. *BMJ Open* 2019;9(10):e026722, Table 1. PMID 31662350. Jafar TH, et al. *N Engl J Med* 2020;382(8):717–726, Table 1. PMID 32074419.

What you will be able to do at the end of today

The test of this session

Take any numerical row in any Table 1. Say which summary was reported, why that one and not the other, and roughly what shape the column must have had.

- Say what the **mean**, the **median** and the **mode** each are, and when each is misleading
- Read a **five-number summary** and the **interquartile range** off a published table
- Say what a **standard deviation** means in the units of the measurement
- Decide, from a printed table alone, whether the mean or the median was the right choice

Act I

Where the middle is

Three middles, and they answer different questions

Summary	What it is	Needs
Mean	Add every value, divide by how many. The balance point.	numerical
Median	Sort the values, take the middle one. Half above, half below.	an order
Mode	The value, or category, that occurs most often.	nothing

Why there are three and not one

The mode works on a blood group, where there is no order at all. The median works on a wealth tertile, where there is an order but no arithmetic. Only the mean needs the gaps between values to be real quantities, and that is exactly the assumption it can get wrong.

The mode, counted rather than computed

The mode is the value, or category, occurring most often. Counting is the whole calculation; there is no sum and no sort.

Number of previous pregnancies, seven women at an antenatal visit:

$$0, \quad 1, \quad 1, \quad 2, \quad 2, \quad 2, \quad 4$$

Value	0	1	2	4
Count	1	2	**3**	1

Mode $= 2$: three of the seven women, more than any other value.

Where it is useless and where it is not

On a waist circumference almost every value is unique, and the mode says nothing. On a small count like this one, real repeats occur, and the mode is the value most patients actually have.

The seven values are illustrative, constructed by us so that a repeated value exists. No anchor paper in this course prints a raw column small enough to show a mode by counting.

One value moves, and only one of them moves with it

Five waist measurements, in cm:

70, 74, 77, 81, 88

Mean 78.0. Median 77.

Now the largest is recorded as 188, a typing error nobody caught:

70, 74, 77, 81, 188

Mean 98.0. Median 77.

A single mistyped digit moved the mean by twenty centimetres and the median not at all.

What the mean is doing

The mean is a balance point, so every value pulls on it in proportion to how far out it sits.

The median only counts how many values are on each side, so it does not care how far away they are.

Five illustrative values, chosen by us to make the arithmetic easy. The arithmetic is ours.

The mean, in the sum it hides

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The same five waist measurements as before:

$$70, \quad 74, \quad 77, \quad 81, \quad 88$$

$$\bar{x} = \frac{70 + 74 + 77 + 81 + 88}{5} = \frac{390}{5}$$

$$\bar{x} = 78.0$$

The same five illustrative values as the previous frame.

Why this is worth doing by hand

Software returns a mean instantly. Being able to check it for five numbers is what lets you catch a formula pointing at the wrong column in a spreadsheet.

The median's rule for an even sample

Sort the values. If n is odd, the median is the middle one. If n is even, it is the average of the two middle values.

Five lengths of stay, days: $2, 3, 4, 5, 41$. Sorted, the middle value is 4.

Six lengths of stay: $2, 3, 3, 4, 5, 41$. The middle two are 3 and 4:

$$\text{median} = \frac{3+4}{2} = 3.5$$

Both sets of values and the arithmetic are ours.

Sort first

A median taken from an unsorted list is not a median, and a spreadsheet will compute one without complaining.

Trimming the ends, and the summary that survives it

Keep the mean, but stop letting the extremes vote. Drop the same number of values from *each* end, then average what is left.

Five waist values	**clean**	**with the typo**
Mean	78.0	**98.0**
Median	77.0	77.0
Trimmed mean	77.3	**77.3**

Drop 70 and 88, average 74, 77 and 81. The typo was the value that got dropped, so it never reached the arithmetic.

You already trim

Three blood pressures, one obviously wild. You quietly work from the other two.

That is a trimmed mean, done by eye and without a name.

What you are buying, and paying

Buying: a summary that uses the size of the values, like a mean, without being hostage to two of them.

Paying: you must say how much you trimmed, because 5% and 20% give different answers.

The five values are the invented ones from the previous frame. All arithmetic here is ours.

The spread that matches the median

The standard deviation is built out of the mean, so it inherits the mean's problem. The median has its own measure of spread:

Median absolute deviation. Take each value's distance from the median, then take the median of those distances.

Distances from 77: 7, 3, 0, 4, 11. Their median is **4**.

	clean	with the typo
Standard deviation	6.9	**50.5**
Median abs. deviation	4.0	**4.0**

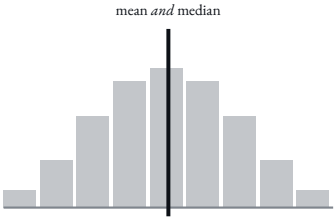
Same five invented values. Every figure on this frame is ours.

Why the typo cannot reach it

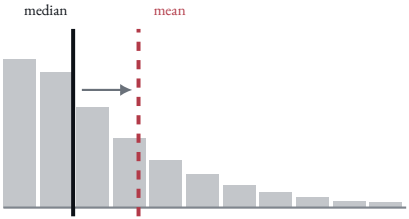
With the typo the distances become 7, 3, 0, 4, 111.

The 111 is only ever *one* value in a list of five, so it moves where the middle of that list sits by nothing at all. Squaring, which is what the SD does, would have let it dominate.

So the gap between them describes the shape



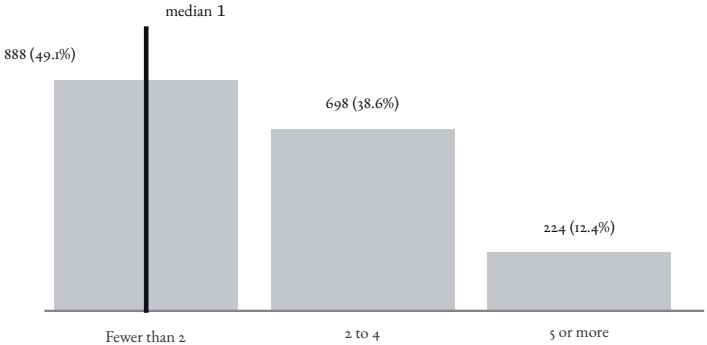
Symmetric
The two coincide. Either one describes the centre, and the mean is the more useful because later methods are built on it.



Skewed to the right
The long tail drags the mean away from the bulk of the data. The median stays where most of the observations are.

Both distributions are illustrative. The rule they carry is not: **the further apart the mean and the median, the more skewed the column**, and the median is the summary that survives the skew.

A column with a floor, and why its mean would be a fiction



Half the sample ate fewer than two servings a day and the reported median is 1. A mean computed from this column would land somewhere between the bars, describing a portion size nobody was recorded as eating, and one participant reporting twenty servings would move it.

Counts, percentages and the median are from the paper. Khanam 2019, Table 1. PMID 31662350.

Reading a median off a table you did not collect

The Sylhet survey reports the median age as 48 years, and separately as 50 for men and 47 for women.

The median of the whole sample is *not* the average of the two. $(50 + 47)/2 = 48.5$, and the reported figure is 48.

Why medians do not average

A median is a position in a sorted list, not a quantity that can be pooled.

Combining two groups means re-sorting all 1,810 values, which only somebody with the data can do.

Means can be pooled if you know both group sizes. Medians cannot be pooled at all.

Medians from Khanam 2019, Table 1. PMID 31662350. The comparison with 48.5 is ours.

Act II

How spread out it is

Means do pool, if you weight them

Two clinics measure the same thing.

	Patients	Mean
Clinic A	80	138
Clinic B	920	148

The simple average of the two means is 143. The correct combined mean is

$$\frac{80 \times 138 \;+\; 920 \times 148}{1000} \;=\; 147.2.$$

Both clinics are constructed by us. The arithmetic is ours.

Why 143 is wrong

Averaging the two means treats a clinic of 80 as though it carried the same weight as a clinic of 920.

Almost everybody in the combined group came from Clinic B, so the answer has to sit close to Clinic B's figure.

A centre on its own describes almost nothing

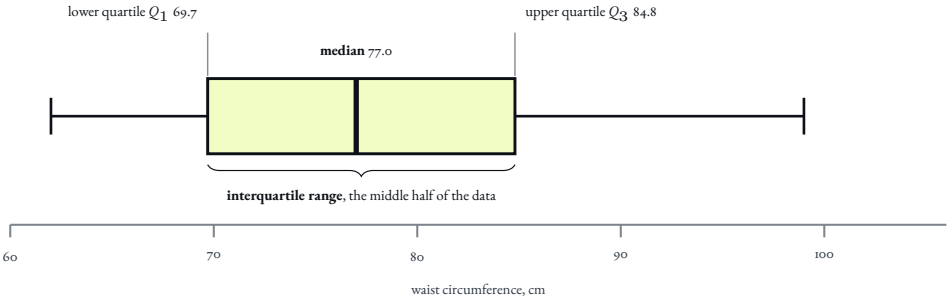
	Ward A	Ward B
Systolic pressures	138, 139, 140, 141, 142	110, 125, 140, 155, 170
Mean	140	140
Median	140	140

The same centre, two different wards

On Ward A nobody needs urgent attention. On Ward B somebody is at 170. Every summary of centre reports these two wards as identical.

Ten illustrative values, constructed by us so that both centres come out equal.

Quartiles, and the middle half of the data



Five numbers, and between them they say where the data sit and how spread out they are. The box holds the middle half; the line inside it is the middle observation. None of the five moves if a single extreme value moves further out.

Lower quartile, median and upper quartile as reported. Khanam 2019, Table 1. PMID 31662350. The whisker ends are illustrative: the paper does not report the minimum or the maximum.

The interquartile range, written as the subtraction it is

$$\text{IQR} \;=\; Q_3 - Q_1$$

Waist circumference, the same column as the previous frame:

$$Q_1 = 69.7, \qquad Q_3 = 84.8$$

$$\text{IQR} \;=\; 84.8 - 69.7 \;=\; 15.1 \text{ cm}$$

Quartiles are positions

Q_1 and Q_3 are not observed values in the usual sense, they are places in the sorted list. Moving the largest observation further out changes none of the five numbers in the summary.

Quartiles from Khanam 2019, Table 1. PMID 31662350. The subtraction is ours.

Quartiles are percentiles, and you already read those

The kth **percentile** is the value below which k per cent of the sample falls.

- The median is the 50th percentile.
- The lower quartile is the 25th, the upper quartile the 75th.

Nothing new has been introduced. Quartiles are three particular percentiles that turned out to be worth naming.

Where you use them every week

A growth chart is a set of percentiles drawn as curves.

Saying a child is on the third centile for weight is saying that three per cent of children of that age weigh less. It is a position in a distribution, not a measurement.

A percentile needs a reference population before it means anything, which is why a growth chart is always labelled with the population it was built from.

The five-number summary, as a reporting standard

Number	What it fixes	Waist, cm
Minimum	the floor, and any impossible value	not reported
Lower quartile	where the bottom quarter ends	69.7
Median	the middle	77.0
Upper quartile	where the top quarter begins	84.8
Maximum	the ceiling, and any impossible value	not reported

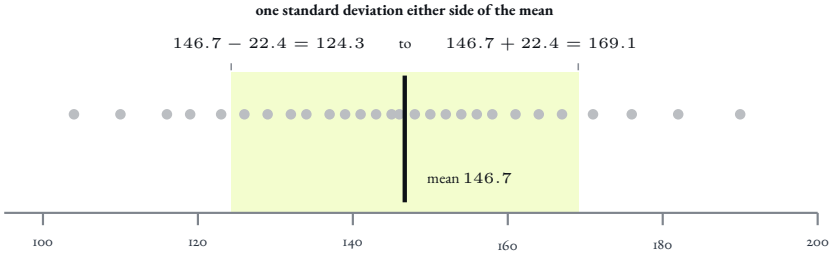
What five numbers buy that two do not

A median and an interquartile range describe the middle half and say nothing about the other half. Adding the minimum and the maximum tells a reader how far the tails reach, which is often the clinically interesting part.

Most papers print three of the five. Printing all five costs nothing and is what a box plot draws.

Quartiles and median from Khanam 2019, Table 1. PMID 31662350. The paper does not report the minimum or the maximum for this column.

The standard deviation, in the units of the thing measured



The standard deviation is a typical distance from the mean. It is in the same units as the measurement, so 22.4 here means mm Hg, and it says that most of these patients sat within about twenty-two points either side of 146.7. Exactly how many is a later session.

Mean and standard deviation as reported for the intervention group at baseline. Jafar 2020, PMID 32074419. Dot positions illustrative.

Where the standard deviation comes from, without the algebra

To measure spread around the mean, take each value's distance from it. Those distances cancel: the ones above are positive and the ones below negative, and they sum to zero by construction.

So square them first, which makes every one positive, and average the squares. That average is the **variance**.

The variance is in *squared* units, so a spread of blood pressures comes out in squared millimetres of mercury, which cannot be put on a blood pressure scale. Taking the square root undoes that, and the result is the standard deviation.

Why you will still meet the variance

Variances of independent quantities add and standard deviations do not, so almost every formula in the rest of this course is written in variances.

Every number a paper reports is a standard deviation, because that is the one in readable units.

Computing the variance, step by step

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The same five waist values, mean 78.0:

x_i	70	74	77	81	88
$x_i - \bar{x}$	-8	-4	-1	3	10
$(x_i - \bar{x})^2$	64	16	1	9	100

$$s^2 = \frac{64 + 16 + 1 + 9 + 100}{5 - 1} = \frac{190}{4} = 47.5$$

Same five illustrative waist values. All arithmetic here is ours.

Squared units

47.5 is in squared centimetres. It cannot yet be placed on the waist-circumference scale; that is the next frame.

Computing the standard deviation from the variance

$$s = \sqrt{s^2}$$

Carrying the variance from the previous frame:

$$s^2 = 47.5 \text{ cm}^2$$

$$s = \sqrt{47.5} \approx 6.9 \text{ cm}$$

Where that earlier table got its 6.9

This is the arithmetic behind "The spread that matches the median". The divisor $n - 1$ is a correction taken up in Session 7, not derived here.

Same five illustrative waist values. All arithmetic here is ours, and it reproduces the 6.9 already shown earlier in the session.

How far is one, two, three standard deviations

Same column, mean 78.0 cm, $s \approx 6.9$ cm:

	lower	upper	width
$\pm 1s$	71.1	84.9	13.8
$\pm 2s$	64.2	91.8	27.6
$\pm 3s$	57.3	98.7	41.4

Each band is just the mean plus or minus a multiple of s: $78.0 \pm 1(6.9)$, $78.0 \pm 2(6.9)$, $78.0 \pm 3(6.9)$.

What this is not

This is arithmetic, not a claim about how many participants fall inside each band. That claim needs an assumption about the column's shape, and it is Session 5.

Same five illustrative waist values. All arithmetic here is ours.

COBRA-BPS, systolic pressure at baseline:

	Mean	SD
Intervention	146.7	22.4
Control	144.7	21.0

The two means are close. So are the two standard deviations, and that second agreement is the one usually passed over.

What a matching spread says

Randomisation is supposed to produce two groups that resemble each other, and resembling each other means more than having the same average.

Two arms with the same mean and very different spreads would not be comparable, whatever the means said.

A Table 1 in a trial is not describing the patients for its own sake. It is evidence about whether the randomisation worked.

Jafar TH, et al. *N Engl J Med* 2020;382(8):717–726, Table 1. PMID 32074419.

Spread as a share of size

A standard deviation is in the units of the thing measured, so two columns measured in different units cannot be compared for spread. Divide it by its own mean and the units cancel:

$$\text{CV} = \frac{s}{\bar{x}} \times 100\%$$

COBRA-BPS baseline	mean ± SD	**CV**
Systolic BP	146.7 ± 22.4	15.3%
Diastolic BP	89.1 ± 14.7	16.5%
Age, years	58.8 ± 11.5	**19.6**%

Blood pressures in mm Hg.

Diastolic looks far less variable than systolic. Relative to its own size it is slightly more.

Why this is obvious to you already

Five mm Hg of wobble in a blood pressure is nothing.

Five kilograms of wobble in a baby's weight is everything.

Same number, and you have never once been confused by it.

Where it breaks

Only for measurements on a scale with a true zero, and only where the mean is comfortably above zero. A CV of a temperature in Celsius is meaningless.

Means and SDs from Jafar 2020 (COBRA-BPS), Table 1. PMID 32074419. All three CVs are **ours**.

Standard deviation is not standard error

Standard deviation

How spread out the **patients** are.

A property of the population being described. It does not shrink if you measure more people, because the people are as different as they are.

Standard error

How uncertain the **mean** is.

A property of the estimate. It does shrink as the study gets bigger, because a larger sample pins the mean down better.

Why it matters in print

Both are written with a plus-or-minus after a mean, and the standard error is always the smaller of the two.

A table that says "mean ±" without saying which one it means cannot be read, and reporting the standard error makes a study look more precise than it is.

Act III

Choosing, and the answer to the puzzle we opened with

A change is a column too, and it has its own centre and spread

COBRA-BPS does not report the pressures at 24 months. It reports the *fall*:

- Intervention: fell by 9.0 mm Hg
- Control: fell by 3.9 mm Hg
- Difference: 5.2 mm Hg (95% CI 3.2 to 7.1)

Each participant contributed one number, their own change, and the trial summarised that column exactly as this session has summarised any other.

Why the change is its own column

A change is a value per person, computed before any summarising. It is not the difference between two group means.

The two often come out similar and they are not the same quantity, because only the first one keeps track of who moved.

Jafar TH, et al. *N Engl J Med* 2020;382(8):717–726. PMID 32074419. The interval is the paper's.

You can audit the choice from the printed table alone

Column	Median (IQR), as printed	Lower half	Upper half	What that says
Waist circumference, cm	77.0 (69.7 to 84.8)	7.3	7.8	almost symmetric
Body mass index	20.3 (18.1 to 22.9)	2.2	2.6	mildly right-skewed
Age, years	48 (41 to 59)	7	11	right-skewed
Years of schooling	2 (1 to 5)	1	3	strongly right-skewed
Servings per day	1 (0 to 1)	1	0	piled against a floor

Nothing here needed the raw data. Subtract the lower quartile from the median, and the median from the upper quartile. Equal distances mean the middle half of the column is symmetric; unequal distances mean it is not, and the longer side is the direction of the tail.

Medians and quartiles as printed in the paper. The two distances and the verdict in the last column are our arithmetic.

So why did the Sylhet paper never print a mean?

Of the five numerical columns in its Table 1, by the check on the previous slide:

- two are close to symmetric
- two are clearly right-skewed
- one is piled against a floor

Reporting a mean for some rows and a median for others would make the table impossible to read down. So they chose one convention and applied it to everything.

And COBRA-BPS?

Its numerical columns are blood pressures and ages in a trial of people already known to have hypertension.

Those are symmetric, the mean is the right summary, and every method the trial uses to compare its groups is built on the mean.

Our reading of Khanam 2019, Table 1, using the quartiles the paper reports. PMID 31662350.

Reconstructing a mean you were never given

Khanam prints medians. COBRA-BPS prints means. The two cannot be compared, and neither can be pooled with the other. There is a published way out.

n, median, Q_1, Q_3 $\longrightarrow$ estimated mean and SD

Sylhet waist: $n = 1{,}810$, median 77.0, IQR 69.7 to 84.8.

The kind of estimate this is

You estimate a patient's weight from their build every week, and you act on it.

Close enough to use, with conditions, and never confused with having weighed them.

Method: Wan et al., *BMC Med Res Methodol* 2014;14:135, PMID 25524443, CC BY. Quartiles from Khanam 2019, Table 1. Both estimates are **ours**.

Reconstructing a mean you were never given

Khanam prints medians. COBRA-BPS prints means. The two cannot be compared, and neither can be pooled with the other. There is a published way out.

n, median, Q_1, Q_3 $\longrightarrow$ estimated mean and SD

Sylhet waist: $n = 1{,}810$, median 77.0, IQR 69.7 to 84.8.

Estimated mean **77.2** cm, estimated SD **11.2** cm.

The kind of estimate this is

You estimate a patient's weight from their build every week, and you act on it.

Close enough to use, with conditions, and never confused with having weighed them.

Method: Wan et al., *BMC Med Res Methodol* 2014;14:135, PMID 25524443, CC BY. Quartiles from Khanam 2019, Table 1. Both estimates are **ours**.

Reconstructing a mean you were never given

Khanam prints medians. COBRA-BPS prints means. The two cannot be compared, and neither can be pooled with the other. There is a published way out.

n, median, Q_1, Q_3 $\longrightarrow$ estimated mean and SD

Sylhet waist: $n = 1{,}810$, median 77.0, IQR 69.7 to 84.8.

Estimated mean **77.2** cm, estimated SD **11.2** cm.

Servings: $n = 1{,}810$, median 1, IQR 0 to 1. Estimated mean 0.67, SD 0.74. **Do not use this one.**

The kind of estimate this is

You estimate a patient's weight from their build every week, and you act on it.

Close enough to use, with conditions, and never confused with having weighed them.

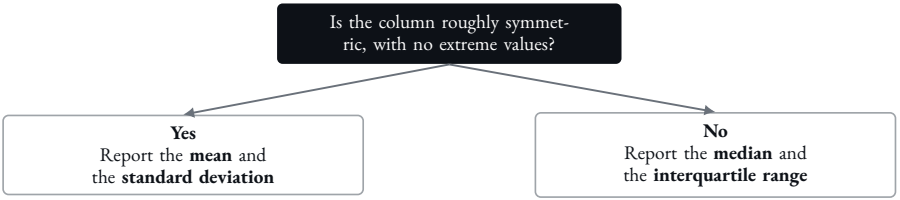
The condition, in the authors' own words

Nearly unbiased for *normal* data, and biased for *skewed* data.

Waist is near-symmetric, so the estimate is usable. Servings is piled on a floor of zero, so it is not.

Method: Wan et al., *BMC Med Res Methodol* 2014;14:135, PMID 25524443, CC BY. Quartiles from Khanam 2019, Table 1. Both estimates are **ours**.

The rule, and it is shorter than the discussion



Systolic blood pressure in COBRA-BPS, 146.7 ± 22.4 mm Hg. Symmetric, and the later methods of this course are built on the mean.

Years of schooling in the Sylhet survey, median 2, interquartile range 1 to 5. One long tail and a floor at zero.

And when in doubt, report both. They cost one line each, and a reader who has the mean, the standard deviation, the median and the quartiles can reconstruct most of the shape of the column without ever seeing it.

The question at the top is answerable from a published table alone, by comparing the two halves of the interquartile range. It does not require the raw data, and it does not require the normal distribution, which is a later session.

What to print, and in what form

If you report	Write it as	Example
Mean and SD	mean (SD), or mean ± SD, *said*	146.7 (22.4) mm Hg
Median and IQR	median (lower to upper quartile)	77.0 (69.7 to 84.8) cm
Either	with the unit, every time	mm Hg, cm, years

Two ambiguities worth avoiding

77.0 (69.7–84.8) could be an interquartile range, a full range, or a confidence interval. Say which.

146.7 ± 22.4 could be a standard deviation or a standard error. Say which.

Precision, units, and the digits you are entitled to

The rule

One decimal place more than the raw measurement, and no more.

A tape read to 0.1 cm supports a mean to 0.01 cm. It does not support 77.043.

And the unit, every time

146.7 is not a blood pressure. 146.7 mm Hg is.

A column heading carrying the unit once is enough, and is better than repeating it in every cell.

Where the extra digits come from

Software. A mean computed from 1,810 values arrives with fourteen decimal places and none of them were measured.

Rounding is the last step of the analysis, not the first, so compute at full precision and round only when printing.

The same discipline as Session 3's one decimal place on a percentage, one level along.

Act IV

When one number will not do the job

Stat.BD

Summaries that cannot be true

Some combinations of printed numbers are impossible whatever the study found. Four checks, none needing the raw data:

1. The median lies between Q_1 and Q_3.
2. Q_1 and Q_3 lie inside the range.
3. A bounded measurement's mean lies inside its bounds.
4. Bands and quantiles tell the same story.

The move, not the technique

A potassium of -2 does not get interpreted. It gets queried.

The check comes before the meaning.

Bands, medians and the sex breakdown from Khanam 2019, Table 1. The argument from 49.1% to a median of at least two is **ours**.

Summaries that cannot be true

Some combinations of printed numbers are impossible whatever the study found. Four checks, none needing the raw data:

1. The median lies between Q_1 and Q_3.
2. Q_1 and Q_3 lie inside the range.
3. A bounded measurement's mean lies inside its bounds.
4. Bands and quantiles tell the same story.

Check four, on the Sylhet servings column. The bands say 49.1% ate fewer than two servings. Fewer than half are below two, so the middle observation sits at two or above, so the median must be at least 2.

The printed median is **1**.

The move, not the technique

A potassium of -2 does not get interpreted. It gets queried.

The check comes before the meaning.

Bands, medians and the sex breakdown from Khanam 2019, Table 1. The argument from 49.1% to a median of at least two is **ours**.

Summaries that cannot be true

Some combinations of printed numbers are impossible whatever the study found. Four checks, none needing the raw data:

1. The median lies between Q_1 and Q_3.
2. Q_1 and Q_3 lie inside the range.
3. A bounded measurement's mean lies inside its bounds.
4. Bands and quantiles tell the same story.

Check four, on the Sylhet servings column. The bands say 49.1% ate fewer than two servings. Fewer than half are below two, so the middle observation sits at two or above, so the median must be at least 2.

The printed median is **1**.

The move, not the technique

A potassium of −2 does not get interpreted. It gets queried.

The check comes before the meaning.

And now the honest part

We cannot say which figure is wrong without the raw data.

The Methods say fruit and vegetables were combined before banding, so the two rows may simply count different things. By sex, the male rows agree and the female rows do not.

Bands, medians and the sex breakdown from Khanam 2019, Table 1. The argument from 49.1% to a median of at least two is **ours**.

When the summary is the wrong object entirely

Two clinics, both reporting a median waiting time of 30 minutes.

Clinic A: almost everybody waits between 25 and 35 minutes.
Clinic B: half the patients are seen in under 10 minutes and half wait more than an hour.

The second is not one distribution with a middle. It is two groups, and the median sits in the gap between them, describing neither.

The general warning

Every summary in this session assumes the column describes one population.

Where it describes two, the summary lands between them and is worse than useless, because it looks reasonable.

The only way to see this is to look at the distribution, which is Sessions 5 and 6.

Both clinics are constructed by us to make the point, and are not from any study.

The third option, and it is no longer deferred

A right-skewed column has a third move, besides mean and median: **change the scale**. Work with the logarithm of each value, average *that*, then convert back.

$$\text{geometric mean} \;=\; \exp\left(\frac{1}{n}\sum_{i} \ln x_i\right)$$

Antibody titres 2, 4, 8, 16	
Arithmetic mean	7.5
Geometric mean	**5.66**

On a $\log_2$ scale the titres are $1, 2, 3, 4$, evenly spaced. Their mean is 2.5, and $2^{2.5} = 5.66$.

You already read on a log scale

You write titres as 1:2, 1:4, 1:8, 1:16 and treat each step as one step.

Those are not evenly spaced numbers. They are evenly spaced *ratios*, which is what a log scale is.

Never label it a mean

A geometric mean is always smaller than the arithmetic mean of the same data, and papers do not always say which they report.

The four titres are illustrative. All arithmetic here is ours.

Common traps, collected

Reading

A mean quoted for a skewed column.
± after a mean with no word saying which.
A median averaged with another median.
A range compared between studies of different sizes.
Any summary of a column that is really two groups.

Writing

Check the two halves of the interquartile range before choosing.
Name the interval: SD, IQR or range.
Give the unit on every row.
Match the precision to the measurement.
When in doubt, report both.

A mean waist of 77.043 cm claims a precision the tape does not have. One decimal place more than the measurement, at most.

Close

Summarise your own columns.

Before the next session

One. Summarise every numerical column in your own study. For each one give a centre, a spread, the unit, and one line saying why you chose that pair rather than the other.

Two. Audit a published table. For every numerical row, apply the two-halves check to the quartiles, and say whether the summary the authors printed was the right one.

What the second task is for

You will find a row where a mean was printed for a column that is plainly skewed.

Bring it. It is the commonest reporting error in descriptive statistics and it is almost never deliberate.

Next: Session 5, Understanding Distributions & Outliers. The shape behind the summary.

References & sources

1. Khanam R, Ahmed S, Rahman S, Kibria GMA, Syed JRR, Khan AM, Moin SMI, Ram M, Gibson DG, Pariyo G, Baqui AH. Prevalence and factors associated with hypertension among adults in rural Sylhet district of Bangladesh: a cross-sectional study. *BMJ Open* 2019;9(10):e026722. PMID 31662350. PMC6830635. doi:10.1136/bmjopen-2018-026722. CC BY-NC: numbers re-typeset here as facts, text not reproduced, PDF not redistributed.
2. Jafar TH, Gandhi M, de Silva HA, Jehan I, Naheed A, Finkelstein EA, Turner EL, Morisky D, Kasturiratne A, Khan AH, Clemens JD, Ebrahim S, Assam PN, Feng L. A community-based intervention for managing hypertension in rural South Asia. *N Engl J Med* 2020;382(8):717–726. PMID 32074419. doi:10.1056/NEJMoa1911965.
3. Altman DG, Bland JM. Presentation of numerical data. *BMJ* 1996;312(7030):572. PMID 8595293. PMC2350327. doi:10.1136/bmj.312.7030.572.
4. Daniel WW. *Biostatistics: A Foundation for Analysis in the Health Sciences*, 9th ed. Wiley. Sections 2.4 and 2.5, on measures of central tendency and of dispersion.

Our arithmetic, and our constructions. The five waist values and both means; the two wards and their ten pressures; the two clinics; the comparison of 48.5 against the reported median age of 48; the two halves of every interquartile range and the verdicts drawn from them; and the reading of why each paper chose the convention it did. Everything else is as printed by the paper cited beside it.