Stat.BD

CLASS NOTES • CHAPTER 4

Describing Numerical Data

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

Applied Biostatistics & Research Methodology for Clinical Research
Stat.BD mentorship programme

Mentor: Muhtasim Munif Fahim, MSc

Chapter 4

Describing Numerical Data

Two published studies of blood pressure in South Asia describe their participants in opposite ways, and both are right. Working out how they each decided is the whole of this chapter.

Chapter 3 described the columns that name a group. A count in each category, a total that matches n, and a denominator stated plainly: that is a complete description, and nothing further can be extracted from a column of names.

This chapter takes the other family. A numerical column supports arithmetic, and that opens summaries the previous chapter could not use. It also opens a decision the previous chapter did not have. There is more than one way to say where a column sits, and more than one way to say how spread out it is. The choice between them is not free.

The decision is easier to see than to state, so the chapter is organised around a comparison. The survey used throughout Chapters 2 and 3 reports a median and an interquartile range for every numerical row and never once a mean. The trial used in Chapter 1 reports a mean and a standard deviation for every one and never a median. Neither is careless. Section 3 explains what each of them saw in their own data, and shows that the same judgement can be made by a reader from the printed table alone.

What this chapter establishes

1. That the mean, the median and the mode answer different questions, and that the mean is the only one of the three which assumes the gaps between adjacent values are equal quantities.
2. That the median is unaffected by how far an extreme value lies from the rest, and the mean is not, so the distance between them describes the shape of the column.
3. That a summary of centre without a summary of spread is half a description, and the less useful half.
4. That the interquartile range and the standard deviation measure spread in different ways, both in the units of the measurement.
5. That the standard deviation describes the participants while the standard error describes the estimate, and that the two are routinely confused because both are printed after a plus-or-minus sign.
6. That whether a mean or a median was the right choice can be judged from a published table alone, by comparing the two halves of the interquartile range.

Notation

n	the number of observations
x_i	the ith observation, $i = 1, \dots, n$
$\bar{x}$	the sample mean
s	the sample standard deviation
s^2	the sample variance
μ, σ	the corresponding population quantities, never observed
Q_1, Q_3	the lower and upper quartiles

Chapter 2 established the convention these symbols follow. Greek letters name parameters, which describe a population and are never seen; roman letters name statistics, which are computed from a sample and differ between samples.

1 Measures of centre

Three summaries answer the question of where a column sits, and they ask progressively more of the data.

Definition 1.1 • Mean, median and mode

The *mean* is the sum of the observations divided by how many there are. It requires that the values be numbers whose differences are meaningful quantities.

The *median* is the middle observation when the values are sorted, or the average of the two middle observations when n is even. It requires only that the values can be put in order.

The *mode* is the value or category occurring most often. It requires nothing at all.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.1}$$

$\bar{x}$ the sample mean, in the same units as the measurement
x_i the ith observation
n the number of observations

Robust
Not much affected by a small number of extreme values. The median is robust; the mean is not.

The ladder matters more than the definitions. Chapter 3 established that a nominal column admits only a mode, and an ordinal column a mode and a median. The mean is the top rung, and the extra thing it asks for is the one that can be wrong: it treats the distance from 1 to 2 as the same quantity as the distance from 2 to 3. For a wealth tertile that is false, which is why a mean tertile is meaningless however easily a computer will produce one.

Intuition

The mean is a balance point. Every observation pulls on it in proportion to how far away it sits, so one value a long way out can move it a long way. The median only counts how many observations lie on each side, so it does not notice how far away any

of them are.

1.1 Computing a mode

Count how often each value occurs. The mode is the value with the highest count; nothing is sorted and nothing is summed.

Example 1.1 • A repeated value

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

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

Counting each value gives 0 once, 1 twice, 2 three times and 4 once. The value 2 occurs most often, so the mode is 2.

The seven values are constructed here so that a repeated value exists. A continuous measurement such as waist circumference rarely repeats at all, which is why a mode computed from one is close to useless: almost every recorded value occurs exactly once.

Watch out

Two values can tie for the highest count, giving a column two modes rather than one. A mode is reported as however many values achieve the highest count, not forced to a single number.

1.2 Computing a median

Sort the values, then take the middle one. Where n is odd there is a single middle observation; where n is even there is not, and the median is the average of the two values either side of the centre.

Example 1.2 • Both cases

Five lengths of stay, in days: 2, 3, 4, 5, 41. Sorted, the third value is the middle one, so the median is 4 days.

Six lengths of stay: 2, 3, 3, 4, 5, 41. The third and fourth values are 3 and 4, so the median is $(3 + 4)/2 = 3.5$ days.

Two things are worth noticing. The median of an even-sized sample can be a value that appears nowhere in the data, and 3.5 days is a perfectly good median. And in both cases the 41 has no influence at all, which is Section 1.3 arriving from another direction. Both sets of values and the arithmetic are ours.

Watch out

Sort first. A median taken from an unsorted list is not a median, and a spreadsheet will compute one without complaining. This is a real error rather than a pedantic one: it happens whenever a column is sorted for display and the median is taken from a

different range.

1.3 What a single extreme value does

Example 1.3 • One mistyped digit

Five waist measurements, in centimetres, constructed here so that the arithmetic is easy to follow:

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

The sum is 390, so by Equation 1.1 the mean is $390/5 = 78.0$ cm. The middle value of the sorted list is 77, so the median is 77 cm. The two agree closely, as they will whenever a column is roughly symmetric.

Now suppose the last value was entered as 188 rather than 88, a transposition of the kind that occurs in every dataset:

$$70, \quad 74, \quad 77, \quad 81, \quad 188.$$

The sum becomes 490 and the mean becomes 98.0 cm. The median is still 77 cm, because the largest value is still the largest value and the middle one has not moved.

A single keystroke moved the mean by twenty centimetres and the median not at all. Both the values and the arithmetic here are ours; they are constructed to isolate the property.

The point

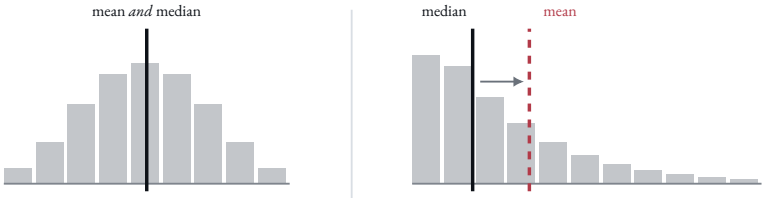
Data entry errors are the commonest reason a real dataset contains an impossible value, and they affect the mean without limit and the median not at all. This is a practical argument for looking at the minimum and maximum of every column before computing anything.

1.4 The gap between mean and median

Skewed
Having a longer tail on one side than the other. A column with a long right tail is right-skewed, and its mean exceeds its median.

Figure 1 carries a rule a reader can apply immediately: the further apart a reported mean and a reported median, the more skewed the column, and the mean lies on the side of the tail. Where a paper prints both, the shape of the column can be inferred without ever seeing it.

Columns with a floor and no ceiling are skewed to the right as a matter of course, and clinical data is full of them: length of stay, waiting time, number of admissions, cost of an episode, time to an event. None of these can fall below zero, all of them can be arbitrarily large, and a mean computed from one of them sits above the experience of most of the people it describes.



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.

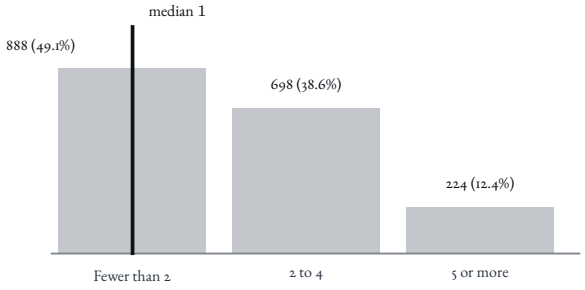
Figure 1: Two illustrative distributions. Where a column is symmetric the mean and the median coincide; where it has a long tail on one side, the mean is pulled towards the tail and the median is not.

Example 1.4 • A column against its floor

The Sylhet survey recorded servings of fruit and vegetables per day. It reports 888 participants (49.1%) eating fewer than two, 698 (38.6%) eating two to four, and 224 (12.4%) eating five or more, with a median of 1 and an interquartile range of 0 to 1.

A mean of this column would sit somewhere between the bars, describing a number of servings nobody was recorded as eating. Worse, the column has a hard floor at zero and no ceiling: one participant reporting twenty servings pulls the mean upward, and nobody can pull it back, because nobody can report a negative number of servings.

Nothing here is a defect of the mean as a quantity. It is the wrong summary for this column, and the paper's authors evidently thought so too.



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.

Figure 2: The same column drawn as counts, with the reported median marked. Counts, percentages and the median are from reference 1..

1.5 Medians cannot be pooled

A mean can be combined across groups if the group means and the group sizes are both known, because the mean is a quantity built by adding. A median cannot, because it is a position in a sorted list rather than a quantity.

Example 1.5 • Why the arithmetic fails

The Sylhet survey reports a median age of 50 years among men and 47 among women. The average of the two is 48.5. The reported median for the whole sample is 48.

The two do not agree, and there is no reason they should. Obtaining the median of the combined sample requires sorting all 1,810 ages and taking the middle one, which only somebody holding the data can do. The comparison with 48.5 is ours; the paper prints the three medians.

Example 1.6 • The pooling that does work

Two clinics measure the same quantity. Clinic A has 80 patients with a mean of 138; Clinic B has 920 with a mean of 148.

The average of the two means is 143, and it is wrong, because it treats a clinic of 80 as carrying the same weight as a clinic of 920. The combined mean is obtained by weighting each group mean by its size:

$$\frac{80 \times 138 \;+\; 920 \times 148}{80 + 920} \;=\; \frac{11{,}040 + 136{,}160}{1{,}000} \;=\; 147.2.$$

The answer sits close to 148 because almost everybody in the combined group came from Clinic B. Both clinics and the arithmetic are constructed here.

Watch out

This is one reason trials reporting medians are harder to combine in a systematic review than trials reporting means, and it is a reason to report both when a column is borderline. Meta-analysis is the subject of Chapter 29.

Weighted mean
A mean in which each value contributes in proportion to a weight, here the size of the group it came from.

2 Measures of spread

A centre on its own describes almost nothing.

In practice

Two wards, five systolic pressures each, constructed here to make the point.

Ward A: 138, 139, 140, 141, 142. Ward B: 110, 125, 140, 155, 170.

Both have a mean of 140 and a median of 140. On the first ward nobody needs urgent attention; on the second somebody is at 170. Every measure of centre reports the two wards as identical.

2.1 The range

The *range* is the largest value minus the smallest. On Ward B it is $170 - 110 = 60$ mm Hg.

It has two defects and both are visible immediately. It uses exactly two of the observations and discards everything between them, so it is determined entirely by the two most unusual members of the sample. And it grows with sample size: a larger study finds a more extreme value simply by looking at more people, so the ranges of two studies of different sizes cannot be compared.

The point

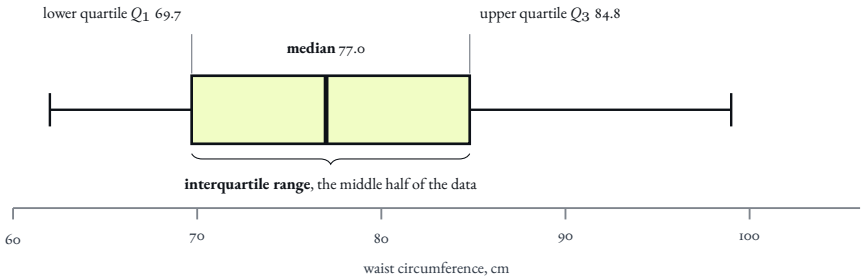
The minimum and the maximum are nevertheless worth inspecting for every column, not as a summary but as a check. An age of -3 or a systolic pressure of 600 is a data entry error, and the range is where it becomes visible.

2.2 Quartiles and the interquartile range

Definition 2.1 • Quartiles and the interquartile range

The *lower quartile* Q_1 is the value below which a quarter of the observations fall; the *upper quartile* Q_3 is the value below which three quarters fall. The median is the value below which half fall.

The *interquartile range* is $Q_3 - Q_1$, the width of 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.

Figure 3: The five-number summary of a real column. The quartiles and the median are as reported in reference 1.; the whisker ends are illustrative, because the paper does not report a minimum or a maximum.

Example 2.1 • Waist circumference in the Sylhet survey

The survey reports a median waist circumference of 77.0 cm with an interquartile range of 69.7 to 84.8 cm.

The interquartile range is therefore $84.8 - 69.7 = 15.1$ cm; this subtraction is ours. Read into the room: take the 1,810 participants, discard the slimmest quarter and the widest quarter, and the 905 who remain span fifteen centimetres.

Like the median, the quartiles are positions. Moving the largest observation further out changes none of the five numbers in the summary.

2.3 Percentiles

Percentile
The value below which a stated percentage of the sample falls. The median is the 50th; the quartiles are the 25th and the 75th.

Quartiles are three particular percentiles that proved worth naming, and nothing new is introduced by the general term. Clinicians read percentiles fluently already: a growth chart is a set of percentile curves, and a child on the third centile for weight is one below whom three per cent of children of that age and sex fall.

That familiar example also carries a warning from Chapter 2. A percentile is meaningless without the reference population it was computed from. That is why a growth chart is always labelled with the population it was built on, and why a WHO chart and a national chart give different centiles for the same child.

2.4 The five-number summary

Taken together, the minimum, the lower quartile, the median, the upper quartile and the maximum are the *five-number summary*, and they are exactly what a box plot draws.

A median with an interquartile range gives three of the five and describes the middle half. Adding the minimum and the maximum tells the reader how far the tails reach, which is often the clinically interesting part, and it is where an impossible value becomes visible. Most papers print three of the five; printing all five costs nothing.

2.5 Comparing spread between groups

A spread is also evidence about whether two groups resemble each other, which matters most in a randomised trial.

In practice

COBRA-BPS reports baseline systolic pressures of 146.7 ± 22.4 mm Hg in the intervention group and 144.7 ± 21.0 in the control group.

The two means are close, and so are the two standard deviations. The second agreement is usually passed over and is doing real work: two arms with identical means but standard deviations of, say, 22 and 8 would not be comparable groups, whatever the means said. A trial's descriptive table is not describing its patients for their own sake. It is evidence about whether the randomisation produced two groups that resemble each other.

Watch out

Testing baseline differences between trial arms for statistical significance is discouraged, and the reasons are the subject of Chapter 13. Reading the table is not the same as testing it.

2.6 Robust summaries: trimming, and the MAD

Section 1.3 showed one mistyped digit moving a mean by twenty centimetres and leaving the median untouched. That is the whole case for robustness, and it leaves a practical question. The median ignores the sizes of the values entirely. Is there anything between a summary

hostage to its extremes and one that discards them?

There are two, and both are used in clinical research.

Definition 2.2 • Trimmed mean

Order the n values, discard the lowest k and the highest k, and take the ordinary mean of what remains.

$$\bar{x}_{\text{tr}} = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} x_{(i)} \tag{2.1}$$

$x_{(i)}$ the ith value once the sample has been put in order, so $x_{(1)}$ is the smallest

k how many values are discarded from *each* end. Usually set as a percentage: a 20% trimmed mean discards the lowest and highest fifth

$n-2k$ what is left, and the divisor

A trimmed mean is not one quantity but a family indexed by k, so an unqualified "trimmed mean" is not a complete description. The trim must be stated.

Definition 2.3 • Median absolute deviation

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

$$\text{MAD} = \text{median}_i\left(|x_i - \tilde{x}|\right) \tag{2.2}$$

$\tilde{x}$ the median of the sample

$|x_i - \tilde{x}|$ how far value i sits from that median, without regard to direction

MAD the middle of those distances, and a measure of spread on the same footing as the median is a measure of centre

Example 2.2 • The same five values, and the same typo

The five invented waist measurements of Example 1.3 were 70, 74, 77, 81 and 88 cm, with the largest later mistyped as 188.

A trim of one value from each end leaves 74, 77 and 81, whose mean is 77.3 cm. The mistyped value is one of the two discarded, so it never enters the arithmetic, and the trimmed mean is 77.3 cm with the typo present or absent.

For the MAD, the median is 77. The distances are 7, 3, 0, 4 and 11; sorted, their median is 4. With the typo the distances become 7, 3, 0, 4 and 111, whose median is still 4, because a single large value is still only one entry in a list of five.

	clean	with the typo
Mean	78.0	98.0
Trimmed mean	77.3	77.3
Median	77.0	77.0
Standard deviation	6.9	50.5
MAD	4.0	4.0

The standard deviation rises more than sevenfold. It is built from squared distances,

Example 2.2 • *continued*

and an error eleven times too large becomes more than a hundred times too large once squared. The MAD takes a middle rather than a sum, so no single value can dominate it. Every figure in this example is computed here; the five values are invented.

Watch out

Software frequently reports a *scaled* MAD, multiplied by about 1.4826 so that it matches the standard deviation for normally distributed data. A printed MAD may therefore not be the raw median of the distances, and the documentation has to be checked before the number is compared with anything.

2.7 Spread relative to size: the coefficient of variation

A standard deviation carries the units of the thing measured, which is what makes it readable and also what prevents two columns measured differently from being compared. Dividing by the mean removes the units.

Definition 2.4 • Coefficient of variation

$$\mathrm{CV} = \frac{s}{\bar{x}} \times 100\% \tag{2.3}$$

s the sample standard deviation, in the units of the measurement
$\bar{x}$ the sample mean, in the same units, which is why the ratio has none
CV spread expressed as a percentage of the average size of the thing being measured

Example 2.3 • Three columns from COBRA-BPS

Baseline measurement	mean ± SD	**CV**
Systolic blood pressure	146.7 ± 22.4	15.3%
Diastolic blood pressure	89.1 ± 14.7	16.5%
Age, years	58.8 ± 11.5	19.6%

Blood pressures are in mm Hg. Read the standard deviations alone and systolic pressure is much the most variable of the three, at 22.4 against 14.7 and 11.5. Relative to its own size it is the least variable. The means and standard deviations are reported by the trial; all three coefficients are computed here.

Intuition

Five units of variation means nothing until the size of the thing is known. Five mm Hg of movement in a blood pressure is negligible; five kilograms of movement in an infant's weight is not. The coefficient of variation is that judgement made arithmetic.

Watch out

The coefficient of variation requires a measurement scale with a true zero, and a mean comfortably away from it. A temperature in degrees Celsius has an arbitrary zero, so its coefficient of variation changes if the same temperatures are expressed in Fahrenheit, and the quantity is meaningless. Where the mean approaches zero the ratio becomes unstable whatever the scale.

2.8 Variance and standard deviation

The quartiles measure spread by position. The standard deviation measures it by distance from the mean, and it is the summary of spread that the rest of this course is built on.

The construction has three steps and each has a reason. Take each observation's distance from the mean. Those distances sum to zero by construction, since the ones above the mean are positive and the ones below negative, so squaring them first makes every one positive. The average of the squares is the *variance*. The variance is in squared units, which cannot be placed on the original measurement scale, so taking the square root returns the summary to the units of the data. That square root is the *standard deviation*.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad s = \sqrt{s^2} \tag{2.4}$$

s^2 the sample variance, in squared units of the measurement

s the sample standard deviation, in the units of the measurement

$x_i - \bar{x}$ one observation's distance from the mean

$n-1$ the divisor. It is $n-1$ rather than n because the mean was itself estimated from the same data; the reason is taken up in Chapter 7

Intuition

The standard deviation is a typical distance from the mean, expressed in the units of the thing measured. A standard deviation of 22.4 mm Hg means that a typical patient sat about twenty-two points of mercury away from the average.

Example 2.4 • Computing a standard deviation from raw values

The same five waist measurements as Section 1.3, mean 78.0 cm:

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

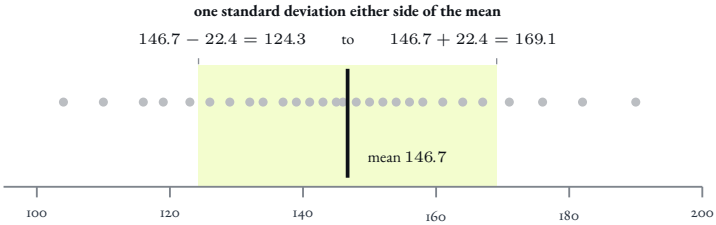
Each value's distance from the mean, then that distance squared:

$$-8, -4, -1, 3, 10 \qquad \longrightarrow \qquad 64, 16, 1, 9, 100.$$

The squares sum to 190. By Equation 2.4, with $n = 5$:

$$s^2 = \frac{190}{5-1} = \frac{190}{4} = 47.5, \qquad s = \sqrt{47.5} \approx 6.9 \text{ cm}.$$

This 6.9 is the same figure reported for the clean column in Section 2.6's spread comparison; this is the arithmetic behind it. All five values and the arithmetic are ours.



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.

Figure 4: A reported mean and standard deviation, drawn on the scale of the measurement. Mean and standard deviation as reported for the intervention group at baseline in reference 3.; the dot positions are illustrative.

Example 2.5 • Baseline blood pressure in COBRA-BPS

The trial reports a baseline systolic blood pressure of 146.7 ± 22.4 mm Hg in the intervention group.

One standard deviation either side of the mean spans $146.7 - 22.4 = 124.3$ to $146.7 + 22.4 = 169.1$ mm Hg; both subtractions are ours. Most of the participants sat within that band.

How many is most, expressed as a percentage, cannot be said from these two numbers alone. It follows from an assumption about the shape of the distribution, and that assumption is the subject of Chapter 5. It is left unstated here deliberately, because quoting a figure for it now would attach a rule to a case that has not been shown to satisfy it.

Watch out

The variance is not a competing description of a dataset. It carries exactly the same information as the standard deviation and is in units that cannot be read off a measurement scale. It survives in the literature because variances of independent quantities add and standard deviations do not, so the algebra of later chapters is written in variances while the tables are written in standard deviations.

2.9 Standard deviation and standard error

Two quantities are routinely printed after a mean and a plus-or-minus sign, and they describe different things.

Definition 2.5 • Standard deviation against standard error

The *standard deviation* describes how spread out the observations are. It is a property of the population being described, and it does not shrink as the sample grows, because

Definition 2.5 • *continued*

the participants are as different from one another as they are.

The *standard error* describes how uncertain an estimate is. It is a property of the estimate, and it does shrink as the sample grows, because a larger sample locates the mean more precisely.

The point

The standard error is always the smaller of the two, so reporting it in place of a standard deviation makes a study appear more precise than it is. Where a table prints a mean with a plus-or-minus and does not say which quantity follows, it cannot be read.

The relationship between the two, and the reasoning that makes the standard error the foundation of every interval in this course, is the subject of Chapter 7. This chapter establishes only that they are different quantities describing different things.

3 Choosing a summary

Everything so far has described the tools. The decision is which pair to report, and it can be made, and audited, from a published table.

3.1 A check that needs no raw data

A median divides a column in half, and the quartiles divide each half again. If the distance from Q_1 up to the median equals the distance from the median up to Q_3, the middle half of the column is symmetric. If the two distances differ, the column is skewed, and the longer side points in the direction of the tail.

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.

Figure 5: The check applied to every numerical column of one published table. Medians and quartiles as printed in reference 1.; the two distances and the verdict in the final column are ours.

Example 3.1 • Reading five columns from their quartiles

Waist circumference, median 77.0 with quartiles 69.7 and 84.8: the lower half spans 7.3 cm and the upper half 7.8 cm. Almost equal, so the column is close to symmetric and a mean would have been a defensible summary.

Age, median 48 with quartiles 41 and 59: the halves are 7 and 11 years. The upper half is half as long again, so the column is skewed to the right.

Years of schooling, median 2 with quartiles 1 and 5: the halves are 1 and 3 years. Three times as far on the upper side, and with a floor at zero, so strongly right-skewed.

Servings per day, median 1 with quartiles 0 and 1: the halves are 1 and 0, because the median and the upper quartile are the same number. That is what a column piled against a floor looks like in a printed table.

All four subtractions are ours. The medians and quartiles are the paper's.

Watch out

This is a reader's check and not a formal test of skewness. It uses only the middle half of the data and says nothing about the tails beyond the quartiles. Its virtue is that it needs nothing except what was printed.

3.2 Why each paper chose what it chose

Of the five numerical columns in the Sylhet survey's descriptive table, two are close to symmetric, two are clearly right-skewed, and one is piled against a floor. Reporting a mean for the symmetric rows and a median for the rest would make the table impossible to read down, since a reader moving down a column would have to check the convention at every line. Choosing one convention and applying it throughout is the ordinary solution, and the convention that works for all five columns is the median.

The trial reports means for a different and stronger reason. Its numerical columns are ages and blood pressures among people already known to have hypertension, and those are close to symmetric. More importantly, its entire analysis is a comparison of mean blood pressures between two groups, so a descriptive table in medians would describe one quantity and analyse another.

The point

Report the mean and the standard deviation when the column is roughly symmetric and free of extreme values, and the median and the interquartile range otherwise. When the case is borderline, report both: it costs two lines and it makes the study usable by somebody combining it with others later.

This reading of the two papers is ours. Neither explains its choice, and almost none do.

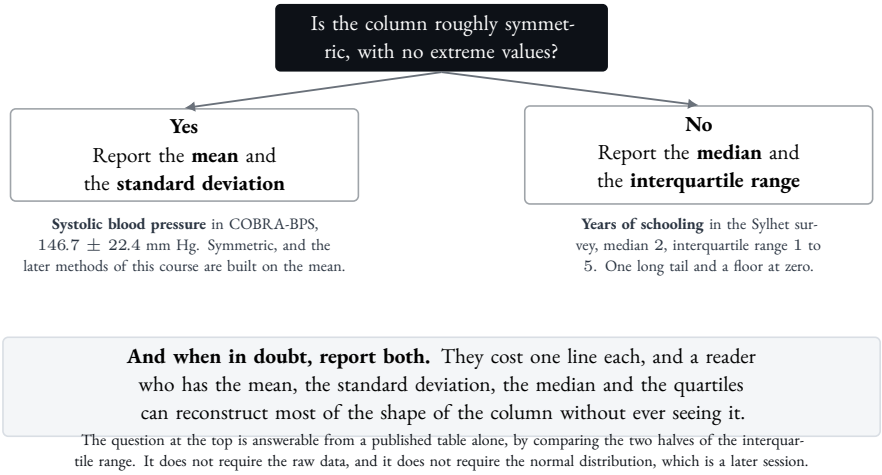


Figure 6: The decision, and the real column that goes down each branch.

3.3 Describing a change

A change is a numerical column like any other, and it is summarised the same way. What is worth being careful about is where it comes from.

Example 3.2 • What COBRA-BPS actually reports

The trial does not report blood pressures at 24 months. It reports the fall: systolic pressure fell by 9.0 mm Hg in the intervention group and by 3.9 in the control group, a difference of 5.2 mm Hg (95% CI 3.2 to 7.1). All four figures are the paper's.

Each participant contributed one number to that summary, their own change, computed before any summarising took place. In the language of Chapter 2, the change is a constructed column: it was built by subtracting one measured column from another, row by row.

Watch out

A mean change is not the same quantity as the difference between two group means. The two usually agree numerically and they answer different questions, because only the first follows individuals. Which to use, and what to do about the correlation between a person's two measurements, are the subjects of Chapters 20 and 26.

3.4 Recovering a mean from a median

The two anchor papers of this course cannot be compared. Khanam reports medians with quartiles; COBRA-BPS reports means with standard deviations. Section 1.5 established that medians cannot be pooled and Section 3.2 established why each paper chose as it did, and both left the practical problem untouched.

A published estimator closes part of the gap. Given a sample size, a median and both quartiles, it returns an approximate mean and standard deviation.

Definition 3.1 • Estimating a mean and SD from a five-number summary

For a sample of size n reported as median m with quartiles q_1 and q_3,

$$\bar{x} \approx \frac{q_1 + m + q_3}{3} \tag{3.1}$$

$$s \approx \frac{q_3 - q_1}{\eta(n)}, \qquad \eta(n) = 2\,\Phi^{-1}\!\left(\frac{0.75n - 0.125}{n + 0.25}\right) \tag{3.2}$$

m, q_1, q_3 the reported median and lower and upper quartiles

n the sample size, which is why it must be reported alongside the quartiles

Φ^{-1} the inverse standard normal distribution function. It is tabulated and built into every statistical package, and nothing here requires evaluating it by hand

$\eta(n)$ a divisor that depends only on the sample size. It approaches 1.35 as n grows, so for any sizeable sample $s \approx \text{IQR}/1.35$

The estimator's justification is that a five-number summary constrains where the mean and spread can lie, tightly if the distribution is close to symmetric and loosely otherwise. That condition is not a technicality, and the authors state it themselves: the method gives a nearly unbiased estimate of the standard deviation for normal data and a biased one for skewed data.

Example 3.3 • One column where it works and one where it does not

Both columns are from the Sylhet survey, where $n = 1{,}810$, so $\eta(n) = 1.348$.

Waist circumference, median 77.0 cm with quartiles 69.7 and 84.8. The two halves of the interquartile range are 7.3 and 7.8 cm, so the column is close to symmetric, and Section 3.1 classified it as such. Then

$$\bar{x} \approx \frac{69.7 + 77.0 + 84.8}{3} = 77.2 \text{ cm}, \qquad s \approx \frac{84.8 - 69.7}{1.348} = 11.2 \text{ cm}.$$

Servings of fruit and vegetables, median 1 with quartiles 0 and 1. The same arithmetic returns an estimated mean of 0.67 and a standard deviation of 0.74. These should not be used. The column is piled against a floor of zero, which Section 1.4 showed makes a mean a fiction, and the condition the estimator needs is exactly the one this column fails.

Every figure in this example is computed here. The paper reports the quartiles and the sample size only.

The point

The estimate is worth having and is not a measurement. It occupies the same position as estimating a patient's weight from their build: close enough to act on, stated as an estimate, and never written down as though the thing had been measured. A paper using it says so, and says which estimator.

Pooling results across studies, which is what this estimator exists to make possible, is the subject of Chapter 29.

3.5 Changing the scale

A third option exists for a right-skewed column, besides reporting a mean and reporting a median.

Working with the logarithm of each value instead of the value itself often turns a column with a long right tail into one that is close to symmetric, at which point the mean becomes usable again. The summary is then in log units, which nobody can read, and converting it back gives a quantity called the *geometric mean*.

Definition 3.2 • Geometric mean

$$\text{GM} = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\ln x_i\right) \tag{3.3}$$

x_i the ith value, which must be strictly positive. A column containing a zero has no geometric mean, since the logarithm of zero is undefined

$\ln x_i$ its natural logarithm. Any base works provided the same base is used to convert back

$\exp(\cdot)$ the conversion back to the original units, so that the result is readable

Intuition

Some quantities move by multiplication rather than addition. Antibody titres are written 1:2, 1:4, 1:8, 1:16 and each step is treated as one step, although the gaps between the numbers double every time. Those are evenly spaced *ratios*, and a log scale is the scale on which evenly spaced ratios become evenly spaced numbers.

Example 3.4 • Four titres

Take titres of 2, 4, 8 and 16. Their arithmetic mean is 7.5.

On a base-2 log scale they are 1, 2, 3 and 4, evenly spaced, with mean 2.5. Converting back, $2^{2.5} = 5.66$, which is the geometric mean.

The two answers differ because the arithmetic mean is pulled towards 16 while the geometric mean treats each doubling as one step. The four titres are illustrative and the arithmetic is ours.

Watch out

Viral loads, antibody titres, hormone concentrations, costs and durations are routinely analysed on a log scale, and papers do not always say so. Where a paper reports a mean for a column that is plainly skewed, it is worth asking which mean. Immunogenicity studies report geometric mean titres as a matter of course, usually abbreviated to GMT and rarely defined.

Deciding *when* a column is skewed enough to justify a transformation belongs to Chapter 5, which takes up shape, and transformation in the service of a regression belongs to Chapter 24. This section covers what the quantity is and how to read it.

Geometric mean
The mean of the log-transformed values, converted back. Always smaller than the arithmetic mean of the same data unless every value is identical, and not interchangeable with it.

3.6 Reporting

If you report	Write it as	Example
Mean and SD	mean (SD), with SD named	146.7 (22.4) mm Hg
Median and IQR	median (Q_1 to Q_3)	77.0 (69.7 to 84.8) cm
Either	with the unit, every time	mm Hg, cm, years

Two ambiguities are common and both are removed by one word. A bracketed pair after a median may be an interquartile range, a full range, or a confidence interval, and the typography does not distinguish them. A plus-or-minus after a mean may be a standard deviation or a standard error. Name the quantity in the column heading, as Chapter 3 recommended for the direction of a percentage.

Precision follows the measurement. A mean waist circumference of 77.043 cm claims a resolution the tape does not have; one decimal place more than the raw measurement is the usual limit.

3.7 Summaries that cannot be true

Some combinations of printed summary statistics are impossible whatever the study found, and noticing them requires no raw data and no software. Four checks cover most of it.

1. The median lies between the first and third quartiles.
2. Both quartiles lie inside the reported range.
3. The mean of a bounded measurement lies inside its bounds. A percentage cannot average -3, and a count cannot average less than zero.
4. Where a column is reported twice, once as bands and once as quantiles, the two tell the same story.

The first three are arithmetic and rarely fail in a published paper. The fourth is a consistency check between two rows, and it is the one that finds things.

Example 3.5 • The servings column, reported twice

The Sylhet survey reports servings of fruit and vegetables per day in two ways. As bands: 49.1% took fewer than two servings, 38.6% took two to four, and 12.4% took five or more. As a summary: median 1, interquartile range 0 to 1.

Take the bands. If 49.1% of the sample lies below two servings, then fewer than half the sample lies below two, so the middle observation cannot lie below two. The median must therefore be at least 2.

The reported median is 1.

By sex the picture differs. For males the two agree: 52.8% below two and a median of 0, which is consistent, because more than half being below two puts the middle observation below two. For females, 45.7% below two against a median of 1, and for the total, the two do not reconcile.

What follows from this, and what does not. It does not follow that either figure

Example 3.5 • *continued*

is wrong. Without the raw data neither can be checked. The Methods record that the fruit and vegetable data were combined before banding, which leaves the likeliest explanation open: the two rows may count slightly different things. What does follow is that the two rows cannot both be read as summaries of one column, and that a reader who quotes the median beside the bands is quoting two different quantities.

The argument from 49.1% to a median of at least two is ours. The bands, the medians and the figures by sex are the paper's.

The point

The check is the same move as querying a potassium of -2 rather than interpreting it: a value that cannot be true is not evidence about the patient, it is evidence about the record. Arriving at a question rather than a verdict is the normal outcome, and a reader who expects verdicts will stop checking.

3.8 When no summary is appropriate

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

In practice

Two clinics, both reporting a median waiting time of 30 minutes. In the first, almost everybody waits between 25 and 35 minutes. In the second, half the patients are seen within ten minutes and half wait more than an hour.

The second is not one distribution with a middle. It is two groups, and the median falls in the gap between them, describing neither. Both clinics are constructed here to make the point.

Watch out

A summary compresses a column into one or two numbers, and compression assumes there is a single thing to compress. Where two populations have been mixed, every summary in this chapter returns a plausible and misleading answer. Nothing taught here would reveal it; only looking at the distribution would, which is the subject of Chapter 5, with the displays that make it visible in Chapter 6.

4 Chapter summary

1. The mode requires nothing of a column, the median requires an order, and the mean requires that the gaps between values be equal quantities.
2. The mean is a balance point and is moved without limit by an extreme value. The median is a position and is not moved at all.

3. Where a mean and a median are both reported, the distance between them describes the skew, and the mean lies towards the tail.
4. Columns with a floor at zero and no ceiling are right-skewed as a matter of course, and clinical data is full of them.
5. Means can be pooled across groups by weighting each group mean by its size. Medians cannot be pooled at all.
6. A change is a column in its own right, computed per person before any summarising, and it is not the same quantity as a difference between two group means.
7. A centre without a spread is half a description. The range is a data cleaning check rather than a summary.
8. The interquartile range is the width of the middle half. The standard deviation is a typical distance from the mean, in the units of the measurement.
9. The variance carries the same information as the standard deviation in squared units, and survives because the algebra of later chapters needs it.
10. The standard deviation describes the participants; the standard error describes the estimate. Both are printed after a plus-or-minus and a table must say which.
11. Whether the right summary was chosen can be judged from a printed table by comparing the two halves of the interquartile range.

5 Key terms

Term	Meaning
Mean, $\bar{x}$	The sum of the observations divided by how many there are.
Median	The middle observation of the sorted values.
Mode	The value or category occurring most often.
Robust	Not much affected by a small number of extreme values.
Trimmed mean	The mean of what is left after discarding an equal number of the largest and smallest values. A family of quantities, so the trim must be stated.
Median absolute deviation	The median of the distances of the values from their median. Stands to the median as the standard deviation stands to the mean.
Coefficient of variation, CV	The standard deviation as a percentage of the mean. Unitless, so it compares spread across different measurements. Needs a true zero.
Geometric mean, GM	The mean of the logged values, converted back. Always smaller than the arithmetic mean of the same data, and never labelled as one.
Skewed	Having a longer tail on one side. Right-skewed means the mean exceeds the median.
Range	Largest minus smallest. A check, not a summary.
Quartiles, Q_1, Q_3	The values below which a quarter and three quarters of the observations fall.
Interquartile range	$Q_3 - Q_1$, the width of the middle half.
Five-number summary	Minimum, Q_1, median, Q_3, maximum.
Percentile	The value below which a stated percentage falls.
Variance, s^2	The average squared distance from the mean, in squared units.
Standard deviation, s	The square root of the variance, in the units of the measurement.
Standard error	How uncertain an estimate is. Shrinks as the sample grows. Chapter 7.
Weighted mean	A mean in which each value contributes in proportion to a weight, usually a group size.
Geometric mean	The mean of log-transformed values, converted back. Always smaller than the arithmetic mean of the same data.

6 Exercises

THE NUMBERS YOU WILL NEED

From the Sylhet survey (reference 1.), medians with interquartile ranges, whole sample of 1,810: age 48 (41 to 59) years; body mass index 20.3 (18.1 to 22.9); waist circumference 77.0 (69.7 to 84.8) cm; years of schooling 2 (1 to 5); servings of fruit and vegetables per day 1 (0 to 1). Median age was 50 in men and 47 in women. Servings were also reported in bands: fewer than two 888 (49.1%), two to four 698 (38.6%), five or more 224 (12.4%).

From COBRA-BPS (reference 3.), means with standard deviations, 2,645 adults in 30 villages: age 58.8 ± 11.5 years; baseline systolic blood pressure 146.7 ± 22.4 mm Hg in the intervention group and 144.7 ± 21.0 in the control group; baseline diastolic

89.1 ± 14.7 and 87.8 ± 13.8.

Part A. Centre

1. For each of the following, state which of the mean, median and mode can be computed: blood group; wealth tertile; waist circumference in cm; years of schooling recorded in bands.

2. Six lengths of stay, in days: 2, 3, 3, 4, 5, 41. Compute the mean and the median. Then state which of the two you would report to a hospital manager, and why.

3. A paper reports a mean length of stay of 8.4 days and a median of 4 days. Without any further information, describe the shape of that column and say roughly what is producing it.

4. Two wards report median waiting times of 20 and 30 minutes. A colleague concludes that the median across both wards is 25 minutes. Explain the error and state what information would be needed to obtain the true figure.

Part B. Spread

5. Compute the interquartile range of waist circumference and of body mass index from the figures in the box above. State what each one means in words.

6. Give the interval spanned by one standard deviation either side of the mean baseline systolic pressure in the COBRA-BPS control group. State what can, and what cannot, be said about how many participants fall inside it.

7. Two trials each report a mean baseline systolic pressure of 146.7 mm Hg. The first reports a standard deviation of 22.4 and the second of 6.0. Describe how the two sets of participants differ.

8. A table reports "systolic blood pressure 146.7 ± 1.2 mm Hg" for a trial of 2,645 participants. Which quantity is almost certainly being reported, and how can you tell?

Part C. Choosing

9. Apply the two-halves check to the age column and to the years of schooling column. Give both distances for each, and state which is the more skewed.

10. For each of the five numerical columns in the box above, say whether a mean or a median is the better summary, and give your reason in one clause.

11. Section 3.7 runs the bands-against-quantiles check on the servings column and finds the two rows do not reconcile. Run the same check on the *education* column, which is reported as three bands (18.4, 62.2 and 19.4 per cent) and as a median of 2 years with an interquartile range of 1 to 5. Do the two rows agree? Show the reasoning that settles it.

12. Rewrite each of these so that it cannot be misread: ”age 48 (41–59)”; ”systolic pressure 146.7 ± 22.4”; ”waist circumference 77.043”.

Part D. Appraisal

13. Take Table 1 of any paper in your own field. Find one numerical row where you believe the wrong summary was reported, and justify your view using only what the table prints.

14. Describe a measurement in your own practice whose distribution you would expect to have two peaks. State why a single mean or median would describe nobody.

15. A reviewer asks an author to report both the mean and the median for a skewed column. Give two distinct reasons why that is a reasonable request.

7 Solutions

Part A

1. Blood group: mode only, since the categories have no order.

Wealth tertile: mode and median, since the categories are ordered but the gaps between them are not equal quantities.

Waist circumference in cm: all three, though the mode is close to useless because almost every recorded value occurs once.

Years of schooling in bands: mode and median. A mean cannot be computed from banded data at all, since the bands do not record how many years each participant had.

2. The sum is $2 + 3 + 3 + 4 + 5 + 41 = 58$, so the mean is $58/6 = 9.67$ days. With six values the median is the average of the third and fourth, $(3 + 4)/2 = 3.5$ days.

Report the median, or both. The mean of 9.67 days is longer than five of the six stays and describes none of them; it is being carried by the single stay of 41 days. A manager planning bed occupancy needs both figures, because the long stay is real and consumes beds, but a mean presented alone would misrepresent the typical patient.

3. The mean is more than twice the median, so the column is strongly skewed to the right, with the mean pulled towards a long upper tail.

Length of stay has a floor at one day and no ceiling. A minority of patients with complications, or awaiting placement, stay for weeks, and those few observations carry the mean upwards while leaving the median where the bulk of patients sit.

4. Medians cannot be averaged. A median is a position in a sorted list rather than a quantity that can be combined arithmetically, so the average of two medians has no interpretation.

Obtaining the true combined median requires the individual waiting times from both wards, sorted together. Knowing only the two medians and the two ward sizes is not sufficient, which is a difference from the mean: a pooled mean can be computed from group means and group sizes by weighting.

Part B

5. Waist circumference: $84.8 - 69.7 = 15.1$ cm. The middle half of the sample, once the slimmest quarter and the widest quarter are set aside, spans fifteen centimetres.

Body mass index: $22.9 - 18.1 = 4.8$ kg/m², with the same interpretation.

6. $144.7 - 21.0 = 123.7$ to $144.7 + 21.0 = 165.7$ mm Hg.

What can be said is that this is the band within one typical distance of the mean, and that most participants lie inside it. What cannot be said, from these two numbers alone, is what percentage do. Attaching a percentage requires an assumption about the shape of the distribution, which is the subject of Chapter 5.

7. The centres are identical, so a table of means alone would report the two trials as having recruited the same patients.

The first trial's participants are widely spread: a typical patient sits about twenty-two points from the mean, so the trial contains people near 120 and people near 170. The second trial's participants are tightly clustered within about six points of 146.7, which suggests a narrow entry criterion, a more homogeneous population, or both.

8. Almost certainly the standard error. Blood pressure varies between people by far more than 1.2 mm Hg, so 1.2 cannot be describing the spread of participants. It is instead describing how precisely the mean has been located, which improves with sample size, and 2,645 participants is a large sample.

The general tell is the same: a plus-or-minus that is small relative to the biological variability of the measurement is a standard error, and a table that does not say which quantity it prints cannot be read.

Part C

9. Age: $48 - 41 = 7$ years below the median, $59 - 48 = 11$ years above. The upper half is half as long again, so the column is right-skewed.

Years of schooling: $2 - 1 = 1$ year below, $5 - 2 = 3$ years above. The upper half is three times the lower.

Years of schooling is by some way the more skewed of the two, which is what would be expected of a column with a hard floor at zero where a large minority of the sample sits.

10. Waist circumference: mean, because the two halves of the interquartile range are almost equal.

Body mass index: mean, for the same reason, though the difference of 2.2 against 2.6 makes it a borderline case where reporting both would be defensible.

Age: median, because the upper half of the interquartile range is half as long again as the lower.

Years of schooling: median, because the column is strongly right-skewed and has a floor at zero.

Servings per day: median, because the median and the upper quartile are the same number, which means the column is piled against its floor.

11. They agree, and setting out why is the point of the exercise.

Work down the bands as a cumulative distribution. Nobody with no schooling can be above 0 years, so the lowest 18.4% of the sample sits at 0. Adding the next band, $18.4 + 62.2 = 80.6\%$ of the sample has five years of schooling or fewer. The 1 to 5 band therefore spans everything from the 18.4th percentile to the 80.6th.

The median is the 50th percentile and the third quartile is the 75th, and both fall inside that span. So both must lie between 1 and 5 years. The reported median is 2 and the reported third quartile is 5. Both are inside. The first quartile is the 25th percentile, also inside the band, and it is reported as 1. Nothing contradicts anything.

Contrast this with the servings column of Section 3.7, where the same procedure produces a median that must be at least 2 against a reported median of 1. The check is identical; only the outcome differs, and a check that passes is as informative as one that fails.

The cumulative figure 80.6% and the reasoning are ours. The bands, the median and the quartiles are the paper's.

12. "Age, median 48 years (interquartile range 41 to 59)."

"Systolic blood pressure, mean 146.7 mm Hg (standard deviation 22.4)."

"Waist circumference, mean 77.0 cm." The two extra decimal places claim a precision the tape measure does not have, and one decimal place beyond the raw measurement is the usual limit.

Part D

13. No single answer. The expected finding is a mean and standard deviation reported for a count or a duration, most often length of stay, waiting time, number of visits, or cost. The justification available from the table alone is that the standard deviation is comparable to or larger than the mean, which for a quantity with a floor at zero implies a long upper tail and therefore skew.

14. No single answer. Common examples include time to presentation in a service with both walk-in and referred patients, birth weight in a population combining term and preterm deliveries, and any measurement taken on two protocols. In each case the distribution has two peaks and the single summary lands between them, describing a patient who does not exist.

15. Two of several. The mean and the median together describe the shape of the column, because the distance between them indicates the direction and degree of skew, so reporting both tells the reader something neither does alone.

And a systematic review combining this study with others will need whichever summary the other studies reported; a paper that gives only one may be excluded from a meta-analysis years later for want of two lines of table.

8 References and further reading

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. Published under CC BY-NC; its figures are re-typeset here as facts and none of its text is reproduced.
2. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Med Res Methodol* 2014;14:135. PMID 25524443. PMC4383202. doi:10.1186/1471-2288-14-135. Published under CC BY. The source of equations (3.1) and (3.2), and of the condition stated beneath them.
3. 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.
4. Altman DG, Bland JM. Presentation of numerical data. *BMJ* 1996;312(7030):572. PMID 8595293. PMC2350327. doi:10.1136/bmj.312.7030.572.
5. Daniel WW. *Biostatistics: A Foundation for Analysis in the Health Sciences*, 9th ed. Wiley. Section 2.4 on measures of central tendency, book pp. 38–43; Section 2.5 on measures of dispersion, book pp. 43–53.

Figures computed or constructed in this chapter rather than reported by a paper. The five waist values and both means; the five and six lengths of stay and their medians; the two clinics for the weighted mean, and the figure 147.2; the two wards and their pressures; the two waiting-time clinics; the interquartile ranges 15.1 and 4.8; the comparison of 48.5 against the reported median age of 48; the bands 124.3 to 169.1 and 123.7 to 165.7; the two halves of every interquartile range and the verdicts drawn from them; the consistency check on the servings column; and the reading of why each paper chose the convention it did.

Added in the September 2026 revision, and all computed here: the trimmed means 77.3, the median absolute deviations 4.0, and the standard deviations 6.9 and 50.5 of Section 2.6; all three coefficients of variation in Section 2.7; the reconstructed mean 77.2 and standard deviation 11.2 for waist and 0.67 and 0.74 for servings in Section 3.4, with the divisor $\eta(1810) = 1.348$; the four illustrative titres of Section 3.5 and the geometric mean 5.66; and the cumulative figure 80.6% used in the solutions.

9 For students in the mentorship programme

Everything above stands on its own. This section is the only part of the chapter that assumes a session.

Before the next session

One. Summarise every numerical column in your own study: 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. Expect to find a mean reported for a column that plainly has a floor at zero; bring that row.

Reading

Daniel sections 2.4 and 2.5 (reference 5.) cover both families of summary formally. Altman and Bland (reference 4.) is one page and covers how to print them.

Questions this chapter deliberately left open

Question	Where it is settled
What percentage of observations lie within one standard deviation?	Chapter 5, the normal distribution
When is an extreme value an outlier rather than data?	Chapter 5, outliers
How are these summaries drawn rather than tabulated?	Chapter 6, tables and graphs
How is the standard error related to the standard deviation, and why $n - 1$?	Chapter 7, variability and precision
How are two means compared between groups?	Chapter 20, comparing two groups
Once a mean has been reconstructed from a median, how are results from different studies actually combined?	Chapter 29, systematic reviews
When is a column skewed enough to justify working on a log scale?	Chapter 5 for the shape; Chapter 24 where a regression needs it
What should be done about an outlier once it has been identified?	Chapter 19, choosing a method

Next

Chapter 5, Understanding Distributions & Outliers. Every summary in this chapter compressed a column into two numbers. The next asks what the column actually looks like, and when that compression hides something that matters.