Stat.BD

CLASS NOTES • CHAPTER 5

Understanding Distributions & Outliers

What a column looks like, the model it may be close to, and the values out in its tails

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

Mentor: Muhtasim Munif Fahim, MSc

Chapter 5

Understanding Distributions & Outliers

A column of numbers has a shape, and most of what can honestly be said about it depends on that shape. This chapter draws the columns that Chapter 4 could only summarise.

Chapter 4 compressed every numerical column into two numbers, a centre and a spread, and judged five columns' shapes from four printed values each. That judgement was made without ever looking at a column. This chapter looks.

It proceeds in four parts. The first draws a column and names what can be seen in it: symmetry, skew and the mechanisms that produce skew, and the two-peaked shape that no summary can describe. The second introduces the normal distribution as a working model, and gives the share of a column that lies within one and two standard deviations of its mean together with the condition under which that share holds. It then gives the much weaker guarantee that holds with no condition at all. The third turns to the tails of a column: percentiles, reference ranges, and the values that sit far from the rest. The fourth shows how the shape of a column decides which summary to report and, later in the course, which method of analysis to use.

What this chapter establishes

1. That a histogram is the column itself, drawn, and that the two-halves check of Chapter 4 was a reading of its middle.
2. That skew has mechanisms: a floor the measurement cannot pass, a floor the study imposed, a ceiling, and a mixture of two populations.
3. That the normal distribution is a model some columns are close enough to for some purposes, never a fact about biological data.
4. That about 68% of a column lies within one standard deviation of its mean and about 95% within two *if the column is close to normal*, and that with no assumption about shape nothing at all can be guaranteed within one standard deviation.
5. That a reference range is the central 95% of a healthy reference group, so one healthy person in twenty falls outside it on any single test.
6. That an outlier is a question with three possible answers, and that no value is ever removed without a rule written in advance and a statement of what the removal changed.

Notation

$X \sim N(\mu, \sigma^2)$	X is modelled as normal, with mean μ and variance σ^2
$\bar{x}$, s	the sample mean and standard deviation
z	a standardised score: distance from the mean in standard deviations
k	a number of standard deviations
Q_1, Q_3	the lower and upper quartiles; $Q_3 - Q_1$ is the interquartile range
G_1	the sample skewness coefficient, as printed by SPSS
$\Pr(\cdot)$	the probability of the event in brackets

A NOTE ON THE PICTURES IN THIS CHAPTER

No study used in this course publishes the individual measurements behind its tables. The histograms and Q-Q plots in this chapter are therefore drawn from *simulated* columns, generated to reproduce exactly the medians and quartiles that Khanam and colleagues printed for their Sylhet survey (reference 1.), from fixed starting values so that the same pictures are produced every time. The middle of each simulated column agrees with the paper. Everything the paper did not print, above all the tails, was chosen here. The pictures show what such a shape looks like; none of them is evidence about the Sylhet sample.

Contents

1 The shape of a column

1.1 A histogram is the column, drawn

A histogram divides the range of a measurement into bands of equal width and draws, over each band, a bar whose height is the number of observations falling in it. Nothing is summarised and nothing is lost except the position of each observation within its band. The picture is the column: every row of the dataset, stood on its end and sorted into piles. A clinician who writes the ages of every patient on a ward on slips of paper and stacks them in five-year piles has built a histogram without the word.

The horizontal axis carries the measurement and its units; the vertical axis carries a count, or a percentage of the sample if the caption says so. Two things are read from it first. The *centre* is where the bulk of the observations sit. The *shape* is how they are arranged around it: balanced on both sides, or with a longer tail on one side, or piled into two separate heaps.

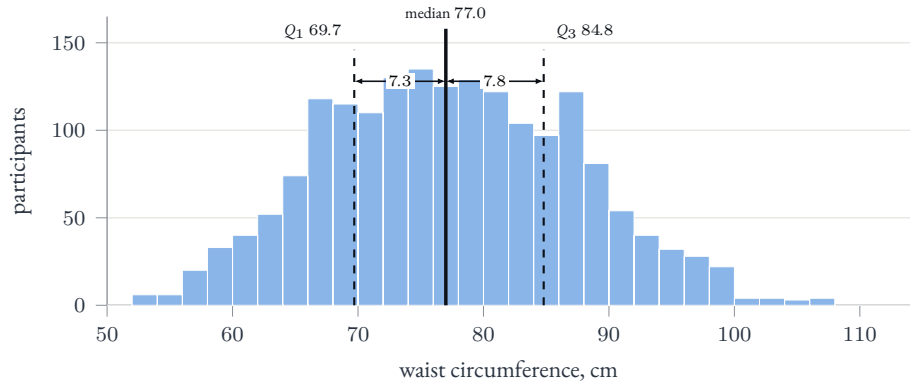

Figure 1: Waist circumference, 1,810 adults. **Simulated** to match the median and quartiles printed in reference 1., Table 1; the tails are our choice. The two arrows are the two halves of the interquartile range, 7.3 and 7.8 cm.

Intuition

The two-halves check of Chapter 4 compared the distance from the lower quartile to the median with the distance from the median to the upper quartile. Figure 1 shows what it was measuring: the width of the two halves of the central heap. Equal widths mean the middle of the column is balanced, and the check needed only four printed numbers to see it.

Example 1.1 • Three Sylhet columns, drawn

Chapter 4 read three columns from their printed quartiles: waist, with halves of 7.3 and 7.8 cm; age, with halves of 7 and 11 years; and years of schooling, with halves of 1 and 3. Its verdicts were almost symmetric, right-skewed, and strongly right-skewed.

Figure 2 draws simulated columns matched to those quartiles. Waist is balanced about its median. Age rises steeply from its lowest value and trails away above 60. Schooling is piled at zero and one and thins out towards twelve years. Every verdict survives. The

Example 1.1 • *continued*

check cannot see the tails, which is its limit; it did not need to see them to name the shape.

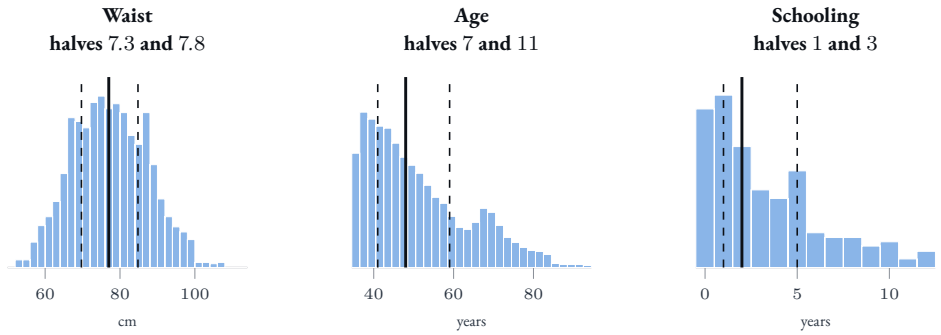

Figure 2: Three columns, **simulated** to match the medians and quartiles of reference 1., Table 1 (schooling also to its printed bands). Dashed lines are the quartiles, the solid line the median.

Watch out

A histogram's shape depends on the width of its bands, which software chooses silently. Too narrow and the picture is noise; too wide and real features vanish. Chapter 6 treats the choice in full.

1.2 Where skew comes from

A column is *skewed* when one of its tails is longer than the other. *Right skew*, a long tail towards high values, is the commoner of the two in clinical data, and it nearly always has an identifiable mechanism. Naming the mechanism is more useful than naming the skew, because the mechanism says what to do about it.

The first mechanism is a **floor the measurement cannot pass**. Counts, concentrations and durations cannot fall below zero and have no ceiling. A clinician who watches length of stay on a medical ward sees most patients discharged in three or four days and one who stays ninety; nobody stays minus ninety days to balance them. That asymmetry is the whole of right skew, and years of schooling in Figure 2 is the same shape for the same reason.

The second is a **floor the study imposed**. The Sylhet survey enrolled adults aged 35 and over, so its age column begins at a wall the investigators chose. A reader who concluded from that column's shape that rural Sylhet has a young population would be reading the design of the study, not the population. Any study with an inclusion threshold on the variable being described does the same: a trial that enrols patients with an HbA1c above 7.5% will have a baseline HbA1c column piled just above 7.5.

The third is a **ceiling**. Scores with a maximum, such as a Glasgow Coma Scale of 15, pile against the top and trail downwards. The resulting *left skew* is less common in clinical columns. Gestational age at delivery is its clearest example. Most births fall near 39 or 40 weeks, with a practical ceiling not far beyond, and preterm births form a long tail to the left. The fourth mechanism, a mixture of two populations, is taken up in Section 1.4.

The point

In every skewed column the mean lies on the side of the tail: above the median under right skew, below it under left skew. The median stays with the bulk of the observations.

1.3 Skewness as a number

Statistical software summarises the lean of a column in a single coefficient. The version printed by SPSS under its descriptive procedures is the adjusted Fisher-Pearson coefficient:

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} \cdot \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}} \tag{1.1}$$

G_1 the sample skewness: zero for a symmetric column, positive for a longer right tail, negative for a longer left tail

$x_i - \bar{x}$ each observation's distance from the mean. Cubing keeps the sign, so distances in the long tail dominate the numerator

n the number of observations; the first factor is a small-sample correction

SPSS prints a standard error beside the coefficient, and for a sample of n it is

$$\mathrm{SE}(G_1) = \sqrt{\frac{6\,n\,(n-1)}{(n-2)(n+1)(n+3)}} \tag{1.2}$$

$\mathrm{SE}(G_1)$ how much the skewness coefficient would vary between samples of this size from a symmetric population. It shrinks as n grows

Example 1.2 • Skewness at a large sample size

At $n = 1{,}810$, equation (1.2) gives a standard error of 0.058 (ours). The simulated columns of Figure 2 have skewness coefficients of 0.14 for waist, 0.84 for age and 1.02 for schooling, and ratios to the standard error of 2.4, 14.7 and 17.7. All three figures are from simulated columns.

A ratio above 2 is often read as "significantly skewed". By that reading the waist column is significantly skewed, although its histogram is visibly close to symmetric. Both statements are true. The ratio answers whether the column is exactly symmetric, and with enough observations no real column is. The useful question is whether it is skewed enough to matter for the purpose in hand.

Watch out

There is no universally agreed threshold for "too skewed", and a threshold quoted without a purpose attached should be treated with suspicion. Formal tests of normality share the same defect at large sample sizes and are discussed in Chapter 19. For describing a column, looking at it is more informative than testing it.

1.4 Two peaks mean two populations

Some columns have two heaps separated by a gap. The pattern almost always means that two different groups have been pooled into one column, and every summary of centre then describes a patient who does not exist.

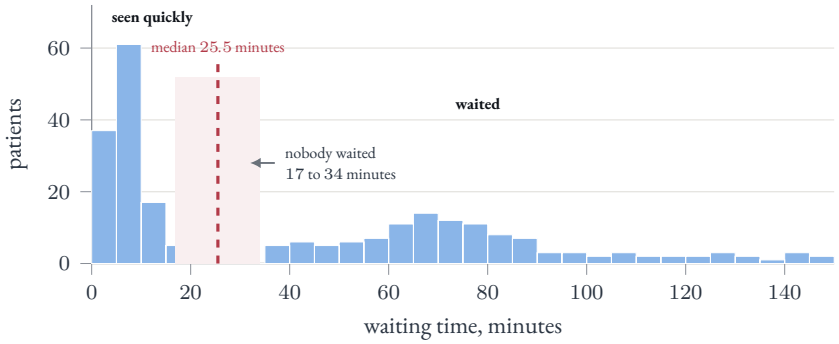

Figure 3: Waiting times at a clinic whose patients are either seen almost at once or wait more than half an hour. **Constructed** for illustration; not from any study.

Example 1.3 • A median in the gap

The clinic of Figure 3 has a median waiting time of 25.5 minutes. No patient waited between 17 and 34 minutes: the median falls in the empty gap between the two groups. The mean, 42 minutes, describes nobody either.

A clinician recognises the pattern at once. Triage produces it: patients who need urgent attention are seen within minutes and the rest wait. The constructive step is to find the variable that separates the two groups, here the triage category, split the column by it, and describe each part separately. Each part then usually has an ordinary shape.

Watch out

Two peaks are invisible in every summary of Chapter 4 and, as Chapter 6 shows, in a box plot. Only a picture of the whole column reveals them, which is a reason to draw every column before summarising it.

2 The normal distribution as a working model

2.1 What the model is

The *normal distribution* is a single symmetric, single-peaked shape, fixed entirely by two numbers: where its centre lies and how spread out it is. Its curve is the familiar bell, falling away equally on both sides and never reaching zero in either tail.

Definition 2.1 • The normal distribution

A quantity X is modelled as normally distributed with mean μ and standard deviation σ, written $X \sim N(\mu, \sigma^2)$, when the relative frequency of its values follows

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2.1}$$

$f(x)$ the height of the curve at the value x; areas under it are proportions of the population
μ the centre of the curve, which is also its mean and median
σ the standard deviation, which sets the width of the bell
σ^2 the variance, the form in which the model is conventionally written

The equation is not needed to use the model; what is needed is to know that the whole shape follows from μ and σ, so a column that is close to normal is fully described by its mean and standard deviation. That is why Chapter 4's decision tree sent symmetric columns to the mean and standard deviation.

Intuition

The normal distribution is a model, not a description of biology. No clinical measurement is exactly normal. Some are close enough for some purposes, and the purpose decides what "close enough" means.

The builders of the charts clinicians use most took exactly this stance. When the World Health Organization constructed its child growth standards, it examined the shape of each indicator rather than assuming it. Length and height for age turned out close to normal, while the other indicators needed their skew modelled before the curves could be drawn (reference 6.).

2.2 Why so many columns come out roughly normal

A single blood pressure reading is the result of many small influences at once: genes, salt, sleep, the cuff, the time of day, the walk to the clinic. None dominates, each pushes the reading a little up or down, and the pushes add. Quantities built from many small, independent additive influences tend to pile into a bell shape. That is the intuition behind the result named in Chapter 7 as the central limit theorem.

Other quantities grow by multiplication. A bacterial count doubles; a drug concentration halves with each half-life; an antibody titre rises in steps of two. Quantities built by many small multiplicative influences come out skewed to the right in their raw units and close to normal on a logarithmic scale, which Section 2.9 draws. Columns piled against a floor are neither.

The point

Before looking at a column, ask how it is made. Additive columns are usually close to normal; multiplicative columns are usually close to normal after logging; floored counts and mixtures are neither.

2.3 The share within one and two standard deviations, and its condition

If a column is close to normal, fixed shares of it lie within fixed distances of the mean, measured in standard deviations.

$$\Pr(\mu - k\sigma \le X \le \mu + k\sigma) \approx \begin{cases} 0.683 & k = 1 \\ 0.954 & k = 2 \\ 0.997 & k = 3 \end{cases} \qquad \text{if } X \sim N(\mu, \sigma^2) \tag{2.2}$$

$\Pr(\cdot)$ the proportion of the population satisfying the condition in brackets
k the number of standard deviations either side of the mean
if $X \sim N$ the condition. Without it, the right-hand side does not hold

The three figures are usually quoted as 68, 95 and 99.7 per cent. The exact multiple of σ that encloses 95% of a normal distribution is 1.96 rather than 2, a figure that returns in Chapter 8.

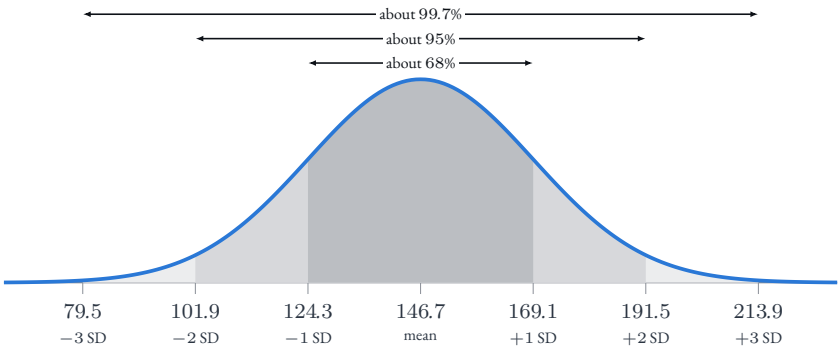

Figure 4: The normal model drawn on the scale of the measurement, for the mean and standard deviation reported by reference 2. for its intervention arm. The band endpoints are ours and assume a normal shape the paper does not report.

Example 2.1 • Bands on a reported mean and standard deviation

COBRA-BPS reports a baseline systolic pressure of 146.7 ± 22.4 mm Hg in its intervention arm of 1,330 adults (reference 2.). The bands, all computed here:

Band	Arithmetic	If normal, contains
±1 SD	146.7 ± 22.4 = 124.3 to 169.1	about 68%
±2 SD	146.7 ± 44.8 = 101.9 to 191.5	about 95%
±1.96 SD	146.7 ± 43.9 = 102.8 to 190.6	95%
±3 SD	146.7 ± 67.2 = 79.5 to 213.9	about 99.7%

Chapter 4 described the first band as containing "most" participants and declined to say how many. The figure is now available, and so is the condition: about two thirds, *provided* the column is close to normal. The paper printed a mean and a standard deviation and made no claim about shape.

2.4 What survives skew and what does not

The condition matters more for some statements than for others, and the difference is worth knowing precisely.

	Within 1 SD	Within 2 SD	Below -2 SD	Above $+2$ SD
Normal model	68%	95%	2.3%	2.3%
Waist (simulated)	66%	96%	1.2%	2.5%
Age (simulated)	65%	95%	0.0%	4.5%
Schooling (simulated)	66%	94%	0.0%	5.9%
Clinic B (constructed)	82%	95%	0.0%	4.6%

Example 2.2 • Reading the table

The first two columns are close to the textbook figures even for the skewed columns: the middle of a single-peaked column is forgiving. The last two columns are not. The normal model places the observations beyond two standard deviations evenly in the two tails. In the schooling column nobody lies more than two standard deviations below the mean, since two standard deviations below a mean of 3.25 years is a negative number of years. All of the outlying observations are in the upper tail. Any statement about a tail, such as "the top two and a half per cent", is badly wrong if taken from the normal model for such a column. The two-peaked clinic breaks even the total within one standard deviation, because both of its heaps sit close to one standard deviation from the mean. Every share in the table is computed here from simulated or constructed columns.

2.5 The guarantee that needs no condition

If nothing at all is assumed about the shape of a column, a much weaker statement still holds for every column that exists.

$$\Pr(\,|X-\mu| \geq k\sigma\,) \;\leq\; \frac{1}{k^2} \qquad \text{for any distribution and any } k > 1 \tag{2.3}$$

$|X-\mu| \geq k\sigma$ lying k or more standard deviations from the mean, in either direction
$1/k^2$ the largest share that can do so, whatever the shape

This is *Chebyshev's inequality*. For $k = 2$ it says that at least $1 - \frac{1}{4} = 75\%$ of any column lies within two standard deviations of its mean; for $k = 3$, at least $1 - \frac{1}{9} = 89\%$. For $k = 1$ the bound is 1, which guarantees nothing.

Intuition

Clinicians make two kinds of statement. "Most patients with this condition recover within a week" is a prognosis: better, and true only for the patients it was built on. "Nobody survives without a heartbeat" is a guarantee: weak, and always true. The normal figures of equation (2.2) are the first kind; Chebyshev's inequality is the second.

The point

This is why Chapter 4 would not say what share lies within one standard deviation. Without an assumption about shape, there is no such figure to give.

2.6 Standardised scores

A *standardised score*, or z-score, expresses a value as its distance from the mean in standard deviations.

$$z = \frac{x - \bar{x}}{s} \tag{2.4}$$

z the standardised score, with no units
x the value being standardised
$\bar{x}, s$ the mean and standard deviation of the reference group

Removing the units is the point. Once columns are expressed in standard deviations, a waist circumference and a blood pressure can be placed on one axis, and a child's height and weight can be compared with each other. Growth charts do exactly this: the lines on them are z-scores (reference 6.).

Example 2.3 • Two standardised scores in clinical use

In COBRA-BPS, a participant with a baseline systolic pressure of 180 mm Hg has $z = (180 - 146.7)/22.4 = 33.3/22.4 = 1.49$ (ours): about one and a half standard deviations above the trial's average participant.

Bone density is reported the same way. The reference standard describes osteoporosis as a bone mineral density 2.5 standard deviations or more below the average for young adult women. It recommends as the reference group women aged 20 to 29 in the NHANES III survey (reference 5.). A T-score is a z-score against that group.

Watch out

A z-score is a distance from a stated reference group, and it changes when the group changes. A seventy-year-old's T-score compares her with women fifty years younger, deliberately. And a z-score converts into a percentile only through the normal model: $z = 1.49$ is about the ninety-third percentile if the column is normal, and says nothing about percentiles if it is not.

2.7 Reading a normal Q-Q plot

A *normal quantile-quantile plot* places each percentile of a column against the same percentile of a normal distribution, both in standard deviations. If the column is normal the two agree and the points lie on a straight line; departures from the line show where the column departs from the model. SPSS draws one when normality plots are requested in its Explore procedure.

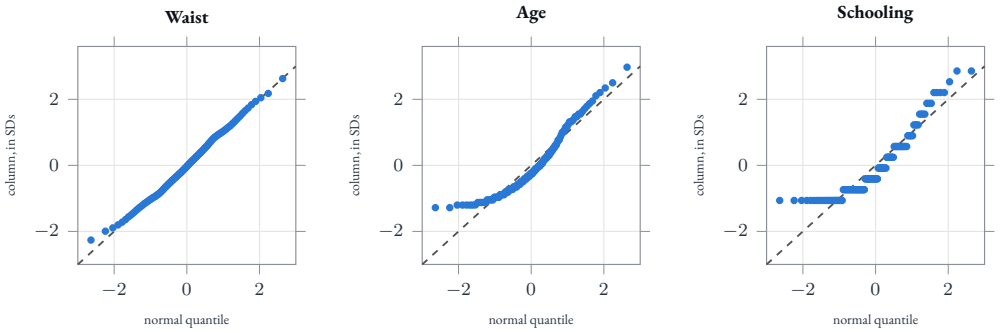

Figure 5: Normal Q-Q plots of three **simulated** columns matched to reference 1., Table 1.

Example 2.4 • Three columns, three patterns

Waist in Figure 5 lies along the line: close to normal for almost any purpose. Age bends above the line at the right: its upper percentiles lie further out than a normal curve would put them, which is a long right tail. It also sits above the line at the left, because its lowest percentiles are held up by the floor at 35. Schooling climbs in steps, because it takes only whole numbers, and bends at both ends for the same two reasons as age.

The point

A Q-Q plot shows the tails, which is where the normal model most often fails. A histogram hides the tails in bars with two or three observations in them.

2.8 A skew check from a mean and a standard deviation

Chapter 4 gave a check on skew that needs only printed quartiles. There is a companion that needs only a printed mean and standard deviation, and it applies to measurements that cannot be negative. A normal distribution has about 2.3% of its values more than two standard deviations below the mean. If two standard deviations below the mean is below zero, those values would have to be negative, which the measurement does not allow. So:

$$\bar{x} < 2s \implies \text{the column is likely to be skewed} \tag{2.5}$$

$\bar{x}, s$ the reported mean and standard deviation of a measurement that cannot be negative

Altman and Bland state the rule in this form, and illustrate it with urinary cotinine levels reported as means and standard errors. Once the standard deviations were recovered, every group's mean was smaller than its standard deviation, and t-tests had been run on the data (reference 3.).

Example 2.5 • Two rows checked

COBRA-BPS reports a mean age of 58.8 years with a standard deviation of 11.5 (reference 2.); $2s = 23.0$, and the check passes. Chapter 4 reconstructed an approximate mean and standard deviation for daily servings of fruit and vegetables in the Sylhet survey, 0.67 and 0.74, and warned that the reconstruction's condition failed; $2s = 1.48$,

Example 2.5 • *continued*

and the check flags the column. Both comparisons are ours.

Watch out

The rule works in one direction only. Failing it is informative; passing it proves nothing, since many skewed columns have a mean well above twice their standard deviation.

2.9 Columns that are normal on another scale

A column produced by multiplicative influences is skewed in its raw units and often close to symmetric after each value is replaced by its logarithm. Such a column is called *log-normal*.

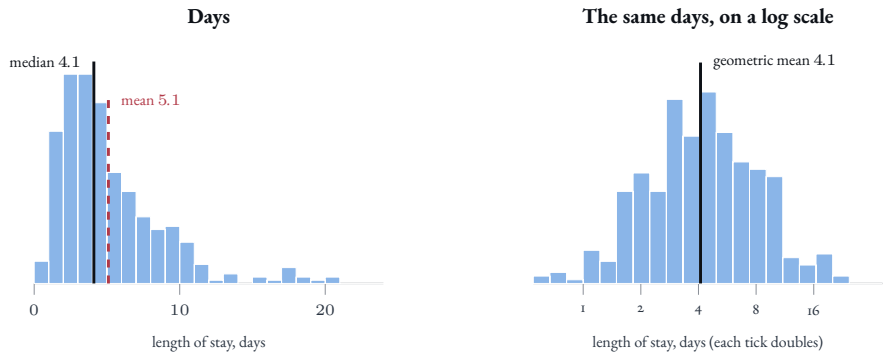

Figure 6: Length of stay for 400 patients, drawn in days and on a logarithmic scale on which each tick doubles. **Constructed**; not from any study.

Example 2.6 • The geometric mean returns

In Figure 6 the raw lengths of stay are right-skewed, with a mean of 5.1 days and a median of 4.1. On the logarithmic scale the same values are close to symmetric. The geometric mean of Chapter 4, the mean of the logged values converted back, is 4.1 days. For a column that is close to normal on the log scale, the geometric mean lands near the median, while the ordinary mean is pulled into the tail. That is why papers report geometric means for such columns. All figures are from the constructed column.

Watch out

When to transform a column before an analysis, and how to report results on a log scale, belong to Chapter 24. Recognising the shape comes first.

3 Tails and outliers

3.1 Percentiles describe tails

The centre of a column is described by its median. A tail is described by the percentiles out in it: the 2.5th and 97.5th, which bound the central 95%; the 5th and 95th; the 3rd and 97th on a growth chart. None of them needs a model. They are read off the sorted column, which makes them the natural language for tails whatever the shape. The patients a clinician worries about are usually in the tails, and a description that stops at the quartiles says nothing about them.

3.2 Reference ranges

A laboratory's normal range is a statement about a tail, and it is usually built as a percentile interval.

Definition 3.1 • Reference interval

A reference interval is usually the central 95% of the values measured in a reference population of healthy individuals: the range from its 2.5th to its 97.5th percentile (reference 4.).

The consequence follows from the definition. If the interval contains 95% of healthy people, then 5% of healthy people fall outside it: one in twenty, on any single test, by construction. Across several tests the chance of at least one result outside its range grows. For m tests that were independent of each other,

$$\Pr(\text{at least one outside}) \;=\; 1 - 0.95^{\,m} \tag{3.1}$$

m the number of tests in the panel
0.95^{m} the chance that all m results fall inside their ranges, if the tests are independent

Example 3.1 • A routine panel on a healthy patient

For $m = 10$ tests, $1 - 0.95^{10} = 1 - 0.599 = 0.40$. For $m = 20$, $1 - 0.95^{20} = 1 - 0.358 = 0.64$. On these figures nearly two in three healthy people would have at least one "abnormal" result on a twenty-test panel. The arithmetic is ours, and it assumes the tests are independent; real panels are correlated, liver enzymes rising together for example, so the true figures are lower. The clinical picture is familiar: a well patient, one marginal enzyme result, and a sequence of investigations that finds nothing.

Watch out

A range computed as the mean plus or minus two standard deviations assumes a normal shape. For a right-skewed analyte, Section 2.4 shows what that does: the lower limit falls towards zero and flags almost nobody, while too many healthy people are flagged high. Percentile intervals need no such assumption.

3.3 Unusual, compared with what?

An "unusual" value can be unusual against three different standards, and they ask different questions.

Compared with	The question	Illustration
The rest of the column	Is it far from the bulk?	a waist of 104 cm in rural Sylhet
The human body	Could a person have this value?	a systolic pressure of 310
The instrument	Could it have been recorded?	a potassium of −2

The second and third are clinical and technical checks, and they come first; a value that is impossible is an error whatever its statistical position. Only the first is statistical, and it asks whether a value is far from where the rest of its column sits, not whether it is true. A value can be far from the bulk and entirely real, and treating distance as evidence of error is the source of most poor decisions about outliers. The illustrative values in the table are not from any study.

3.4 Tukey's fences

The commonest rule for flagging values far from the bulk is built from the quartiles.

$$\text{lower fence} = Q_1 - 1.5\,(Q_3 - Q_1), \qquad \text{upper fence} = Q_3 + 1.5\,(Q_3 - Q_1) \tag{3.2}$$

Q_1, Q_3 the lower and upper quartiles
$Q_3 - Q_1$ the interquartile range
1.5 a conventional multiplier, chosen by Tukey because it works well in practice rather than derived from any theory (reference 7.)

A value beyond either fence is *flagged*. The great virtue of the rule is that it is built from the quartiles, which the flagged values cannot move.

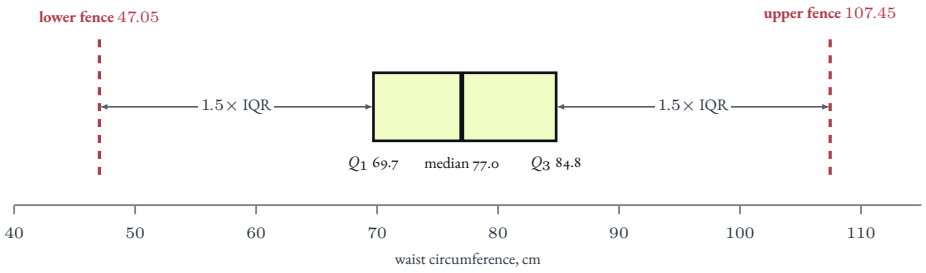

Figure 7: Tukey's fences on the Sylhet waist column. Quartiles and median from reference 1., Table 1; the fences are ours.

Example 3.2 • Fences on the Sylhet waist column

The quartiles are 69.7 and 84.8 cm, so the interquartile range is 15.1 cm and one and a half times it is 22.65 cm. The fences are $69.7 - 22.65 = 47.05$ and $84.8 + 22.65 = 107.45$ cm, both ours. The paper reports neither a minimum nor a maximum, so how

Example 3.2 • *continued*

many participants lay beyond the fences cannot be known from outside the study. In the simulated column of Figure 1, whose tails were chosen here, one value does.

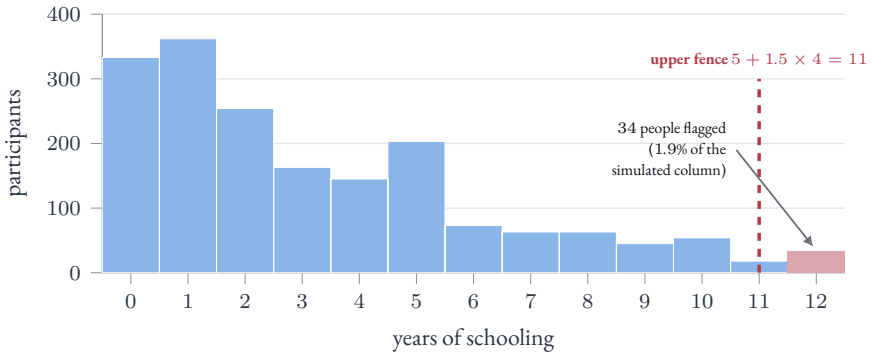

Figure 8: Years of schooling, **simulated** to match the quartiles and bands of reference 1., Table 1; how the upper band splits by year is our choice. The fence is ours.

Example 3.3 • A fence on a skewed column

For schooling the quartiles are 1 and 5 years, the interquartile range 4, and the upper fence $5 + 1.5 \times 4 = 11$ years (ours). In the simulated column of Figure 8, the participants beyond it are those with twelve years of schooling, 1.9% of the column: people who completed higher secondary education, which is unusual in rural Sylhet and entirely real. The fence has flagged the data, not errors in it.

Watch out

The fences assume the two sides of a column are comparable. On a column with a floor and a long upper tail, the upper fence falls inside the genuine tail, and every such clinical column, length of stay above all, will have its most complicated patients flagged. For such a column, the fences are best applied only after drawing it on its natural scale, or on a log scale if it is multiplicative. Even then they give a list of values to check, not a list of values to remove.

3.5 An outlier is a question

A flagged value is a question, and it has three possible answers, each with a different action.

1. **An error.** A waist entered as 188 for 88 cm. Correct it from the source record, or set it to missing, and report the change.
2. **Real, and interesting.** A 35-year-old with a systolic pressure of 230 mm Hg. Keep it: it may be the most informative row in the study.
3. **Real, and outside the question.** Someone who does not belong to the population the study is about. Exclude by a rule written before the data were examined, never because of the value itself.

Example 3.4 • An exclusion stated in advance

The Sylhet survey excluded 14 pregnant women and reported the exclusion in its account of participation (reference 1.). The rule concerned who the study was about, a survey of hypertension in the general adult population, and it was applied as an eligibility criterion rather than as a response to any measured value.

3.6 Why no value is removed quietly

Rules based on the mean and standard deviation, such as "remove values more than three standard deviations from the mean", have a flaw that is easy to show: the values they are looking for inflate the standard deviation used to judge them.

Example 3.5 • Two extremes hiding each other

Ten values: 5, 5, 6, 6, 6, 7, 7, 7, 30, 31. Their mean is 11.0 and their standard deviation 10.3. The value 30 has $z = (30 - 11.0)/10.3 = 1.84$, and 31 has $z = 1.94$. Neither reaches two standard deviations, let alone three. The median, 6.5, is unaffected by either. The values are constructed and the arithmetic is ours. The effect is called *masking*.

The point

Careful work decides the rule for handling outliers before seeing the data, reports how many values the rule affected, and runs the analysis both with and without them. If the conclusion survives, the report says so; if it does not, that is itself a finding. Running the analysis both ways is a *sensitivity analysis*. Which methods to use when outliers are present is taken up in Chapter 19.

4 Why shape matters

4.1 Shape chooses the summary

Chapter 4 chose a summary from four printed numbers. With the picture, the choice is made from the shape itself, and two cases appear that Chapter 4 could not see.

Shape	Report	Example
Close to symmetric	mean (SD)	systolic pressure in COBRA-BPS
Skewed, one peak	median (Q_1 to Q_3)	years of schooling
Symmetric on a log scale	geometric mean	antibody titres, length of stay
Most values at zero	share above zero, then median	servings, cigarettes per day
Two peaks	split, and describe each part	a triaged clinic

For a column in which most people score zero, no single centre describes it; the informative report gives the share of people with any at all, and then describes those who have some.

4.2 Shape will choose the method

Every method of analysis in Chapters 19 to 26 makes an assumption about shape, and the assumption is usually not that the raw column is normal. A comparison of two means depends on how the *means* behave (Chapter 20). A method based on ranks exists for when that fails (Chapter 22). A regression depends on the shape of what is left over once the model has been fitted (Chapter 24). A mistake seen in published Methods sections is to test the raw outcome for normality, find it skewed, and abandon a method whose assumption was never about the raw outcome.

The reason means behave better than the columns they come from is taken up in Chapter 7. Its content, in one sentence: take thirty values at a time from a column piled against zero and average them, and the averages pile up symmetrically around the column's mean.

5 Chapter summary

1. A histogram is the column itself, drawn. Its centre and shape are read first.
2. Skew has mechanisms: a floor the measurement cannot pass, a floor the study imposed, a ceiling, and a mixture. The mean lies on the side of the tail.
3. The skewness coefficient printed by software has a standard error that shrinks with the sample, so at large sample sizes a nearly symmetric column is "significantly" skewed.
4. Two peaks usually mean two populations pooled. No summary describes such a column; it is split by the variable that separates the groups.
5. The normal distribution is fixed by its mean and standard deviation. It is a model, and no clinical column is exactly normal.
6. If a column is close to normal, about 68% lies within one standard deviation of the mean and about 95% within two (1.96 exactly).
7. The shares within one and two standard deviations change little under moderate skew; which tail the remainder lies in is not.
8. With no assumption about shape, at least 75% lies within two standard deviations and nothing is guaranteed within one.
9. A z-score is a distance from a stated reference group, in standard deviations.
10. A Q-Q plot shows departures from normality in the tails, where they matter most.
11. For a measurement that cannot be negative, a mean below twice the standard deviation signals skew.
12. A reference interval is the central 95% of a healthy group, so one healthy person in twenty lies outside it on any one test.
13. Tukey's fences flag values beyond 1.5 interquartile ranges from the quartiles. Flagged is not wrong, and on a skewed column the fence flags the data.
14. An outlier is an error, a real and interesting value, or a value from outside the study's question. Each has a different action, and none is removed quietly.

6 Key terms

Term	Meaning
Histogram	Bars over equal bands of a measurement, each as tall as the count in its band.
Skew (right, left)	A longer tail on one side. Right skew: a long tail to high values.
Skewness coefficient, G_1	A single number for the lean of a column: 0 symmetric, positive for a right tail.
Normal distribution	The symmetric bell-shaped model fixed by μ and σ.
$N(\mu, \sigma^2)$	Shorthand for a normal model with mean μ and variance σ^2.
68–95–99.7 rule	The shares within 1, 2 and 3 SDs of the mean, if normal.
Chebyshev's inequality	At most $1/k^2$ of any column lies k or more SDs from its mean.
z-score	$(x - \bar{x})/s$: distance from the mean in standard deviations.
Q-Q plot	Percentiles of a column against those of a normal distribution.
Log-normal	Skewed in raw units, close to normal after logging.
Percentile	The value below which a stated percentage of observations fall.
Reference interval	Usually the central 95% of a healthy reference population.
Tukey's fences	$Q_1 - 1.5$ IQR and $Q_3 + 1.5$ IQR.
Outlier	A value far from the bulk of its column; a question, not a verdict.
Masking	Extreme values inflating the SD so that none is flagged by an SD rule.
Sensitivity analysis	Repeating an analysis under a different reasonable choice to see whether the conclusion changes.

7 Exercises

THE NUMBERS YOU WILL NEED

From the Sylhet survey (reference 1.), medians with quartiles, $n = 1{,}810$: 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).

From COBRA-BPS (reference 2.), means with standard deviations: baseline systolic pressure 146.7 ± 22.4 mm Hg (intervention, $n = 1{,}330$) and 144.7 ± 21.0 (control); diastolic 89.1 ± 14.7 (intervention); age 58.8 ± 11.5 years.

Part A. Shape

1. A paper reports a mean C-reactive protein of 18 mg/L and a median of 6 mg/L. Describe the shape of the column and name the mechanism most likely to have produced it.

2. For each column, predict the direction of any skew and name its mechanism: Apgar score at five minutes; age in a trial that enrolled only people aged 65 or over; number of hospital admissions last year; haemoglobin in healthy adult men.

3. SPSS reports a skewness of 0.21 with a standard error of 0.04 for a column of 3,000 fasting glucose values. A colleague says the column is significantly skewed and a mean cannot be used. Respond.

4. Describe a measurement in your own practice that you would expect to have two peaks. Name the variable that would separate the two groups.

Part B. The normal model

5. Give the ranges spanned by one and by two standard deviations either side of the mean baseline systolic pressure in the COBRA-BPS control arm. State what share of participants each would contain, and the condition under which that statement holds.

6. Compute the z-score of a COBRA-BPS intervention participant whose baseline systolic pressure is 120 mm Hg, and say in words what it means.

7. Using Chebyshev's inequality, what is the smallest share of any column that can lie within 1.5 standard deviations of its mean?

8. Apply the check of Section 2.8 to each of the following, and state what it shows: COBRA-BPS diastolic pressure; a length of stay reported as 6.2 ± 7.9 days; triglycerides reported as 1.9 ± 1.2 mmol/L.

9. Of the four Sylhet columns in the box above, which would you be willing to model as normal to estimate the share of participants above a clinical threshold? Justify your choice from the printed quartiles.

Part C. Tails and outliers

10. Compute Tukey's fences for body mass index in the Sylhet survey, and state whether a participant with a body mass index of 31 would be flagged.

11. Compute the fences for age in the Sylhet survey. Explain why the lower fence is of no practical interest in this study.

12. A healthy volunteer has twelve independent laboratory tests, each with a 95% reference interval. What is the probability that at least one result falls outside its interval? Why is the true figure likely to be lower?

13. Classify each flagged value as an error, real and interesting, or outside the question, and give the action for each: a birth weight recorded as 35 kg; a serum sodium of 118 mmol/L in a patient on thiazides in a study of hyponatraemia; a participant aged 17 in a survey of adults.

14. Explain, without calculation, why a rule of "exclude values more than three standard deviations from the mean" can fail to flag two extreme values in a small dataset.

Part D. Reading a result

15. A Methods section reads: "Data were not normally distributed (Shapiro-Wilk $p = 0.03$, $n = 2{,}400$) and were therefore analysed with non-parametric tests." Give two criticisms.

16. A normal Q-Q plot of a column bends below the line at its left end and lies on the line elsewhere. Describe the column.

17. A paper reports C-reactive protein as 12 ± 25 mg/L and describes it as normally distributed. What does the printed summary alone tell you?

Part E. Appraisal

18. Take one numerical column from your own work or from a dataset you have. Draw it, name its shape and its mechanism, and say which summary you would report.

19. A healthy patient's routine panel returns one liver enzyme marginally above its reference interval. Using Section 3.2, write two sentences explaining to a colleague why further investigation may not be warranted on that result alone.

20. Write an outlier-handling rule for a study of your choosing that could be placed in a protocol before any data are collected. State what would be reported about the values it affected.

8 Solutions

Part A

1. The mean is three times the median, so the column is strongly right-skewed: most patients have low values and a few have very high ones. The mechanism is a floor at zero with no ceiling, since a concentration cannot be negative; acute inflammation in a minority produces the long upper tail.

2. Apgar at five minutes: left-skewed, piled near the ceiling of 10, with a tail of distressed infants. Age in a trial of people aged 65 and over: right-skewed, from a floor the study imposed at 65. Admissions last year: right-skewed, a count with a floor at zero and most people at zero. Haemoglobin in healthy men: close to symmetric, a measurement shaped by many small additive influences.

3. The ratio $0.21/0.04 \approx 5$ says the column is not exactly symmetric, which with 3,000 values is almost certain for any real column. It does not say the skew matters. A coefficient of 0.21 is a mild lean; the mean and standard deviation will describe such a column well, and the right step is to look at its histogram rather than to rely on the ratio.

4. Answers vary. Examples: time to first antibiotic in an emergency department, separated by triage category; serum creatinine in a clinic that mixes people with and without chronic kidney disease, separated by disease status; birth weight in a unit that serves both a general population and a referral stream of very preterm infants.

Part B

5. One standard deviation: $144.7 \pm 21.0 = 123.7$ to 165.7 mm Hg. Two: $144.7 \pm 42.0 = 102.7$ to 186.7 mm Hg. If the column is close to normal, these contain about 68% and 95% of the participants. The paper reports no shape, so the statement is conditional; without the condition, Chebyshev's inequality guarantees only that at least 75% lie within two standard deviations.

6. $z = (120 - 146.7)/22.4 = -26.7/22.4 = -1.19$. The participant sits a little more than one standard deviation below the average participant of the intervention arm at baseline.

7. $1 - 1/1.5^2 = 1 - 1/2.25 = 1 - 0.444 = 0.556$. At least 55.6% of any column lies within 1.5 standard deviations of its mean.

8. Diastolic pressure: 89.1 against $2 \times 14.7 = 29.4$; the check passes, which shows nothing either way. Length of stay: 6.2 against 15.8; the check fails, and the column is almost certainly right-skewed, so a median and quartiles would describe it better. Triglycerides: 1.9 against 2.4; the check fails, consistent with the right skew typical of lipid concentrations.

9. Waist, whose quartile halves are 7.3 and 7.8 cm, is the only one of the four whose middle is close to symmetric, and it is the best candidate. Body mass index is mildly skewed (2.2 against 2.6) and might serve. Age (7 against 11) and schooling (1 against 3) are not suitable; for schooling in particular a normal model would place a noticeable share of people below zero years. Any tail estimate from a normal model would still be an approximation.

Part C

10. The interquartile range is $22.9 - 18.1 = 4.8$, and $1.5 \times 4.8 = 7.2$. The fences are $18.1 - 7.2 = 10.9$ and $22.9 + 7.2 = 30.1$. A body mass index of 31 lies above the upper fence and would be flagged: a reason to check the record, not a reason to remove it.

11. The interquartile range is $59 - 41 = 18$ years and $1.5 \times 18 = 27$. The fences are $41 - 27 = 14$ and $59 + 27 = 86$ years. Nobody under 35 was eligible, so no participant can lie below the lower fence; the column's floor was imposed by the study.

12. $1 - 0.95^{12} = 1 - 0.540 = 0.46$. The true figure is likely to be lower because laboratory tests are correlated with each other, and equation (3.1) assumes independence.

13. A birth weight of 35 kg is impossible and is an error: correct it from the source or set it to missing, and report the change. A sodium of 118 mmol/L in a thiazide user in a study of hyponatraemia is real and exactly the kind of case the study is about: keep it. A 17-year-old in a survey of adults is outside the question: exclude by the eligibility rule and report the exclusion.

14. Extreme values enlarge the standard deviation, which is the yardstick the rule uses. With few observations, two extremes can enlarge it so much that neither lies three standard deviations from the mean, and the rule fails to flag the very values it was meant to catch.

Part D

15. First, with 2,400 observations a normality test will reject for trivial departures, so the *p*-value says the column is not exactly normal and nothing about whether the departure matters. Second, the decision to abandon methods based on means should depend on how the means behave, or on the shape of the residuals in a model, not on the shape of the raw outcome. At this sample size, means of even a clearly skewed column behave well.

16. The lowest percentiles are lower than a normal curve would place them, so the column has a longer left tail than a normal distribution and is otherwise close to normal: mildly left-skewed.

17. The mean is smaller than the standard deviation, and far smaller than twice it, for a concentration that cannot be negative. The column is almost certainly strongly right-skewed, and the description of it as normally distributed is not credible from the printed summary alone.

Part E

18. Answers vary. A good answer names the shape (for example, right-skewed with a floor), the mechanism (a count, a concentration, an inclusion threshold), and a summary that follows from it (a median and quartiles), and says what the picture showed that a summary would have hidden.

19. A reference interval is built to contain 95% of healthy people, so one healthy person in twenty lies outside it on any single test. On a routine panel of many tests, at least one marginal result in a healthy patient is expected, and such a result should be interpreted in the clinical context rather than investigated on its own.

20. Answers vary. A good rule states in advance which checks are applied (plausibility against physiological limits, verification against the source record), what happens to a value that fails (corrected or set to missing, never silently deleted), and that the primary analysis will be repeated with and without values flagged by a stated statistical rule, with the number affected and any change in the conclusion reported.

9 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. Jafar TH, Gandhi M, de Silva HA, et al. 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. Detecting skewness from summary information. *BMJ* 1996;313(7066):1200. PMID 8916759. PMC2352469. doi:10.1136/bmj.313.7066.1200.
4. Jones G, Barker A. Reference intervals. *Clin Biochem Rev* 2008;29(Suppl 1):S93–S97. PMID 18852866. PMC2556592.
5. Kanis JA, McCloskey EV, Johansson H, Oden A, Melton LJ, Khaltaev N. A reference standard for the description of osteoporosis. *Bone* 2008;42(3):467–475. PMID 18180210. doi:10.1016/j.bone.2007.11.001.
6. WHO Multicentre Growth Reference Study Group. WHO Child Growth Standards based on length-/height, weight and age. *Acta Paediatr Suppl* 2006;450:76–85. PMID 16817681.
7. Tukey JW. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977. A book; no PMID or DOI.
8. Daniel WW. *Biostatistics: A Foundation for Analysis in the Health Sciences*, 9th ed. Wiley. Section 4.6, the normal distribution, book pp. 117–123; Section 4.7, applications, pp. 124–128.

Figures computed, simulated or constructed in this chapter rather than reported by a paper. Computed: every band endpoint on the COBRA-BPS mean, the z-scores, the Chebyshev bounds, the standard error of skewness at $n = 1{,}810$, the fences on waist, body mass index, age and schooling, the probabilities 0.40, 0.46 and 0.64, and the check $\bar{x} < 2s$ on each row it is applied to. Simulated, matched to the medians and quartiles of reference 1.: the waist, age and schooling columns of every histogram and Q-Q plot, their skewness coefficients, and every share within standard deviations. The simulated age column matches its quartiles and its share aged 65 and over but not every printed age band. Constructed: the triaged clinic, the length-of-stay column, and the ten masking values.

10 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. Draw every numerical column in your own study, or in a dataset you have. Name each shape, and say how the column is made: added, multiplied, floored or mixed.

Two. Take one value flagged by Tukey's fences in any column and answer the three questions of Section 3.5: an error, real and interesting, or outside the question.

Reading

Daniel Section 4.6 (reference 8.) covers the normal distribution formally. Altman and Bland (reference 3.) is one page. Jones and Barker (reference 4.) is a short account of how reference intervals are built.

Questions this chapter deliberately left open

Question	Where it is settled
How should a two-peaked column be displayed, and does a box plot show it?	Chapter 6, tables and graphs
Why do means of a skewed column behave better than the column, and by how much?	Chapter 7, the standard error and the central limit theorem
Where does 1.96 come from in a confidence interval?	Chapter 8, confidence intervals
Should normality be tested formally, and what should be done with outliers in an analysis?	Chapter 19, choosing a method
When a method's assumptions fail, what replaces it?	Chapters 20 and 22
Why does a panel of many tests produce false alarms, in general?	Chapter 22, multiple comparisons
When and how should a skewed column be transformed?	Chapter 24, regression

Next

Chapter 6, Tables, Graphs & Describing a Clinical Sample. This chapter asked what a column looks like; the next asks how to show it, and a whole sample, to a reader without misleading them.