Stat.BD

APPLIED BIOSTATISTICS & RESEARCH METHODOLOGY • SESSION 5

Understanding Distributions & Outliers

What a column looks like, and the values out in its tails

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

Session 5 of 30

Session 4 judged five shapes without drawing one

Sylhet column	Median (Q_1 to Q_3)	Halves	Verdict
Waist, 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
Schooling, years	2 (1 to 5)	1 | 3	strongly right-skewed
Servings per day	1 (0 to 1)	1 | 0	piled against a floor

Today's first question

Four printed numbers per column, and a verdict on its shape. If you could see these columns, would the verdicts survive?

Medians and quartiles: Khanam 2019, Table 1, PMID 31662350. The halves and the verdicts are ours, from Session 4.

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

The test of this session

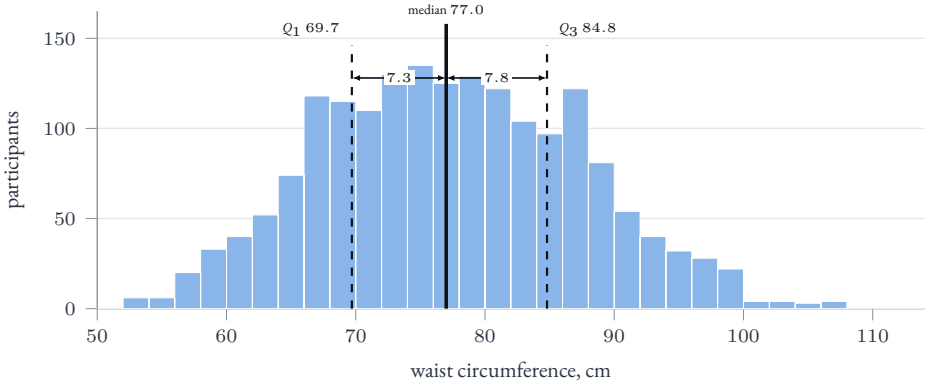
Look at any numerical column, drawn or summarised, and say what shape it has, whether a normal model is close enough for the job in hand, and what to do about the values out in its tails.

- Name a shape: symmetric, skewed, floored, two-peaked
- State what share of a column lies within one and two SDs, **and the condition** under which that holds
- Read a normal Q-Q plot, and the skewness figure SPSS prints
- Flag an unusual value by a stated rule, and then decide what it is

Act I

What a column looks like

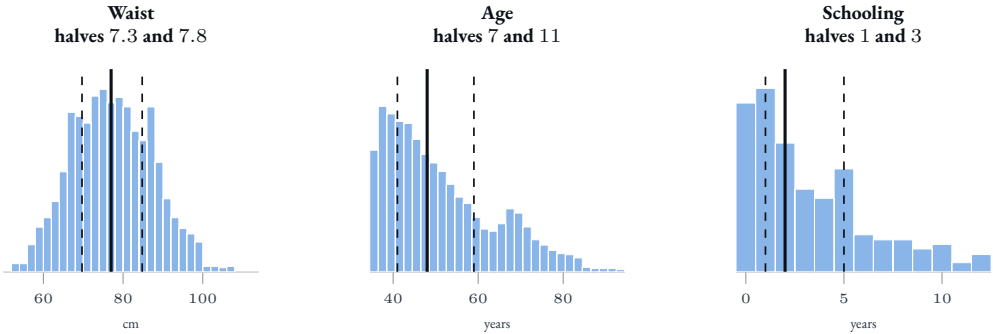
A histogram is the column, drawn



Each bar counts the people whose waist fell in a two-centimetre band. The two halves Session 4 computed are the two arrows.

Simulated to match the median and quartiles printed by Khanam 2019 (Table 1, PMID 31662350); not the study's data. The tails are our choice.

The two-halves check was reading the middle of this picture



Equal halves, and the picture is balanced about its middle.

A longer upper half, and a tail runs out to the right.

Simulated to match the medians and quartiles of Khanam 2019, Table 1 (schooling also to its printed bands). Dashed lines are the quartiles, solid the median.

Where skew comes from

A floor the measurement cannot pass. Counts, concentrations, durations. Nobody has fewer than zero years of schooling, and a long tail can only run upward.

A floor the study built. The Sylhet survey enrolled adults aged 35 and over, so its age column starts at a wall it chose.

A ceiling. Scores with a maximum, such as a Glasgow Coma Scale of 15. The tail runs the other way.

You see this every ward round

Length of stay: most patients go home in three or four days, and one stays ninety. Nobody stays minus ninety to balance them.

That asymmetry is the whole of right skew.

A floor by design is still a floor

The age column's skew says nothing about Sylhet. It says what the inclusion criterion was.

Inclusion criterion: Khanam 2019, Methods. PMID 31662350.

Skew to the left exists, and it is rarer

A long tail to the **left**, the bulk piled near a ceiling:

- oxygen saturation on a general ward
- gestational age at delivery
- scores capped at a maximum

The mean now sits *below* the median, pulled down by the tail.

The rule does not change

The mean is always on the side of the tail.

Right skew pulls it up. Left skew pulls it down. The median stays with the bulk either way.

The number SPSS prints as skewness

Skewness is a single number for the lean: zero for a symmetric column, positive for a tail to the right. SPSS prints it with a standard error beside it.

$n = 1{,}810$	**Skewness**	**Ratio to SE**
Waist	0.14	2.4
Age	0.84	14.7
Schooling	1.02	17.7

The standard error at this sample size is 0.058.

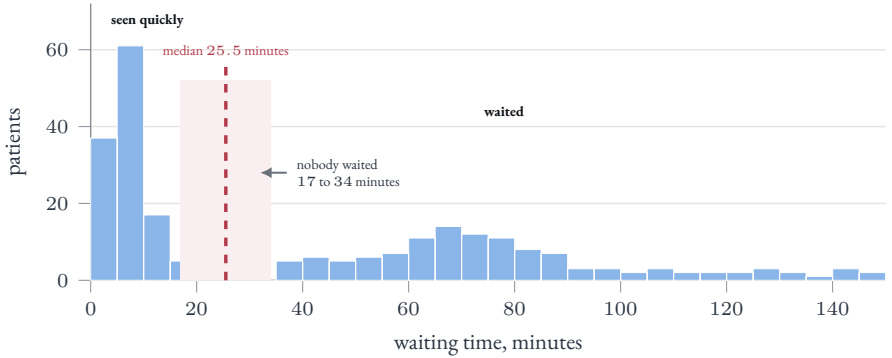
Why the ratio misleads

A ratio above 2 is often read as "significantly skewed". At $n = 1{,}810$ the waist column passes that bar, and its histogram is visibly close to symmetric.

With enough people, every real column is "significantly" skewed. The question worth asking is whether it is skewed *enough to matter*.

Simulated columns, matched to Khanam 2019, Table 1. The skewness figures, their standard errors and the ratios are ours.

Two peaks mean two populations



The median describes a wait nobody had. The fix is not a better summary. It is to find what separates the two groups, and describe each.

Constructed by us: Session 4's Clinic B, drawn. Not from any study.

Act II

The normal model, and its condition

Stat.BD

A model, not a fact about patients

The normal distribution is one symmetric, single-peaked shape, fixed entirely by two numbers: where its centre is and how spread out it is.

$$X \sim N(\mu,\, \sigma^2)$$

Read aloud: the column is modelled as normal, with mean μ and standard deviation σ.

No clinical column is exactly normal. Some are close enough for some purposes, and the purpose decides.

Where you have already met this

The WHO child growth standards found length and height for age close to normal, and had to model the skew in the other indicators before drawing the curves.

Even the reference charts treat normality as a finding to be checked.

WHO Multicentre Growth Reference Study Group. *Acta Paediatr Suppl* 2006;450:76–85. PMID 16817681.

Why so many columns come out roughly normal

Many small causes, added

A blood pressure is genes, salt, sleep, the cuff, the time of day and the walk to the clinic, each nudging it a little up or down. Nudges that add up pile into a bell.

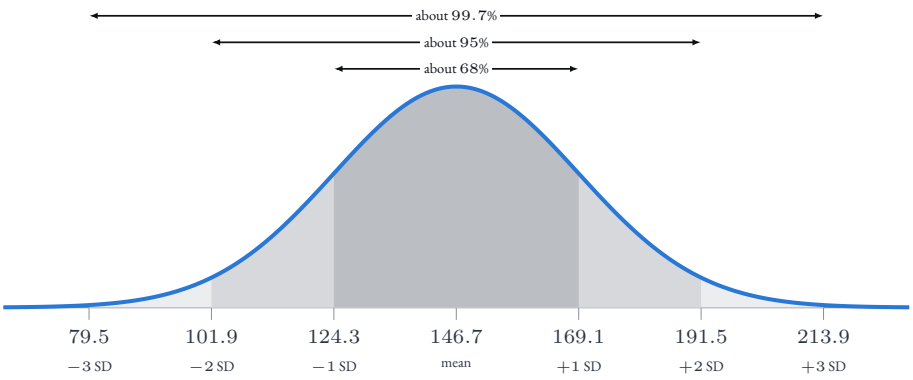
Many small causes, multiplied

A concentration, a titre or a length of stay grows by proportions. Those pile into a shape that is skewed in the raw units and close to a bell on a log scale.

What this buys you

Before looking at a column, ask how it is made. Additive columns are usually close to normal. Multiplicative ones are usually close to normal *after logging*. Floored counts are neither.

The figure Session 4 refused to give, and its condition



The condition, which is the whole point

If the column is close to normal, about 68% of patients sit within one SD of the mean and about 95% within two. If it is not, these shares can be wrong.

Mean and SD: Jafar 2020 (COBRA-BPS), intervention arm at baseline, PMID 32074419. Every band endpoint is **ours**, and assumes a normal shape the paper does not report.

The share is robust, the sides are not

	Within 1 SD	Within 2 SD	Below −2 SD	Above +2 SD
Normal model	68%	95%	2.3%	2.3%
Waist	66%	96%	1.2%	2.5%
Age	65%	95%	0.0%	4.5%
Schooling	66%	94%	0.0%	5.9%
Clinic B	**82**%	95%	0.0%	4.6%

What survives skew

The total inside two SDs comes out near 95% even for a skewed column.

What does not

Where the rest sit. On a skewed column they are all in one tail, and a two-peaked column breaks even the total.

Simulated columns (waist, age, schooling matched to Khanam 2019, Table 1) and a **constructed** Clinic B. Every share here is ours.

With no assumption at all, this is all you are owed

For *any* column, whatever its shape, the share lying k or more SDs from the mean is at most $1/k^2$:

$$\Pr(\,|X-\mu| \geq k\sigma\,) \;\leq\; \frac{1}{k^2}$$

	No assumption	If normal
Within 1 SD	nothing guaranteed	68%
Within 2 SD	at least 75%	95%
Within 3 SD	at least 89%	99.7%

A guarantee against a prognosis

The left column is a guarantee: it holds for every column that exists.

The right column is a prognosis: better, and true only for the patients it was built on.

This is why Session 4 would not give a figure for one SD. Without the shape, there is none.

Chebyshev's inequality, a mathematical result. The comparison is ours.

Distance counted in standard deviations

A z-score says how many SDs a value sits from the mean:

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

A COBRA-BPS patient with a systolic of 180:

$$z = \frac{180 - 146.7}{22.4} = 1.49$$

About one and a half SDs above the trial's average patient.

You read these already

A growth chart line labelled -2 is a z-score.

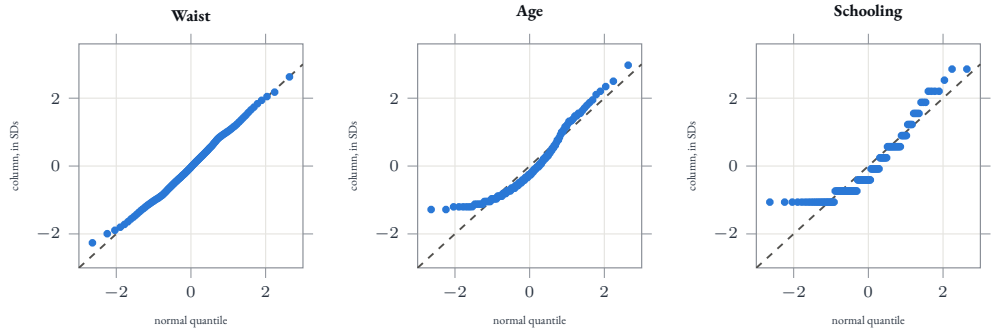
So is a bone density T-score: osteoporosis is 2.5 SDs or more below the average for *young adult women*.

Relative to whom?

A z-score is a distance from *a particular* reference group. The T-score's group is women aged 20 to 29, not the patient's own peers.

Mean and SD: Jafar 2020, PMID 32074419; the 1.49 is ours. T-score definition and reference group: Kanis 2008, *Bone* 42:467–75, PMID 18180210. Growth z-scores: WHO MGRS 2006, PMID 16817681.

Reading a normal Q-Q plot



Each dot is one percentile of the column against the same percentile of a normal curve. **On the line:** close to normal. **Bending up at the right:** a long upper tail. **Stairs:** a column that only takes whole numbers.

Simulated columns, matched to Khanam 2019, Table 1. Both axes in SDs.

A mean smaller than twice its SD

For a measurement that cannot be negative, a normal shape needs room for two SDs below the mean. So if

$$\bar{x} < 2s$$

the column is likely to be skewed, from the printed summary alone.

	$\bar{x}$	$2s$	
COBRA-BPS age	58.8	23.0	passes
Servings, rebuilt	0.67	1.48	**flags**

An audit you can run on any table

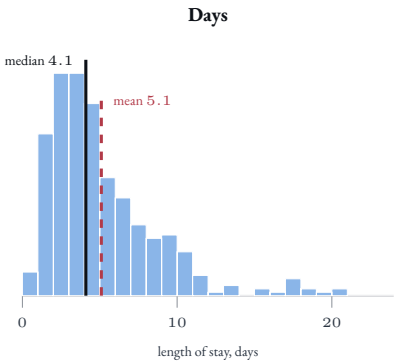
Altman and Bland's own example is urinary cotinine, where each group's mean was smaller than its SD, and *t*-tests had been run on it.

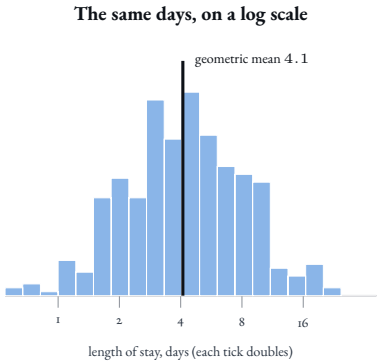
One direction only

Passing proves nothing. Failing is informative.

Rule: Altman DG, Bland JM, *BMJ* 1996;313:1200, PMID 8916759. COBRA-BPS age: Jafar 2020. The rebuilt servings mean and SD are **ours**, from Session 4.

Some skewed columns are normal on another scale





Skewed in days. Close to symmetric in log days, where each step is a doubling.

On such a column the geometric mean, 4.1 days, lands on the median, 4.1. The ordinary mean does not.

Constructed length of stay, 400 patients, not from any study. All figures ours.

Act III

Tails, and the values out in them

Stat.BD

Percentiles are the language of tails

The centre is described by the median. A tail is described by the percentiles out in it:

- the 2.5th and 97.5th, which bound the middle 95%
- the 5th and 95th
- the 3rd and 97th, on a growth chart

None of them needs the normal model. They are read off the sorted column.

Why tails matter clinically

The median patient rarely needs you. The ones in the tails do, and a description of a column that stops at the quartiles says nothing about them.

A reference range is the middle 95% of healthy people

A laboratory's normal range is usually "the central 95% of a reference population".

So **one healthy person in twenty** falls outside it on any one test, by construction.

Across 20 unrelated tests:

$$1 - 0.95^{20} = 0.64$$

Nearly two in three healthy people have at least one "abnormal" result.

You have seen this patient

A well person, a routine panel, one liver enzyme marginally flagged, and a cascade of investigation that finds nothing.

That flag was expected before the blood was drawn.

And on a skewed analyte

A range built as mean $\pm$ 2 SD, rather than from percentiles, flags almost nobody low and too many high.

Definition: Jones G, Barker A, *Clin Biochem Rev* 2008;29(Suppl 1):S93–7, PMID 18852866. The one-in-twenty and the 0.64 are **ours**; the 0.64 assumes the tests are independent, which real panels are not.

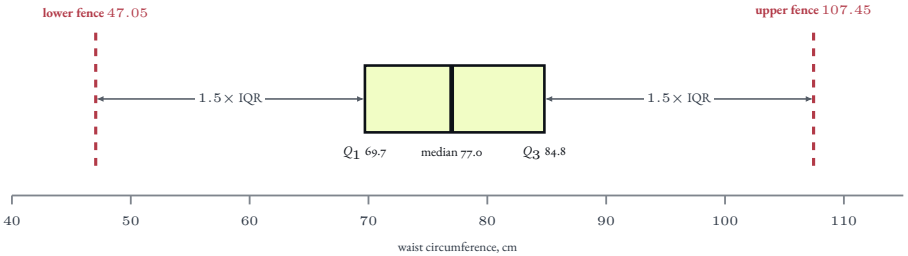
Unusual, compared with what?

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

Only the first is statistics

The second and third are clinical and technical checks, and they come first. A value that is impossible is an error whatever the statistics say, and a value that is far from the bulk may be perfectly real.

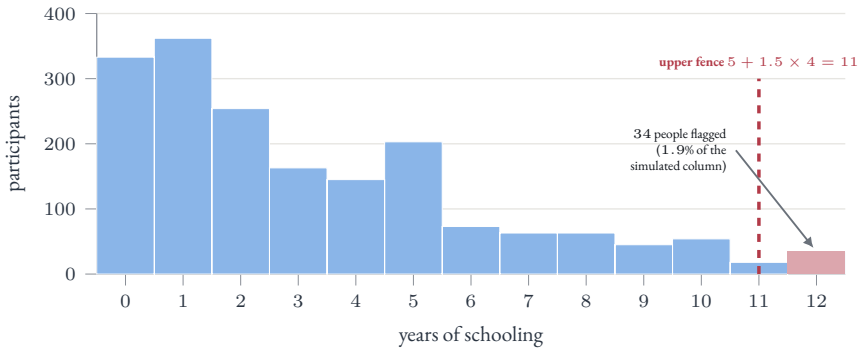
Tukey's fences, on the Sylhet waist column



A value beyond a fence is **flagged**. That is a reason to look, not a verdict. The simulated column puts 1 of 1,810 beyond them; the paper does not say how many real participants did.

Quartiles: Khanam 2019, Table 1, PMID 31662350. Fences **ours**: $69.7 - 1.5 \times 15.1 = 47.05$ and $84.8 + 1.5 \times 15.1 = 107.45$. Rule: Tukey, *Exploratory Data Analysis*, 1977 (a book; no PMID).

On a skewed column, the fence flags the data



Twelve years of schooling is higher secondary completed. Flagging those people as outliers says nothing about them. It says the rule assumed a symmetric column.

Simulated to match the quartiles and bands of Khanam 2019, Table 1; how the upper band splits by year is our choice. The fence is **ours**.

An outlier is a question, and it has three answers

An error

A waist entered as 188 for 88.

Correct it from the source, or set it missing. Say so.

Real, and interesting

A 35-year-old with a systolic of 230.

Keep it. It may be the most informative row in the study.

Real, and outside the question

A pregnant woman in a survey of chronic hypertension.

Exclude by a rule written *before* looking.

The Sylhet survey did the third: it excluded 14 pregnant women, and said so in its flow of participants.

Exclusions: Khanam 2019, participant flow, PMID 31662350. The other two examples are constructed.

Why a value is never deleted quietly

Ten values, two of them far out:

$$5,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 7,\ 30,\ 31$$

Mean 11.0, SD 10.3. The two extremes sit at $z = 1.84$ and 1.94.

A rule of "delete beyond 3 SD" removes neither. The two outliers inflated the very SD being used to judge them.

Masking

Extremes hide each other from any rule built on the mean and SD.

What careful work does

Decides the rule before seeing the data.
Reports how many values it touched.
Runs the analysis with and without them, and says whether the answer changed.

Ten values constructed by us. All arithmetic ours.

Act IV

Why the shape decides what comes next

Stat.BD

Shape chooses the summary: Session 4, run backwards

Shape	Report	Sylhet or trial example
Close to symmetric	mean (SD)	COBRA-BPS systolic
Skewed, one peak	median (Q_1 to Q_3)	schooling
Symmetric on a log scale	geometric mean	titres, length of stay
Most values at zero	share above zero, then median	servings, any count
Two peaks	split, and describe each	Clinic B

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

Shape will choose the method

Every method in Phase III makes an assumption about shape. What it assumes is usually **not** that the raw column is normal:

- comparing two means, about the *means* (Session 20)
- ranks instead of values, for when that fails (Session 22)
- regression, about what is left *after* the model (Session 24)

The commonest mistake in a Methods section

Testing the raw outcome for normality, finding it skewed, and abandoning a method whose assumption was never about the raw outcome.

Means of a skewed column behave better than the column

Take the schooling column, piled against zero. Draw thirty people at random and average them. Do it again, and again.

The averages do not pile against zero. They pile up symmetrically around the column's mean, and more tightly the more people go into each.

The next session but one

Why that happens, how tight the pile is, and what it lets a researcher say about a population from one sample, is Session 7.

Traps, collected

Reading

"Normally distributed" asserted without a picture.
A mean and SD for a column where $\bar{x} < 2s$.
"Significantly skewed" at a large n.
A normal range that assumes a normal analyte.
Outliers removed and not counted.

Doing

Draw every column before summarising it.
Ask how the column is made.
Give a share within SDs only with its condition.
Treat a flagged value as a question.
Write the exclusion rule before looking.

Close

Draw your own columns.

Stat.BD

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 flagged value from any column, by the fences, and answer the three questions: error, real and interesting, or outside the question.

What the first task is for

The next session is about displays, and it starts from your pictures rather than from mine.

Next: Session 6, Tables, Graphs & Describing a Clinical Sample. The same decisions, drawn on purpose.

References & sources

1. Khanam R, Ahmed S, Rahman S, et al. 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. CC BY-NC: numbers re-typeset as facts, text not 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. 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. doi:10.1111/j.1651-2227.2006.tb02378.x.
6. Kanis JA, McCloskey EV, Johansson H, et al. A reference standard for the description of osteoporosis. *Bone* 2008;42(3):467–475. PMID 18180210. doi:10.1016/j.bone.2007.11.001.
7. Tukey JW. *Exploratory Data Analysis*. Addison-Wesley, 1977. A book: no PMID or DOI.
8. Daniel WW. *Biostatistics*, 9th ed. Wiley. Section 4.6, the normal distribution, book pp. 117–123.

Simulated, constructed, and ours. Every histogram and Q-Q plot is drawn from a column **simulated** to match Khanam's printed quartiles, with tails chosen by us; Clinic B, the length of stay and the ten masking values are **constructed**. Ours: the SD bands and z-score on COBRA-BPS figures, every share within SDs, the skewness figures, the fences, the 0.64, and the $\bar{x} < 2s$ check on Session 4's rebuilt servings. Everything else is as printed by the source beside it.