Stat.BD

APPLIED BIOSTATISTICS & RESEARCH METHODOLOGY • SESSION 3

Describing Categorical Data

Counts, and the four things people do to them

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

Session 3 of 30

Your table, and what happens to it next

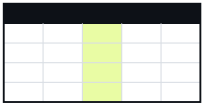
Last session ended with a variable table: one row per variable, with its role, its kind, how a value gets into it, and its unit or levels.

Today takes the columns you marked **binary**, **nominal** and **ordinal**, and asks the only question you can ask of them.

How many, out of how many?

Why the categorical ones come first

A categorical column cannot be averaged. Counting is not a lesser thing you do while waiting for real statistics: for these columns it is the whole of description.



Back to the rectangle. Today is about one column at a time, and then about two columns crossed.

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

The test of this session

Take any Table 1 in any paper. For every categorical row, say what was counted, what it was divided by, and whether the percentage runs down the column or across the row.

- Turn a column of answers into a **frequency distribution**
- Say **ratio**, **proportion**, **percentage** and **rate** and mean four different things
- Find the **denominator** of any percentage in a paper, including the ones that change halfway down a table
- Read a **two-way table** three ways and know which one answers your question

Act I

Counting, and the three things it turns into

Stat.BD

Four words, and they are not synonyms

RATIO	PROPORTION	PERCENTAGE	RATE
$162 : 177$	$339/1810 = 0.187$	$0.187 \times 100 = 18.7\%$	not in this study
One count against another count. The two parts do not have to add up to anything, and the answer has no upper limit.	A part divided by the whole it belongs to. Always between 0 and 1.	The same proportion, in hundredths. Nothing has been added.	A count divided by the *time* people were at risk. Needs a denominator with years in it, so a survey cannot give you one.

The first three are all in this paper and only one of them is called by its right name in the text. The fourth is not here at all, and the commonest error in a first manuscript is to call a proportion a rate.

Only one of the four has time in it. A rate needs a denominator that counts watching, not people:

$$\text{incidence rate} = \frac{\text{new cases}}{\text{person-time at risk}}$$

Two hundred patients followed for six months contribute one hundred person-years. So does one patient followed for a hundred.

Where you already count this way

A ward reporting infections per 1,000 catheter-days is doing exactly this. Two wards with the same number of infections are not comparable until you know how many catheter-days each ran.

Prevalence is a proportion with a place and a time attached

$$\text{prevalence} = \frac{339}{1{,}810} = 18.7\%$$

The full claim the paper is entitled to make:

Among adults aged 35 *and over in Zakiganj and Kanaighat, surveyed between August* 2017 *and January* 2018, 18.7% *had hypertension.*

What gets dropped in the retelling

By the time this reaches a talk it is "hypertension is 18.7% in Bangladesh".

The age floor, the two subdistricts, the season and the definition have all gone, and each one of them changes the number.

Khanam R, Ahmed S, Rahman S, et al. *BMJ Open* 2019;9(10):e026722, Table 2. PMID 31662350.

Missing values, worked

A survey of 500 people asks about smoking. 470 answer, 30 do not. Of those answering, 141 say yes.

$$\frac{141}{470} = 30.0\% \qquad \frac{141}{500} = 28.2\%$$

Both are defensible. They answer different questions and only one of them is usually stated.

Which to choose

Of those who answered is the honest description of what was measured.

Of everybody is the right denominator if you are planning a service, because the 30 who did not answer still exist and still need cigarettes counted.

The version that is not defensible

Reporting 30.0% under a heading that says $N = 500$. That is not a choice between two denominators; it is one denominator printed over another.

The 500, 470, 30 and 141 are constructed by us so the arithmetic is clean.

Collapsing categories is the same trade as a threshold

Education	As collected	Collapsed
No education	333 (18.4%)	333 (18.4%)
1 to 5 years	1,126 (62.2%)	1,477 (81.6%)
6 years or more	351 (19.4%)	

What it buys

One number instead of three, and a comparison a health worker can make.

What it costs, here

62% of this sample had between one and five years of schooling. Collapsed, they are counted as educated alongside the graduates.

Counts and percentages from Khanam 2019, Table 1. PMID 31662350. The collapsed column, 1,477 and 81.6%, is ours.

An order without a distance

Four of the Sylhet columns are **ordinal**: schooling, wealth tertile, physical activity, servings per day. The categories have an order, and no distance.

Schooling	%	cumulative %
None	18.4	18.4
1–5 years	62.2	**80.6**
≥ 6 years	19.4	100.0

So four in five had five years of schooling **or less**. That sentence needs the order to exist, and it is unavailable to a nominal column.

You already own this

ASA grade, tumour stage, a pain score out of ten.

ASA III is worse than ASA II. It is not one unit worse, and the average of a II and a IV is not a III.

What an ordinal column earns, and what it does not

Earns: a middle category, and a cumulative read in either direction.

Does not earn: a mean. There is no arithmetic on the labels.

Percentages from Khanam 2019, Table 1, total column. The cumulative 80.6 is ours, being $18.4 + 62.2$.

Act II

Out of how many? The question the number does not answer

Stat.BD

One table, two denominators, one row header between them

Everybody the survey measured. The denominator of the prevalence: 18.7% of these people had hypertension.

$n = 1810$

Those found to have hypertension. The denominator of everything below the midline of Table 2: controlled 31.3%, uncontrolled 24.2%, newly identified 44.5%.

$n = 339$

Of those, the ones who did not already know. 44.5% of 339. Our arithmetic, not a figure the paper prints.

$n = 151$

Khanam 2019, Table 2. PMID 31662350. The count of 151 is ours.

Percentages inside subgroups

Current smoker	**Men ($n = 864$)**	**Women ($n = 946$)**	**All ($n = 1{,}810$)**
Yes	546 (63.2%)	36 (3.8%)	582 (32.2%)

Three denominators, one row

864, 946 and 1,810. Each percentage is a share of its own column heading, and the heading is where the denominator is declared.

Why the combined figure is not a middle

32.2% does not sit between 63.2 and 3.8 because it is not their average.

It is a weighted average, and the weights are the group sizes.

Counts and column percentages from Khanam 2019, Table 1. PMID 31662350. The observation about weighting is ours.

The same 44.5%, in two sentences

The number in the paper: 44.5% were newly identified.

Right

"Of the people found to have hypertension, 44.5% did not already know."

Denominator 339. About 151 people.

Wrong, and easy to write

"44.5% of adults in rural Sylhet have undiagnosed hypertension."

That reads the same 44.5% against 1,810. It would mean 805 people.

The true figure for everybody is 151 out of 1,810, which is 8.3%. The paper is not at fault: the denominator is in the row header. **A percentage without its denominator is not a result. It is a number.**

151, 805 and 8.3% are our arithmetic on the paper's percentages, not figures the paper prints.

Both sentences are worth saying, and they are for different people

For the clinician

Of the people who have it, 44.5% *do not know.*

This is about the disease. It says the condition is silent, and it argues for asking rather than waiting.

For the health service

8.3% *of all adults have it and do not know.*

This is about the population. It says roughly how many cuffs, how many staff, how many tablets.

So the denominator is a choice, not a technicality

Choose the denominator that matches the decision the reader has to make, and then write it into the sentence so nobody has to reconstruct it.

8.3% is ours: 151 of 1,810. Khanam 2019, Table 2. PMID 31662350.

Where the denominator usually goes missing

- **A subgroup.** Percentages of women, printed in a table whose heading says $N = 1{,}810$.
- **A filter question.** "Of those who had ever smoked" changes n for every row after it.
- **Missing values.** Quietly dropped, so the denominator differs row by row and nothing says so.
- **An abstract.** Where the percentage survives and the n does not.

The habit that catches all four

For every percentage you read, write the two numbers it came from in the margin.

If you cannot recover them from the paper, that is a finding about the paper.

A percentage of a small denominator is mostly noise

Among the 162 men found to have hypertension, 23.5% had it controlled.

That is about 38 men. If four of them had been classified the other way, the figure would read 26.0% instead.

The percentage looks as precise as the 18.7% next to it. It is not.

The reporting rule that fixes it

Print the count beside the percentage, always.

38/162 tells the reader instantly what 23.5% conceals, and it costs one bracket.

23.5% and 162 are from Khanam 2019, Table 2. The 38, the 26.0% and the four reclassified men are ours, to show what the percentage is resting on.

Act III

Two columns at once, and the three percentages in every cell

Stat.BD

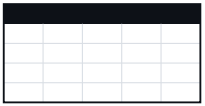
Crossing two categorical columns

Take two categorical columns from the rectangle. Count how many people fall into each combination.

Sex has two levels, hypertension has two, so there are four combinations and four counts. That is a **two-way table**, and every one of the four counts is a plain tally.

Why it is worth the trouble

A one-way table describes a group. A two-way table is the first thing in this course that lets you *compare* two groups, which is what a clinical question almost always is.



Two columns of the rectangle, crossed. Nothing has been added to the data.

One cell, three percentages, all of them correct

	Hypertension	No	
Men	162	702	864
Women	177	769	946
Total	**339**	**1471**	**1810**

Row percentage

162 ÷ 864, the shaded row.
Of the men, this share.

	Hypertension	No	
Men	162	702	864
Women	177	769	946
Total	**339**	**1471**	**1810**

Column percentage

162 ÷ 339, the shaded column.
Of those with it, this share are men.

	Hypertension	No	
Men	162	702	864
Women	177	769	946
Total	**339**	**1471**	**1810**

Percentage of **everybody**

162 ÷ 1810, everybody.
Of the whole survey, this share.

162, 177, 339, 864, 946 and 1,810 are from Khanam 2019, Tables 1 and 2. PMID 31662350. The other cells, the margins and all three percentages are ours.

So which percentage do you print?

The rule that works

Percentage *across* the exposure, so that each group sums to 100%.

Men 18.8%, women 18.7%. Now the two are comparable, because they are shares of their own groups.

Comparing 47.8% against 52.2% would only tell you there are more women in Sylhet.

The tell

Look at which direction adds to 100.

If the two numbers you are comparing come from a set that sums to 100, you are comparing shares of one group, not two groups with each other.

The footnote that decides the sentence

Current smoker	Men ($n = 864$)	Women ($n = 946$)	All
No	318 (36.8%)	910 (96.2%)	1,228
Yes	546 (63.2%)	36 (3.8%)	582

Footnote in the paper: *column percentage*.

What 63.2% means

Of the 864 men, this share smoked.
$546 \div 864$.

What it does not mean

"63.2% of smokers are men."
That figure is $546 \div 582 = 93.8\%$.

Same four counts, two sentences, and one of them is out by thirty percentage points. The only thing that tells you which is right is a footnote in six point type.

Counts and column percentages from Khanam 2019, Table 1. PMID 31662350. The 582, the 93.8% and the wrong sentence are ours.

You already read a table this way: sensitivity and specificity

	Disease	No disease	Total
Test positive	80	90	170
Test negative	20	810	830
Total	100	900	1,000

Sensitivity $= 80/100 = 80\%$, a *column* percentage.

Positive predictive value $= 80/170 = 47\%$, a *row* percentage.

The same cell, twice

80 is one number. Divided down its column it answers "of those with the disease, how many did the test catch". Divided across its row it answers "of those the test flagged, how many really have it".

Every argument you have ever had about a screening test is an argument about which denominator.

A constructed table, chosen by us for round numbers. Diagnostic accuracy proper is Session 27; this frame is about which way the percentage runs.

Act IV

When the comparison is not a fair one

The crude number and the stratum numbers disagree

Two hospitals, one thousand operations each, survival to discharge.

	Hospital A	Hospital B
Everybody	**81.9%**	**96.0%**

Case-mix, which you already argue about

A tertiary centre's raw mortality looks worse than a district clinic's because the sicker patients were sent there.

Nobody concludes the clinic is the better hospital.

These two hospitals are **constructed**, not a study. The counts are chosen so the reversal is exact and the arithmetic checks: $99 + 720 = 819$ and $882 + 78 = 960$.

The crude number and the stratum numbers disagree

Two hospitals, one thousand operations each, survival to discharge.

	Hospital A	Hospital B
Everybody	**81.9%**	**96.0%**
Mild cases	99.0% (99/100)	98.0% (882/900)
Severe cases	80.0% (720/900)	78.0% (78/100)

Case-mix, which you already argue about

A tertiary centre's raw mortality looks worse than a district clinic's because the sicker patients were sent there.

Nobody concludes the clinic is the better hospital.

These two hospitals are **constructed**, not a study. The counts are chosen so the reversal is exact and the arithmetic checks: $99 + 720 = 819$ and $882 + 78 = 960$.

The crude number and the stratum numbers disagree

Two hospitals, one thousand operations each, survival to discharge.

	Hospital A	Hospital B
Everybody	**81.9%**	**96.0%**
Mild cases	99.0% (99/100)	98.0% (882/900)
Severe cases	80.0% (720/900)	78.0% (78/100)

B looks better overall. A is better on **every** kind of patient it treats.

Case-mix, which you already argue about

A tertiary centre's raw mortality looks worse than a district clinic's because the sicker patients were sent there.

Nobody concludes the clinic is the better hospital.

What happened

A takes 900 severe cases; B takes 100. The crude figure is mostly a report on *who walked in*.

These two hospitals are **constructed**, not a study. The counts are chosen so the reversal is exact and the arithmetic checks: $99 + 720 = 819$ and $882 + 78 = 960$.

What a reversal does, and does not, license

What you may say

That the crude comparison is not a fair one, and that the third column explains why.

That is a statement about **the table**, and the table is enough to support it.

What you may not say

That Hospital A *causes* better survival.

Severity was not assigned. Something decided which patients went where, and that something may be doing the work.

The reversal tells you a comparison is unfair. It does not tell you what a fair one would show.

Comparing two populations as if they had the same patients

Hypertension rises steeply with age, so any two populations with different age structures are not directly comparable. **Direct standardisation** fixes the structure and lets the rates vary:

$$p_{\text{std}} = \sum_i w_i \, p_i$$

p_i is prevalence in age band i; w_i is that band's share of a *standard* population, the same one for everybody being compared.

Age	w_i (%)	2011	2018
35–44	36.01	17.46	28.45
45–54	27.44	24.90	38.48
55–64	18.66	30.05	46.75
65–74	11.26	39.30	51.46
75+	6.63	41.66	59.85

Age bands, weights and both prevalence columns are from Chowdhury 2021, Table 1, adults aged 35 and over. PMID 34855768, CC BY. The weights are that paper's own combined 2011–2018 age distribution.

The move you already make

Before comparing two surgeons' outcomes you ask what each was operating on.

Standardisation is that argument done arithmetically instead of verbally.

What the weights do

Every population is re-scored on one common set of patients.

The choice of standard changes the numbers, so it has to be stated. Only comparisons against the *same* standard mean anything.

Did hypertension rise, or did Bangladesh get older?

Adults 35+	2011	2018
Crude prevalence	25.84%	39.40%

A check that returns nothing is still a result

The obvious objection to a rising prevalence is that the country aged.

Here it is answered, in one sum, and the answer is no. That sentence belongs in the paper whether or not the adjustment changed anything.

Crude figures from Chowdhury 2021, Table 1. The standardised figures 25.91 and 39.29 and both ratios are **ours**, computed from that table's age bands and weights.

Did hypertension rise, or did Bangladesh get older?

Adults 35+	2011	2018
Crude prevalence	25.84%	39.40%
Age-standardised	25.91%	39.29%
Ratio, crude	1.52	
Ratio, standardised	1.52	

A check that returns nothing is still a result

The obvious objection to a rising prevalence is that the country aged.

Here it is answered, in one sum, and the answer is no. That sentence belongs in the paper whether or not the adjustment changed anything.

Crude figures from Chowdhury 2021, Table 1. The standardised figures 25.91 and 39.29 and both ratios are **ours**, computed from that table's age bands and weights.

Did hypertension rise, or did Bangladesh get older?

Adults 35+	2011	2018
Crude prevalence	25.84%	39.40%
Age-standardised	25.91%	39.29%
Ratio, crude	1.52	
Ratio, standardised	1.52	

The adjustment moves almost nothing. The rise is real, and it is not an artefact of an ageing population.

A check that returns nothing is still a result

The obvious objection to a rising prevalence is that the country aged.

Here it is answered, in one sum, and the answer is no. That sentence belongs in the paper whether or not the adjustment changed anything.

Crude figures from Chowdhury 2021, Table 1. The standardised figures 25.91 and 39.29 and both ratios are **ours**, computed from that table's age bands and weights.

Comparing a percentage with somebody else's

Sylhet, rural, adults 35+: **18.7%** hypertensive.
Nationally, adults 35+, the same period: **39.40%**.
Sylhet division: **39.08%**.

18.7% and the age structure from Khanam 2019; 39.40% and 39.08% and the age-specific rates from Chowdhury 2021, Table 1. The expected 39.5% and the ratio 0.47 are **ours**.

Comparing a percentage with somebody else's

Sylhet, rural, adults 35+: **18.7%** hypertensive.
Nationally, adults 35+, the same period: **39.40%**.
Sylhet division: **39.08%**.

Is Sylhet's sample simply younger? Apply the *national* age-specific rates to *Khanam's* age structure:

Expected, if Sylhet followed the nation	39.5%
Observed	**18.7%**
Observed / expected	**0.47**

18.7% and the age structure from Khanam 2019; 39.40% and 39.08% and the age-specific rates from Chowdhury 2021, Table 1. The expected 39.5% and the ratio 0.47 are **ours**.

Comparing a percentage with somebody else's

Sylhet, rural, adults 35+: **18.7%** hypertensive. Nationally, adults 35+, the same period: **39.40%**. Sylhet division: **39.08%**.

Is Sylhet's sample simply younger? Apply the *national* age-specific rates to *Khanam's* age structure:

Expected, if Sylhet followed the nation	39.5%
Observed	**18.7%**
Observed / expected	**0.47**

Age explains **none** of the gap.

So what is left

Who was counted. Two subdistricts, rural only, against a whole nation.

What counted as a case. Khanam counts anyone on medication, whatever the reading.

How it was measured. The average of the second and third readings, against whatever the other survey did.

And sampling. A 2016 census list, in two subdistricts.

The end point is honest, not tidy

We have ruled one explanation out and cannot choose between the rest.

18.7% and the age structure from Khanam 2019; 39.40% and 39.08% and the age-specific rates from Chowdhury 2021, Table 1. The expected 39.5% and the ratio 0.47 are **ours**.

Act V

What a percentage hides, and how to check

Stat.BD

Reading a Table 1, in order

1. Read the **title**. Who is in this table?
2. Read the **column headings**. What is *N* for each column?
3. Read the **footnotes**. Which way do the percentages run?
4. Now, and only now, read a **number**.

Do it now, out loud, on Table 1 of the Sylhet paper.

Why in that order

Everybody reads the numbers first, and then reads the headings only if a number surprises them. That is exactly backwards, because a number that does not surprise you is the one that will be quoted wrongly.

Khanam 2019, Table 1. PMID 31662350.

Auditing a table before you believe it

Four checks, none of which needs the raw data:

1. Every count column sums to its n.
2. Every percentage reproduces from its own count.
3. A count and its percentage imply **one** denominator. Find it.
4. Anything ordered runs in the direction it should, or the paper says why not.

The habit, not the technique

You already check a drug chart before you sign it.

Not because you expect an error, but because signing means you looked.

Servings percentages from Khanam 2019, Table 1; the sum to 100.1 is ours. BMI rows from Chowdhury 2021, Table 1. Neither is presented as an error.

Auditing a table before you believe it

Four checks, none of which needs the raw data:

1. Every count column sums to its n.
2. Every percentage reproduces from its own count.
3. A count and its percentage imply **one** denominator. Find it.
4. Anything ordered runs in the direction it should, or the paper says why not.

The Sylhet servings column reads $49.1 + 38.6 + 12.4 = 100.1$. That is three honest round-ups, and it is what a correct table looks like.

The habit, not the technique

You already check a drug chart before you sign it.

Not because you expect an error, but because signing means you looked.

Servings percentages from Khanam 2019, Table 1; the sum to 100.1 is ours. BMI rows from Chowdhury 2021, Table 1. Neither is presented as an error.

Auditing a table before you believe it

Four checks, none of which needs the raw data:

1. Every count column sums to its n.
2. Every percentage reproduces from its own count.
3. A count and its percentage imply **one** denominator. Find it.
4. Anything ordered runs in the direction it should, or the paper says why not.

The Sylhet servings column reads $49.1 + 38.6 + 12.4 = 100.1$. That is three honest round-ups, and it is what a correct table looks like.

The habit, not the technique

You already check a drug chart before you sign it.

Not because you expect an error, but because signing means you looked.

When check four fires

In the 2011 national data, prevalence among the *overweight* was 41.1% and among the *obese* 28.0%.

Out of order, unexplained. That is a reason to read the methods, not a finding.

Servings percentages from Khanam 2019, Table 1; the sum to 100.1 is ours. BMI rows from Chowdhury 2021, Table 1. Neither is presented as an error.

A percentage on its own is a claim with the evidence removed

What you can see

87.9% of this sample used smokeless tobacco.

What you cannot see, from that alone

How many people. How many were asked. Whether it was self-reported. Whether the question was asked the same way of men and of women.

The three that travel together

Print the **count**, the **denominator** and the **percentage**.

Any two of the three give you the missing one. Printing only the percentage is the one choice that loses information permanently.

87.9% from Khanam 2019, Table 1, self-reported. 1,591 of 1,810. PMID 31662350.

The same percentage, from three different studies

	With it	Total	Percentage
A village survey	9	48	18.8%
The Sylhet survey	162	864	18.8%
A national programme	9,411	50,059	18.8%

The percentage cannot tell them apart

Three studies, one number, and they do not deserve the same confidence. The percentage is identical because it has thrown away the one thing that distinguishes them.

The Sylhet row is real: 162 of 864 men, Khanam 2019. The other two rows are constructed by us to make the point, and are not from any study.

Common traps, collected

Reading

Percentages without denominators.
A denominator that changes mid-table.
Column percentages read as row percentages.
A percentage of fewer than about thirty people, quoted as though it were solid.

Writing

Count, denominator and percentage, together.
N in the column heading.
Direction of the percentages in the heading, not the footnote.
One decimal place, and no more.

One decimal place is not a style preference. 18.73% of 1,810 implies you could distinguish two people, and you cannot.

Close

Audit a table.

Before the next session

One. Build the frequency table for every categorical column in your own study. Count, denominator, percentage. *N* in the heading.

Two. Audit somebody else's. Take Table 1 of any paper in your field and, for every percentage in it, write the two numbers it came from.

What the second task is really for

You will find at least one percentage whose denominator you cannot recover from the paper.

Bring it. That is the interesting one.

Next: Session 4, Describing Numerical Data. The columns you could not count.

References & sources

1. Khanam R, Ahmed S, Rahman S, Kibria GMA, Syed JRR, Khan AM, Moin SMI, Ram M, Gibson DG, Pariyo G, Baqui AH. Prevalence and factors associated with hypertension among adults in rural Sylhet district of Bangladesh: a cross-sectional study. *BMJ Open* 2019;9(10):e026722. PMID 31662350. PMC6830635. doi:10.1136/bmjopen-2018-026722. CC BY-NC: numbers re-typeset here as facts, text not reproduced, PDF not redistributed.
2. Daniel WW. *Biostatistics: A Foundation for Analysis in the Health Sciences*, 9th ed. Wiley. Section 2.3, book pp. 22–30 (PDF pp. 36–44), on the frequency distribution; relative frequencies at book pp. 24–26.
3. Altman DG, Bland JM. Presentation of numerical data. *BMJ* 1996;312(7030):572. PMID 8595293. PMC2350327. doi:10.1136/bmj.312.7030.572.

Our arithmetic, not the paper's. The count 151 and the share 8.3%; the cells 702, 769 and 1,471 and every margin of the two-way table; the three percentages 18.8, 47.8 and 9.0; the 582 smokers and the 93.8%; the 38 controlled men; the exact percentages 49.06, 38.56 and 12.38; and both invented rows of the three-studies table.